Robust ECG Heartbeat Classification using Deep Residual Networks

Sanskar Jadhav
Cornell Tech
sgj32@cornell.edu

Alif Madani
Cornell Tech
aim57@cornell.edu

Abstract

This project investigates automated classification of electrocardiogram (ECG) heartbeats from grayscale waveform images under realistic deployment constraints. Each sample corresponds to a single cardiac cycle represented as a 128x128 image, reflecting scenarios where only image-based ECG records are available. The dataset presents significant challenges, including severe class imbalance and distribution shift between training and test sets caused by noise, amplitude scaling, baseline drift, and temporal warping.

We evaluate a progression of models, starting from image-based convolutional baselines and CNN–LSTM architectures and culminating in a signal-based deep residual network. By reconstructing one-dimensional ECG signals from images, applying domain-specific augmentation, and using focal loss with cross-validation and ensemble inference, the final system achieves a Kaggle test score of 0.83678, substantially outperforming earlier baselines. These results highlight the importance of representation choice, augmentation design, and robust training strategies for ECG classification under distribution shift.

1. Introduction

Electrocardiogram (ECG) analysis plays a critical role in the diagnosis and monitoring of cardiovascular conditions. Traditional automated ECG classification systems typically operate on raw electrical signals acquired directly from sensors. In many real-world scenarios, however, raw signal data may be unavailable. Instead, ECGs are often stored as scanned documents, printed charts, or rasterized images in legacy hospital systems. This motivates the study of heartbeat classification from image-based ECG representations, where the temporal signal must be inferred indirectly from visual information.

In this project, we address a five-class ECG heartbeat classification problem using grayscale waveform images. Each image represents a single cardiac cycle plotted as a black waveform on a white background and resized to 128x128 pixels. The task is to assign one of five heartbeat categories to each image, with labels provided for the training set and withheld for the test set. This setup closely resembles realistic deployment conditions, where models must generalize to unseen data distributions and operate without access to clean, structured signal inputs.

The dataset presents two major challenges. First, the class distribution is extremely imbalanced. Over 80% of the training samples belong to the normal heartbeat class, while certain abnormal classes constitute less than 1% of the data. Naive accuracy-driven optimization under this imbalance leads to models that over-predict the majority class while failing to recognize rare but clinically important events. Second, the test set exhibits a deliberate distribution shift relative to the training data. The shift includes amplitude scaling, baseline drift, additive Gaussian noise, and temporal warping, simulating ECG recordings collected from different devices, hospitals, or acquisition settings.

These challenges make the problem non-trivial and rule out purely off-the-shelf solutions. Models that perform well on clean training data may fail catastrophically under even mild distortions, as we have observed with our baseline model performance. Consequently, this project emphasizes robustness and generalization rather than just training accuracy. We adopted an iterative experimental workflow, beginning with simple image-based baselines to establish reference performance and progressively introducing architectural, representational, and optimization improvements guided by empirical observations.

First, we evaluated image-domain models, including a baseline convolutional neural network and a CNN–LSTM hybrid, to understand the limitations of treating ECGs purely as images. Second, we introduced a signal-reconstruction preprocessing step that converts waveform images into one-dimensional ECG signals, enabling the use of domain-specific data augmentation. Third, we incorporated imbalance-aware loss functions, stratified cross-validation, and ensemble inference to improve stability under distribution shift. Together, these components form what we believe is a principled approach to ECG classification from image-based data under realistic constraints.

2. Related Work

Automated ECG classification has been a focus of significant research in recent years, particularly with the adoption of deep learning methods. While early work on ECG arrhythmia detection relied on handcrafted features and classical machine learning models, recent approaches have shifted toward end-to-end neural architectures that directly learn from raw data or engineered representations. A systematic review by Wu and Guo highlights the proliferation of convolutional and recurrent architectures in ECG analysis for cardiovascular disease diagnosis, emphasizing the importance of model selection and preprocessing for robust performance [5].

Within signal-based ECG classification, deep convolutional networks and hybrid models combining convolutional and recurrent layers dominate the literature. For example, hybrid CNN–LSTM frameworks have been proposed to jointly capture local waveform morphology and temporal dependencies, showing improved performance over single-modality models on benchmark arrhythmia datasets [3]. Similarly, hybrid CNN–BLSTM architectures have demonstrated enhanced arrhythmia detection accuracy by coupling feature extraction with bidirectional temporal modelling [6]. Advanced transformer-based methods that fuse multiple views of ECG data have also been introduced to explicitly model uncertainty and noise in ECG signals, further improving robustness [2].

Although many signal-based methods operate on raw ECG recordings, image-based ECG classification, where the model ingests waveform plots rather than direct time series, has gained attention due to practical constraints in clinical and archival settings. Several recent studies use visual representations of ECG data as input to deep networks. For instance, a 2024 study leverages continuous wavelet transforms to generate time–frequency ECG spectrograms followed by transfer learning with pretrained image CNNs, highlighting the utility of image-domain features for arrhythmia classification [4]. Likewise, an ensemble of pretrained CNNs fine-tuned on ECG images has been shown to improve diagnostic accuracy across multiple cardiac conditions [1].

Despite progress in both domains, relatively little work explicitly addresses *distribution shifts* in ECG classification. Realistic variation in ECG morphology can arise from differences in acquisition devices, patient conditions, and signal noise, posing challenges for models trained on clean datasets. Research in related medical domains suggests that domain-aware augmentation and model ensembling can help mitigate such shifts, but systematic studies focusing on ECG image distributions remain limited. The present work extends this line of inquiry by combining signal reconstruction from ECG images with domain-specific augmentation and imbalance-aware optimization to enhance generalization under distribution shift.

Thus, recent advancements in ECG classification span a range of deep learning techniques from CNN, RNN, and hybrid models to image-based approaches for waveforms. However, there is still a gap in methods that integrate signal-level transformations with visual representation learning and robustness-oriented training strategies, a gap that this project aims to address.

3. Methods

This project follows an iterative, experiment-driven methodology in which model complexity and robustness are progressively increased based on empirical observations. As depicted in Figure 1, we begin with simple image-based baselines to establish reference performance and then introduce architectural, representational, and optimization improvements motivated by the structure of ECG data and the challenges posed by class imbalance and distribution shift. All experiments are conducted under realistic computational constraints, with models trained primarily on CPU resources unless otherwise noted.

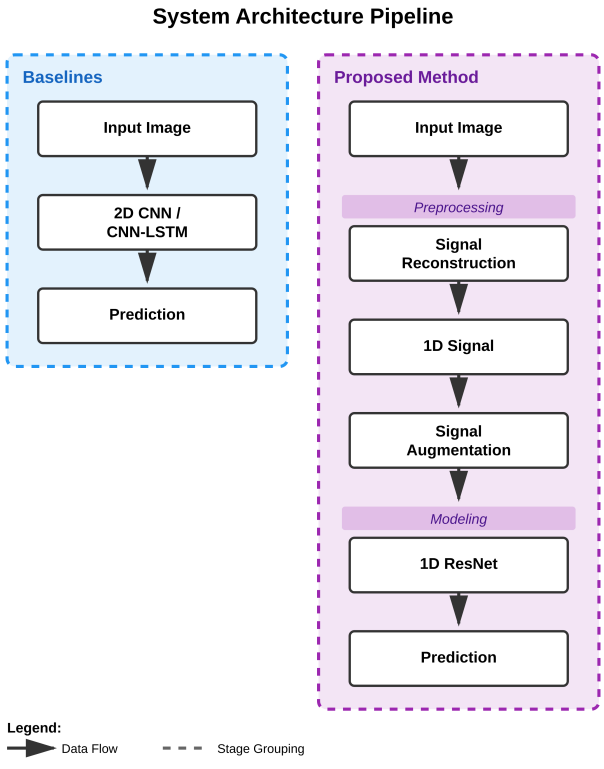

Figure 1. Schematic overview of the experimental methodology. The pipeline evaluates baseline image-based CNNs (left) before transitioning to the proposed signal-domain approach (right).

3.1. Baseline Image-Based Models

The initial experiments are implemented in a Jupyter notebook to rapidly prototype and evaluate baseline pipelines. The first baseline is a standard convolutional neural network trained directly on the 128×128 grayscale ECG images. This model consists of stacked convolutional and pooling layers followed by fully connected layers for classification. While this approach captures local spatial patterns in waveform plots, it treats ECGs purely as images and does not explicitly model temporal dependencies.

To better incorporate sequential structure, we next evaluate a CNN–LSTM hybrid architecture. In this setup, convolutional layers extract local features from the image, which are then reshaped into a sequence and passed through an LSTM module before classification. This model improves over the baseline CNN by modeling longer-range dependencies but introduces additional complexity and sensitivity to hyperparameters. These early pipelines serve as reference points for later improvements and establish the limitations of purely image-domain modeling.

3.2. Signal Reconstruction and Caching

A key design decision in this project is to reconstruct one-dimensional ECG signals from waveform images as shown in Figure 2. Since each image represents a plotted ECG trace, we extract the waveform by identifying the foreground signal and mapping pixel coordinates back to a fixed-length time series. This transformation produces a 1D signal representation that more closely matches the physical nature of ECG data.

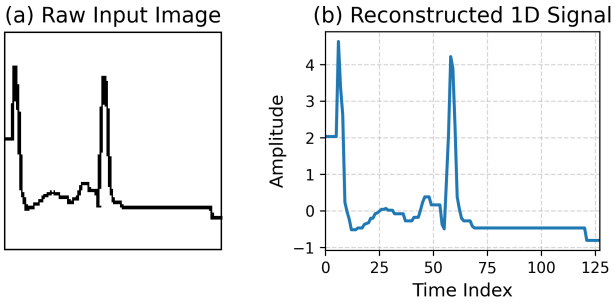

Figure 2. 1D Signal Reconstruction process. Foreground pixels are identified from the 128×128 grayscale input image (left) and mapped to vertical coordinates to recover a fixed-length time series array (right).

To improve computational efficiency, extracted signals are cached as NumPy arrays and reused across experiments. This preprocessing step significantly reduces training overhead and enables the application of signal-domain transformations that are difficult to express directly in pixel space. The separation of preprocessing and training also ensures consistent input representations across all models.

3.3. Data Augmentation

Given the distribution shift present in the test set, data augmentation is a central component of the training pipeline. Rather than using generic image augmentations, we apply domain-specific transformations in the signal space, visualized in Figure 3. These include amplitude scaling, temporal stretching and compression, baseline perturbation, and additive Gaussian noise. Each transformation is applied probabilistically during training to prevent overfitting to any single distortion.

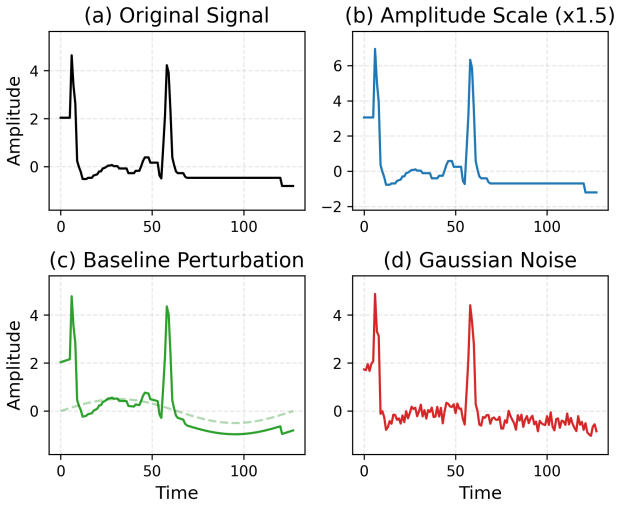

Figure 3. Visualization of domain-specific data augmentations applied in the signal space. (a) The original reconstructed ECG signal. (b-d) Transformations including amplitude scaling, baseline perturbation, and additive Gaussian noise.

Augmentation is applied exclusively to the training data, while validation data remain unaugmented. This design choice ensures that validation metrics reflect true generalization performance rather than robustness to artificial noise. The augmentation strategy is intentionally aligned with the known test-time distortions, encouraging the model to learn invariances that are directly relevant to deployment conditions.

3.4. Deep Residual Models

Building on the reconstructed signal representation, we evaluate deep residual networks adapted for one-dimensional inputs. Residual connections enable stable optimization of deeper architectures by mitigating vanishing gradient issues and allowing the network to learn incremental refinements over lower-level features. We first experiment with a 1D ResNet-18 architecture and then adopt a deeper 1D ResNet-34 model to increase representational capacity, detailed in Figure 4.

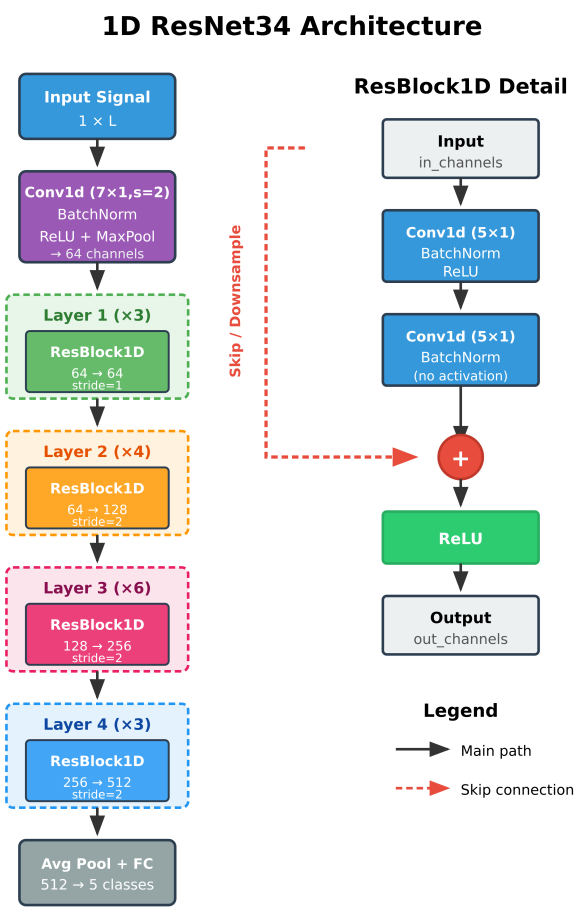

Figure 4. Schematic of the proposed 1D ResNet-34 architecture. The network consists of an initial stem (Conv1d, k=7) followed by four residual stages containing 3, 4, 6, and 3 residual blocks respectively. Each block utilizes two 5×5 1D convolutions with residual connections. Feature map resolution is halved at the start of stages 2, 3, and 4, while channel dimension doubles, culminating in a global average pooling layer and a fully connected output head.

The final model operates directly on reconstructed ECG signals and outputs probabilities over five heartbeat classes. Compared to image-based CNNs, this approach explicitly models temporal morphology and demonstrates improved robustness under distribution shift.

3.5. Loss Function and Optimization

Class imbalance is addressed using focal loss, which down-weights easy, majority-class examples and emphasizes harder, minority-class samples during training. This choice is motivated by the extreme skew in class distribution, where naive cross-entropy loss leads to degenerate solutions biased toward the dominant class.

Models are trained using the AdamW optimizer with weight decay for regularization. A learning rate scheduler based on validation performance is employed to improve convergence stability. All model parameters are initialized using standard PyTorch defaults, and random seeds are fixed to ensure reproducibility across runs.

3.6. Cross-Validation and Ensemble Inference

To reduce variance and improve robustness, we employ stratified k-fold cross-validation during training. Each fold preserves the class distribution of the original dataset, ensuring that minority classes are represented in both training and validation splits. The best-performing model from each fold is saved based on validation macro-F1 score.

At inference time, predictions from multiple cross-validation folds are combined via probability averaging to form an ensemble. This ensemble strategy mitigates overfitting to any single training split and yields more stable predictions under distribution shift. The final Kaggle submission is generated using this ensemble approach, representing the most robust configuration evaluated in this project.

4. Results

This section presents the empirical results obtained from successive model pipelines and highlights key insights gained through experimentation. Model performance is evaluated using Kaggle test scores for submitted models, with validation metrics such as macro-F1 used internally to guide model selection during training. Given the severe class imbalance in the dataset, macro-F1 is emphasized during development, while Kaggle scores provide a consistent external benchmark for comparison.

4.1. Baseline and Intermediate Results

We begin by evaluating image-based models implemented in the Jupyter notebook to establish reference performance. A baseline convolutional neural network trained directly on ECG images achieves a Kaggle score of 0.43042, indicating limited generalization under distribution shift. Introducing temporal modeling via a CNN–LSTM hybrid improves performance to 0.58519, suggesting that incorporating sequential structure helps capture ECG dynamics but remains insufficient when operating purely in the image domain.

Transitioning to deeper architectures yields further improvements. A ResNet-18 model trained with data augmentation achieves a score of 0.64082, demonstrating the benefit of residual connections and domain-aware perturbations. Applying test-time augmentation and ensembling to the ResNet-18 model increases performance to 0.67266, highlighting the effectiveness of variance reduction techniques even when the underlying representation remains image-based.

Table 1. Kaggle test performance of evaluated models.

Model Pipeline	Macro F1 Score
Baseline CNN (image-based)	0.43042
CNN–LSTM (image-based)	0.58519
ResNet-18 + augmentation	0.64082
ResNet-18 ensemble + TTA	0.67266
1D ResNet-34 + CV ensemble	**0.83678**

4.2. Signal-Based Models and Final Submission

The most significant performance gains are observed after introducing signal reconstruction and training models directly on the extracted one-dimensional ECG signals. A 1D ResNet-34 trained using focal loss and stratified cross-validation substantially outperforms all previous pipelines. The final ensemble, which averages predictions across cross-validation folds, achieves a Kaggle test score of 0.83678. Table 1 summarizes the performance of all major model pipelines evaluated in this project.

This improvement reflects multiple contributing factors: (1) the alignment of the input representation with the underlying ECG signal structure, (2) augmentation strategies tailored to known test-time distortions, and (3) robustness gained through cross-validation and ensemble inference. Notably, this performance increase is achieved without relying on excessive model complexity or specialized hardware, underscoring the effectiveness of principled design choices.

5. Discussion

This project explored ECG heartbeat classification from grayscale waveform images under realistic constraints, emphasizing robustness to class imbalance and distribution shift. Rather than relying on a single static model, we adopted an iterative experimental workflow in which design choices were guided by empirical observations from successive pipelines. Beginning with simple image-based CNN baselines and progressing toward signal-based deep residual models, each stage revealed limitations that motivated the next refinement.

A central insight from this work is the importance of **representation choice**. Models trained directly on ECG images were able to capture coarse waveform patterns but struggled to generalize under distribution shift, even when augmented or ensembled. Reconstructing one-dimensional ECG signals from images aligned the input representation more closely with the underlying physical process, enabling the use of domain-specific augmentations and substantially improving performance. This shift, combined with residual architectures, proved more impactful than increasing model complexity in the image domain alone.

Handling class imbalance was another critical factor. The extreme skew in label distribution caused standard cross-entropy training to favor the majority class. Incorporating **focal loss** improved minority-class sensitivity without requiring aggressive resampling, leading to more balanced predictions and higher macro-F1 scores during validation. **Cross-validation** further stabilized training by reducing variance across splits, while ensemble inference improved robustness by aggregating complementary decision boundaries learned from different folds.

Despite strong results, this project has several limitations. First, the signal reconstruction process is heuristic and may discard subtle morphological details present in the original images. Second, the models operate on single-lead ECG representations; multi-lead information, which is commonly available in clinical settings, could provide richer diagnostic context. Additionally, while domain-specific augmentation mitigates known distribution shifts, it does not guarantee robustness to unforeseen perturbations.

Future work could explore end-to-end architectures that jointly learn image-to-signal transformations, reducing reliance on handcrafted preprocessing. Domain adaptation or self-supervised pretraining on unlabeled ECG images may further improve generalization across acquisition settings. Finally, extending the approach to multi-lead ECGs and evaluating clinical interpretability would be valuable steps toward real-world deployment.

Overall, the results demonstrate that principled integration of representation design, augmentation, imbalance-aware optimization, and ensemble learning can significantly improve ECG classification performance under realistic constraints, even without access to raw signal data.

References

[1] Ahmed Alsayat et al. Enhancing cardiac diagnostics: a deep learning ensemble approach for precise ecg image classification. *Journal of Big Data*, 12, 2025. 2

[2] Mohd Ashhad, Sana Rahmani, Mohammed Fayiz, Ali Etemad, and Javad Hashemi. Uncertainty-aware multi-view arrhythmia classification from ecg. *arXiv preprint*, 2025. arXiv:2506.06342. 2

[3] Alaa Eleyan and Ebrahim Alboghbaish. Electrocardiogram signals classification using deep-learning-based incorporated convolutional neural network and long short-term memory framework. *Computers*, 13(2):55, 2024. 2

[4] Pinjala N Malleswari, Venkata Krishna Odugu, T. J. V. Subrahmanyeswara Rao, and T. V. N. L. Aswini. Deep learning-assisted arrhythmia classification using 2-d ecg spectrograms. *EURASIP Journal on Advances in Signal Processing*, 2024, 2024. 2

[5] Zhenyan Wu and Caixia Guo. Deep learning and electrocardiography: systematic review of current techniques in cardiovascular disease diagnosis and management. *BioMedical Engineering OnLine*, 24:23, 2025. 2

[6] Yuguang Ye, Kavimbi Chipusu, Muhammad Awais Ashraf,

Bijiao Ding, Yifeng Huang, and Jianlong Huang. Hybrid cnn-blstm architecture for classification and detection of arrhythmia in ecg signals. *Scientific Reports*, 15:17671, 2025. 2