

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

Álvaro García-Casas

S1. DETAILED CLASS-LEVEL AUROC BREAKDOWN ON PTB-XL

Table S1 presents the granular breakdown of Area Under the Receiver Operating Characteristic (AUROC) across all five SCP-ECG diagnostic superclasses on the canonical held-out PTB-XL Fold 10 benchmark ($N = 2,163$ records from 1,881 unique patients). Performance is disaggregated for both the shallow 1D CNN baseline and the deep clinical ResNet-18 architecture, under standard training and Stochastic All-Pass Phase Augmentation (AP-Aug) regimes.

TABLE S1
CLASS-SPECIFIC AUROC UNDER ALL-PASS PROFILE 1 ($\tau_g = 13.49$ SAMPLES / 134.9 MS) ON HELD-OUT PTB-XL FOLD 10.

Diagnostic Superclass	1D CNN (Std)	1D CNN (Aug)	ResNet (Std)	ResNet (Aug)
Normal ECG (NORM)	0.8705	0.9213	0.8486	0.9270
Myocardial Infarction (MI)	0.7882	0.8666	0.7745	0.8655
ST/T Changes (STTC)	0.8957	0.9230	0.8629	0.9261
Conduction Defect (CD)	0.6180	0.8833	0.6247	0.8929
Hypertrophy (HYP)	0.8551	0.8938	0.8392	0.8980
Macro-Averaged AUROC	0.8055	0.8976	0.7900	0.9019

**Note: Models evaluated on held-out Fold 10. Conduction Defect (CD) exhibits acute sensitivity to non-linear group delay dispersion ($\Delta \approx -0.29$ in ResNet), which is effectively regularized by AP-Aug (restoring CD AUROC to 0.8929).*

Electrophysiologically, normal ventricular depolarization manifests as a narrow QRS complex (80–100 ms duration), whereas intraventricular conduction defects (such as Bundle Branch Blocks) prolong QRS duration beyond 120 ms with high-frequency morphological notches (e.g., rSR' morphology in Lead V1). Because All-Pass Profile 1 introduces a peak group delay of $\tau_g = 13.49$ samples (134.9 ms) at 10.06 Hz, localized frequency components in the QRS wave packet are shifted asymmetrically across time, smearing the diagnostic notch into the subsequent ST-segment. This localized wave-packet dispersal causes acute diagnostic failure in Conduction Defect detection, whereas broader, low-frequency abnormalities (such as myocardial infarction or ST/T segment changes) exhibit comparatively greater resilience.

Supplementary Material for Manuscript No. TBME-02871-2026 submitted to IEEE Transactions on Biomedical Engineering.

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).

S2. COMPREHENSIVE CLINICAL OPERATING DECISION METRICS ACROSS THRESHOLD PROTOCOLS

While AUROC evaluates aggregate discriminative capability across all conceivable decision boundaries, clinical deployment in patient telemetry, intensive care units, and ambulatory monitors depends on fixed classification thresholds. Table S2 details the operational metrics—Sensitivity, Specificity, Precision, F1-Score, and Area Under the Precision-Recall Curve (AUPRC)—across Chapman-Shaoxing, CPSC-2018, and PTB-XL test cohorts.

A. False Alarm Surge Mechanics

In prospective clinical practice, decision thresholds are calibrated during model validation (τ_{base}) and frozen prior to hospital deployment. When unadapted models encounter phase-distorted signals, output prediction distributions shift toward higher values on negative samples. On Chapman-Shaoxing RBBB classification, Specificity collapses from 97.3% to 24.4% under the locked threshold, precipitating a false alarm surge of 75.6% and reducing Precision to 6.5%. In clinical telemetry, such high false alarm rates degrade trust, induce alarm fatigue among healthcare personnel, and compromise patient safety. Under AP-Aug, Specificity is restored to 88.7% under the locked threshold (96.5% under re-calibrated threshold), suppressing unwarranted clinical alarms while preserving high sensitivity (91.8%–93.9%).

S3. THEORETICAL GROUP DELAY MODELING ACROSS STANDARD TEXTBOOK FILTER TOPOLOGIES

To contextualize all-pass probe profiles within physical biomedical instrumentation, Table S3 characterizes group delay characteristics across standard 4th-order causal IIR topologies meeting AHA/ACC clinical monitoring guidelines ([0.5, 40] Hz): Butterworth, Bessel, and Chebyshev Type II.

A. Analytical Formulation of Group Delay Dispersion

Let $H(z)$ be an N -th order causal discrete-time transfer function:

$$H(z) = \frac{\sum_{k=0}^M b_k z^{-k}}{1 + \sum_{k=1}^N a_k z^{-k}} = K \frac{\prod_{k=1}^M (1 - z_k z^{-1})}{\prod_{k=1}^N (1 - p_k z^{-1})}. \quad (1)$$

TABLE S2
CLINICAL OPERATING DECISION METRICS ACROSS EVALUATED DIAGNOSTIC CONDITIONS AND DECISION THRESHOLD PROTOCOLS.

Cohort	Diagnosis	Condition	Threshold Protocol	Thresh (τ)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score	AUPRC
Chapman-Shaoxing ($N_{\text{test}} = 926, N_{\text{pos}} = 49$)	RBBB	Butterworth Base (Cond A)	Locked Baseline	0.023	98.0%	97.3%	66.7%	0.7934	0.9594
		All-Pass P1 (Standard, Cond C)	Locked Baseline	0.023	93.9%	24.4%	6.5%	0.1214	0.0742
		All-Pass P1 (Standard, Cond C)	Re-calibrated (Youden)	0.166	85.7%	40.5%	7.5%	0.1370	0.0742
		All-Pass P1 (AP-Aug, Cond C)	Locked Baseline	0.023	93.9%	88.7%	31.7%	0.4742	0.9093
		All-Pass P1 (AP-Aug, Cond C)	Re-calibrated (Youden)	0.072	91.8%	96.5%	59.2%	0.7200	0.9093
	LBBB	Butterworth Base (Cond A)	Locked Baseline	0.038	87.5%	97.7%	40.0%	0.5490	0.6975
		All-Pass P1 (Standard, Cond C)	Locked Baseline	0.038	87.5%	81.9%	7.8%	0.1436	0.5785
		All-Pass P1 (Standard, Cond C)	Re-calibrated (Youden)	0.060	87.5%	88.2%	11.6%	0.2044	0.5785
		All-Pass P1 (AP-Aug, Cond C)	Locked Baseline	0.038	81.2%	96.7%	30.2%	0.4407	0.6767
		All-Pass P1 (AP-Aug, Cond C)	Re-calibrated (Youden)	0.011	100.0%	87.6%	12.4%	0.2207	0.6767
CPCSC-2018 ($N_{\text{test}} = 592, N_{\text{pos}} = 184$)	RBBB	Butterworth Base (Cond A)	Locked Baseline	0.442	93.5%	97.8%	95.0%	0.9425	0.9805
		All-Pass P1 (Standard, Cond C)	Locked Baseline	0.442	94.0%	70.6%	59.0%	0.7254	0.8903
		All-Pass P1 (Standard, Cond C)	Re-calibrated (Youden)	0.705	91.3%	81.9%	69.4%	0.7887	0.8903
		All-Pass P1 (AP-Aug, Cond C)	Locked Baseline	0.442	91.3%	97.3%	93.9%	0.9256	0.9706
		All-Pass P1 (AP-Aug, Cond C)	Re-calibrated (Youden)	0.472	91.3%	98.0%	95.5%	0.9333	0.9706
PTB-XL ($N_{\text{test}} = 2,163, N_{\text{pos}} = 498$)	CD	Butterworth Base (Cond A)	Locked Baseline	0.379	77.9%	89.8%	69.7%	0.7355	0.8262
		All-Pass P1 (Standard, Cond C)	Locked Baseline	0.379	98.4%	2.9%	23.3%	0.3763	0.3990
		All-Pass P1 (Standard, Cond C)	Re-calibrated (Youden)	0.848	63.5%	55.0%	29.7%	0.4044	0.3990

*Note: N_{pos} denotes confirmed positive test cases. Locked Baseline fixes decision threshold τ from the baseline validation set; Re-calibrated re-tunes τ via Youden's J index directly on the test partition. Minor subclasses (e.g., LBBB with $N_{\text{pos}} = 16$) carry wider binomial confidence intervals.

TABLE S3
THEORETICAL GROUP DELAY DISPERSION ACROSS STANDARD
TEXTBOOK IIR FILTER TOPOLOGIES IN THE DIAGNOSTIC PASSBAND
([1, 35] Hz).

Filter Topology	Nominal Band	min τ_g	max τ_g	Passband $\Delta\tau_g$
Butterworth 4th-order (Ref)	[0.5, 40] Hz	0.64	20.61	19.98 samples
Bessel 4th-order (Prototype)	[0.5, 40] Hz	0.78	22.69	21.91 samples
Chebyshev-II 4th-order (20 dB)	[0.5, 40] Hz	0.53	21.54	21.01 samples
All-Pass Profile 1 (Dispersion)	Flat ($ H \equiv 1$)	0.71	13.49	12.78 samples

*Note: Bilinear transformation of continuous-time analog prototypes introduces non-linear frequency warping that distorts the maximally flat delay property of continuous Bessel filters. Passband group delay variations in all bandpass topologies are dominated by transition-band dispersion near the 1 Hz lower edge.

The continuous phase response $\theta(\omega) = \arg H(e^{j\omega})$ is evaluated by summing phase contributions of individual zeros and poles:

$$\theta(\omega) = \arg(K) + \sum_{k=1}^M \arg(e^{j\omega} - z_k) - \sum_{k=1}^N \arg(e^{j\omega} - p_k) - (M-N)\omega. \quad (2)$$

Differentiating with respect to normalized radian frequency ω yields the closed-form group delay:

$$\begin{aligned} \tau_g(\omega) &= -\frac{d\theta(\omega)}{d\omega} \\ &= \sum_{k=1}^N \frac{1 - r_{p,k} \cos(\omega - \theta_{p,k})}{1 - 2r_{p,k} \cos(\omega - \theta_{p,k}) + r_{p,k}^2} \\ &\quad - \sum_{k=1}^M \frac{1 - r_{z,k} \cos(\omega - \theta_{z,k})}{1 - 2r_{z,k} \cos(\omega - \theta_{z,k}) + r_{z,k}^2}, \end{aligned} \quad (3)$$

where $p_k = r_{p,k} e^{j\theta_{p,k}}$ and $z_k = r_{z,k} e^{j\theta_{z,k}}$. For an all-pass section where $z_k = \frac{1}{r_{p,k}} e^{j\theta_{p,k}}$, the zero and pole contributions reinforce each other, producing localized peaks in group delay centered at $\omega = \theta_{p,k}$ with amplitude proportional to $\frac{1+r}{1-r}$.

S4. NUMERICAL STABILITY AND TRANSIENT BOUNDARY DECAY ANALYSIS

To rigorously guarantee that startup transients at segment boundaries do not confound downstream deep learning evaluation, we analyze the homogeneous natural response of a second-order recursive section:

$$\begin{aligned} y[n] - 2r \cos(\theta_0) y[n-1] + r^2 y[n-2] = \\ r^2 x[n] - 2r \cos(\theta_0) x[n-1] + x[n-2]. \end{aligned} \quad (4)$$

The homogeneous solution $y_h[n]$ satisfies $y_h[n] = A r^n \cos(\theta_0 n + \phi)$, where A and ϕ depend on the initial filter state $z_0 = [y[-1], y[-2]]^T$. Because $r = 0.85 < 1$, the transient envelope decays exponentially at a geometric rate of $(0.85)^n$.

For a continuous lead-in stabilization buffer of $L_{\text{lead}} = 200$ samples (2.0 seconds at $f_s = 100$ Hz), the residual transient amplitude is bounded by:

$$|y_h[200]| \leq |A| \times (0.85)^{200} \approx |A| \times 7.712 \times 10^{-15}. \quad (5)$$

Because IEEE 754 single-precision floating-point numbers allocate 24 bits to the significand (providing an effective precision of $\epsilon \approx 1.19 \times 10^{-7}$), a residual decay factor of 10^{-15} guarantees that startup transients are completely extinguished before the evaluation window commences in Condition C. Furthermore, in practical finite-window processing without lead-in buffers (Condition B), empirical tracking demonstrates that boundary discrepancy decays below 5×10^{-4} amplitude units within 50 samples (0.5 s). Over a standard 7.5-second diagnostic window (750 samples), global average pooling aggregates activations across the entire recording, rendering boundary transients statistically undetectable ($\Delta_{C-B} = +0.00004$, 95% CI $[-0.00024, +0.00034]$ on PTB-XL).