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

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

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

Automated classification of cardiac arrhythmias and related abnormalities from 12-lead electrocardiograms (ECGs) is a clinically critical task. However, deep learning models trained on a single dataset fail to generalize across different clinical settings, limiting their real-world applicability. Although data augmentation has been explored to improve model robustness, its effect on cross-dataset generalization in 12-lead ECG classification remains insufficiently studied.

We trained a 1D ResNet-based deep learning model to classify nine classes (SNR, AF, IAVB, LBBB, RBBB, PAC, PVC, STD, and STE) and compared multiple augmentation strategies: Baseline (no augmentation), Cutout, CutMix, MixUp, Random, and others. For in-domain evaluation, models were trained and evaluated using 10-fold cross-validation on a dataset collected in China (6,877 records), covering all nine classes. For cross-dataset evaluation, models trained on the China dataset were evaluated on an external dataset collected in the United States. Only records whose diagnostic labels exclusively matched seven shared classes (excluding STD and STE due to label inconsistency) were retained (4,175 records).

In the in-domain evaluation, all augmentation strategies outperformed the baseline in F1 score, with Random (0.757), MixUp (0.753), Cutout (0.752), and CutMix (0.752) compared to Baseline (0.736). In cross-dataset evaluation, Cutout (0.588), CutMix (0.587), and Random (0.580) improved over the baseline (0.575), whereas MixUp (0.573) showed slight performance degradation, suggesting that its mixing strategy may not consistently transfer across datasets.

These findings indicate that the effectiveness of data augmentation is dependent on the underlying data distribution. Gains achieved in in-domain settings may not consistently generalize across datasets, highlighting the importance of incorporating cross-dataset evaluation when assessing model robustness in ECG classification.

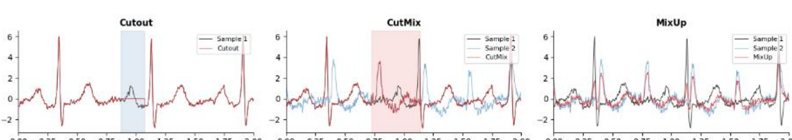

Figure 1. Examples of ECG augmentation using Cutout, CutMix, and MixUp.