Enhancing Cross-Dataset Generalization of 12-Lead ECG Classification Using Advanced Augmentations

Bomin Kim[1], Juhee Lee[2], Minseo Go[3], Yujin Lee[4], Dillon J Dzikowicz[5], Ben Bailey[5], Junmo An[5]

[1]Korea University, Seoul, Republic of Korea
[2]Soonchunhyang University, Asan, Republic of Korea
[3]Chosun University, Gwangju, Republic of Korea
[4]Inha University, Incheon, Republic of Korea
[5]Philips, Cambridge, MA, USA

Abstract

Automated classification of cardiac abnormalities from 12-lead electrocardiograms (ECGs) using deep learning may show reduced generalization across datasets collected in different clinical settings. This study evaluated whether data augmentation improves in-domain and cross-dataset generalization of ECG classification. A 1D ResNet34-based model was trained for multi-label classification using CutOut, CutMix, MixUp, and Random. In-domain performance was evaluated by 10-fold cross-validation on the China Physiological Signal Challenge 2018 dataset (6,877 ECGs, nine classes), while cross-dataset performance was evaluated on an independent U.S. dataset containing 4,175 ECGs from seven shared classes. All augmentation methods significantly improved the in-domain F1-score, with Random augmentation achieving the highest F1-score of 0.7566, a 2.81% relative improvement over baseline. Cross-dataset performance was substantially lower and augmentation effects were method-dependent. CutOut achieved the highest external F1-score of 0.5876 (2.25% improvement), while CutOut and CutMix significantly improved over baseline; Random and MixUp did not. These findings show that in-domain augmentation gains do not necessarily translate to external datasets and emphasize the importance of cross-dataset evaluation when assessing ECG model robustness.

1. Introduction

Accurate detection and classification of cardiac abnormalities from 12-lead electrocardiograms (ECGs) are important for clinical diagnosis and decision-making. Automated ECG classification can further support this workflow, and deep learning (DL) models, particularly end-to-end DL approaches that directly process raw ECG waveforms, have demonstrated promising performance in automated 12-lead ECG classification, often outperforming feature-based machine learning models and rule-based algorithms [1], [2], [3], [4].

However, despite strong performance in internal validation, DL models often exhibit degraded performance when applied to external datasets collected in different clinical settings. ECG datasets may differ in patient population, recording devices, acquisition protocols, signal duration, sampling frequency, noise characteristics, and diagnostic labeling criteria. These differences can result in distribution shift and may cause models to learn dataset-specific patterns rather than robust and transferable ECG representations. Therefore, achieving reliable cross-dataset generalization is important for the real-world application of end-to-end DL models for ECG classification.

Data augmentation has been widely investigated as an approach to improve model robustness and generalization [1], [5], [6]. Advanced augmentation methods such as CutOut, CutMix, MixUp, as well as strategies that randomly select among augmentation methods, may promote the learning of more robust representations [7]. However, augmentation methods that improve in-domain performance may not necessarily improve performance on external datasets.

Previous studies of data augmentation have primarily focused on performance within the same dataset [1], [2], [3], [4]. More recent work has begun to examine whether augmentation-related improvements transfer across datasets, including nonlinear ECG augmentation strategies [8]. However, systematic comparisons of commonly used augmentation methods under both in-domain and cross-dataset evaluation remain limited. In particular, it remains unclear whether performance gains observed during internal cross-validation consistently translate to improved performance on independent external datasets.

In this study, we evaluate the effects of multiple data augmentation strategies, CutOut, CutMix, MixUp, and a randomized augmentation strategy, on both in-domain and cross-dataset performance in 12-lead ECG classification.

A one-dimensional (1D) ResNet-based model was trained for multi-label classification across nine diagnostic classes using the China Physiological Signal Challenge 2018 (CPSC2018) dataset. In-domain performance was assessed using 10-fold cross-validation, while cross-dataset generalization was evaluated using the independent Georgia 12-Lead ECG Challenge Database (G12EC). The objective was to determine whether performance gains achieved using these augmentation strategies in in-domain evaluation would translate into improved generalization to an independent external dataset.

2. Methods

2.1. Dataset and Preprocessing

Two independent 12-lead ECG datasets collected in China and the United States were used for in-domain and cross-dataset evaluation. CPSC2018 was used for in-domain 10-fold cross-validation and for model training in the cross-dataset evaluation. The dataset comprises 6,877 12-lead ECG recordings collected from 11 hospitals [9]. The recordings were sampled at 500 Hz and ranged from 6 to 60 seconds in duration. Each recording was annotated with one or more of nine diagnostic classes: normal sinus rhythm (NSR), 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).

G12EC, collected at Emory University in Atlanta, Georgia, USA, was used for external testing in the cross-dataset evaluation [10]. The dataset contains 10,344 10-second, 12-lead ECG recordings sampled at 500 Hz and annotated with one or more of 67 SNOMED-CT diagnostic codes. Nine diagnostic classes overlap with those in CPSC2018. For cross-dataset evaluation, seven shared classes (NSR, AF, I-AVB, LBBB, RBBB, PAC, PVC) were included. STD and STE were excluded from the cross-dataset analysis because their classification performance was substantially lower than that of the other shared classes. After applying the class-selection criteria, 4,175 G12EC recordings were retained for cross-dataset evaluation.

Baseline wander and low-frequency components were attenuated using a 0.5-Hz high-pass filter. Each lead was independently normalized using z-score normalization to have zero mean and unit variance. For recordings longer than 10 seconds, the final 5,000 samples (10 seconds) were retained, whereas recordings shorter than 10 seconds were zero-padded to 5,000 samples. Consequently, each ECG recording was represented as a 12 × 5,000 input tensor.

2.2. Model Architecture

A 1D ResNet34-based deep learning model, shown in Figure 1, was used for ECG classification. The model receives 12-lead ECG waveforms as input and produces diagnostic class probabilities. The task was formulated as multi-label classification. We trained the model using the Adam optimizer with a batch size of 64 for 30 epochs and a learning rate of 0.0001.

The input to the deep learning model passes through one-dimensional convolution (Conv1D), batch normalization (BatchNorm), a rectified linear unit (ReLU) activation, and max pooling (MaxPool1D). The resulting feature map is then processed through four series of residual blocks. Each residual block consists of two consecutive sequences of convolution, batch normalization, and rectified linear unit. After the residual stages, it is reduced by adaptive average pooling (AvgPool1D) and passed through a fully connected (FC) layer to produce the output logits. Binary cross-entropy with logits is used as the loss function, aligning with the multi-label nature of the classification task.

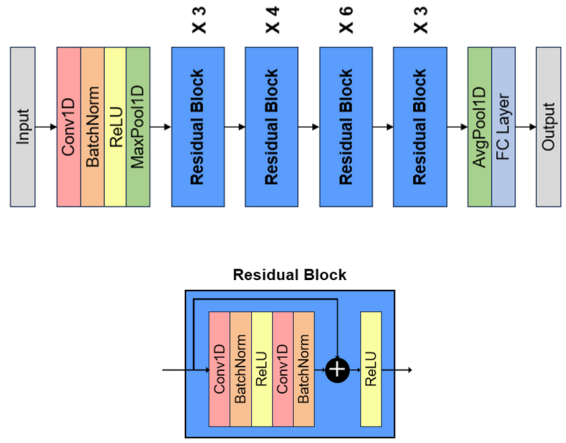

Figure 1. The architecture of the ResNet34-based deep learning model.

2.3. Data Augmentation

Figure 2 illustrates examples of the CutOut, CutMix, and MixUp methods applied to ECG signals. Data augmentation was applied only during model training.

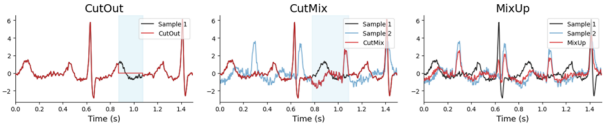

Figure 2. Examples of ECG augmentation using CutOut, CutMix, and MixUp (applied to lead II).

The baseline uses the original ECG signals without data augmentation. CutOut randomly masks a single contiguous temporal segment of an ECG signal by setting its amplitude values to zero. The length of the segment is randomly selected, ranging from 2 milliseconds to 1

second. CutMix generates an augmented ECG by replacing a randomly selected temporal segment with the corresponding segment from another ECG in the same batch. The shaded regions in Figure 2 indicate the temporal segments affected by CutOut or CutMix. MixUp generates an augmented sample by linearly interpolating two ECG signals and their corresponding multi-label targets. The mixing coefficient, λ, was randomly sampled from a Beta distribution, Beta(α, α), with$\alpha = 0.2$. MixUp was defined as:

$$x_{mix} = \lambda * x + (1 - \lambda) * x'$$
$$y_{mix} = \lambda * y + (1 - \lambda) * y'$$

where x and x′ denote two ECG samples, and y and y′ denote their corresponding multi-label targets.

The Random augmentation strategy selected MixUp with a probability of 0.4 or CutMix with a probability of 0.6 for each augmented batch. For MixUp, α was randomly sampled from 0.4 to 0.8, whereas for CutMix, α was randomly sampled from 0.8 to 1.2, with the replaced segment constrained to 5-30% of the signal length.

2.4. Training and Evaluation

For in-domain evaluation, 10-fold cross-validation was performed on the CPSC2018 dataset. The samples were randomly divided into 10 folds without stratification by diagnostic label. In each iteration, eight folds were used for training, one fold for validation, and one fold for testing. The validation fold was used to select the model with the highest F1-score over 30 epochs, and the selected model was subsequently evaluated on the held-out test fold. This procedure was repeated until each fold had served as the test fold once, and performance metrics were averaged across the 10 test folds.

For cross-dataset evaluation, only CPSC2018 samples corresponding to the seven shared diagnostic classes (NSR, AF, I-AVB, LBBB, RBBB, PAC, PVC) were retained. In each iteration, eight CPSC2018 folds were used for training and two folds for validation. The G12EC samples corresponding exclusively to the same seven classes were used as a fixed independent external test set. This procedure was repeated across the 10 CPSC2018 folds, and performance on the external G12EC test set was averaged across the 10 trained models.

Sensitivity, specificity, positive predictive value (PPV), accuracy, and F1-score were calculated for each diagnostic class. Class-wise metrics were macro-averaged across the nine classes for in-domain evaluation and across the seven classes for cross-dataset evaluation. Macro-averaging assigns equal weight to each diagnostic class regardless of class prevalence and was therefore used as the primary aggregation approach given the class imbalance in the datasets.

2.5. Implementation Details

All experiments were conducted in a Google Colaboratory (Colab) environment with an NVIDIA T4 GPU (16 GB VRAM). The deep learning models were implemented using PyTorch, with data preprocessing and numerical computation performed using pandas and NumPy, respectively. ECG data were read and processed using the WFDB library.

3. Results

Table 1. Performance of different augmentation strategies in in-domain 10-fold cross-validation. Bold indicates the best value for each metric. Statistical significance was assessed against the baseline using a paired-sample t-test (****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$).

Augmentation	Baseline	CutMix	CutOut	MixUp	Random
Sens.	0.7413	**0.7739**	0.7657	0.7611	0.7598
Spec.	0.9692	0.9678	0.9698	**0.9705**	0.9700
PPV	0.7488	0.7409	0.7547	0.7564	**0.7664**
Acc.	0.9492	0.9504	**0.9518**	0.9517	0.9513
F1	0.7359	0.7515	0.7524	0.7534	**0.7566**
Relative F1 Improvement (%)	-	2.1198 ***	2.2448 **	2.3821 ***	**2.8088** ****

As shown in Table 1, all augmentation methods improved the F1-score over the baseline in the in-domain 10-fold cross-validation. Random augmentation achieved the highest F1-score (0.7566), corresponding to a 2.8088% relative improvement over the baseline. All augmentation methods showed statistically significant F1-score improvements compared with the baseline (paired-sample t-test, $p < 0.05$).

Table 2. Performance of different augmentation strategies in cross-dataset evaluation. Bold indicates the best value for each metric. Statistical significance was assessed against the baseline using a paired-sample t-test (**$p < 0.01$; *$p < 0.05$; n.s., not significant).

Augmentation	Baseline	CutMix	CutOut	MixUp	Random
Sens.	0.6227	0.6325	**0.6506**	0.6210	0.6337
Spec.	0.9415	0.9417	0.9356	**0.9426**	0.9388
PPV	0.5929	**0.5976**	0.5919	0.5940	0.5934
Acc.	0.8874	**0.8895**	0.8866	0.8876	0.8868
F1	0.5746	0.5866	**0.5876**	0.5729	0.5802
Relative F1 Improvement (%)	-	2.0918 **	**2.2502** *	-0.3028 (n.s.)	0.9641 (n.s.)

As shown in Table 2, overall performance was lower in the cross-dataset evaluation than in the in-domain evaluation. CutOut achieved the highest F1-score (0.5876), representing a 2.2502% relative improvement over the baseline. The F1-score improvements with CutMix and CutOut were statistically significant, whereas those with Random augmentation and MixUp were not. Random augmentation produced a modest, non-significant

improvement, while MixUp resulted in a slight, non-significant decrease in F1-score.

4. Discussion and Future Work

Data augmentation generally improved in-domain ECG classification performance; however, its effects differed in the cross-dataset evaluation. The relative performance of the augmentation strategies was not consistent between the in-domain and cross-dataset settings, suggesting that improvements observed within a single dataset may not directly translate to external data with a different underlying distribution. Rather than demonstrating the general superiority of a particular augmentation method, these findings highlight that the effectiveness of data augmentation may depend on the dataset and evaluation setting.

These results also demonstrate the importance of evaluating model robustness beyond internal cross-validation. Performance was substantially lower in the cross-dataset evaluation than in the in-domain evaluation, indicating a considerable generalization gap between the two datasets. Therefore, augmentation strategies should not be evaluated solely based on improvements in internal cross-validation; their effects under distribution shifts should also be considered when assessing model generalizability.

This study has several limitations and opportunities for future work. First, the experiments were conducted using two datasets and a single model architecture. The primary objective of this study was not to achieve state-of-the-art classification performance, but to investigate how the effects of different augmentation strategies vary between in-domain and cross-dataset evaluation. Therefore, the observed effects should not be assumed to generalize to other datasets, populations, or model architectures. Future studies should evaluate augmentation strategies across multiple ECG datasets representing greater diversity in patient demographics, including age and race/ethnicity, as well as geographic regions, countries, institutions, recording devices, and acquisition settings. Evaluation across different deep learning architectures would further help determine whether the observed augmentation effects are model-dependent.

In addition, the external evaluation was limited to seven diagnostic classes shared between CPSC2018 and G12EC. Future work should include broader diagnostic coverage and a wider range of augmentation strategies, including different combinations, probabilities, parameter settings, and application schemes. Larger-scale multi-institutional studies integrating diverse datasets and architectures may provide a more comprehensive understanding of how augmentation influences cross-dataset generalization. Building on these findings, future research can further focus on improving absolute cross-dataset performance toward state-of-the-art generalization, potentially in combination with approaches such as domain adaptation, self-supervised pre-training, and label harmonization.

References

[1] J. An, R. E. Gregg, and S. Borhani, "Effective Data Augmentation, Filters, and Automation Techniques for Automatic 12-Lead ECG Classification Using Deep Residual Neural Networks," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC),* Glasgow, U.K., Jul. 2022, pp. 1283-1287.

[2] J. An, B. Bailey, and R. E. Gregg, "Paradigm Shift from Feature-Based Machine Learning to End-to-End Deep Residual Neural Networks for Pediatric Age Classification from 12-Lead ECG," in *Proc. Comput. Cardiol. (CinC),* Nov. 2023.

[3] J. An, R. Gregg, B. Bailey, Y.-H. Zhang, and D. J. Dzikowicz, "Left Bundle Branch Block Detection in 12-Lead ECG Using End-to-End Deep Learning with Explainability," in *Proc. Comput. Cardiol. (CinC),* Dec. 2024.

[4] J. An *et al.*, "Automated Pediatric Ventricular Pre-Excitation (Wolff-Parkinson-White Syndrome) Detection in 12-Lead ECG Using Deep Learning and SHAP-Based Interpretability," in *Proc. Comput. Cardiol. (CinC),* Dec. 2025.

[5] N. Nemati, "Comparative Analysis of Data Augmentation for Clinical ECG Classification with STAR," arXiv:2510.24740, Oct. 2025.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," arXiv:1710.09412, Apr. 2018.

[7] P. Guo, H. Yang, and A. Sano, "Empirical Study of Mix-based Data Augmentation Methods in Physiological Time Series Data," in *Proc. IEEE 11th Int. Conf. Healthcare Informat. (ICHI),* Houston, TX, USA, Jun. 2023, pp. 206–213.

[8] J. Lee *et al.*, "Nonlinear Compression Augmentation for Generalizable 12-Lead ECG Classification Across Multi-Site Datasets," in *Proc. Comput. Cardiol. (CinC),* Sep. 2026.

[9] F. Liu *et al.*, "An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection," *J. Med. Imag. Health Inform.,* vol. 8, no. 7, pp. 1368–1373, Sep. 2018.

[10] E. A. P. Alday *et al.*, "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," *Physiol. Meas.,* vol. 41, no. 12, Art. no. 124003, Dec. 2020.

Address for correspondence:

Junmo An, PhD
222 Jacobs St, Cambridge, MA 02141, USA
junmo.an@philips.com