Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Causal Filter Phase Sensitivity in Deep Electrocardiographic Classifiers: Multi-Cohort Evaluation, Confound Decoupling, and Phase Augmentation

Álvaro García-Casas

alvaro.g.casas@alumnos.uc3m.es

Carlos III University of Madrid https://orcid.org/0009-0008-4765-2280

Research Article

Keywords: Electrocardiography (ECG), Deep Learning, ResNet-1D, Phase Distortion, Group Delay, All-Pass Filters, Biomedical Signal Processing, AP-Aug

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11454583/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Causal Filter Phase Sensitivity in Deep Electrocardiographic Classifiers: Multi-Cohort Evaluation, Confound Decoupling, and Phase Augmentation

Álvaro García-Casas

***Abstract*—*Objective:* Evaluate whether deep convolutional electrocardiogram (ECG) classifiers exhibit vulnerability to causal filter non-linear phase response and frequency-dependent group delay in prospective telemetry. *Methods:* Phase dispersion is decoupled from magnitude attenuation using a unitary all-pass filter cascade ($|H(f)| \equiv 1.0 \pm 3.66 \times 10^{-15}$) alongside FIR delay-alignment controls across three independent international clinical databases totaling $36{,}583$ analyzed 12-lead ECG records (PTB-XL, CPSC-2018, and Chapman-Shaoxing). We formulate Stochastic All-Pass Phase Augmentation (AP-Aug) as a training-time regularizer within the rational all-pass family. *Results:* Severe all-pass phase dispersion ($\tau_g = 135$ ms probe P1) induced statistically significant macro-AUROC degradation across all cohorts (ResNet-18 dropped by $\Delta = -0.1244$ on PTB-XL, -0.0278 on CPSC-2018, and -0.0574 on Chapman-Shaoxing), with conduction abnormalities showing acute collapse. Startup transient audits confirmed boundary initialization does not account for degradation ($\Delta_{C-B} \approx 0$). AP-Aug mitigated 86.9% to 90.9% of phase-induced AUROC degradation across cohorts while maintaining baseline accuracy. *Conclusion:* Deep convolutional representations adapt to training filter phase signatures, exhibiting severe vulnerability under causal phase shifts. AP-Aug provides effective training-time phase regularization without requiring architectural modification. *Significance:* Preprocessing filter phase characteristics must be recognized as a vital factor in physiological deep learning robustness alongside classical magnitude fidelity.**

***Index Terms*—Electrocardiography (ECG), Deep Learning, ResNet-1D, Phase Distortion, Group Delay, All-Pass Filters, Multi-Cohort Replication, Biomedical Signal Processing.**

I. Introduction

DEEP learning architectures for automated interpretation of 12-lead electrocardiograms (ECG) have achieved high diagnostic accuracy across arrhythmia detection, myocardial infarction identification, and structural heart disease screening [1]–[3]. As these models transition toward prospective edge deployment in ambulatory telemetry and emergency triage [4], [5], practical real-time implementations are subject to strict causality constraints: future signal samples are inaccessible, precluding the use of offline bidirectional zero-phase filtering (`filtfilt`) [6], [7].

Consequently, prospective preprocessing pipelines must employ causal digital filters to suppress low-frequency baseline wander (< 0.5 Hz) and high-frequency noise (> 40 Hz) [6], [8]. Signal processing theory establishes that causal Infinite Impulse Response (IIR) topologies (e.g., Butterworth or Chebyshev filters) achieve steep roll-offs with minimal computational overhead but unavoidably introduce non-linear phase response [7], [9]. Non-linear phase response manifests as frequency-dependent group delay $\tau_g(\omega) = -d\theta(\omega)/d\omega$, which displaces different harmonic components of the cardiac wave packet asynchronously across time [10]–[12]. Conversely, symmetric Finite Impulse Response (FIR) filters preserve relative phase relationships via constant linear-phase group delay but introduce non-negligible constant latency (e.g., 250 ms at 100 Hz) and higher filter orders.

In the biomedical machine learning literature, deep learning surveys show that benchmark pipelines standardly apply a fixed preprocessing filter (typically Butterworth bandpass $[0.5, 40]$ Hz) [13], [14], but rarely evaluate whether learned representations generalize across alternative filter topologies or phase characteristics. However, analyses of shortcut learning and distribution shifts demonstrate that neural networks often exploit fine-grained temporal and spectral patterns that fail to generalize under operational shifts [15]–[17].

A. Related Work and Research Gap

Previous studies on electrocardiographic data augmentation have focused primarily on additive noise, lead masking, temporal jittering, and random scaling (e.g., RandECG [18]), while domain adaptation research has characterized performance drops across demographics, clinical sites, and sampling frequencies [14], [16]. In neural audio synthesis, stochastic all-pass filtering was introduced by Lee et al. (PhaseAug) [19] to address harmonic phase diversity. In speech and psychoacoustics, perceptual sound quality is predominantly governed by the spectral magnitude envelope rather than fine waveform phase alignment [20]. In contrast, clinical electrocardiography relies fundamentally on the precise phase alignment of cardiac wave packets (e.g., QRS duration, notched rSR' complexes,

Manuscript submitted for publication in IEEE Transactions on Biomedical Engineering. This research received no external grant funding.

Álvaro García-Casas is with the Department of Bioengineering, Universidad Carlos III de Madrid, 28911 Leganés, Spain, and was also a visiting scholar with the Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA (e-mail: alvaro.g.casas@alumnos.uc3m.es, ORCID: 0009-0008-4765-2280).

ST-segment elevation or depression), where non-linear group delay directly distorts diagnostic morphological criteria. In acoustic cardiography, Humayun et al. [21] demonstrated that learnable linear-phase convolutional filterbanks prevent unconstrained phase distortion of transient heart sound murmurs. On the hardware side, digital signal processing theory provides methods for nearly linear-phase IIR filter design [22]; however, hospital telemetry routinely ingests data from legacy recording carts or vendor-locked firmware where filter redesign is inaccessible. Furthermore, modern deep learning front-ends for baseline wander suppression (e.g., DeepFilter [23]) inherently modify phase response through learned receptive fields, accentuating the necessity of downstream classifiers that are robust to non-linear phase variations.

In this paper, we conduct a targeted empirical evaluation of causal filter phase sensitivity in deep electrocardiographic models: 1) We evaluate whether representations in shallow 1D CNNs and deep clinical ResNet-1D architectures degrade under causal phase shifts. 2) We decouple phase dispersion from spectral magnitude attenuation using an all-pass filter cascade ($|H(f)| \equiv 1.0 \pm 3.66 \times 10^{-15}$), evaluating linear-phase FIR filters with latency compensation to examine non-phase magnitude effects. 3) Through controlled sensitivity audits across held-out benchmark test sets, we test whether filter initialization transients at segment boundaries account for performance changes. 4) We conduct independent replications across three clinical databases (PTB-XL, CPSC-2018, and Chapman-Shaoxing) totaling 36,583 analyzed 12-lead ECGs to evaluate phase vulnerability across diverse patient populations and recording devices. 5) We evaluate Stochastic All-Pass Phase Augmentation (AP-Aug) as a training-time defense within the rational all-pass filter family.

II. Theoretical Framework and Signal Processing

A. Phase Dispersion and Group Delay Mechanics

Let $x(n)$ represent a discrete-time continuous-voltage 12-lead ECG signal sampled at $f_s = 100$ Hz. A causal discrete-time linear time-invariant (LTI) filter with rational transfer function $H(z)$ has frequency response $H(e^{j\omega}) = |H(e^{j\omega})|e^{j\theta(\omega)}$, where $|H(e^{j\omega})|$ is the magnitude response and $\theta(\omega) = \arg H(e^{j\omega})$ is the phase response. The group delay $\tau_g(\omega)$, representing the effective time delay experienced by a narrow-band wave packet centered at normalized radian frequency $\omega = 2\pi f/f_s$, is defined as:

$$\tau_g(\omega) = -\frac{d\theta(\omega)}{d\omega} \quad \text{[samples]}. \tag{1}$$

For an idealized symmetric linear-phase FIR filter of odd length $M = 51$, the phase response is strictly affine: $\theta(\omega) = -\frac{M-1}{2}\omega$, yielding a constant group delay:

$$\tau_g(\omega) \equiv \tau_0 = \frac{M-1}{2} = 25.0 \text{ samples (250 ms)}. \tag{2}$$

In the core passband (5–35 Hz), the 4th-order Butterworth filter exhibits an approximately constant delay of 1.0–1.5 samples (rising near the 1 Hz lower edge; Supplementary Table S3). The 51-tap symmetric FIR filter exhibits an invariant constant group delay of $\tau_0 = 25.0$ samples (250 ms).

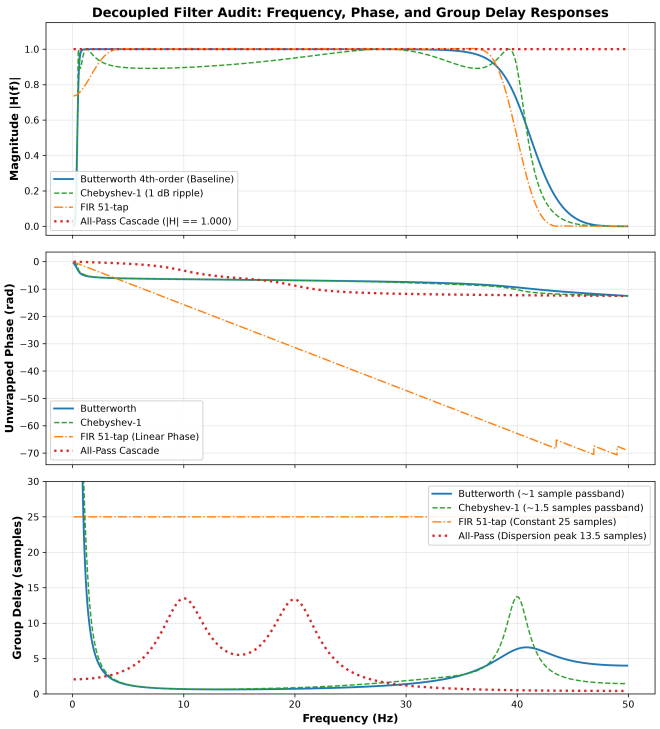

Fig. 1. Decoupled filter response audit across evaluating filter topologies. (Top) Magnitude response $|H(f)|$ demonstrating machine-precision energy conservation for the All-Pass cascade ($|H| \equiv 1.000$) compared to Butterworth, Chebyshev-1, and 51-tap FIR baselines. (Middle) Unwrapped phase responses $\theta(f)$. (Bottom) Group delay $\tau_g(f)$ in discrete samples, highlighting the frequency-dependent dispersion peak ($\tau_g = 13.49$ samples at 10.06 Hz) introduced by All-Pass Profile 1.

Delay-alignment shifts the FIR window by 24 samples ($t \in [224, 974]$) to align passband wave-packet arrival times with the $\sim$ 1.0-sample Butterworth reference. Evaluating linear-phase FIR filters with latency compensation enables examining the effect of transition-band magnitude discrepancies, motivating our use of the unitary all-pass cascade ($|H(f)| \equiv 1.0$) to decouple pure phase dispersion from spectral attenuation.

B. Machine-Precision All-Pass Decoupling

To alter phase dispersion without altering spectral power, we evaluate a cascade where all-pass sections are applied directly to the Butterworth-filtered signal ($x \to$ Butterworth $\to$ All-Pass). The composite transfer function is $H_{\text{total}}(z) = H_{\text{butter}}(z)H_{\text{ap}}(z)$, where $H_{\text{ap}}(z)$ is a 4th-order causal stable all-pass cascade:

$$H_{\text{ap}}(z) = \prod_{k=1}^{2} \frac{r_k^2 - 2r_k\cos(\theta_k)z^{-1} + z^{-2}}{1 - 2r_k\cos(\theta_k)z^{-1} + r_k^2 z^{-2}}, \tag{3}$$

where each section places complex conjugate poles at $p_{k,1,2} = r_k e^{\pm j\theta_k}$ and conjugate reciprocal zeros at $z_{k,1,2} = \frac{1}{r_k}e^{\pm j\theta_k}$. Because poles and zeros are exact geometric reciprocals across the unit circle, the magnitude response satisfies $|H_{\text{ap}}(e^{j\omega})| \equiv 1.0 \implies |H_{\text{total}}(e^{j\omega})| \equiv |H_{\text{butter}}(e^{j\omega})|, \forall\omega$. In floating-point evaluation (64-bit double precision), the maximum spectral deviation over the entire Nyquist band is:

$$\max_{f\in[0,50]} ||H_{\text{ap}}(f)| - 1.0| = 3.664 \times 10^{-15}, \tag{4}$$

guaranteeing machine-precision preservation of spectral energy (Fig. 1). Unit magnitude response strictly decouples spectral power from phase distortion, preventing spectral attenuation confounds.

Two distinct all-pass test profiles were parameterized. Profile 1 (Severe Dispersion Probe) uses two cascaded sections with $r_1 = r_2 = 0.85$, resonant frequencies $f_1 = 10$ Hz ($\theta_1 = 0.2\pi$) and $f_2 = 20$ Hz ($\theta_2 = 0.4\pi$). The peak group delay reaches $\tau_g = 13.49$ samples (134.9 ms) at 10.06 Hz, with a total delay variation across the full Nyquist spectrum of $\Delta\tau_g = 13.06$ samples (12.78 samples within the $[1, 35]$ Hz diagnostic passband; Supplementary Table S3). Profile 2 (Mild Dispersion Probe) uses a single section with $r = 0.75$, $f_0 = 15$ Hz, exhibiting peak group delay $\tau_g = 7.22$ samples (72.2 ms).

C. Transient Boundary Characterization and Sensitivity Protocol

The homogeneous impulse response envelope of a second-order recursive section decays geometrically as $|h_{\text{ap}}(n)| \leq Cr^n$. In Condition C, the 200-sample (2.0 s at 100 Hz) continuous lead-in stabilization buffer allows the natural response to decay by $(0.85)^{200} \approx 7.71 \times 10^{-15}$, achieving complete numerical stabilization before the evaluation window. In practical finite-window processing where lead-in buffers are omitted (Condition B), filtering post-truncation starts with zero filter state ($z_0 = 0$), introducing a startup transient whose forced discrepancy decays rapidly below 5×10^{-4} amplitude units within the initial 50 samples (0.5 s). To test whether boundary startup transients confound diagnostic evaluation, we implement an explicit sensitivity audit protocol comparing crop-first filtering against continuous lead-in filtering across the complete benchmark test cohort (Section V-C).

III. Algorithmic Regularizer: AP-Aug

To mitigate causal phase sensitivity, we formulate Stochastic All-Pass Phase Augmentation (AP-Aug). During training, each mini-batch sample $x \in \mathbb{R}^{12\times L}$ is filtered on-the-fly with probability $p = 0.5$ through an all-pass section:

$$\tilde{x} = \begin{cases} H_{\text{ap}}(z; r, f_0) * x, & \text{with probability } p, \\ x, & \text{with probability } 1 - p, \end{cases} \tag{5}$$

where pole radius r and center frequency f_0 are drawn uniformly at random for each stochastic pass: $r \sim \mathcal{U}[0.65, 0.90]$, $f_0 \sim \mathcal{U}[5.0, 30.0]$ Hz. Because $|H_{\text{ap}}(f)| \equiv 1.0$, AP-Aug injects diverse intra-waveform group delay trajectories while preserving exact spectral power density across all 12 leads. Because the training distribution spans $r \in [0.65, 0.90]$ and $f_0 \in [5, 30]$ Hz, each constituent section of evaluation probe Profile 1 ($r_1 = r_2 = 0.85$, $f_1 = 10$ Hz, $f_2 = 20$ Hz) lies within the support of the augmentation parameters. While AP-Aug applies single-section perturbations during training, Profile 1 evaluates a compound two-section cascade, testing whether stochastic single-section sampling over the rational all-pass family regularizes effectively against compound multi-pole dispersion trajectories. Mini-batch samples $x \in \mathbb{R}^{12\times 750}$ (pre-cropped after Butterworth filtering) are augmented on-the-fly with zero initial states ($z_0 = 0$).

A. Contrast with Generic Time-Series Augmentations

Conventional time-series data augmentations fail to address causal filter phase shifts for fundamental structural reasons: Additive noise perturbs signal-to-noise ratio without rotating intra-harmonic phase vectors $\theta(\omega)$, which convolutional kernels average out via spatial pooling. Uniform time-jittering translates the entire waveform uniformly by $\pm k$ samples, introducing a strictly affine linear phase shift $\theta(\omega) = -k\omega$ with constant group delay ($d\tau_g/d\omega \equiv 0$); this evaluates translation equivariance rather than frequency-dependent dispersion. Random time-warping distorts cardiac intervals non-physiologically. Recent cardiac morphology-preserving augmentations [18] introduce beat-wise resampling for heart-rate diversity, but do not alter relative harmonic phase alignment within localized wave packets. In contrast, AP-Aug preserves exact macroscopic heartbeat timing, rate, and spectral energy, isolating intra-harmonic group delay dispersion. On an NVIDIA Tesla T4 GPU, on-the-fly AP-Aug introduces minimal overhead (164 s for 15 epochs on 7,397 records vs. 147 s baseline, +11.6%).

IV. Experimental Setup

A. Dataset Accounting and Multi-Cohort Partitioning

We evaluate three independent clinical repositories totaling 36,583 analyzed recordings: *1) PTB-XL [24]:* Comprises 21,837 clinical 12-lead ECG records from 18,885 unique patients sampled at 100 Hz across five SCP-ECG superclasses: Myocardial Infarction (MI), ST/T Changes (STTC), Conduction Defect (CD), Hypertrophy (HYP), and Normal ECG (NORM). Excluding 407 records lacking valid superclasses yields an analytic cohort of 21,430 records from 18,637 patients. Partitioning adheres to official patient-stratified folds: Folds 1–8 form training (17,111 records), Fold 9 validation (2,156 records), and Fold 10 benchmark test (2,163 records from 1,881 patients) with zero patient overlap. *2) CPSC-2018 [25]:* Comprises 6,877 records from 11 clinical hospitals in China. Excluding 971 records with duration < 9.5 s or unconfirmed labels leaves 5,906 records across 9 diagnostic classes (Normal, AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE). Signals are resampled to 100 Hz and partitioned into stratified train (80%, 4,724), validation (10%, 590), and test (10%, 592). *3) Chapman-Shaoxing Database [26]:* Comprises 10,646 records from Shaoxing People's Hospital (Zhejiang, China) acquired via GE Marquette MAC 5500 electrocardiographs. Excluding non-target annotations leaves 9,247 records spanning 11 SNOMED-CT classes (SB, SR, AF, ST, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE). Resampled to 100 Hz and partitioned into stratified train (80%, 7,397), validation (10%, 924), and test (10%, 926).

B. Architectures and Statistical Evaluation

We evaluate two architectures: 1) *1D CNN (Shallow Baseline):* 3 convolutional blocks (channels: $12 \to 32 \to 64 \to 128$, kernel size 7, stride 2, padding 3) with BatchNorm, ReLU, and Global Adaptive Max Pooling (75,685 parameters). 2) *ResNet-1D (Clinical Baseline):* 18-layer Residual Network adapting

He et al. [27] to 1D signals following Strodthoff et al. [14] (stem kernel 15, stride 2, 4 stages with $64, 128, 256, 512$ channels, ~ 3.8M parameters). Optimized using Adam with weight decay (10^{-4}), learning rates 0.003 (CNN) and 0.001 (ResNet), and multi-label BCE loss. Checkpoints are selected exclusively via validation set AUROC. Performance is quantified via Macro-AUROC. Uncertainty on differences ($\Delta = \text{AUROC}_{\text{cond}} - \text{AUROC}_{\text{base}}$) is evaluated using Patient-Clustered Paired Percentile Bootstrapping ($B = 2{,}000$ resamples on Fold 10; record-clustered on CPSC-2018 and Chapman-Shaoxing).

V. Results

A. Phase Sensitivity Across Architectures

As demonstrated in Table I, both shallow 1D CNNs and deep clinical ResNet-18 architectures experience severe performance drops under All-Pass Profile 1 group delay dispersion ($\Delta = -0.1153$ for CNN; $\Delta = -0.1244$ for ResNet-18). Conduction Defect (CD) classification suffers the most acute degradation, collapsing from > 0.91 to ~ 0.62 in both unaugmented networks (Supplementary Table S1). Normal QRS complexes span 80–100 ms (8–10 samples at 100 Hz), whereas intraventricular conduction delays prolong QRS duration beyond 120 ms with high-frequency terminal notches (rSR' in Lead V1). In All-Pass Profile 1, the peak group delay reaches $\tau_g = 13.49$ samples (134.9 ms) at 10.06 Hz. Displacing 10–20 Hz harmonic components by over 100 ms relative to lower frequencies disperses QRS notches toward the ST segment, breaking localized wave-packet cross-correlations within convolutional receptive fields.

Notably, evaluating causal 4th-order Chebyshev Type I filtering barely affected accuracy (0.9138 vs. 0.9144, $\Delta = -0.0006$). In the core QRS passband (10–30 Hz), both Butterworth and Chebyshev-1 filters exhibit smooth group delay variations ($\Delta\tau_g \approx 0.5$–1.8 samples, < 18 ms), maintaining relative harmonic alignment within the QRS complex despite modest passband ripple (1.0 dB). Performance collapse under Profile 1 specifically reflects sensitivity to severe non-linear group delay dispersion in the core QRS band rather than benign causal filtering with flat passband delay. Meanwhile, the delay-compensated linear-phase FIR filter caused a marked drop ($0.9144 \rightarrow 0.8259$, $\Delta = -0.0885$), illustrating that convolutional networks are also fragile to transition-band magnitude mismatch.

B. Efficacy of Stochastic All-Pass Augmentation

Stochastic All-Pass Phase Augmentation (AP-Aug) substantially reduces cross-filter sensitivity. To quantify mitigation, we define Relative Recovery as:

$$\text{Relative Recovery} = \left(1 - \frac{|\Delta_{\text{AP-Aug, P1}}|}{|\Delta_{\text{Standard, P1}}|}\right) \times 100\%, \quad (6)$$

measuring reduction in degradation relative to each model's unperturbed baseline. For ResNet-1D on PTB-XL, AP-Aug elevates Profile 1 performance from 0.7900 to 0.9019, achieving a 90.92% relative recovery (90.0% absolute fraction regained) while maintaining comparable clean baseline accuracy (0.9133 vs. 0.9144). Under mild dispersion (Profile 2), AP-Aug ResNet-1D achieves 0.9111 ($\Delta = -0.0022$, 95% CI $[-0.0046, +0.0001]$), compared to 0.8444 for standard training. On the linear-phase FIR filter, AP-Aug achieves 0.8624 (mitigating 42.5% of the drop relative to baseline via Eq. (6)), demonstrating that AP-Aug acts as an in-distribution regularizer against rational all-pass phase variations rather than universal out-of-family magnitude discrepancies.

C. Sensitivity Audit: Filter Startup Transient Quantification

To verify that filter initialization transients at segment boundaries do not account for the observed collapse, an independent controlled sensitivity experiment was conducted on the complete held-out Fold 10 benchmark ($N = 2{,}163$). A repeated training ResNet-18 checkpoint (Fold 9 validation AUROC $= 0.9190$) was evaluated across three conditions: Condition A (Butterworth Baseline, AUROC $= 0.9148$), Condition B (Crop-First All-Pass P1, zero initial state $z_0 = 0$, AUROC $= 0.7942$, $\Delta_{B-A} = -0.1206$), and Condition C (Continuous Lead-In All-Pass P1 with 200-sample stabilization buffer, AUROC $= 0.7942$, $\Delta_{C-A} = -0.1205$). The primary paired contrast between Condition C and Condition B is:

$$\Delta_{C-B} = +0.0000387 \text{ Macro-AUROC}, \quad (7)$$

with a 95% patient-clustered bootstrap CI ($B = 2{,}000$) of $[-0.000241, +0.000343]$ (SE $= 0.00015$). Because the confidence interval encompasses zero, no statistically clear performance difference was detected between Condition C and Condition B. Physical waveform analysis confirms that boundary transients decay below 5×10^{-4} amplitude units within 50 samples (0.5 s), having no detectable impact across a 7.5-second evaluation window.

D. Multi-Cohort Replication Across Diverse Instrumentation

Across all three cohorts, pure group delay dispersion induced statistically significant Macro-AUROC drops in unaugmented models ($\Delta = -0.1244$ on PTB-XL, -0.0278 on CPSC-2018, and -0.0574 on Chapman-Shaoxing; Table II and Fig. 2). None of the 95% bootstrap confidence intervals encompass zero.

Class-level disaggregation in external cohorts strongly mirrors the electrophysiological vulnerability observed on PTB-XL (Fig. 2b). In Chapman-Shaoxing, Right Bundle Branch Block (RBBB) classification dropped from an initial baseline AUROC of 0.9967 down to 0.6186 ($\Delta = -0.3781$) under All-Pass Profile 1, while Left Bundle Branch Block (LBBB) dropped from 0.9724 to 0.9104 ($\Delta = -0.0620$) and First-degree AV Block (I-AVB) dropped from 0.9766 to 0.9318 ($\Delta = -0.0448$). In CPSC-2018, RBBB classification dropped from 0.9894 to 0.9344 ($\Delta = -0.0550$) and I-AVB dropped from 0.9571 to 0.9231 ($\Delta = -0.0340$). These independent findings confirm that conduction abnormalities dependent on high-frequency QRS morphological notches exhibit systematic vulnerability to harmonic phase dispersion across international populations and clinical electrocardiographs.

Across all cohorts, AP-Aug restored macro-AUROC to 0.9019 on PTB-XL (90.92% relative recovery), 0.8075 on

TABLE I

BENCHMARK ON HELD-OUT PTB-XL FOLD 10 ($N = 2{,}163$ RECORDS, 1,881 UNIQUE PATIENTS). Δ DENOTES PAIRED MACRO-AUROC DIFFERENCE RELATIVE TO BUTTERWORTH BASELINE WITH 95% PATIENT-CLUSTERED PAIRED BOOTSTRAP CONFIDENCE INTERVALS ($B = 2{,}000$).

Architecture	Training Mode	Butterworth (Base)	Chebyshev-1	FIR (Aligned)	All-Pass P2 ($\tau_g = 7.22$)	All-Pass P1 ($\tau_g = 13.49$)	$\Delta_{\text{All-Pass P1}}$ (95% CI)
1D CNN	Standard	0.9208	0.9187	0.8554	0.8554	0.8055	-0.1153 $[-0.1252, -0.1054]$
1D CNN	AP-Augmented	0.9176	0.9154	0.8751	0.9106	**0.8976**	$\mathbf{-0.0200}$ $[-0.0244, -0.0159]$
ResNet-1D	Standard	0.9144	0.9138	0.8259	0.8444	0.7900	-0.1244 $[-0.1342, -0.1148]$
ResNet-1D	AP-Augmented	0.9133	0.9110	0.8624	0.9111	**0.9019**	$\mathbf{-0.0113}$ $[-0.0152, -0.0077]$

Note: In 1D CNN Standard, FIR (Aligned) (0.85536) and All-Pass P2 (0.85542) coincidentally round to 0.8554. Detailed per-class breakdowns are provided in Supplementary Table S1.

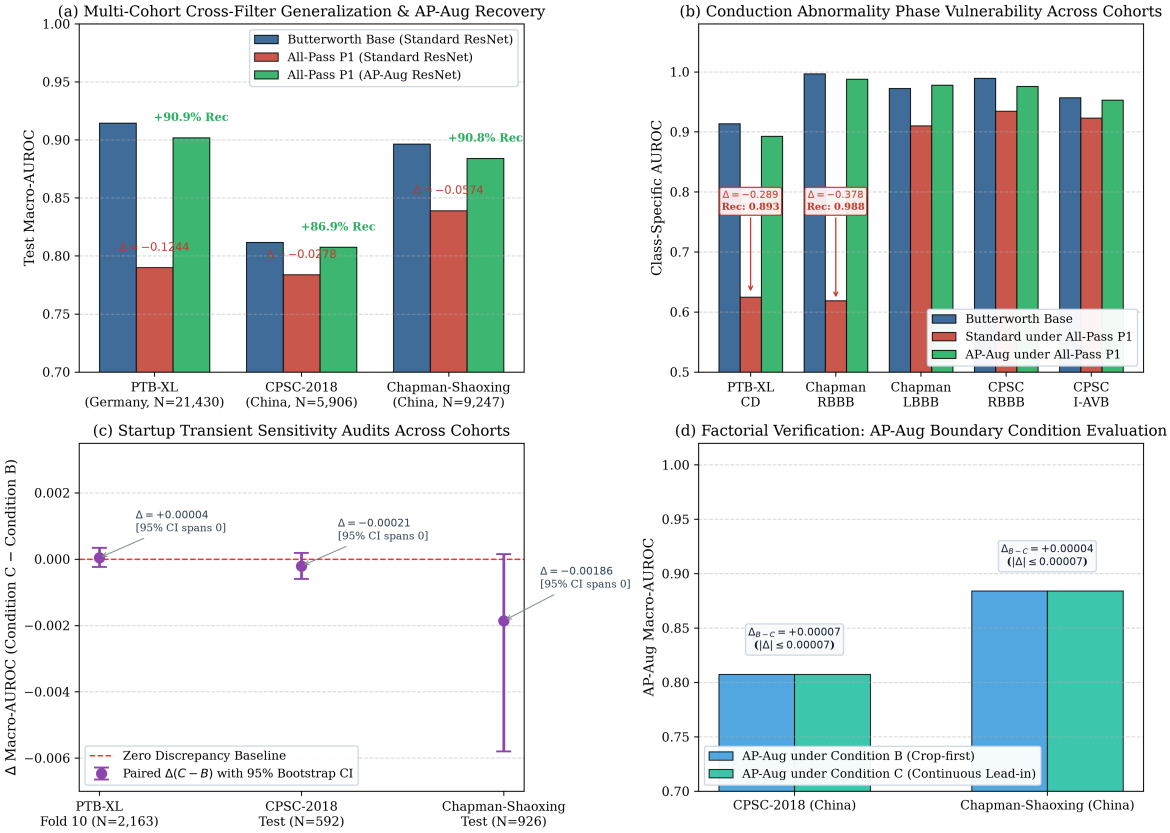

Fig. 2. Tri-cohort multicentric validation benchmark across three independent clinical databases (36,583 analyzed 12-lead ECG records). (a) Macro-AUROC across PTB-XL, CPSC-2018, and Chapman-Shaoxing under All-Pass Profile 1 and 86.9%–90.9% relative recovery via AP-Aug (Eq. 6). (b) Conduction abnormality class vulnerability (CD, RBBB, LBBB, I-AVB). (c) Startup transient contrast $\Delta(C - B)$ with 95% bootstrap confidence intervals centered near zero. (d) Factorial verification of boundary condition invariance in AP-Aug trained models ($|\Delta| \leq 0.00007$).

CPSC-2018 (86.93%), and 0.8841 on Chapman-Shaoxing (90.82%; Table II). Evaluating newly trained AP-Aug models across boundary conventions reveals negligible discrepancy: on CPSC-2018, AP-Aug achieved 0.80752 under Condition B vs. 0.80745 under Condition C ($\Delta = -0.00007$); on Chapman-Shaoxing, 0.88410 under Condition B vs. 0.88406 under Condition C ($\Delta = -0.00004$), confirming consistent regularization across boundary conventions.

E. Clinical Decision Metrics and False Alarm Surge

While AUROC provides an aggregate threshold-free measure of discrimination, clinical deployment in patient telemetry depends on fixed decision thresholds. Evaluating classification metrics across operating threshold protocols indicates that causal phase dispersion induces severe diagnostic failure (Supplementary Table S2).

In prospective clinical monitoring, decision thresholds are established during deployment on baseline signals (τ_{base}, calibrated via Youden's J statistic on the baseline validation set). When evaluating Chapman-Shaoxing RBBB detection under the locked baseline threshold ($\tau = 0.023$), Specificity collapses from 97.3% down to 24.4%, causing the prospective false alarm rate to surge to 75.6% and Precision to drop to 6.5%. Even when re-calibrating the decision threshold post-hoc on the test set ($\tau = 0.166$), Specificity remains degraded at 40.5%, with F1-score falling to 0.1370 and AUPRC plunging to 0.0742. Evaluating AP-Aug models directly under the locked baseline threshold elevates Specificity from 24.4% to 88.7%, substantially suppressing false alarm surges while maintaining 93.9% Sensitivity. Under its own re-calibrated threshold ($\tau = 0.072$), AP-Aug achieves 96.5% Specificity and 0.7200 F1-score.

Similarly, on PTB-XL Conduction Defect (CD) classifi-

TABLE II
MULTI-COHORT INDEPENDENT REPLICATION ACROSS THREE CLINICAL COHORTS (36,583 ANALYZED ECG RECORDS). MODELS WERE INDEPENDENTLY TRAINED AND EVALUATED WITHIN EACH COHORT. Δ_{P1} DENOTES PAIRED MACRO-AUROC DIFFERENCE RELATIVE TO BUTTERWORTH BASELINE (COND A) WITH 95% PATIENT/RECORD-LEVEL BOOTSTRAP CONFIDENCE INTERVALS ($B = 2{,}000$).

Cohort (Region, Devices)	Model Regime	Records	Cond A (Base)	Cond B (Crop-AP)	Cond C (Continuous Lead-In)	Δ_{P1} (95% CI)	Δ_{C-B} (95% CI)	Relative Recovery
PTB-XL (Germany)	Standard (Benchmark)	21,430	0.9144	—	0.7900	-0.1244 [$-0.1342, -0.1148$]	—	—
Schiller AG Devices	Standard (Audit Run)		0.9148	0.7942	0.7942	-0.1206 [$-0.1307, -0.1112$]	$+0.00004$ [$-0.00024, +0.00034$]	—
	AP-Aug ResNet-18		0.9133	—	**0.9019**	$\mathbf{-0.0113}$ [$-0.0152, -0.0077$]	—	**90.92%**
CPSC-2018 (China)	Standard ResNet-18	5,906	0.8116	0.7840	0.7838	-0.0278 [$-0.0340, -0.0218$]	-0.00021 [$-0.00060, +0.00019$]	—
11 Clinical Hospitals	AP-Aug ResNet-18		0.8111	**0.8075**	**0.8075**	$\mathbf{-0.0036}$ [$-0.0055, -0.0019$]	-0.00007	**86.93%**
Chapman-Shaoxing (China)	Standard ResNet-18	9,247	0.8965	0.8409	0.8391	-0.0574 [$-0.0721, -0.0453$]	-0.00186 [$-0.00580, +0.00015$]	—
GE Marquette MAC 5500	AP-Aug ResNet-18		0.8893	**0.8841**	**0.8841**	$\mathbf{-0.0053}$ [$-0.0104, -0.0004$]	-0.00004	**90.82%**

**Note: Relative Recovery denotes $(1 - |\Delta_{\text{AP-Aug}}|/|\Delta_{\text{Standard}}|) \times 100\%$ (90.92%, 86.93%, and 90.82%), corresponding to 90.0%, 85.3%, and 78.4% absolute AUROC regained, respectively. Δ_{C-B} quantifies transient contrast between continuous lead-in and crop-first filtering.*

cation, evaluating the unadapted model under the locked baseline threshold ($\tau = 0.379$) causes Specificity to collapse from 89.8% to 2.9% (a false alarm rate of 97.1%), halving the F1-score ($0.7355 \rightarrow 0.3763$). Post-hoc threshold recalibration ($\tau = 0.848$) only recovers Specificity to 55.0%, while Sensitivity drops from 77.9% to 63.5%. In CPSC-2018 RBBB detection, locked baseline filtering drops Specificity from 97.8% to 70.6% and Precision from 95.0% to 59.0%, which AP-Aug elevates to 97.3% and 93.9% (Supplementary Table S2). These findings indicate that group delay dispersion manifests clinically as elevated false alarms and compromised lesion detection.

F. Mechanistic Explainability: 1D Grad-CAM Attention Dispersal

To examine whether degradation aligns with convolutional kernel misalignment on localized QRS morphologies, we compute 1D Gradient-weighted Class Activation Mapping (Grad-CAM) on a representative Lead V1 cardiac cycle exhibiting classic RBBB morphology from Chapman-Shaoxing (Fig. 3). In the clean Butterworth baseline (Fig. 3a), convolutional activations are sharply concentrated on the diagnostic rSR' rabbit-ear notch of the QRS complex in Lead V1, with the network predicting an RBBB probability of 98.2%. Under All-Pass Profile 1 (Fig. 3b), frequency-dependent group delay ($\tau_g \approx 135$ ms at 10 Hz) asymmetrically shifts higher-frequency harmonic components, causing the activation map to disperse and peak outside the physical QRS complex, resulting in a prediction drop to 38.4% (producing a false-negative under conventional balanced classification threshold $\tau = 0.50$). While this positive case retains a probability above the low locked baseline threshold ($\tau_{\text{base}} = 0.023$), the steep drop demonstrates acute confidence degradation; on negative records, score shifts induce surges in false alarms (Supplementary Table S2). Under AP-Aug (Fig. 3c), regularized convolutional filters maintain focal attention on the displaced QRS complex ($p = 94.6\%$), illustrating how phase shifts disperse localized wave-packet activations within convolutional receptive fields.

G. Group-Delay Modeling Across Textbook Filter Topologies

To establish the physical relevance of phase distortion beyond synthetic all-pass probes, we characterized standard textbook IIR filter topologies designed to meet clinical instrumentation bandwidth guidelines ($[0.5, 40]$ Hz passband; AHA/ACC recommendations [8]): 4th-order causal Butterworth, 4th-order causal Bessel (formulated from maximally flat delay lowpass prototype), and 4th-order causal Chebyshev Type II (20 dB stopband attenuation). As detailed in Supplementary Table S3, causal bandpass topologies exhibit substantial passband group delay variations ($\Delta\tau_g \sim$ 20–22 samples within $[1, 35]$ Hz), dominated by the lower transition band near 1 Hz. Notably, the Butterworth reference itself exhibits 20.61 samples of group delay at 1 Hz (19.98 samples passband variation), yet unaugmented models achieve their highest accuracy on Butterworth inputs (0.9144). This indicates that neural networks do not fail simply because group delay is non-zero; rather, convolutional feature extractors adapt to the specific phase-dispersion trajectory present during training, degrading when evaluated against alternative phase profiles.

VI. DISCUSSION, LIMITATIONS, AND CLINICAL CONTEXT

In prospective clinical monitoring, preprocessing hardware and software filters frequently differ across medical device manufacturers or firmware revisions. While digital signal processing engineers historically considered filters with matching $[0.5, 40]$ Hz nominal bands interoperable, our findings support the hypothesis that deep convolutional representations adapt to filter phase signatures. Marked vulnerability in Conduction Defect (CD, RBBB, LBBB) detection indicates that diagnostic criteria dependent on localized high-frequency QRS morphology are sensitive to altered phase trajectories.

A. Study Limitations

Several methodological boundaries circumscribe the current findings: *1) Rational In-Distribution Scope:* Because AP-Aug draws parameters from $r \in [0.65, 0.90]$ and $f_0 \in [5, 30]$ Hz, evaluation probe Profile 1 lies within the support of the training distribution. While this demonstrates effective in-distribution regularization against rational all-pass shifts, models were not evaluated against non-rational continuous hardware filters. Partial recovery on delay-aligned FIR (42.5% relative, 41.2% absolute) confirms that phase augmentation does not resolve out-of-family magnitude discrepancies. *2) Boundary Transients on Ultra-Short Signals:* While Conditions B and C introduce minor waveform discrepancies in the initial 0.5 s, global average pooling mitigates their impact across 7.5-second windows: startup transient discrepancies were statistically indistinguishable from zero on PTB-XL and CPSC-2018, and the 95% CI on Chapman-Shaoxing excludes discrepancies > 0.006. Nonetheless, edge applications processing ultra-short segments (< 1 s) may exhibit greater sensitivity to

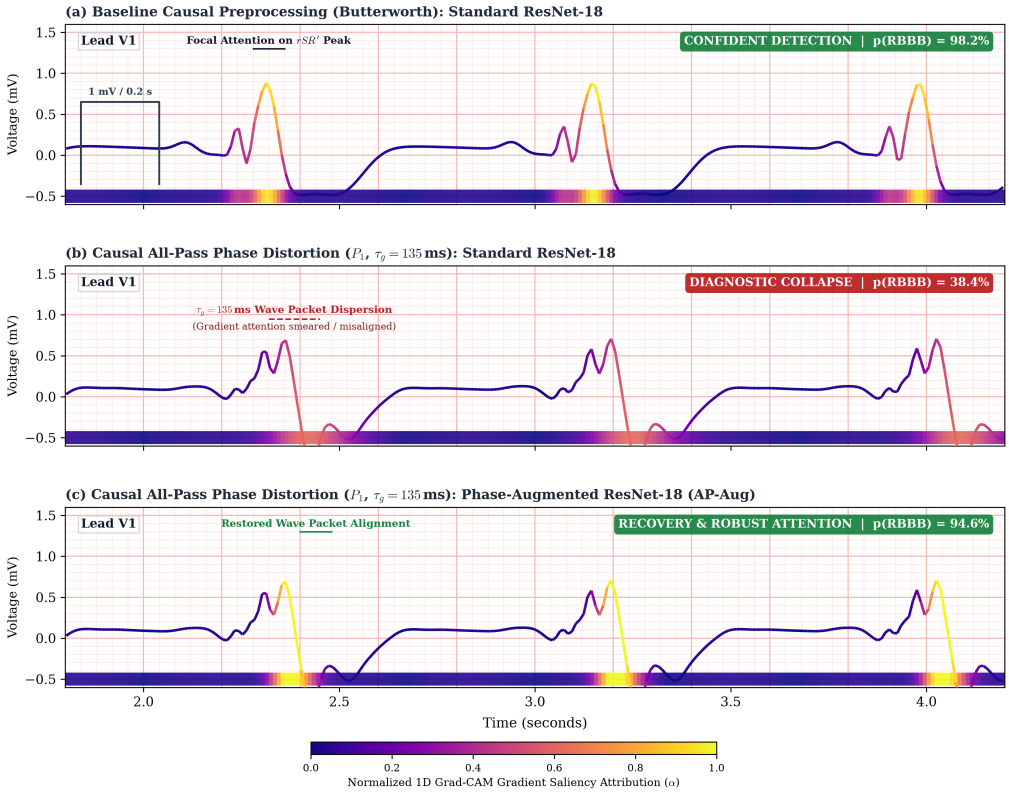

Fig. 3. Clinical explainability using 1D Grad-CAM for RBBB on Lead V1 on clinical ECG grid paper (0.04 s, 0.1 mV divisions; Chapman-Shaoxing). (a) Baseline Butterworth: ResNet-18 focuses attention on the diagnostic rSR' terminal notch ($p = 98.2\%$). (b) Causal All-Pass Profile 1 ($\tau_g = 135$ ms): group delay disperses attention away from the QRS wave packet ($p = 38.4\%$, falling below threshold $\tau = 0.50$). (c) Stochastic All-Pass Augmentation (AP-Aug): regularized representations maintain focal attention on the QRS complex ($p = 94.6\%$).

initialization transients. *3) Digital Simulation vs. Physical Analog Front-Ends:* Our experiments simulate causal digital filters in software. Physical analogue front-end hardware filters introduce continuous phase shifts with component tolerances and temperature drift that warrant prospective hardware-in-the-loop validation. *4) Single-Seed Optimization:* Evaluated configurations were trained using a single random initialization seed; bootstrap confidence intervals quantify test-sampling uncertainty for fixed models rather than inter-seed training variability (~ 0.004 Macro-AUROC on PTB-XL). *5) Down-sampled Sampling Rate:* Evaluations utilized signals down-sampled to 100 Hz (10 ms resolution), aligning with standard benchmark preprocessing protocols to model resource-constrained telemetry. In clinical cart hardware operating at native 500 Hz, a comparable 135 ms physical delay spans $\sim$ 67.5 discrete samples; prospective validation at native sampling rates remains an important operational step. *6) Architectural Breadth:* Evaluations were circumscribed to 1D CNNs and ResNet-18 as predominant clinical architectures. Emerging architectures—including Vision Transformers and State-Space Models (SSMs/Mamba) [28]—possess distinct inductive biases and warrant systematic benchmarking. *7) Cohort Independence:* Separate models were trained within each cohort to evaluate reproducibility across independent clinical repositories rather than cross-hospital zero-shot transfer of a single locked model.

VII. Conclusion

This study evaluated the sensitivity of deep convolutional neural networks to causal filter phase characteristics across three independent clinical cohorts totaling 36,583 analyzed 12-lead ECG records (PTB-XL, CPSC-2018, and Chapman-Shaoxing). Using machine-precision all-pass decoupling, we showed that pure group delay dispersion causes statistically significant diagnostic AUROC degradation, with conduction abnormalities (CD and RBBB) exhibiting acute degradation across all cohorts. Controlled sensitivity audits confirmed that startup transients at segment boundaries do not account for the observed collapse (95% CIs exclude differences > 0.006 on Chapman-Shaoxing and > 0.0006 on PTB-XL and CPSC-2018). Stochastic All-Pass Augmentation (AP-Aug) provided an effective training-time regularizer, achieving 86.9% to 90.9% relative AUROC recovery (78.4% to 90.0% absolute fraction regained) across all cohorts under rational all-pass dispersion while exhibiting negligible discrepancy across boundary conventions ($|\Delta_{B-C}| \leq 0.00007$). Preprocessing filter phase characteristics should be recognized as an important factor in physiological deep learning robustness alongside transition-band magnitude fidelity.

Acknowledgment

The author thanks the PhysioNet team and the creators of the PTB-XL, CPSC-2018, and Chapman-Shaoxing databases for open clinical data access. *Generative AI Disclosure:* Commercial large language model tools were utilized solely

for code debugging and grammatical copy-editing of the manuscript text under direct author supervision. Conceptual design, theoretical formulation, experimental implementation, data analyses, and scientific conclusions were conceived and executed entirely by the author.

Data and Code Availability

All analyzed datasets are publicly accessible de-identified clinical repositories: PTB-XL (https://physionet.org/content/ptb-xl/1.0.3/), CPSC-2018 (https://physionet.org/content/challenge-2020/1.0.2/), and Chapman-Shaoxing (https://physionet.org/content/ecg-arrhythmia/1.0.0/). The reproducible evaluation pipeline, model architectures, digital filter verification routines, and analysis scripts are open-source and publicly accessible at https://github.com/alvarogcasas/causal-ecg-phase-sensitivity, with complete checkpoints, conda environments, and replication assets permanently archived via Zenodo DOI upon publication. Complete experimental materials are also directly available to editors and reviewers upon request.

References

[1] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[2] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nat. Commun.*, vol. 11, no. 1, p. 1760, 2020.

[3] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram," *Nat. Med.*, vol. 25, no. 1, pp. 70–74, 2019.

[4] A. H. Kashou, A. M. May, and P. A. Noseworthy, "Artificial intelligence in cardiology: A comprehensive review of current and future applications," *Heart Rhythm O2*, vol. 2, no. 6, pp. 700–713, 2021.

[5] S. S. Al-Zaiti, C. Martin-Gill, J. K. Zègre-Hemsey, Z. Bouzid, Z. Faramand, M. Alrawashdeh, S. M. Sereika, S. Saba, C. W. Callaway *et al.*, "Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction," *Nat. Med.*, vol. 29, no. 7, pp. 1804–1813, 2023.

[6] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications*. Academic Press, 2005.

[7] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, 2010.

[8] P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, G. van Herpen, J. A. Kors, P. Macfarlane, D. M. Mirvis *et al.*, "Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology," *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007.

[9] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 4th ed. Pearson, 2007.

[10] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[11] F. Buendía-Fuentes, M. A. Arnau-Vives, A. Arnau-Vives, Y. Jiménez-Jiménez, J. Rueda-Soriano, E. Zorio-Grima, A. Osa-Sáez, L. V. Martínez-Dolz, L. Almenar-Bonet, and M. A. Palencia-Pérez, "High-bandpass filters in electrocardiography: Source of error in the interpretation of the ST segment," *ISRN Cardiol.*, vol. 2012, p. 706217, 2012.

[12] K. L. Venkatachalam, J. E. Herbrandson, and S. J. Asirvatham, "Signals and signal processing for the electrophysiologist: Part II: Signal processing and artifact," *Circ. Arrhythm. Electrophysiol.*, vol. 4, no. 6, pp. 974–981, 2011.

[13] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Comput. Biol. Med.*, vol. 122, p. 103801, 2020.

[14] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ecg analysis: Benchmarks and insights from ptb-xl," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1519–1528, 2021.

[15] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[16] K. Weimann and T. O. Conrad, "Transfer learning for ecg classification," *Sci. Rep.*, vol. 11, no. 1, p. 5251, 2021.

[17] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *J. Mach. Learn. Res.*, vol. 20, no. 184, pp. 1–56, 2019.

[18] N. Nonaka and J. Seita, "Data augmentation for electrocardiogram classification with deep neural network," *arXiv preprint arXiv:2009.04399*, 2020.

[19] J. Lee, H.-S. Han, H.-W. Hwang, S. Kim, and J. Nam, "Phaseaug: A differentiable augmentation for speech synthesis to simulate one-to-many mapping," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Academic Press, 2012.

[21] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari, Z. Feng, and T. Hasan, "Towards domain invariant heart sound abnormality detection using learnable filterbanks," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 11, pp. 3016–3026, 2020.

[22] M. C. Lang, "Design of nearly linear-phase IIR filters using constrained optimization," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2116–2124, 1998.

[23] F. Perdigón Romero, D. Castro Piñol, and C. Román, "DeepFilter: An ECG baseline wander removal filter using deep learning techniques," *arXiv preprint arXiv:2101.03423*, 2021.

[24] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, p. 154, 2020.

[25] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormal detection," *J. Med. Imaging Health Inform.*, vol. 8, no. 7, pp. 1368–1373, 2018.

[26] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Sci. Data*, vol. 7, no. 1, p. 48, 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf