Embedding ECG Signals into 2D Image with Preserved Spatial Information for Chagas Disease Classification

Sung-Eun Kim[1], Hong-Cheol Yoon[1], Hyun-Seok Kim[2], Woo-Young Seo[3], Sung-Hoon Kim[1,4]

[1]Department of Anesthesiology and Pain Medicine, Asan Medical Center, Brain Korea 21 Project, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Department of Anesthesiology and Pain Medicine, University of Ulsan College of Medicine, Seoul, Republic of Korea
[3] Biomedical Engineering Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea
[4] Signal House Co., Ltd, Seoul, Republic of Korea

Abstract

The 12-lead electrocardiogram (ECG) provides multiple spatial perspectives of cardiac activity, but its format complicates direct use in deep learning. We propose a framework that embeds ECG signals into structured 2D images while preserving inter-lead dependencies and aligns the image encoder with a pretrained ECG foundation model using cosine similarity. EfficientNetV2-S was employed for Chagas disease classification, with preprocessing steps, structured image construction, and physiologically motivated augmentations. Experiments used PTB-XL, SaMi-Trop, and CODE-15% datasets in the PhysioNet Challenge 2025 setting. The model achieved a score of 0.507 in 5-fold cross-validation and 0.369 on the hidden validation set, ranking 34st of 368 submissions. The score reflects the fraction of Chagas patients prioritized in the top 5% of the cohort. These results indicate that physiologically grounded image construction with foundation model alignment enables robust ECG classification, offering a scalable approach that generalizes beyond Chagas disease.

1. Introduction

The standard 12-lead electrocardiogram (ECG) records cardiac activity through twelve one-dimensional signals, each offering a distinct spatial perspective. This format poses challenges for deep learning, as spatial relationships between leads are difficult to capture, limiting the integration of recent advances in computer vision, where pretrained models and established architectures have driven substantial progress.

Transforming ECGs into structured 2D images that preserve inter-lead dependencies offers a way to bridge this gap, enabling the use of pretrained vision models and reducing reliance on large domain-specific datasets. To further enhance these representations, we align image encoder features with those from a foundation ECG model [1] using cosine similarity, a lightweight step inspired by REPA [2], which embeds explicit physiological knowledge while maintaining compatibility with vision backbones.

We evaluate this framework in the context of Chagas disease, a parasitic condition often underdiagnosed due to nonspecific symptoms and limited access to serological testing. Applied to the George B. Moody PhysioNet Challenge 2025, our method demonstrates accurate and efficient classification. Beyond Chagas, the framework provides a general strategy for integrating ECG foundation models with computer vision, supporting scalable and physiologically grounded ECG diagnostics.

2. Methods

In this section, we describe the methodology used in our study. Figure 1 illustrates the overall framework. We first introduce the datasets used, followed by the preprocessing steps, the architecture of our deep learning model, and the experimental settings employed for evaluation.

2.1. Dataset

In this study, we used three publicly available 12-lead ECG datasets: PTB-XL [3], SaMi-Trop [4], and CODE-15% [5]. Chagas labels for the CODE-15% dataset were obtained from the PhysioNet Challenge 2025. The PTB-XL dataset contains 21,799 recordings collected in Europe between 1989 and 1996, each 10 seconds long

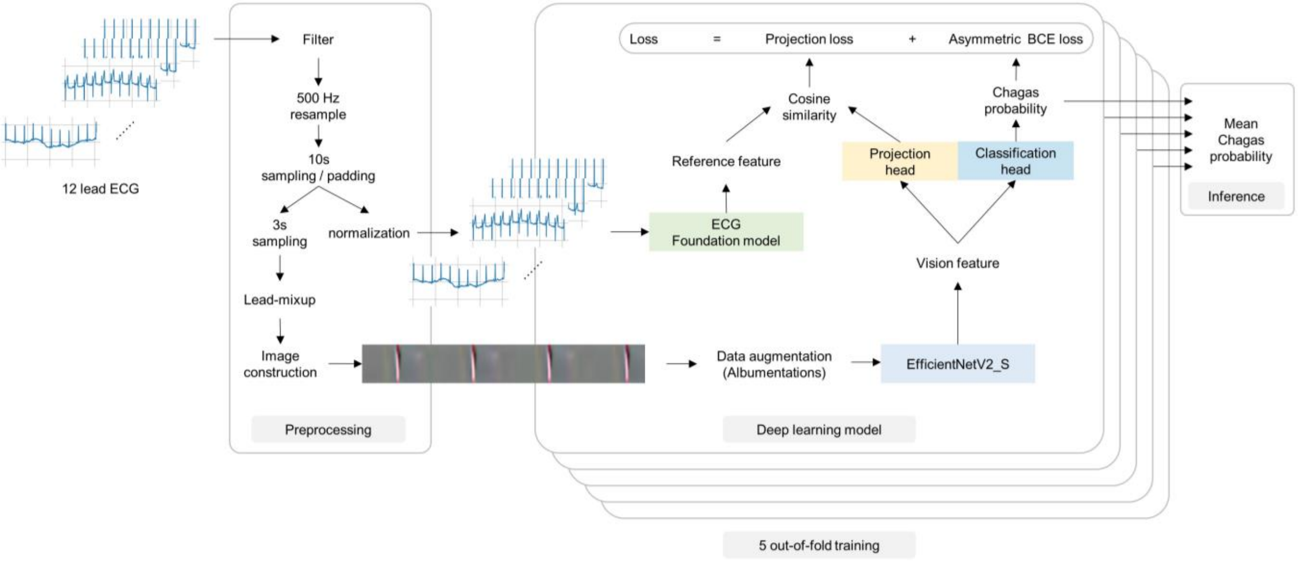

Figure 1. Overall architecture of the proposed framework. Raw 12-lead ECG signals undergo preprocessing before being embedded into structured 2D images. Physiologically motivated augmentations are applied at both signal and image levels. EfficientNetV2-S, initialized with pretrained weights, is used as the encoder, with alignment to an ECG foundation model via cosine similarity. For the PhysioNet Challenge hidden validation set, inference was performed by averaging probabilities from 5 models trained with 5-fold cross-validation.

with a sampling frequency of 500 Hz. Since PTB-XL patients are presumably non-Chagas, all Chagas labels are assumed negative. The SaMi-Trop dataset comprises 1,631 recordings collected from Chagas patients in Brazil between 2011 and 2012, with durations of 7.3 or 10.2 seconds and a sampling frequency of 400 Hz. All labels in this dataset are positive and validated by serological tests. The CODE-15% dataset includes over 300,000 recordings collected in Brazil from 2010 to 2016, with durations of 7.3 or 10.2 seconds and a sampling frequency of 400 Hz. Chagas labels in CODE-15% are self-reported, reflecting regional disease prevalence, and include both positive and negative cases

2.2. Preprocessing

Our framework required two distinct inputs: one for the image classification model and another for the ECG foundation model. To reduce artifacts and baseline drift, we first estimated and subtracted the baseline from each signal. This was achieved using a moving-average filter, which can be mathematically interpreted as a sinc-function approximation of a low-pass filter. For computational efficiency, baseline estimation was implemented with cumulative sums, a stable and significantly faster alternative to direct convolution.

Because the datasets differed in sampling frequency, all signals were resampled to 500 Hz using the librosa.resample function, which internally relies on the high-quality SoX resampling algorithm. To ensure uniform input length, signals shorter than 10 seconds were zero-padded, while longer signals were truncated to 10 seconds.

For the foundation model input, signals were standardized using z-score normalization, consistent with prior preprocessing protocols in foundation ECG models. For the image classification model, a 3-second segment was randomly sampled from the unnormalized 10-second recording. Following lead-mixup augmentation (Section 2.3), the twelve 3-second signals were embedded into a structured 2D image representation.

2.2.1. Image construction

Constructing physiologically meaningful images from ECG signals requires preserving inter-lead spatial relationships. A straightforward method is to arrange one-dimensional auxiliary signals along the temporal axis, stacking them to form a 2D representation. A key insight is that the standard 12-lead ECGs are not independent signals but rather twelve heuristic projections of a dynamic body-surface potential map, recorded at nine electrode sites. Thus, an effective image representation should approximate this underlying potential distribution while maintaining compatibility with vision-based models.

To achieve this, we constructed a 3-channel image representation analogous to RGB channels in natural images. Each channel corresponded to a distinct contour on the body surface, defined with respect to a reference electrode: right arm (RA), left arm (LA), or left leg (LL). Signals were ordered along each contour, and since all augmented limb and precordial leads share Wilson's central terminal (WCT) as a common reference, the reference augmented limb lead signal was subtracted to obtain potentials relative to RA, LA, or LL. The resulting configuration can be summarized as: Channel 1 (LL, V1–V6, LA / ref = RA), Channel 2 (RA, V1–V6, LL / ref = LA), Channel 3 (RA, V1–V6, LA / ref = LL).

This contour-based construction provides a structured approximation of the body-surface potential map, encoding spatial dependencies between leads while maintaining interpretability for both clinical and computational analysis. Recent approaches that arrange auxiliary signals along a hexaxial reference system [1], though less directly interpretable physically, can be regarded as a special case of this method, where potentials across the triangular contour defined by RA, LA, and LL are linearly approximated.

Finally, the resulting signals were clipped to the range of −3 to 3 and each limit was linearly mapped to a 0–255 scale. The image was then resized from an original resolution of 8 × timestamps to 24 × 2048, producing the final structured input for the image classification model

2.3. Deep learning model

The architecture of our framework is illustrated in Figure 1. For image-based ECG classification, we employed EfficientNetV2-S[6] as the encoder backbone. This CNN architecture combines Fused-MBConv and MBConv layers, optimized via neural architecture search to improve both training speed and accuracy. Unlike transformer-based models such as ViT [7], EfficientNetV2-S does not require fixed input dimensions, enabling direct processing of our unconventional 24 × 2048 ECG images. The network was initialized with pretrained weights and finetuned using asymmetric binary cross-entropy loss [8], which is well-suited for handling class imbalance in binary classification tasks.

Training solely on a single binary label poses a challenge, as the model may focus on spurious features rather than physiologically meaningful patterns. To address this, we leveraged features from a pre-trained ECG foundation model built on a RegNet [9] architecture trained on over 10 million recordings. Since the feature representations of the vision and ECG foundation models differ, we employed a projection module—comprising a depthwise convolution, SiLU activation, and an MLP—to map the vision features to the same dimensional space as the ECG features. During training, cosine similarity loss was applied to align the features of the image encoder with those of the ECG foundation model, following the approach inspired by REPA. The aligned features were then passed through a classification head to predict Chagas disease logits.

2.3.1. Data augmentation

To improve model generalizability and robustness, we employed data augmentation techniques both before and after constructing the image representation. Prior to image construction, we applied lead-mixup augmentation, where each of the nine patch signals was perturbed by linearly interpolating between it and eight other patches, with the interpolation coefficients sampled from a Gaussian distribution. This reflects the physiological insight that slight variations in electrode positions do not alter the diagnostic information.

After constructing the image, we applied a series of augmentations using the Albumentations [10] library. The augmentation pipeline was carefully curated to ensure that transformations remained physiologically meaningful. It included random grid shuffling, coarse dropout, small shift-scale-rotation transformations, Gaussian noise, blurring or downscaling, and mild color perturbations.

2.4. Experimental settings

We conducted extensive experiments to identify optimal training configurations for our framework. For the backbone encoder, we evaluated EfficientNet [11], EfficientNetV2, ConvNeXt [12], and ConvNeXtV2 [13]. Among these, EfficientNetV2-S achieved the best balance between performance and computational efficiency and was therefore adopted in all subsequent experiments. The model was trained for 20 epochs using the AdamW optimizer with a constant learning rate of 2×10^{-5}.

Regarding input resolution, we systematically tested multiple image sizes and found that 24 × 2048 yielded the most stable and accurate results. For lead-mixup augmentation, interpolation coefficients were sampled from a Gaussian distribution with a standard deviation of 0.1, ensuring perturbations remained physiologically realistic. Training employed asymmetric binary cross-entropy loss with parameters $\gamma^{+} = 0$, $\gamma^{-} = 2$ and a positive-class weighting factor of 10 to mitigate class imbalance.

To integrate physiological knowledge, we experimented with projection alignment across multiple intermediate layers of the encoder. However, applying the projection loss only at the final feature layer produced superior performance. The projection loss was weighted by a coefficient of 0.5 to maintain balance with the primary classification loss. We also tested strategies to rebalance the dataset by oversampling positive Chagas cases, but performance was consistently superior when using the original dataset distribution without resampling.

All experiments were rigorously validated using 5-fold cross-validation to ensure robustness and generalizability of the results.

3. Results

The performance of our framework is summarized in Table 1. In internal evaluation with given train set, the model achieved a challenge score of 0.507 across 5-fold cross-validation. When applied to the hidden validation set of the PhysioNet Challenge 2025, the model obtained a score of 0.369, corresponding to a ranking of 31st among 368 submissions. The challenge score is defined as the fraction of Chagas patients correctly prioritized within the top 5% of the cohort, as specified by the competition rules.

Dataset	Challenge Score	Rank
Train set (5-fold cross-validation)	0.507	-
Hidden validation set	0.369	34/368

Table 1. Performance of our final model on the training and official validation sets of the PhysioNet Challenge 2025.

4. Discussion and conclusions

In this study, we introduced a framework integrating conventional ECG analysis with computer vision through structured image construction. By embedding ECG signals into physiologically informed 2D representations, our approach preserves inter-lead dependencies and enables the use of pretrained vision backbones. Aligning image encoder features with a foundation ECG model using cosine similarity (REPA-inspired projection) further incorporated domain knowledge, improving learning efficiency and diagnostic relevance.

Our results demonstrate effectiveness for Chagas disease detection, achieving competitive performance in the PhysioNet Challenge 2025. The ability to leverage pretrained models across computer vision and ECG modeling underscores the flexibility of the approach and its potential for broader biomedical applications.

Several limitations remain. We prioritized computational efficiency over spectral precision, using moving-average baseline removal instead of FFT-based filtering. The datasets were also heterogeneous in labeling quality, with strong labels from SaMi-Trop and PTB-XL but weaker self-reported labels from CODE-15%, likely contributing to the performance gap between cross-validation and the hidden validation set. Additional high-quality data or semi-supervised strategies could improve robustness to unseen distributions.

In conclusion, the framework offers a scalable and physiologically grounded direction for ECG-based diagnostics by bridging ECG foundation models with computer vision. Future work should refine preprocessing, explore semi-supervised training, and validate across diverse cohorts to enhance generalizability.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), which is funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0026).

References

[1] Li, Jun, et al. "An Electrocardiogram Foundation Model Built on over 10 Million Recordings." *NEJM AI* 2.7 (2025): AIoa2401033.

[2] Yu, Sihyun, et al. "Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think." *The Thirteenth International Conference on Learning Representations*.

[3] Wagner, Patrick, et al. "PTB-XL, a large publicly available electrocardiography dataset." *Scientific data* 7.1 (2020): 1-15.

[4] Cardoso, Clareci Silva, et al. "Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile." *BMJ open* 6.5 (2016): e011181.

[5] Ribeiro, Antônio H., et al. "Automatic diagnosis of the 12-lead ECG using a deep neural network." *Nature communications* 11.1 (2020): 1760.

[6] Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller models and faster training." *International conference on machine learning*. PMLR, 2021.

[7] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[8] Ridnik, Tal, et al. "Asymmetric loss for multi-label classification." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[9] Xu, Jing, et al. "RegNet: self-regulated network for image classification." *IEEE Transactions on Neural Networks and Learning Systems* 34.11 (2022): 9562-9567.

[10] Buslaev, Alexander, et al. "Albumentations: fast and flexible image augmentations." *Information* 11.2 (2020): 125.

[11] Tan, Mingxing. "EfficientNet: Rethinking model scaling for convolutional neural networks." *arXiv preprint arXiv:1905.11946* (2019): 6105-6114.

[12] Liu, Zhuang, et al. "A convnet for the 2020s." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2022.

[13] Woo, Sanghyun, et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

Address for correspondence:

Sung-Hoon Kim
88, Olympic-ro 43-gil, Songpa-gu, Seoul, Republic of Korea
shkimans@gmail.com