Robust AI-ECG for Predicting Left Ventricular Systolic Dysfunction in Pediatric Congenital Heart Disease

Yuting Yang, PhD[1,2], Lorenzo Peracchio, MSc[3], Joshua Mayourian, PhD, MD[2,4], John K. Triedman, MD[2,4], Timothy Miller, PhD[1,2,*], William G. La Cava, PhD[1,2,*]

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA; [2]Department of Pediatrics, Harvard Medical School, Boston, MA; [3]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy; [4]Department of Cardiology, Boston Children's Hospital, Boston, MA

[*]Co-senior

Abstract

Artificial intelligence-enhanced electrocardiogram (AI-ECG) has shown promise as an inexpensive, ubiquitous, and non-invasive screening tool to detect left ventricular systolic dysfunction in pediatric congenital heart disease. However, current approaches rely heavily on large-scale labeled datasets, which poses a major obstacle to the democratization of AI in hospitals where only limited pediatric ECG data are available. In this work, we propose a robust training framework to improve AI-ECG performance under low-resource conditions. Specifically, we introduce an on-manifold adversarial perturbation strategy for pediatric ECGs to generate synthetic samples that better reflect real-world signal variations. Building on this, we develop an uncertainty-aware adversarial training algorithm that is architecture-agnostic and enhances model robustness. Internal and external evaluation on real-world pediatric (n=178,495) and adult (n=100,000) datasets demonstrates that our method enables low-cost and reliable detection of left ventricular systolic dysfunction, highlighting its potential for deployment in resource-limited clinical settings.

Introduction

Congenital heart disease (CHD) refers to structural or functional heart abnormalities present at birth. It is one of the most common birth defects, affecting approximately 1% of live births worldwide[1]. Electrocardiogram (ECG) is a rapid, standardized, and cost-effective tool widely used for diagnosing cardiovascular diseases and initial cardiac screening[2]. Existing studies have shown that Artificial intelligence-enhanced electrocardiogram (AI-ECG) can reliably detect early markers of cardiovascular dysfunction, including left ventricular systolic dysfunction (LVSD) in the general adult population[3,4], which is commonly associated with heart failure and adverse cardiovascular outcomes.

However, AI-ECG applications in pediatric cardiology remain largely unexplored. Pediatric ECGs differ significantly from adult ECGs in both epidemiology and characteristics, which may limit the generalizability of adult AI-ECG models[5]. Existing work in pediatric congenital heart disease[6,7] requires large amounts of labeled training data. Privacy concerns and regulatory restrictions make data sharing challenging, and large-scale publicly available pediatric ECG datasets are lacking. Consequently, hospitals with limited ECG data face challenges in developing reliable, site-specific models, highlighting the need for models that are robust in data-scarce scenarios.

To address this challenge, this study proposes a robust AI-ECG approach, incorporating the principles of adversarial training, where the model is exposed to slightly perturbed inputs to improve its robustness[8]. We introduce an adversarial training algorithm to finetune the existing AI-ECG models with generated adversarial perturbations on the model's most uncertain samples (those near the model's decision boundary). This uncertainty-aware adversarial training focuses the model's learning on its most vulnerable regions. It enables the model to learn more robust and intrinsic features, thereby achieving better generalization even in low-sample scenarios. In addition, we propose an on-manifold adversarial example generation algorithm that generates perturbations constrained by the latent manifold learned by an autoencoder[9]. Compared to perturbations in the raw signal domain, embedding-space perturbations tend to remain closer to the manifold of physiologically plausible ECGs, leading to more realistic variations. Extensive experiments on real-world pediatric dataset and external validation on an adult cohort demonstrate that our model exhibits enhanced robustness and can achieve competitive performance using only 10% of the original dataset, especially in specific lesion subgroups (e.g., patients with pacemakers).

In summary, our contributions are the following:

- We introduce on-manifold adversarial perturbation generation for pediatric ECGs, enabling the synthesis of synthetic samples that more closely resemble real-world signals.

- We propose an uncertainty-aware adversarial training algorithm, which is not limited to specific model architectures and can be used to enhance model robustness under limited data conditions.
- Validation on real-world dataset shows that our method can achieve low-cost and reliable detection for left ventricular systolic dysfunction in pediatric patients.

Methods

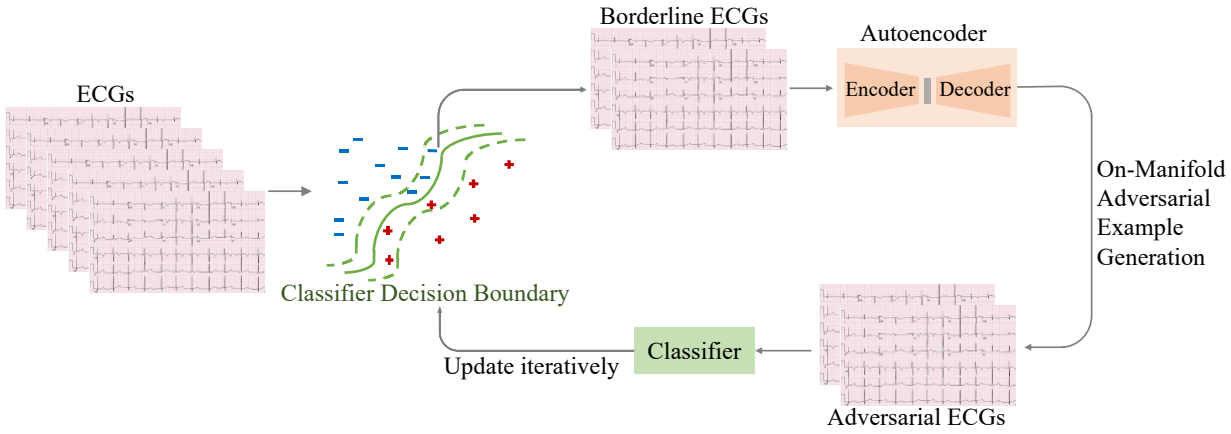

Figure 1. The overall framework of the proposed approach. It identifies "Borderline ECGs", i.e., those near the classification boundary, and augments them with on-manifold adversarial perturbations. The training follows an iterative process, where adversarial ECGs are generated, model parameters are updated, and new adversarial ECGs are regenerated, repeating this loop until convergence.

On-Manifold Adversarial Example Generation

Let X denote the set of training ECGs and $\mathcal{Y}$ their corresponding labels. The training dataset can be written as $\mathcal{D} = \{(x, y) | \, x \in X, y \in \mathcal{Y}\}$. The objective of an AI-ECG model is to learn a predictive model $f_\theta: X \to \mathcal{Y}$, parameterized by θ, which maps ECG inputs to task-specific outputs.

The purpose of adversarial example generation is to perturb a normal input x to generate an adversarial example $x_{adv} = x + \delta$ for a target model (e.g., a LVSD detector), so that x_{adv} preserves the semantics of x while misleading the target model f_θ into making incorrect predictions:

$$f_\theta(x + \delta) \neq f_\theta(x) \tag{1}$$

The loss function of generating adversarial examples ($\mathcal{L}_{adv}$) is:

$$\mathcal{L}_{adv}(x, y, \delta) = \ell(f_\theta(x + \delta), y) - \lambda * d(x + \delta, x) \tag{2}$$

where ℓ is cross-entropy loss function and d is a regularizer that constrains the perturbation δ to preserve the original semantics of x after adding perturbation. λ is used to balance these two losses; we set it to 0.1 by default. d can be cosine similarity function between original ECG signal x and perturbed signal $x + \delta$:

$$d(x + \delta, x) = \frac{\langle x, x + \delta \rangle}{|x|_2 \, |x + \delta|_2} \tag{3}$$

Then, the optimization objective of adversarial example generation is:

$$\max_{\delta} \mathcal{L}_{adv}(x, y, \delta) \tag{4}$$

We use Projected Gradient Descent (PGD[10]) to optimize δ in Equation (2), which iteratively maximizes the loss function based on the gradients of the input $\nabla_\delta \mathcal{L}_{adv}$. After T steps, we get the optimal perturbation δ^T, which can mislead the original prediction of x. As in Smooth Adversarial Perturbations[11], we employ convolution to smooth the

generated signal, which takes the weighted average of one position of the signal and its neighbors. We denote $G(s, \sigma)$ to be a Gaussian kernel with size s and standard deviation σ. The resulting adversarial example can be written as:

$$x_{\text{adv}} = x + \frac{1}{M}\sum_{m=1}^{M} \delta^T \circledast G(s[i], \sigma[i]) \tag{6}$$

M is the number of Gaussian kernels.

We observed that perturbing raw ECG signals often produces physiologically implausible waveforms, such as abnormal QRS shapes or unrealistic amplitudes, which do not correspond to any clinically plausible heart rhythms. In contrast, if one has a high-quality manifold/embedding of ECGs, perturbing directly on that embedding space can preserve the ECG's semantic and structural properties, as the perturbations are constrained to be close to the distribution of plausible patterns. To obtain latent representations of ECGs, we pre-train an autoencoder (ViT-MAE[9]) for pediatric ECGs. Instead of perturbing the raw signal x directly[11], we encode each ECG sequence into a continuous latent representation $z = Enc(x)$ and then apply perturbations in z. Thus, the objective in Equation (2) becomes:

$$\mathcal{L}_{adv}(x, y, \delta) = \ell\big(f_\theta\big(Dec(Enc(x) + \delta)\big), y\big) - \lambda * d(Dec(Enc(x) + \delta), x) \tag{7}$$

where Dec decodes a latent representation into the original input space. Then same optimization process with PGD is applied to optimize $\mathcal{L}_{adv}$ in Equation (7), applying the same Gaussian kernel smoothing from in Equation (6) in the embedding space. Specifically, we first generate the perturbation for each embedding, then smooth this perturbation across the embedding dimensions before adding it back to the original embedding. The autoencoder maps ECGs onto low-dimensional embeddings, where each dimension of the embedding corresponds to different feature patterns. Gaussian smoothing of the perturbations reduces abrupt changes in individual dimensions, making the overall embedding change gradual. As a result, the perturbations induce smooth, semantic changes (e.g., gradually modifying QRS width) rather than noisy, non-physiological alterations (e.g., sudden widening).

Uncertainty-Aware Adversarial Training

Adversarial training is designed to improve the model's resistance to adversarial perturbations. Madry A et al.[10] proposed a min-max optimization training algorithm which formalizes robustness enhancement problem as a saddle point problem:

$$\min_\theta \left[E_{(x,y)\sim\mathcal{D}} \max_\delta \mathcal{L}_{adv}\,(x, y, \delta)\right] \tag{8}$$

This is an *inner maximization* problem and an *outer minimization* problem. The inner maximization problem is a process of generating adversarial examples, aiming to find a perturbation δ that fools the victim model, i.e. achieves a high training loss. The outer minimization problem forces the model to learn generate accurate predictions under the perturbations. The goal of the outer minimization problem is to find model parameters so that the "adversarial loss" given by the inner problem is minimized. When the parameters θ yield a (nearly) vanishing risk, the corresponding model is robust to attacks. The inner maximization objective is optimized by the adversarial perturbation generation algorithm.

Since deep models exhibit varying vulnerabilities across different regions, we propose an uncertainty-aware adversarial training strategy. At each training iteration, we estimate the model's uncertainty for all training samples and retain only those with high uncertainty. Adversarial perturbations are then generated on these uncertain samples to encourage the model to attend to vulnerable regions and to learn smoother decision boundaries. We quantify the model's predictive uncertainty using the entropy of its output distribution. For an input x, let $\mathrm{p_\theta(y|x)}$ denote the predicted probability. The uncertainty $\mathcal{U}(\mathrm{x})$ is computed as:

$$\mathcal{U}(x) = -\sum_{y\in y} p_\theta(y|x) \log p_\theta\,(y|x) \tag{9}$$

Then the training samples $\mathcal{D}_u$ for each iteration is:

$$\mathcal{D}_u = TopK(x \in \mathcal{D}, \mathcal{U}(x), k) \tag{10}$$

k denotes the number of most uncertain samples retained for adversarial training. Then, the final training loss function $\mathcal{L}$ is:

$$\mathcal{L} = E_{(x,y)\sim\mathcal{D}_u} \max_\delta \mathcal{L}_{adv}\,(x, y, \delta) \tag{11}$$

As the model's parameters is optimized in each iteration, the most uncertain samples and its corresponding adversarial examples are also different during the training process. This allows the training process to continually

explore additional vulnerable regions of the model, forcing the learned decision boundary to become smoother and ultimately leading the model to capture more intrinsic and robust features.

Datasets

Internal cohort

We used patient data from Boston Children's Hospital, collected up to January 2023. The patient inclusion criterion was the availability of at least one echocardiogram with a recorded left ventricular ejection fraction (LVEF), a standard measure of heart pump function; low LVEF indicates left ventricular systolic dysfunction. To enrich the training set with overlapping pathophysiology relevant to pediatric heart failure and to broaden applicability across the heterogeneous patient population encountered in pediatric cardiology, we also included patients with cardiomyopathy as well as patients without congenital heart disease.

All raw ECG signals were retrieved from the MUSE ECG data management system (GE Healthcare, Chicago, IL, USA). CHD lesions were identified according to the institutional Fyler coding system. Paced patients were identified based on ECG diagnoses of dual-chamber or ventricular pacing. ECG recordings shorter than 10 seconds or missing lead information were excluded. Fewer than 2% of ECGs failed quality control, typically due to random issues such as unintentionally disconnected leads. The remaining ECGs were resampled to 250 Hz, high-pass filtered, and truncated to 2048 samples (approximately 8 seconds) to facilitate use with convolutional neural networks. Each ECG has 12 leads. Additional details of quality control and preprocessing have been described previously[12]. The training cohort comprised 124,265 ECGs (49,158 patients; median age 10.5 years (IQR (interquartile range) 3.5-16.8); 46.5% patients were female and 53.5% were male). The testing cohort comprised 54,230 ECGs (21,068 patients; median age 10.9 years (IQR 3.7-17.0); 46.6% patients were female and 53.4% were male). 24.1% patients had congenital heart disease in the overall testing cohort. The characteristics of the dataset are summarized in Table 1.

Table 1. Characteristics of cohorts.

	Internal Cohort		External Cohort	
	Training	Testing	Training	Testing
Patients (n)	49,158	21,068	26,218	5,442
ECGs (n)	124,265	54,230	72,475	5,442
Age (years, median (IQR)	10.5 (3.5-16.8)	10.9 (3.7-17.0)	63.0 (52.0-73.0)	64.0 (52.0-74.0)
Outcomes (n (%))				
LVEF ≤ 50%	8,525 (6.9%)	3,674 (6.8%)	19,069 (26.3%)	1,096 (20.1%)
LVEF ≤ 40%	3,381 (2.7%)	1,473 (2.7%)	13,607 (18.8%)	717 (13.2%)
LVEF ≤ 30%	1,490 (1.2%)	598 (1.1%)	8,777 (12.1%)	428 (7.9%)

LVEF values were extracted from echocardiogram reports, with the left ventricle consistently corresponding to the morphological left ventricle. LVEF was determined using the bullet method [13]. The primary outcome was LVEF of 40% or less (quantitatively at least moderate dysfunction). Secondary outcomes included LVEF of 50% or less (quantitatively at least mild dysfunction) and LVEF of 30% or less (quantitatively severe dysfunction). The reports also provide outcomes of LVEF ≤ 45% and LVEF ≤ 35%. The median LVEF was 62.0% (IQR 57.4%-66.0%) in training set and 62.0% (IQR 57.6%-66.0%) in test set, where 2.7% ECGs had an LVEF of 40% or less.

There are 18 CHD lesion subgroups in both training and testing cohorts. In structural lesions, such as Coarctation of the aorta or Ventricular septal defect, ECG abnormalities (e.g., chamber hypertrophy, axis deviation, repolarization changes) more directly reflect the underlying hemodynamic burden. These lesion-specific signatures allow AI-ECG to capture physiologically meaningful features associated with impaired ventricular function, potentially yielding

more consistent performance. In contrast, for patients with pacemakers, the ECG is dominated by pacing artifacts and non-physiologic ventricular activation, which may obscure native conduction and repolarization patterns linked to ventricular function. As a result, LVEF prediction in this cohort represents a greater challenge, but also serves as an important proof-of-concept for the robustness and generalizability of AI-ECG. Thus, we evaluate model performances on overall cohort as well as pacemaker subgroup.

External cohort

The external cohort contains de-identified collection of 100,000 12-lead ECGs with paired structural heart disease (SHD) labels derived from echocardiography, collected at Columbia University Irving Medical Center[14]. All data were retrospectively collected from adult patients (age ≥18 years), providing a distinct cohort compared with pediatric population in the BCH dataset. The training cohort comprised 72,475 ECGs from 26,218 patients and the testing cohort comprised test 5,442 ECGs from 5,442 patients. Echocardiographic data were extracted from the Syngo Dynamics (Siemens) and Xcelera (Philips) systems including LVEF. Using the same set of thresholds, we derived five binary outcomes ranging from LVEF ≤ 50% to LVEF ≤ 30%.

The purpose of this external validation is not to test a specific model's performance across different cohorts, but to assess the generalizability of the proposed adversarial training method. As a task- and model-agnostic approach, adversarial training should theoretically improve robustness regardless of patient age or disease distribution, and validating it on an external cohort allows us to assess its applicability beyond the internal pediatric cohort.

Baselines

ResNet: We use the state-of-the-art AI-ECG model for pediatric LVSD detection as the baseline model[12]. The model is based on ResNet[15], a convolutional neural network originally designed for image recognition, which can process ECG signals by capturing hierarchical temporal features through residual connections. More specifically, the ResNet consisted of a convolutional layer followed by 4 residual blocks with 2 convolutional layers per block. The convolutional layers start with 64 filters for the first layer and residual block, with a filter increase and subsampling. The output of each convolutional layer is rescaled using batch normalization and fed into a rectified linear activation unit, with subsequent dropout at a rate of 0.2. Max pooling and convolutional layers with filter length 1 are included in the skip connections to match main branch signal dimensions. The output of the last block is fed into a fully connected layer with a sigmoid activation function given that outcomes are not mutually exclusive.

ResNet+DA: We further construct an augmented baseline by applying data augmentation (DA) to ResNet. Specifically, for each training sample, we construct an augmented sample by introducing Gaussian noise that stimulates the real-life noise. We add random Gaussian noise at four real-world ECG noise frequency ranges: 3-12 Hz (motion artifact during tremors), 12-50 Hz (lower-frequency muscle activation artifact), 50-100 Hz (electrode motion noise), and 100-150 Hz (higher-frequency muscle activation artifact)[16,17]. Powerline interference noise is further captured within the 50-100 Hz and 100-150 Hz frequency ranges[18]. Then ResNet is trained with both original and augmented samples.

Metrics

Due to data privacy concerns, it is challenging to obtain existing pediatric ECG datasets from smaller hospitals to validate the effectiveness of our approach. As an alternative, we compared the performance of the prediction model when trained on the full dataset versus a smaller subset consisting of our data. To emulate real-world scenarios in hospitals of varying sizes, we randomly selected 10% of the original training set to form smaller training subsets. We apply adversarial training algorithm while training and test on the whole test set. Given the imbalanced dataset, both area under the receiver operating curve (AUROC[19]) and area under the precision-recall curve (AUPRC[20]) are computed to evaluate the model's performance.

Implementation Details

While training, we use Adam optimizer. A maximum of 100 epochs was used with early stopping on the basis of validation loss. Final hyperparameters were kernel size 17, batch size 64, and learning rate 0.001. For adversarial training, we keep Top K=30% uncertain samples for each training iteration. This proportion was chosen to balance focusing on informative hard examples while preserving sufficient training diversity. Empirically, selecting too few samples led to unstable adversarial updates, whereas selecting too many reduced the benefit of uncertainty-based

selection. Inner optimization steps T is 20, α=0.001 and clamping bound δ is 0.5. s is set to [5, 7, 11, 15, 19] and σ is [1, 3, 5, 7, 10].

To provide image-based inputs suitable for the ViT-MAE, ECGs were transformed into spectrograms. Specifically, each of the 12 ECG leads was converted using the Short-Time Fourier Transform (STFT[21]), retaining both the real and imaginary components of the resulting spectra. This process yielded 24 channels per ECG recording, which were treated as the input image channels for the model. Since the pretrained base ViT-MAE was originally designed for three-channel RGB images, we adapted it to handle 24-channel spectrograms by replicating the weights of the first convolutional projection layer across the additional channels, while leaving the remainder of the encoder-decoder architecture unchanged. The model was then adapted to the ECG domain via self-supervised continual learning on the spectrograms, using the mean squared error loss of the reconstructions. This strategy leveraged the pretrained backbone as a strong initialization, while enabling the model to progressively refine its latent representations for ECGs. Source code and supplemental materials are available on GitHub[1].

Results

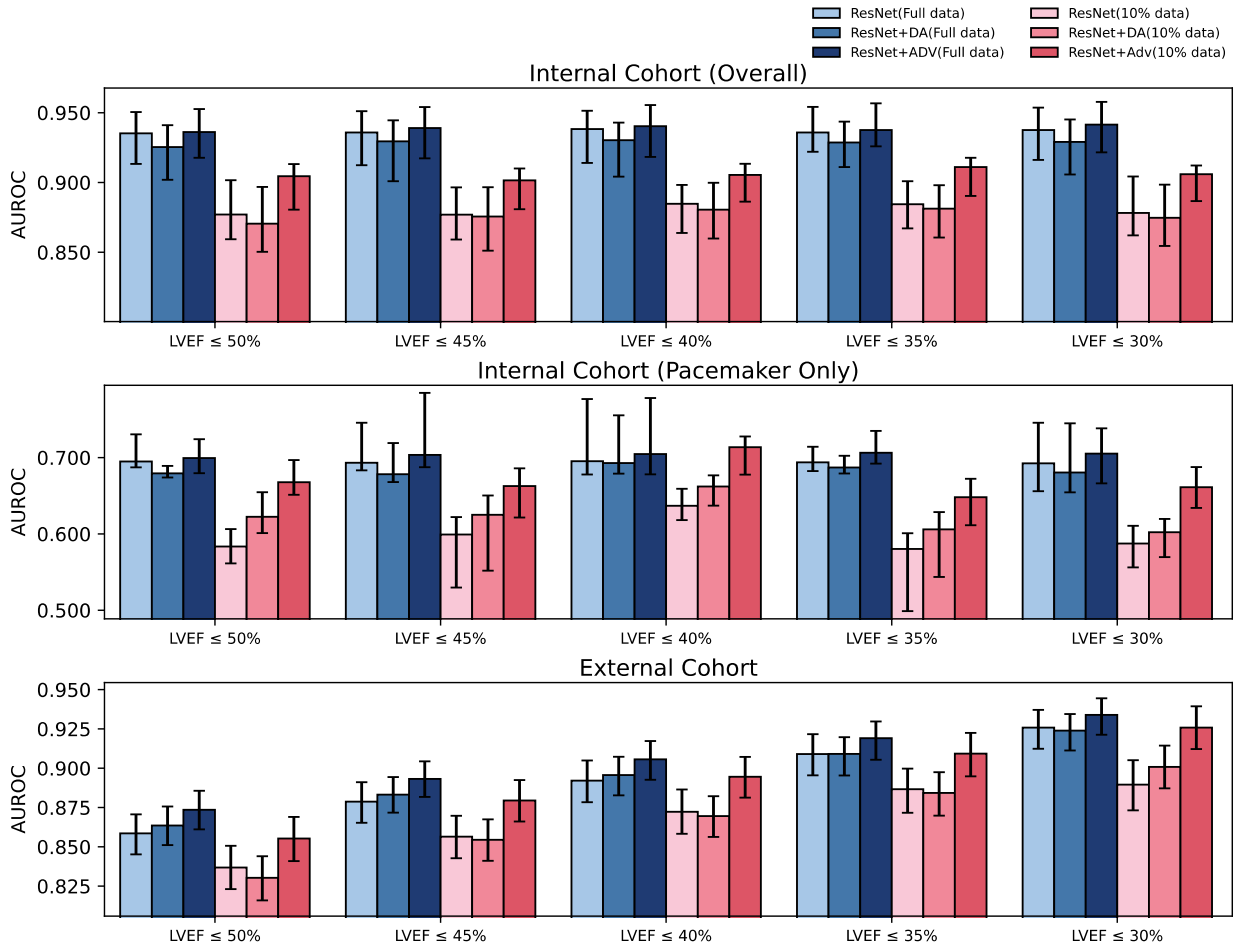

Figure 2. Model performance of ResNet, ResNet+DA (data augmentation), and ResNet+ADV (adversarial training) on the internal cohort test set and the external cohort under full and 10% training data. 95% CIs are shown using bootstrapping and indicated by error bars.

Figure 2 presents the model performance on the internal cohort (overall and pacemaker-only subgroup) as well as on the overall external cohort. We compare three models, ResNet, ResNet+DA and ResNet+ADV (ours), across varying training sizes (full and 10% data) in LVEF outcomes with different thresholds. The performance of "ResNet(Full

[1] https://github.com/cavalab/robust-AI-ECG

data)" represents a performance upper-bound benchmark for the task, such large-scale data collections are rarely available in most hospitals. Confidence intervals (CIs) were obtained via resampling with 1000 bootstraps. We find that models trained on the full dataset ("ResNet(Full data)") consistently outperform those trained on only 10% of the data ("ResNet(10% data)"). For the internal overall cohort, performance decreased from AUROC = 0.94(0.87-0.95) to 0.88(0.82-0.89) in predicting LVEF $\leq$ 40%, corresponding to an absolute drop of 0.06 median AUROC. The decline was larger for pacemaker lesion with LVEF $\leq$ 35% or LVEF $\leq$ 30%, where AUROC decreased by about 0.11($\simeq$ 16%) subgroups further complicated by highly scare positive samples. Specifically, patients with pacemakers account for only 2.1% of the overall cohort, representing a more data-scarce scenario. In addition, compared with LVEF $\leq$ 40%, patients reaching LVEF $\leq$ 35% or $\leq$ 30% are even fewer, as these thresholds reflect more severe left ventricular dysfunction. These findings highlight current AI-ECG model's sensitivity to data scarcity and suggest potential limitations in resource-constrained clinical settings.

We then evaluate the effect of adversarial training ("ResNet+ADV"). Under ideal conditions with abundant training data ("Full data"), adversarial training achieves comparable or slightly improved performance relative to the baseline models for both internal or external cohort. In contrast, under data-scarce settings (e.g., using only 10% of the training data), adversarial training yields more substantial performance gains. With adversarial training, ResNet can achieve comparable performance with that trained with full data: the difference of "ResNet(Full data)" and "ResNet+ADV(10%)" is within a margin of 0.03 median AUROC across all outcomes for overall cohort. Replication of these results in the external cohort highlights the robustness and generalizability of our method across different clinical settings and patient groups.

The improvement is particularly significant for the highly underrepresented internal pacemaker lesion. For example, in predicting LVEF $\leq$ 30%, "ResNet+ADV(10% data)" achieves an AUROC of 0.67 (0.64-0.70), significantly outperforming "ResNet(10% data)" (AUROC=0.59 (0.56-0.61)) by 0.08 and nearly matching "ResNet(Full data)" (AUROC=0.69 (0.65-0.74) with only a 0.02 difference in the median. Given that pacemaker ECGs are dominated by pacing-induced patterns rather than native conduction, these results indicate that adversarial training can enhance model robustness specifically in lesion groups with atypical ECG characteristics. Overall, these findings demonstrate that our approach maintains comparable or slightly improved performance under data-rich conditions, while yielding substantial gains under data-scarce conditions.

Ablation Studies

We perform ablation experiments to evaluate the impact of different components in our approach. As shown in Table 2, compared with ResNet+ADV, "w/o uncertainty" generates adversarial examples for all inputs without leveraging uncertainty $\mathcal{U}(\mathrm{x})$ to select borderline samples. "w/o on-manifold" uses the same adversarial example generation algorithm but generates perturbations directly on the input ECG signals, rather than the latent space learned via autoencoder.

All ablation experiments show a decrease in AUROC and AUPRC compared with the full adversarial model ("ResNet+ADV"), indicating that each component contributes to enhancing the model's robustness. Meanwhile, these ablated models still outperform the baseline without adversarial training ("ResNet"), further demonstrating the effectiveness and robustness of our proposed adversarial training framework. Moreover, the impact of individual modules on the final performance varies: replacing the on-manifold perturbation module with perturbations applied directly in the ECG signal space produces the largest performance drop, particularly in data-scarce scenarios such as

LVEF ≤ 30%. This further underscores the importance of generating realistic augmentations to improve model robustness.

Table 2. Comparison of training strategies on pacemaker subgroup. Variants share the same training framework but differ in one component: "w/o uncertainty" generates adversarial examples for all samples; "w/o on-manifold" applies perturbations directly to input ECG signals. Values are presented as median (95% confidence intervals).

	LVEF ≤ 50%		LVEF ≤ 40%		LVEF ≤ 30%	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
ResNet	0.58 (0.56-0.61)	0.23 (0.09-0.40)	0.64 (0.62-0.66)	0.22 (0.09-0.41)	0.59 (0.56-0.61)	0.19 (0.07-0.39)
ResNet+ADV	**0.68 (0.66-0.71)**	**0.29 (0.14-0.47)**	**0.73 (0.69-0.74)**	**0.28 (0.13-0.46)**	**0.67 (0.64-0.70)**	**0.26 (0.09-0.46)**
w/o uncertainty	0.65 (0.64-0.68)	0.26 (0.11-0.44)	0.69 (0.67-0.69)	0.23 (0.09-0.44)	0.65 (0.61-0.67)	0.25 (0.06-0.44)
w/o on-manifold	0.65 (0.63-0.69)	0.26 (0.11-0.45)	0.68 (0.66-0.69)	0.24 (0.10-0.46)	0.63 (0.62-0.65)	0.22 (0.06-0.43)

Understanding Robustness via Distributional Alignment

To better understand why adversarial training enhances model robustness under data-scarcity scenarios, we analyzed the distributional discrepancies between the training and testing data. Intuitively, a lower discrepancy between training and testing distributions facilitates better generalization of the model. To this end, we measured the discrepancy between the testing data and three training scenarios: the original training set ("Org" in Table 3), the adversarially perturbed training set ("Adv"), and a mixed dataset combining both original and adversarial samples ("Combined"). We employed two categories of discrepancy measures to quantify both global and local relationships between datasets in the latent space. For global similarity, we use "Center (pos)" and "Center (neg)", which measure the distance between the centroids of positive and negative class samples, respectively, as well as the Maximum Mean Discrepancy (MMD[22]). MMD evaluates the difference between two distributions by comparing the mean embeddings of samples in a reproducing kernel Hilbert space, and has been widely used for distribution alignment. For local discrepancy, we exploit the Jensen–Shannon Divergence (JSD[23]) and Kullback–Leibler Divergence (KLD[24]). JSD symmetrizes and smooths KL divergence to quantify the overlap between two probability distributions, while KLD measures the relative entropy, i.e., how one distribution diverges from another.

Table 3. Distributional discrepancies between testing cohorts and three training cohorts: original training set ("Org"), adversarially perturbed training set ("Adv"), and a mixed dataset combining both original and adversarial samples ("Combined").

	Global Discrepancy (↓)			Local Discrepancy (↓)	
	Center (pos)	Center (neg)	MMD	JSD	KLD
Org	0.6257	0.5623	0.0038	0.3950	8.0197
Adv	**0.5721**	**0.4964**	**0.0034**	0.4020	8.1413
Combined	0.5934	0.5293	0.0035	**0.3290**	**6.1033**

As shown in Table 3, adversarial training markedly reduces the global distributional gap between the training and test sets, as indicated by lower center distances and MMD values. However, it simultaneously introduces small bias at the local distribution level, reflected by marginally higher JSD and KLD. In contrast, combining the original and adversarial data balances these effects: the mixture substantially improves local similarity while maintaining global alignment. This hybrid strategy leads to richer and more diverse representations, resulting in better overall alignment with the test distribution. These findings suggest that while adversarial training is effective in capturing global structures, integrating it with original data is crucial to alleviating local distributional bias and achieving more robust generalization.

Discussion

As our proposed method represents an effective training framework that is not limited to specific architectures or tasks, it can be readily applied to other healthcare scenarios. For example, it could be extended to Echocardiography

(Echo) for assessing cardiac function, or to more complex, multi-modal settings that integrate ECG, Echo, and other cardiac-related data.

In future work, we plan to extend this approach to a broader range of clinical tasks to further evaluate its generalization capability, assess its clinical utility, and examine its impact on clinical decision-making and workflow integration. Our goal is to establish it as a generalizable and convenient tool that can be adopted by different institutions, enabling the deployment of site-specific, effective AI models and promoting the democratization of AI in healthcare.

Conclusion

In this work, we propose a robust training framework to improve robustness in low-resource settings. Our approach combines an on-manifold adversarial perturbation strategy for pediatric ECGs with an uncertainty-aware adversarial training algorithm, which identifies borderline samples near the classification boundary and augments them to enhance model robustness. Evaluation on real-world datasets demonstrates reliable and cost-effective detection of left ventricular systolic dysfunction, highlighting its potential for deployment in resource-limited clinical environments.

Acknowledgements Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM012973 and R01LM014300. This work was partially supported by the Kostin Innovation Fund at Boston Children's Hospital. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Van Der Linde, D. *et al.* Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *Journal of the American College of Cardiology* 58, 2241–2247 (2011).
2. Saarel, E. V. *et al.* Electrocardiograms in healthy North American children in the digital age. *Circulation: Arrhythmia and Electrophysiology* 11, e005808 (2018).
3. Attia, Z. I. *et al.* Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine* 25, 70–74 (2019).
4. Adedinsewo, D. *et al.* Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. *Circulation: Arrhythmia and Electrophysiology* 13, e008437 (2020).
5. Siontis, K. C. *et al.* Detection of hypertrophic cardiomyopathy by an artificial intelligence electrocardiogram in children and adolescents. *International Journal of Cardiology* 340, 42–47 (2021).
6. Mayourian, J. *et al.* Deep learning-based electrocardiogram analysis predicts biventricular dysfunction and dilation in congenital heart disease. *Journal of the American College of Cardiology* 84, 815–828 (2024).
7. Mayourian, J. *et al.* Electrocardiogram-based deep learning to predict left ventricular systolic dysfunction in paediatric and adult congenital heart disease in the USA: a multicentre modelling study. *The Lancet Digital Health* 7, e264–e274 (2025).
8. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. Preprint at https://doi.org/10.48550/arXiv.1412.6572 (2015).
9. He, K. *et al.* Masked autoencoders are scalable vision learners. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 16000–16009 (2022).
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (2018).
11. Han, X. *et al.* Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine* 26, 360–363 (2020).
12. Mayourian, J. *et al.* Pediatric ECG-based deep learning to predict left ventricular dysfunction and remodeling. *Circulation* 149, 917–931 (2024).
13. O'Dell, W. G. Accuracy of Left Ventricular Cavity Volume and Ejection Fraction for Conventional Estimation Methods and 3D Surface Fitting. *Journal of the American Heart Association* 8, e009124 (2019).
14. Elias, P. & Finer, J. EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs. PhysioNet https://doi.org/10.13026/3YKD-BF14.
15. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, USA, 2016). doi:10.1109/CVPR.2016.90.

16. Dhingra, L. S. *et al.* Artificial Intelligence–Enabled Prediction of Heart Failure Risk From Single-Lead Electrocardiograms. *JAMA cardiology* 10, 574–584 (2025).
17. Khunte, A. *et al.* Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices. *npj Digital Medicine* 6, 124 (2023).
18. Friesen, G. M. *et al.* A comparison of the noise sensitivity of nine QRS detection algorithms. *IEEE Transactions on biomedical engineering* 37, 85–98 (1990).
19. Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* 27, 861–874 (2006).
20. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, e0118432 (2015).
21. Huang, J., Chen, B., Yao, B. & He, W. ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. *IEEE access* 7, 92871–92880 (2019).
22. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *The journal of machine learning research* 13, 723–773 (2012).
23. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 145–151 (2002).
24. Kullback, S. & Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* 22, 79–86 (1951).