Nonlinear Compression Augmentation for Generalizable 12-Lead ECG Classification Across Multi-Site Datasets

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

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

Abstract

Nonlinear amplitude compression may improve ECG classification, but whether in-domain gains generalize to external datasets remains unclear. Using a one-dimensional ResNet-34, we evaluated eight training-time compression functions: sigmoid, tanh, arctangent, softsign, logarithmic, power-law, Huber-style soft clipping, and adaptive tanh. In-domain performance was assessed by 10-fold cross-validation on CPSC2018 (6,877 recordings; nine classes). Cross-dataset performance was assessed by training and validating on CPSC2018 and testing on 4,175 G12EC recordings from seven shared classes. F1 score was the primary metric. All eight compression strategies outperformed the in-domain baseline (0.7369); tanh performed best (0.7554; +2.51%), followed by power-law (0.7543; +2.36%) and Huber-style clipping (0.7527; +2.15%). Externally, sigmoid achieved the highest F1 (0.5904; +1.74% versus 0.5803) and was the only strategy to significantly outperform the baseline. Tanh reached 0.5863 (+1.03%), whereas Huber-style clipping slightly underperformed the baseline (0.5797; −0.10%). Thus, compression benefits under dataset shift are function-dependent and require external validation.

1. Introduction

Deep neural networks have demonstrated strong performance in the automated analysis of standard 12-lead electrocardiograms (ECGs), including the detection and classification of cardiac abnormalities across different age groups, highlighting the potential of end-to-end models to support automated ECG interpretation [1], [2], [3]. However, most studies have relied on internal validation using data obtained from the same source as the training data. Differences in acquisition systems, preprocessing procedures, patient populations, disease prevalence, and annotation practices can introduce dataset shifts and substantially reduce model performance on external data. Cross-dataset evaluation is therefore necessary to assess model generalizability.

Training-time signal augmentation is commonly used to reduce overfitting by exposing models to plausible variations in ECG morphology and signal quality. Common transformations include amplitude scaling, baseline-wander addition, noise injection, signal dropout, and temporal modification [2], [3], [4], [5]. Previous work showed that sigmoid compression improved the classification performance of a one-dimensional residual network (1D ResNet) on the China Physiological Signal Challenge 2018 (CPSC2018) dataset [2]. However, that study evaluated only a single compression function in an in-domain setting. More recent work has also shown that performance improvements achieved with ECG augmentation methods, including CutMix, MixUp, CutOut, and a randomized combination strategy, do not necessarily persist under cross-dataset evaluation [6]. It therefore remains unclear whether alternative nonlinear compression functions provide consistent benefits in both in-domain and cross-dataset settings.

Nonlinear amplitude compression attenuates large-amplitude values more strongly than values near the signal baseline while preserving the temporal structure of the ECG signal. This approach may reduce model sensitivity to extreme amplitudes and amplitude-related differences between datasets. However, excessive compression may distort diagnostically relevant amplitude relationships and morphological features. The effectiveness of nonlinear compression must therefore be evaluated empirically across different data distributions.

Accordingly, this study systematically evaluates eight nonlinear compression strategies: sigmoid, tanh, arctangent, softsign, logarithmic, power-law, Huber-style soft clipping, and adaptive tanh. All strategies are evaluated using a common 1D ResNet framework. In-domain performance is assessed through 10-fold cross-validation on CPSC2018 using nine diagnostic classes. Cross-dataset performance is assessed by training the models on CPSC2018 and testing them on an external

subset of the Georgia 12-Lead ECG Challenge Database (G12EC) comprising seven diagnostic classes shared between the two datasets.

2. Methods

2.1. Dataset and Preprocessing

Two publicly available 12-lead ECG datasets were used: the China Physiological Signal Challenge 2018 (CPSC2018) dataset and the Georgia 12-Lead ECG Challenge Database (G12EC). CPSC2018 contains 6,877 recordings collected in China, with durations ranging from 6 to 60 s. G12EC contains 10,344 10-s recordings collected in the United States. Both datasets were sampled at 500 Hz.

The in-domain evaluation included nine CPSC2018 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).

For cross-dataset evaluation, both datasets were restricted to seven shared classes: NSR, AF, I-AVB, LBBB, RBBB, PAC, and PVC. The external test set comprised 4,175 G12EC recordings labeled exclusively with one of these classes. STD and STE were excluded from the cross-dataset task. No G12EC recordings were used for training or model selection.

A 0.5-Hz high-pass filter was applied to reduce baseline drift and other low-frequency components. Each lead was independently standardized to zero mean and unit variance using Z-score normalization. Recordings were truncated to the first 10 s or right-padded with zeros as needed, resulting in a 12 × 5000 input tensor.

2.2. Model Architecture

A 1D ResNet-34 was used for multi-label ECG classification, as illustrated in Figure 1. The model received a 12 × 5000 ECG tensor as input. The output layer contained nine units for the in-domain evaluation and seven units for the cross-dataset evaluation, corresponding to the diagnostic classes included in each experimental setting.

The input was initially processed by a one-dimensional convolutional layer with 64 output channels, a kernel size of 7, a stride of 2, and a padding size of 3. This layer was followed by batch normalization, a rectified linear unit activation, and one-dimensional max pooling with a kernel size of 3 and a stride of 2. The resulting feature maps were processed through four residual stages containing 3, 4, 6, and 3 residual blocks, respectively. The four stages used 64, 128, 256, and 512 output channels, respectively. Each residual block contained two one-dimensional convolutional layers with a kernel size of 3 and a shortcut connection linking the block input to its output.

After the final residual stage, adaptive average pooling reduced the temporal dimension to a fixed-length feature representation. The pooled features were flattened and passed through a fully connected layer to generate the output logits. Binary cross-entropy with logits was used as the loss function for the multi-label classification task.

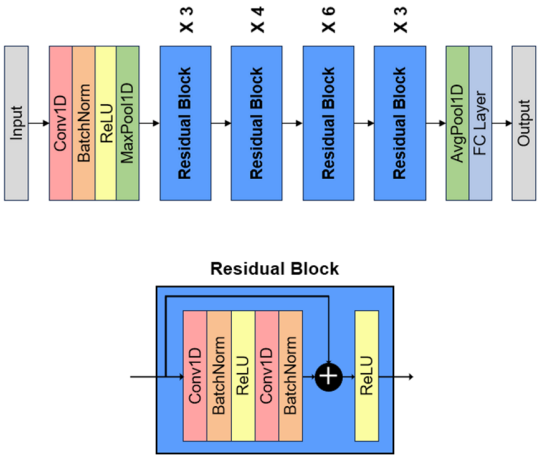

Figure 1. The architecture of the one-dimensional ResNet34-based deep learning model.

2.3. Nonlinear Compression Strategies

Eight nonlinear amplitude compression functions were evaluated separately: sigmoid, tanh, arctangent, softsign, logarithmic, power-law, Huber-style soft clipping, and adaptive tanh. Each function was evaluated as a separate experimental condition, while the model architecture, training hyperparameters, data partitions, and evaluation procedures were held constant. An uncompressed baseline using the preprocessed ECG signals without any additional nonlinear transformation was included for comparison.

For each compression condition, the selected function was applied pointwise to every sample of every lead in the preprocessed training ECGs. The validation and test ECGs underwent the standard preprocessing procedure but remained uncompressed. Thus, model selection and final evaluation were performed using signals that had not undergone nonlinear compression.

The mathematical definitions and parameter settings of the eight nonlinear amplitude compression functions are summarized in Table 1. Here, x denotes a Z-score-normalized ECG sample, and *f(x)* denotes its transformed amplitude. Figure 2 presents the corresponding input–output mappings. These transformations modify the signal dynamic range in function-specific ways and generally reduce the relative influence of large-amplitude components.

Table 1. Mathematical definitions and parameter settings of the eight nonlinear amplitude compression functions.

Compression Function	Definition
Sigmoid	$f(x) = 1 / (1 + e^{(-x)})$
Tanh	$f(x) = \tanh(x/s), \quad s = 1$
Arctan	$f(x) = (2/\pi) \arctan(x/s), \quad s = 1$
Softsign	$f(x) = x / (s + \lvert x \rvert), \quad s = 1$
Logarithmic	$f(x) = \mathrm{sgn}(x) \ln(1 + \lvert x \rvert / s), \quad s = 1$
Power	$f(x) = \mathrm{sgn}(x)(\lvert x \rvert + \varepsilon)^{\gamma}, \quad \gamma = 0.7, \quad \varepsilon = 10^{-6}$
Huber-style Soft Clipping	$f(x) = x, \quad \text{if } \lvert x \rvert \leq \delta$ $f(x) = \delta\, \mathrm{sgn}(x)[1 + \ln(\lvert x \rvert / \delta)], \quad \text{if } \lvert x \rvert > \delta$ $\delta = 1$
Adaptive Tanh	$r_t = \sqrt{(1/W) \sum_{j \in \mathcal{W}_t} x_j^2}$ $f(x_t) = \tanh[x_t / (k\, r_t + \varepsilon)]$ $k = 2, \quad W = 200 \text{ samples}, \quad \varepsilon = 10^{-6}$

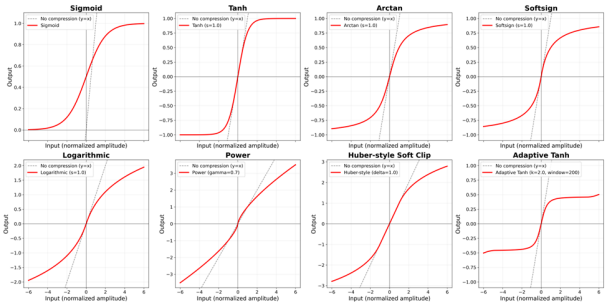

Figure 2. Input–output mapping curves of the eight nonlinear compression functions.

2.4. Training and Evaluation

A separate model was trained for the uncompressed baseline and each of the eight nonlinear compression conditions. Each model was trained for 30 epochs using the Adam optimizer, a batch size of 32, and a learning rate of 0.0001. The model checkpoint with the highest validation macro-averaged F1 score was retained for final evaluation. All experiments were conducted in Google Colaboratory using an NVIDIA T4 GPU with 16 GB of GPU memory. The models were implemented using PyTorch. ECG recordings were loaded and processed using the WFDB library, and pandas and NumPy were used for data preprocessing and numerical computation.

For the in-domain evaluation, CPSC2018 was partitioned into 10 folds. In each cross-validation run, eight folds were used for training, one for validation and checkpoint selection, and one for testing. Each fold served as the test fold once. The same fold assignments were used across the baseline and all compression conditions.

For the cross-dataset evaluation, 10 training runs were performed using CPSC2018 restricted to the seven shared classes. In each run, eight CPSC2018 folds were used for training and two for validation. The selected model was then evaluated on the same fixed external subset of 4,175 G12EC recordings. The same training and validation partitions were used across all experimental conditions.

Output probabilities were obtained by applying a sigmoid function to the model logits. Performance was primarily evaluated using the macro-averaged F1 score, which was calculated by computing the F1 score independently for each diagnostic class and then averaging the class-specific values across all classes. For each nonlinear compression condition, the relative percentage change in the macro-averaged F1 score compared with the uncompressed baseline was calculated. The in-domain and cross-dataset results were then compared to determine whether performance improvements observed during in-domain evaluation were maintained under cross-dataset testing.

3. Results

Table 2 summarizes the in-domain 10-fold cross-validation results. All eight nonlinear compression strategies achieved higher mean macro-F1 scores than the uncompressed baseline (0.7369 ± 0.0203). Tanh compression achieved the highest macro-F1 score of 0.7554 ± 0.0181, corresponding to a relative improvement of 2.51% over the baseline. Seven of the eight compression strategies significantly outperformed the baseline, whereas adaptive tanh did not (paired-sample t-test; $p < 0.05$).

Table 2. Performance scores of different compression techniques on in-domain 10-fold cross-validation. Statistical significance was assessed against the baseline using a paired-sample t-test (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; n.s., not significant).

Compression Strategy	F1 Score (Mean ± SD)	Relative F1 Improvement (%)	Significance
Identity	0.7369 ± 0.0203	-	-
Sigmoid	0.7487 ± 0.0174	+1.60	*
Tanh	0.7554 ± 0.0181	**+2.51**	***
Arctan	0.7496 ± 0.0147	+1.72	**
Softsign	0.7486 ± 0.0133	+1.59	**
Logarithmic	0.7490 ± 0.0193	+1.65	*
Power	0.7543 ± 0.0213	+2.36	**
Huber-style Soft Clipping	0.7527 ± 0.0161	+2.15	**
Adaptive Tanh	0.7455 ± 0.0159	+1.17	n.s.

Table 3 summarizes the cross-dataset evaluation results. Sigmoid compression achieved the highest macro-F1 score of 0.5904 ± 0.0085, representing a relative improvement of

1.74% over the baseline (0.5803 ± 0.0109). Sigmoid was the only compression strategy that significantly outperformed the baseline (paired-sample t-test, $p < 0.05$). None of the remaining strategies differed significantly from the baseline, and Huber-style soft clipping produced a slight performance decrease.

Table 3. Performance scores of different compression techniques on cross-dataset 10-fold cross-validation. Statistical significance was assessed against the baseline using a paired-sample t-test (*$p < 0.05$; n.s., not significant).

Compression Strategy	F1 Score (Mean ± SD)	Relative F1 Improvement (%)	Significance
Identity	0.5803 ± 0.0109	-	-
Sigmoid	0.5904 ± 0.0085	**+1.74**	*
Tanh	0.5863 ± 0.0129	+1.03	n.s.
Arctan	0.5813 ± 0.0102	+0.16	n.s.
Softsign	0.5833 ± 0.0139	+0.51	n.s.
Logarithmic	0.5809 ± 0.0103	+0.10	n.s.
Power	0.5814 ± 0.0099	+0.19	n.s.
Huber-style Soft Clipping	0.5797 ± 0.0083	-0.10	n.s.
Adaptive Tanh	0.5863 ± 0.0097	+1.02	n.s.

4. Discussion and Future Work

This study evaluated eight nonlinear compression strategies for 1D ResNet-based ECG classification under in-domain and cross-dataset settings. Most strategies significantly improved in-domain performance, with tanh producing the largest improvement. However, these gains did not consistently transfer to the external dataset. Sigmoid compression was the only strategy that significantly improved cross-dataset performance, whereas Huber-style soft clipping produced a slight decrease relative to the baseline. These findings indicate that the effectiveness of nonlinear compression depends on the evaluation setting and are consistent with previous studies showing that augmentation gains observed during internal validation may not persist across datasets.

Nonlinear compression may reduce model sensitivity to extreme amplitudes and dataset-specific signal variations. However, excessive compression may also distort diagnostically relevant amplitude and morphological relationships among the P wave, QRS complex, ST segment, and T wave. Differences between CPSC2018 and G12EC in patient populations, acquisition systems, disease prevalence, and annotation practices may have contributed to the inconsistent effects across the two evaluation settings. Therefore, performance improvements observed during in-domain evaluation should not be interpreted as improved model generalization without validation using an independent external dataset.

This study was limited to two public datasets, a single model architecture, and seven diagnostic classes in the cross-dataset analysis. Furthermore, cross-dataset evaluation was performed only in one direction, from CPSC2018 to G12EC, and only one parameter configuration was evaluated for each compression function. Class- and lead-specific morphological effects were also not examined. Future work should evaluate additional multi-institutional datasets, model architectures, transfer directions, and compression parameter settings, while examining how compression affects class-specific and lead-specific ECG morphology. Overall, nonlinear compression may improve ECG classification, but its benefits under dataset shift remain method-dependent.

References

[1] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nat. Commun.*, vol. 11, no. 1, p. 1760, 2020.

[2] 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 *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 1283–1287.

[3] 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 *2023 Computing in Cardiology (CinC)*, IEEE, 2023, pp. 1–4.

[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 *2025 Computing in Cardiology (CinC)*, 2025, pp. 1–4.

[5] J. An, 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 *2024 Computing in Cardiology (CinC)*, 2024, pp. 1–4.

[6] B. Kim *et al.*, "Enhancing Cross-Dataset Generalization of 12-Lead ECG Classification Using Advanced Augmentations," in *2026 Computing in Cardiology (CinC)*, 2026, pp. 1–4.

Address for correspondence:

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