IMPROVING ELECTROCARDIOGRAM CLASSIFICATION USING TRANSFER LEARNING AND LIGHTWEIGHT DENSENET-BILSTM

Bui Thi Hanh
Phenikaa University

ARTICLE INFO

Received: 19/5/2025
Revised: 29/6/2025
Published: 30/6/2025

KEYWORDS

Electrocardiogram
Arrhythmias
Transfer learning
BiLSTM
Convolutional Networks

ABSTRACT

The availability of affordable and user-friendly electrocardiogram monitors has improved healthcare for patients with periodic heart arrhythmias. However, effectively diagnosing electrocardiogram records remains challenging, even for experienced medical professionals. This work introduced a transfer learning-based algorithm for electrocardiogram classification using lightweight Densely Connected Convolutional Networks (DenseNets) integrated with Bidirectional Long Short-Term Memory (BiLSTM). We first pre-trained our model on the Icentia11K dataset, the largest public dataset of continuous electrocardiogram records, then fine-tuned it on the CPSC2018 dataset. Our model demonstrated performance comparable to state-of-the-art methods, obtaining an F_1 score of 0.839 without pre-training. With pre-training, the F_1 score further improved to 0.849. The proposed network structure outperformed existing methods in various metrics, including Area Under the Curve, F_{max}, $F_{\beta=2}$, and $G_{\beta=2}$. The Area Under the Curve and F_{max} values were 0.986 and 0.886, respectively for CPSC2018 dataset.

NÂNG CAO HIỆU SUẤT PHÂN LOẠI ĐIỆN TÂM ĐỒ DỰA TRÊN HỌC CHUYỂN GIAO VÀ MẠNG DENSENET-BILSTM NHẸ

Bùi Thị Hạnh
Đại học Phenikaa

THÔNG TIN BÀI BÁO

Ngày nhận bài: 19/5/2025
Ngày hoàn thiện: 29/6/2025
Ngày đăng: 30/6/2025

TỪ KHÓA

Điện tâm đồ
Rối loạn nhịp tim
Học chuyển giao
Mạng bộ nhớ ngắn hạn hai chiều
Mạng tích chập

TÓM TẮT

Sự xuất hiện của các thiết bị theo dõi điện tâm đồ với giá cả phải chăng và thân thiện với người dùng đã góp phần cải thiện dịch vụ chăm sóc sức khỏe cho những bệnh nhân mắc chứng rối loạn nhịp tim. Tuy nhiên, việc chẩn đoán hiệu quả các bản ghi điện tâm đồ vẫn là thách thức, ngay cả với các chuyên gia y tế giàu kinh nghiệm. Nghiên cứu này đề xuất thuật toán phân loại điện tâm đồ dựa trên học chuyển giao, sử dụng mạng tích chập kết nối dày đặc nhẹ kết hợp mạng bộ nhớ ngắn hạn hai chiều. Trước tiên, chúng tôi huấn luyện sơ bộ mô hình trên tập dữ liệu Icentia11K – tập dữ liệu công khai lớn nhất về các bản ghi điện tâm đồ liên tục – sau đó tinh chỉnh nó trên tập dữ liệu CPSC2018. Mô hình cho thấy hiệu suất tương đương với các phương pháp hiện đại, đạt điểm F_1 là 0,839 mà không cần huấn luyện sơ bộ. Khi được huấn luyện sơ bộ, điểm F_1 tiếp tục cải thiện lên 0,849. Cấu trúc mạng được đề xuất vượt trội hơn các phương pháp hiện tại qua nhiều chỉ số, bao gồm diện tích dưới đường cong, F_{max}, $F_{\beta=2}$, và $G_{\beta=2}$. Giá trị của diện tích dưới đường cong và F_{max} lần lượt là 0,986 và 0,886 cho tập dữ liệu CPSC2018.

DOI: https://doi.org/10.34238/tnu-jst.12831

Email: hanh.buithi@phenikaa-uni.edu.vn

1. Introduction

Electrocardiogram (ECG) classification is challenging when done manually, even for experts. Therefore, automated methods that can analyze and classify ECG signals quickly and accurately are essential. Researchers have developed various approaches, with the most widely used being classic machine learning (ML), deep learning (DL), and hybrid models. ML methods have shown promise in ECG classification [1]. One study introduced a novel technique to reduce 12-lead ECG classification to a single lead through a teacher-student model, achieving significant compression with only a slight accuracy drop of 0.81% [2]. Another approach enhanced classification and arrhythmia analysis through three stages: signal quality improvement, wavelet-based feature extraction, and classification using a hidden Markov model, reaching 99.7% accuracy with high sensitivity and predictive value [3]. DL, a subset of ML, has also delivered remarkable results by effectively analyzing arrhythmias and cardiac abnormalities. The Depthwise Separable Convolutional Neural Network (CNN) with Focal Loss (DSC-FL-CNN) improved performance on imbalanced datasets, achieving a strong F_1 score on the MIT-BIH (Massachusetts Institute of Technology (MIT) and Beth Israel Hospital (BIH)) arrhythmia database [4]. Other notable DL models including Deform-CNN [5] and DMSFNet [6] achieved high accuracy across datasets. DenseNet architectures have also demonstrated strong performance in ECG classification, as they efficiently extract and transmit detailed feature representations throughout the network layers [7], [8]. Other than DL methods, Bidirectional Long Short-Term Memory (BiLSTM) models have also demonstrated good performance by capturing temporal context from both directions. The IB-LSTM framework [9] achieved high accuracy in classifying atrial fibrillation (AF) signals on public databases. Hybrid models that integrate different DL techniques have further improved classification accuracy. For example, a CNN-LSTM hybrid [10] addressed data imbalance and achieved high sensitivity (97.87%) and specificity (99.29%) on the MIT-BIH dataset.

In addition to method development, the use of computational techniques like transfer learning has enhanced ECG classification by leveraging pre-trained models on large-scale datasets, improving performance, and reducing training time on limited data [11], [12]. In this study, we introduce a transfer learning-based model combining lightweight Dense Convolutional Networks and BiLSTM. The model was pre-trained on an upstream dataset, then it was fine-tuned on downstream datasets. Experiments showed that integrating a 1D dense convolutional network for local feature extraction with a BiLSTM and global max pooling for global representation yielded the best results. Comparative analysis was conducted with other models on the downstream dataset to evaluate the performance of our model.

2. Materials and methods

In the work [8], we introduced a compact model design derived from Densely Connected Convolutional Networks, comprising 37 convolutional layers, referred to as DenseNet-37. It was shown that the DenseNet-37 model had comparable performance with other models but required less computational cost. In this study, we employed transfer learning to enhance the performance of multi-label classification of 12-lead ECG signals. First, DenseNet-37 model was trained on the upstream dataset (ICENTIA11K [13]) to acquire the pre-trained weights. Next, a deep learning network based on DenseNet-37 combined with a BiLSTM [14] layer was fine-tuned on the China Physiological Signal Challenge 2018 (CPSC2018) dataset [15]. Figure 1 illustrates the flow chart of the proposed method.

2.1. Experimental datasets

ICENTIA11K: Our model underwent pre-training on the ICENTIA11K [13] dataset, comprising continuous raw ECG signals with a 16-bit resolution sampled at 250 Hz. This dataset, collected using the CartioSTAT^{TM} device, consists of 2 billion labeled beats from 11 thousand patients and serves as a valuable resource for beat and rhythm classification tasks through representation learning. The dataset was annotated initially using proprietary analysis tools by technologists of Icentia, followed by a review conducted by senior technologists.

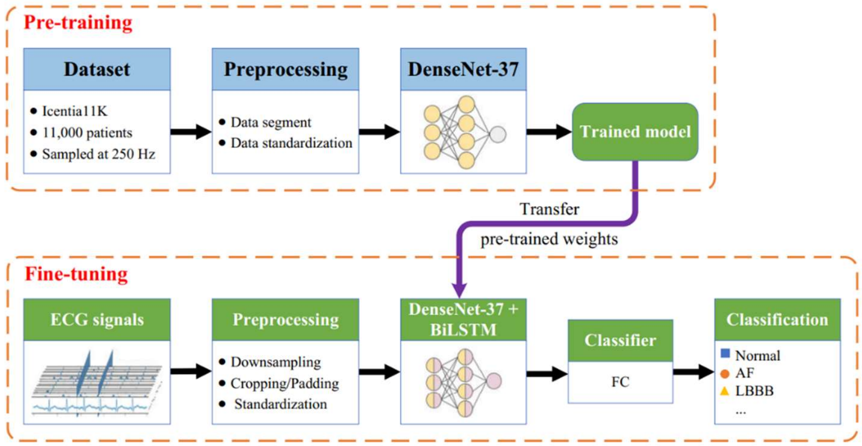

Figure 1. *Flow chart diagram of transfer learning procedure for ECG classification*

China Physiological Signal Challenge 2018 (CPSC2018): The CPSC2018 [15] dataset was gathered from 9458 patients across 11 different hospitals, which includes 9831 standard 12-lead ECG records with lengths ranging from 6 seconds to 60 seconds and sampled at 500 Hz. It consists of 12-lead ECG recordings labeled with nine types of cardiac conditions. The nine types are: Normal rhythm (NORM), Atrial fibrillation (AF), First-degree atrioventricular block (I-AVB), Left bundle branch block (LBBB), Right bundle branch block (RBBB), Premature atrial contraction (PAC), Premature ventricular contraction (PVC), ST-segment depression (STD), and ST-segment elevation (STE). The training set (6877 records) of this dataset is publicly available, while the test set remains undisclosed. Each record in the dataset can be assigned up to three labels, which include one normal sinus rhythm and eight abnormal ECG types. Out of the 6877 recordings in the CPSC2018 dataset, 476 recordings have two or three different labels assigned to them.

2.2. Architecture of proposed network

In this study, we proposed a network structure to improve the multi-label 12-lead ECG classification. Our network structure, DenseNet-BiLSTM, was built by combining the DenseNet-37 [8] and BiLSTM [16] networks. Firstly, ECG signals were used as the input for DenseNet-37 model, which worked as the local feature learning component of the network architecture. DenseNet-37 is a lightweight model developed with 3 dense blocks and some improvements, such as larger kernel filters, growth rate, and the number of input channels that were also changed compared to the original DenseNets network [8]. This model enables extracting local features and compressing the lengthy ECG record into a more concise series of local feature vectors derived from input segments. The feature vectors extracted from the data were then fed into the BiLSTM [16] layer, consisting of forward and backward LSTM components, to process sequential features. The LSTM layers captured local features from surrounding time steps, producing local-focused global feature vectors of length 64, aligned with the number of LSTM units. These were passed through a GMP layer to form a single global feature representation for classification. A classifier then categorized ECG signals using these features, consisting of a dense layer with 9 cells (for the 9 classes) and a sigmoid activation layer. The dense layer performed class recognition, while the activation layer output the probability distribution across the 9 classes.

2.3. Experimentation details

2.3.1. Pre-processing

The signals from the CPSC2018 dataset were downsampled from 500 Hz to 250 Hz to ensure consistency with the sampling rate of the ICENTIA11K dataset. As convolutional neural networks require inputs of consistent length, we cropped or padded the downsampled ECG signals to a fixed length of 30 seconds. That means the shorter signals were extended to 30 seconds by appending zero-valued data at the beginning, while longer signals were truncated to the first 30 seconds.

2.3.2. Pre-training

Pre-training is to obtain favorable initial weights to enhance model learning. The DenseNet-37 model was pre-trained on the Icentia11K dataset, using 80% for training and 20% for validation. Short ECG frames were sampled and standardized using the overall mean and standard deviation. Each frame was analyzed for abnormalities; those without abnormalities were labeled as regular beats, while frames with multiple abnormalities were labeled by the most frequent beat type. We applied the Cyclic Learning Rate technique [17], adjusting the rate from 10^{-6} to 10^{-2} via the triangular2 policy. Model weights were checkpointed at each epoch, and we reverted to the checkpoint with the lowest validation loss.

2.3.3. Fine-tuning

For fine-tuning, the CPSC2018 dataset was partitioned at random into 75% for training, 5% for validation, and 20% for testing, with class distribution maintained across all subsets. As a multi-label classification task, the model used a sigmoid-activated output layer and binary cross-entropy loss. Training employed a mini-batch size of 8, considering GPU memory and trainable parameters. CNNs were trained for up to 200 epochs with early stopping if the training F_1 score did not increase for 30 epochs. The Cyclic Learning Rate technique [17] adjusted the learning rate automatically, and the Adam optimizer was used with default settings.

After each epoch, F_1 score on the validation set was recorded. Upon completion of training, the model was restored to the checkpoint with the highest validation macro F_1 score. Output values between 0 and 1 were considered abnormal if they exceeded a threshold, which was selected by maximizing the F_1 score on the precision-recall curve. The fine-tuning procedure was conducted over 10 iterations, each time drawing fresh training and validation subsets at random from the 80% training partition. The final model used for evaluation was the one whose average precision was just above the median across all runs. This model's macro-averaged scores were then evaluated on the test set.

3. Results and discussion

3.1. Evaluation metrics

The effectiveness of our model was assessed using several evaluation metrics, including Area Under the Curve (AUC), F_1 score, F_β, G_β. F_β and G_β were computed according to formulas (1) and (2) respectively, with $\beta = 2$. The value of $\beta = 2$ indicates a higher emphasis on recall compared to precision, prioritizing the ability of the model to identify positive instances correctly.

$$F_\beta = \frac{(1+\beta^2)\times TP}{FP+\beta^2\times FN+(1+\beta^2)\times TP} \quad (1)$$

$$G_\beta = \frac{TP}{TP+FP+\beta\times FN} \quad (2)$$

Where TP and NP are True and False Positive, respectively. TN and FN are True and False Negative, respectively.

3.2. Performance on the downstream datasets

First, we evaluated the performance of our model against other leading architectures on the CPSC2018 dataset considering F_1 score. The results were summarized in Table 1. We presented the results for both cases, with and without pre-training. With pre-training, there was a significant improvement in the F_1 score. However, without pre-training, the evaluation metrics were already comparable to those achieved by other methods. Our proposed model achieved the highest F_1 scores for 3 out of 9 types, specifically Normal, PAC, and STE. The model proposed by Li et al. [18] attained the highest F_1 score for AF of 0.949. In this study, they introduced a model structure called DSE-ResNet, designed for the automatic classification of normal rhythm and 8 cardiac arrhythmias using two-dimensional ECG data. In order to improve the classification performance of the model, hyper-parameter optimization is performed using an orthogonal experiment method. On the other hand, the work by Zhang et al. [19] achieved the highest average F_1 score across all types. This work presented MLBF-Net, an architecture for arrhythmia classification using multi-lead ECG data. MLBF-Net incorporated multi-loss optimization to learn the integrity and diversity of the ECG signals jointly. The architecture consisted of lead-specific branches, cross-lead features fusion, and multi-loss co-optimization.

Table 1. *Comparison of the proposed model performance on our test set with related works for the CPSC2018 dataset* [15]. *The maximal score in each column is bolded*

Authors	**Evaluation metrics**									
	Normal	**AF**	**I-AVB**	**LBBB**	**RBBB**	**PAC**	**PVC**	**STD**	**STE**	**F_1**
Jeong et al, [4]	0.770	0.860	0.800	0.890	0.850	0.530	0.640	0.760	0.520	0.740
Qin et al, [5]	0.805	0.931	0.893	0.900	**0.948**	0.663	0.871	0.800	0.667	0.831
Zhang et al, [20]	0.805	0.919	0.864	0.866	0.926	0.735	0.851	0.814	0.535	0.813
Zhang et al, [21]	0.812	0.875	**0.923**	**0.929**	0.776	0.753	0.793	**0.837**	0.900	0.844
Zhang et al. [19]	0.847	0.934	0.884	0.896	0.939	0.822	0.878	0.818	0.677	**0.855**
Li et al. [18]	0.787	**0.949**	0.870	0.970	0.935	0.764	**0.897**	0.748	0.667	0.843
Hanh et al. [8]	0.824	0.900	0.834	0.718	0.867	0.732	0.834	0.809	0.912	0.826
This work										
Non pre-training	0.883	0.902	0.857	0.682	0.878	0.828	0.814	0.812	0.899	0.839
Pre-training	**0.886**	0.922	0.821	0.685	0.889	**0.830**	0.852	0.823	**0.932**	0.849

Table 2. *Comparison of our model with the others for the CPSC2018* [15] *dataset using different metrics. The maximal score in each column is bolded. ND means not determined*

Authors	**Evaluation metrics**			
	AUC	F_{max}	$F_{\beta=2}$	$G_{\beta=2}$
Strodthoff et al. [12]	0.974	0.855	0.819	0.602
Jeong et al. [22]	0.850	ND	ND	ND
Qin et al. [5]	0.969	ND	ND	ND
Zhang et al. [20]	0.970	ND	ND	ND
Weimann et al. [11]	0.961	0.854	0.814	0.591
Li et al. [23]	0.974	ND	ND	ND
Hanh et al. [8]	0.969	0.860	0.804	0.594
This work				
Non pre-training	0.985	0.881	0.823	0.628
Pre-training	**0.986**	**0.886**	**0.837**	**0.647**

Next, we analyzed the performance of our proposed model using four additional metrics: AUC, F_{max}, $F_{\beta=2}$, $G_{\beta=2}$. The results, presented in Table 2, showed a better performance of our model over all others in all four metrics, even without pre-training. Notably, pre-training did lead to improvements in $F_{\beta=2}$, $G_{\beta=2}$, but it did not have a significant impact on AUC and F_{ma} . In the study [8], the DenseNet-37 architecture was employed as a standalone model for ECG signal classification and yielded relatively good results compared to other models. In this work, we utilized the DenseNet-37 architecture; however, the weights were pre-trained on the Icentia11K dataset before being transferred to the CPSC2018 and PTB-XL datasets. Furthermore, the output of DenseNet-37 served as the input for the BiLSTM layer, which is effective for processing time-dependent data like ECG signals. Consequently, the DenseNet-BiLSTM model exhibited a significant performance improvement compared to the DenseNet-37 model.

In order to assess the robustness of the model, we also conducted experiments by varying the size of the train set while keeping the validation and test sets unchanged. The train set sizes were 50% and 25%, randomly selected from a pool comprising 75% of the original train set. As depicted in Figure 2, the smaller train set sizes corresponded to lower evaluation metrics. However, even with the smallest train set size of 25%, the evaluation metrics remain reasonably good, suggesting the robustness of the proposed model.

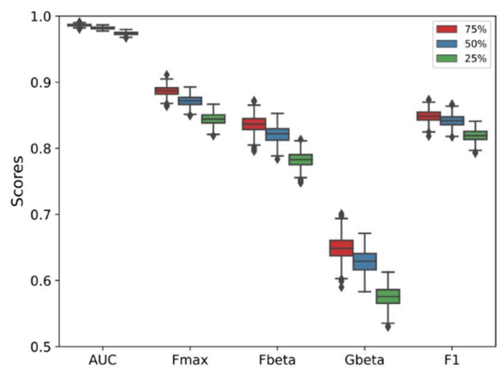

Figure 2. ***(Bootstrapped scores)*** *Empirical distribution of AUC, F_{max}, $F_{\beta=2}$, $G_{\beta=2}$ and F_1 score on the test set of the CPSC2018. Outliers are shown as black diamonds*

4. Conclusion

We introduced a transfer learning-based approach to enhance the classification of multi-label 12-lead ECGs. Our proposed network combined DenseNet for local feature learning and BiLSTM with Global Maximum Pooling for global feature learning. We pre-trained the model using the Icentia11K dataset and fine-tuned it on the CPSC2018. The performance evaluation of our proposed model on the CPSC2018 dataset demonstrated its good performance even without pre-training. With pre-training, the performance of the proposed model improved slightly. Our model achieved the highest $\mathrm{F_1}$ score for three out of nine classes in the CPSC2018 dataset compared to other state-of-the-art models. The overall $\mathrm{F_1}$ score of our model is comparable to other models. Additionally, our proposed model outperformed other models in terms of AUC, $\mathrm{F_{max}}$, $\mathrm{F_{\beta=2}}$, and $\mathrm{G_{\beta=2}}$ on the CPSC2018 dataset.

REFERENCES

[1] T. Anbalagan, M. K. Nath, D. Vijayalakshmi, and A. Anbalagan, "Analysis of various techniques for ECG signal in healthcare, past, present, and future," *Biomedical Engineering Advances*, vol. 6, 2023, Art. no. 100089.

[2] S. Majid and A.-M. Fardin, "A novel method for reducing arrhythmia classification from 12-lead ECG signals to single-lead ECG with minimal loss of accuracy through teacher-student knowledge distillation," *Information Sciences*, vol. 593, pp. 64-77, 2022.

[3] A. K. Sangaiah, M. Arumugam, and G. B. Bian, "An intelligent learning approach for improving ECG signal classification and arrhythmia analysis," *Artificial Intelligence in Medicine,* vol. 103, 2020, Art. no. 101788.

[4] Y. Lu, M. Jiang, L. Wei, J. Zhang, Z. Wang, B. Wei, and L. Xia, "Automated arrhythmia classification using depthwise separable convolutional neural network with focal loss," *Biomedical Signal Processing and Control*, vol. 69, 2021, Art. no. 102843.

[5] L. Qin, Y. Xie, X. Liu, X. Yuan, and H. Wang, “An end-to-end 12-leading electrocardiogram diagnosis system based on deformable convolutional neural network with good antinoise ability,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-13, 2021.

[6] R. Wang, J. Fan, and Y. Li, “Deep multi-scale fusion neural network for multi-class arrhythmia detection,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 2461-2472, 2020.

[7] J. K. Kim, S. Jung, J. Park, and S. W. Han, “Arrhythmia detection model using modified DenseNet for comprehensible Grad-CAM visualization,” *Biomedical Signal Processing and Control*, vol. 173, 2022, Art. no. 103408.

[8] T. H. Bui, M. T. Pham, *et al.,* “Automatic varied-length ECG classification using a lightweight DenseNet model,” *Biomedical Signal Processing and Control*, vol. 82, 2023, Art. no. 104529.

[9] K. Feng and Z. Fan, “A novel bidirectional LSTM network based on scale factor for atrial fibrillation signals classification,” *Biomedical Signal Processing and Control*, vol. 76, 2022, Art. no. 103663.

[10] G. Petmezas, K. Haris, Stefanopoulos, *et al*., “Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets,” *Biomedical Signal Processing and Control*, vol. 63, 2021, Art. no. 102194.

[11] K. Weimann and T. O. Conrad, “Transfer learning for ECG classification,” *Scientific Reports*, vol. 11, pp. 1-12, 2021.

[12] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, “Deep learning for ECG analysis: Benchmarks and insights from PTB-XL,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1519-1528, 2020.

[13] S. Tan, G. Androz, A. Chamseddine, P. Fecteau, A. Courville, Y. Bengio, and J. P. Cohen, “Icentia11k: An unsupervised representation learning dataset for arrhythmia subtype discovery,” *arXiv preprint arXiv:1910.09570*, 2019.

[14] T. Chen, R. Xu, Y. He, and X. Wang, “Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN,” *Expert Systems with Applications*, vol. 72, pp. 221-230, 2017.

[15] F. Liu, C. Liu, *et al*., “An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection,” *Journal of Medical Imaging and Health Informatics,* vol. 8, pp. 1368-1373, 2018.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735-1780, 1997.

[17] L. N. Smith, “Cyclical learning rates for training neural networks,” in *2017 IEEE winter conference on applications of computer vision (WACV)*, 2017, pp. 464-472.

[18] J. Li, S. P. Pang, F. Xu, P. Ji, S. Zhou, and M. Shu, “Two-dimensional ECG-based cardiac arrhythmia classification using DSE-ResNet,” *Scientific Reports,* vol. 2022, 12, Art. no. 14485.

[19] J. Zhang, D. Liang, A. Liu, M. Gao, X. Chen, X. Zhang, and X. Chen, “MLBF-Net: a multi-lead-branch fusion network for multi-class arrhythmia classification using 12-Lead ECG,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1-11, 2021.

[20] D. Zhang, S. Yang, X. Yuan, and P. Zhang, “Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram,” *Iscience,* vol. 24, 2021, Art. no. 102373.

[21] H. Zhang, C. Liu, Z. Zhang, *et al*., “Recurrence Plot-Based Approach for Cardiac Arrhythmia Classification Using Inception-ResNet-v2,” *Frontiers in Physiology*, vol. 12, pp. 648950, 2021.

[22] D. U. Jeong and K. M. Lim, “Convolutional neural network for classification of eight types of arrhythmia using 2D time--frequency feature map from standard 12-lead electrocardiogram,” *Scientific Reports*, vol. 11, 2021, Art. no. 20396.

[23] Z. Li and H. Zhang, “Automatic Detection for Multi-Labeled Cardiac Arrhythmia Based on Frame Blocking Preprocessing and Residual Networks,” *Frontiers in Cardiovascular Medicine*, vol. 8, 2021, Art. no. 616585.