

Physics-Informed Deep Learning for False Ventricular Tachycardia Alarm Reduction in the ICU

Athanasios Papastathopoulos-Katsaros^{1,2}, Alexandra Stavrianidi^{3,4}, Zhandong Liu^{1,2}

¹ Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA

² Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital, Houston, TX, USA

³ Institute for Analysis and Numerics, University of Münster, Germany

⁴ Department of Mathematics, Reed College, Portland, Oregon, USA

Abstract

False ventricular tachycardia (VT) alarms are a leading contributor to alarm fatigue in intensive care units. We propose a deep learning framework combining a 1D SE-ResNet with ICU-realistic data augmentations and a physics-informed auxiliary reconstruction task based on the three-element Windkessel hemodynamic model, implemented as a differentiable forward simulation. By requiring the network’s latent representation to produce physiologically plausible arterial pressure waveforms, artifact-driven ECG patterns are penalized while true VT remains coherent across modalities. Evaluated on the VTaC benchmark under a strict real-time protocol (10 s pre-alarm window), our method achieves a Challenge Score of 85.08 ± 1.65 , a 5-point improvement over prior state-of-the-art. Ablation studies confirm that the physics-informed objective is the primary performance driver, providing gains in accuracy, $\sim 2\times$ label efficiency, and more localized and clinically meaningful ECG segments.

1. Introduction

Ventricular tachycardia (VT) is a life-threatening arrhythmia characterized by anomalous ventricular beats exceeding 100 bpm [1]. Because prolonged VT can rapidly lead to sudden cardiac death, ICU monitors are tuned for high sensitivity, making VT alarms among the most prone to false positives [2, 3]. This contributes to alarm fatigue, a critical patient-safety concern [4, 5]. Reducing false VT alarms is a critical problem at the intersection of machine learning and healthcare.

A key difficulty in detecting false alarms is that sensor detachment, patient movement, and electrical interference produce ECG artifacts that closely resemble true arrhythmias, confounding data-driven classifiers. The VTaC benchmark [6] provides over 5,000 multi-institutional ICU recordings with ECG, photoplethysmography (PLETH),

and arterial blood pressure (ABP), enabling evaluation. Prior real-time methods, including supervised CNNs, contrastive models [7], and cross-modal VAEs [8], achieve Challenge Scores up to 80.08 [9]. Foundation models [10] and retrospective approaches [11] report higher AUCs, but the former rely on large external pretraining corpora and the latter use post-alarm information, while still remaining vulnerable to severe sensor noise and previously unseen artifact patterns.

We address this gap by embedding physiological structure into the learning process. Our architecture jointly optimizes classification with cross-modal physics-informed reconstruction of ABP (via a Windkessel simulation) and of PLETH via a data-driven decoder. This forces the network to ensure that any ECG pattern classified as VT produces a plausible hemodynamic response, thereby penalizing artifact-driven predictions. Our method achieves a Challenge Score of 85.08 ± 1.65 , a ~ 5 -point improvement over prior state-of-the-art, operating strictly within a 10 s real-time window without external pretraining data.

2. Methods

2.1. Dataset and Preprocessing

We use the VTaC dataset [6]: 5,037 expert-annotated alarm events ($\sim 29\%$ true) from three geographically distinct U.S. hospitals with different monitors and lead configurations, with the official patient-level 80-10-10 split. Following the real-time protocol, only the final 10 s before each alarm is used (250 Hz, $T=2500$). Fourteen canonical channels (ECG leads I, II, III, aVR, aVL, aVF, V1–V6, PLETH, ABP) are mapped to fixed slots; unoccupied slots are zero-filled. A binary availability mask $\mathbf{M} \in \{0, 1\}^{14 \times T}$ is concatenated channel-wise and each channel is independently z-normalized per segment.

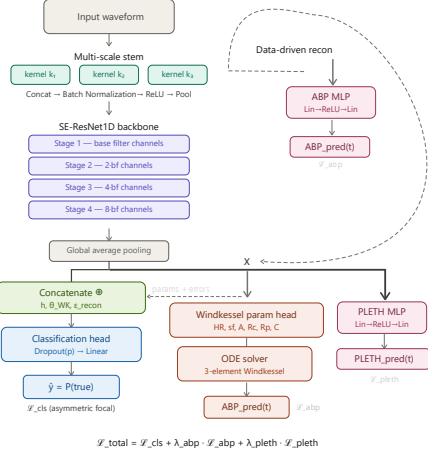


Figure 1. Multi-task architecture. The backbone features are shared across a classification head, a physics-informed Windkessel ODE reconstruction head for ABP, and a data-driven MLP decoder for PLETH. The dashed path shows the data-driven ablation variant.

2.2. Architecture

Our backbone is a 1D Squeeze-and-Excitation Residual Network (SE-ResNet1D) [12, 13]. To handle varying heart rates and arrhythmia cycle lengths, we replace the standard initial convolution with a Multi-Scale Stem comprising three parallel 1D convolutional branches (kernel sizes 15, 51, and 201 at stride 2), whose outputs are concatenated, batch-normalized, activated with ReLU, and max-pooled. The core feature extractor consists of four sequential stages, each containing two SE-ResNet blocks (kernel size 7, SE reduction ratio 16), with progressive spatial downsampling and channel doubling at stages 2–4. The output is aggregated via global average pooling, followed by dropout and a linear classification layer. Depending on the inclusion of auxiliary heads, the model contains 4–7M trainable parameters. The multi-task architecture is illustrated in Figure 1. Hyperparameters were optimized using Optuna’s TPE sampler to maximize the Challenge Score on the validation set [14]. The search included base filter width $\{32, 48, 64\}$, learning rate $[10^{-4}, 10^{-2}]$, weight decay $[10^{-7}, 10^{-3}]$, classification dropout $[0.1, 0.5]$, the focal-loss penalty, augmentation intensities, auxiliary-loss weights $\lambda_{abp}, \lambda_{pleth} \in [10^{-3}, 1]$, and auxiliary target length $\{25, 50, 100, 250, 500\}$.

2.3. Data Augmentation and Class Imbalance Strategy

We apply five augmentations targeting documented ICU degradation modes [15]: temporal jitter (± 50 samples), additive Gaussian noise, per-channel dropout emulating

sensor detachment [2], amplitude scaling, and baseline wander (0.1–0.5 Hz sinusoid).

We address the class imbalance of the dataset through Asymmetric Focal Loss:

$$L_{cls} = -w_{fn} \cdot y(1-\hat{p})^{\gamma^+} \log \hat{p} - (1-y) \hat{p}^{\gamma^-} \log(1-\hat{p}) \quad (1)$$

where $\gamma^+ = \gamma^- = 2.0$ [16] and w_{fn} is a tunable false-negative weight motivated by the asymmetric $5\times$ false-negative penalty of the Challenge Score.

2.4. Physics-Informed Auxiliary Regularization

To enforce physiological consistency and reduce overfitting to electrical artifacts, we introduce an auxiliary ABP reconstruction task grounded in the three-element Windkessel model. Unlike the PINN paradigm [17], we do not enforce ODE residuals; instead, we embed a forward physiological model within the supervised model as an auxiliary reconstruction target.

A projection head (64-unit hidden layer, ReLU) predicts six parameters: heart rate HR , systolic fraction sf , pulse amplitude amp , proximal resistance R_c , peripheral resistance R_p , and compliance C , with HR and sf bounded via scaled sigmoids and amp , R_c , R_p , and C constrained to be positive via shifted softplus activations. These parameters drive a differentiable forward simulation in 3 stages:

Stage 1: A synthetic ejection proxy models blood flow as a half-sine pulse during systole and zero during diastole:

$$Q(t) = \begin{cases} amp \cdot \sin\left(\frac{\pi t_{mod}}{sf \cdot T_{cyc}}\right) & t_{mod} < sf \cdot T_{cyc} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $T_{cyc} = 60/HR$ and $t_{mod} = t \bmod T_{cyc}$.

Stage 2: The three-element Windkessel ODE

$$C dP_{wk}/dt = Q - P_{wk}/R_p$$

is integrated via an exact exponential scheme:

$$P_{wk}[n+1] = \gamma P_{wk}[n] + R_p(1-\gamma)Q[n], \quad \gamma = e^{-\Delta t/(R_p C)} \quad (3)$$

which is unconditionally stable for positive R_p, C , ensuring well-behaved gradients. The total pressure is $P(t) = R_c Q(t) + P_{wk}(t)$, mapped to the target space via the learnable parameters. The simulation executes in under 3 ms.

The objective is not to predict ABP—the mapping from ECG to peripheral pressure is ill-posed. The physics-informed reconstruction acts as *structured regularization* of the shared backbone. Artifact-driven ECG encodings cannot produce coherent hemodynamic waveforms and incur high reconstruction error, propagating corrective gradients through the shared representation.

PLETH is reconstructed via a data-driven MLP decoder (one hidden layer, 128 units, see dashed path in Figure 1) leveraging its availability ($\sim 91\%$ of recordings vs. $\sim 36\%$ for ABP). The total training objective is:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_{\text{abp}} L_{\text{abp}} + \lambda_{\text{pleth}} L_{\text{pleth}} \quad (4)$$

where L_{abp} and L_{pleth} are masked MSE losses computed only when the respective channels are available. The predicted Windkessel parameters and reconstruction errors are concatenated with the latent features before the final classification layer.

3. Results

The primary metric is the PhysioNet 2015 Challenge Score = $(TP + TN) / (TP + TN + FP + 5 \cdot FN)$, which penalizes missed true alarms ($5\times$ weight). All experiments use 5 random seeds; Challenge Scores are reported as mean $\pm$ SD.

3.1. Main Results

Table 1 compares our models against prior work on the VTaC test set.

Table 1. Performance on VTaC. “Aug”= augmentations; “DD”= data-driven reconstruction; “Phys”= physics-informed reconstruction. All our results: mean $\pm$ SD over 5 seeds. Some models do not report Challenge Score.

Method	Window	Score	AUC	TPR	PPV
FCN [9]	10 s	80.08 $\pm$ 2.46	.949	.920	.717
BioCross [8]	10 s	—	.863	.814	.576
CSFM [†] [10]	10 s	—	.967	—	—
FCNN* [11]	6 min	—	.973	.940	.950
Ours: baseline	10 s	80.84 $\pm$ 2.88	.949	.931	.708
+ DD Recon	10 s	81.41 $\pm$ 2.97	.956	.917	.752
+ Phys Recon	10 s	82.26 $\pm$ 2.62	.947	.946	.706
+ Aug	10 s	81.71 $\pm$ 2.24	.953	.944	.700
+ Aug & DD	10 s	84.16 $\pm$ 0.68	.964	.953	.732
+ Aug & Phys	10 s	85.08$\pm$1.65	.961	.958	.742

* Retrospective (uses post-alarm data). [†] Foundation model pretrained on a much larger external corpus.

3.2. Ablation Studies

Reconstruction paradigm: Without augmentations, physics-informed reconstruction improves the baseline by +1.42 Challenge Score points vs. +0.57 for data-driven reconstruction. With augmentations, physics-informed reconstruction reaches 85.08 ± 1.65 vs. the data-driven decoder.

Label efficiency: Table 2 (top) evaluates all variants at 5%, 10%, 20%, and 50% label fractions. The physics-informed model at 5% labels (64.74 ± 2.94) nearly matches the augmentation-only baseline at 10% (65.04 ± 1.98), demonstrating approximately $2\times$ label efficiency. It is the

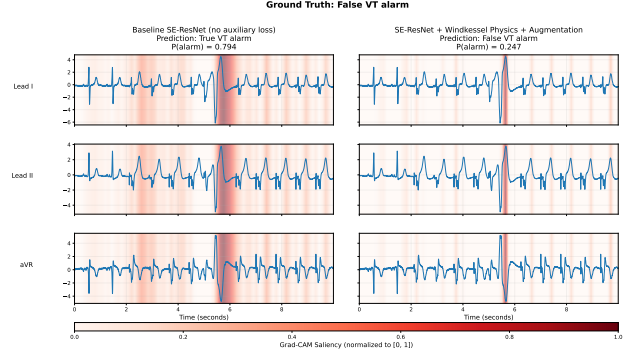


Figure 2. Grad-CAM comparison for a ground-truth false VT alarm. The baseline SE-ResNet (left) is confused by the noisy signal, with diffused saliency across the artifact, leading to an incorrect True VT prediction. In contrast, the physics-informed model (right) has learned to robustly identify non-physiological noise; it perfectly localizes the sharp movement artifact and uses this precise detection to correctly reject the false alarm.

only multi-task variant that improves over baseline at every evaluated fraction.

Robustness to missing modalities: Table 2 (bottom) simulates sensor failures on the physics-informed model trained with 100% of the labels and evaluated over five seeds. Sensor dropout reduces the Challenge Score by -1.33 for ABP and -4.85 for PLETH. Signal replacement with uncorrelated Gaussian noise (mask retained) is more damaging: -4.98 for ABP and -6.42 for PLETH. Thus, corrupted signals are more damaging than missing signals.

Interpretability: Quantitative localization analysis with Grad-CAM [18] was restricted to true-alarm samples ($n = 128$), because these contain a genuine VT transition against which temporal localization can be meaningfully assessed. The physics-informed model produced more localized saliency in 89.1% of cases, with mean Gini coefficient increasing from 0.52 ± 0.11 to 0.65 ± 0.09 . Figure 2 shows a false alarm misclassified by the baseline but correctly rejected by the physics-informed model.

Table 2. (Top) Label efficiency: Challenge Score across label fractions. All variants include augmentations. (Bottom) Modality robustness on the physics-informed model for a single run.

Variant	Label Efficiency (Challenge Score)			
	5%	10%	20%	50%
Augs baseline (no recon)	57.12 $\pm$ 2.79	65.04 $\pm$ 1.98	72.18 $\pm$ 3.24	77.61 $\pm$ 1.91
PLETH only (DD)	62.17 $\pm$ 4.15	68.60 $\pm$ 3.20	70.23 $\pm$ 3.26	76.39 $\pm$ 2.29
PLETH + ABP (DD)	62.72 $\pm$ 3.27	66.85 $\pm$ 3.89	69.63 $\pm$ 2.94	76.47 $\pm$ 2.07
PLETH + ABP (Phys)	64.74$\pm$2.94	68.90$\pm$4.28	73.13$\pm$1.36	78.75$\pm$2.75
Modality Robustness (Physics Model; Score difference compared to baseline)				
Condition	Sensor Dropout		Noise Replacement	
ABP only	-1.33		-4.98	
PLETH only	-4.85		-6.42	
Both	-5.18		-10.15	

4. Discussion

The key mechanism underlying the improvement is cross-modal artifact disentanglement. False VT alarms primarily arise from gross signal corruption—electrode detachment, patient movement, electrosurgical interference—that produces wide-complex, high-rate ECG patterns indistinguishable from VT on a single lead [2]. The Windkessel reconstruction head penalizes artifact-dependent features because electrical artifacts have no hemodynamic correlate: if the backbone encodes an artifactual ECG as “VT,” the simulated pressure waveform will be incoherent and incur high reconstruction loss.

Notably, the ABP reconstruction loss decreases by approximately 2–3% over training, consistent with the ill-posed nature of the ECG-to-pressure mapping. The regularization benefit is disproportionate to this reconstruction accuracy; the value lies in shaping the gradient landscape so that an unconstrained MLP decoder would simply absorb artifacts into flexible weights, while the Windkessel’s fixed dynamical structure propagates corrective gradients. Classification requires ≤ 4 ms on GPU, and the Windkessel head is removed at inference. However, limitations include the lumped-parameter simplicity of the Windkessel approximation, ABP availability in only 36% of recordings, and evaluation on a single benchmark.

5. Conclusions

We have shown that embedding a simple hemodynamic forward model as an auxiliary reconstruction task provides a powerful inductive bias for false VT alarm reduction. The physics-informed constraint drives performance gains while improving label efficiency and interpretability, offering a promising strategy for robust clinical classification in high-acuity settings with limited labeled data.

Acknowledgements

APK and ZL were supported by NIH grant 5R01HG011795, CPRIT grant RP240131, Chan Zuckerberg Initiative (2023-332162), the Chao Endowment and the Huffington Foundation. AS was funded by the DFG under Germany’s Excellence Strategy EXC 2044/2–390685587.

References

- [1] Clifford GD, et al. False alarm reduction in critical care. *Physiol Meas* 2016;37(8).
- [2] Drew BJ, et al. Insights into the Problem of Alarm Fatigue with Physiologic Monitor Devices: A Comprehensive Observational Study of Consecutive Intensive Care Unit Patients. *PLoS ONE* 2014;9(10).
- [3] Aboukhalil A, et al. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of Biomedical Informatics* 2008;41(3).
- [4] Fallet S, et al. False arrhythmia alarms reduction in the intensive care unit: a multimodal approach. *Physiol Meas* 2016;37(8).
- [5] Chromik J, et al. Computational approaches to alleviate alarm fatigue in intensive care medicine: A systematic literature review. *Front Digit Health* 2022;4.
- [6] Lehman Lw, et al. VTaC: A Benchmark Dataset of Ventricular Tachycardia Alarms from ICU Monitors, 2024.
- [7] Zhou Y, et al. A contrastive learning approach for icu false arrhythmia alarm reduction. *Scientific Reports* 2022;12.
- [8] Wang M, et al. BioCross: A cross-modal framework for unified representation of multi-modal biosignals with heterogeneous metadata fusion. *Information Fusion* 2025;123.
- [9] Lehman LwH, et al. Vtac: a benchmark dataset of ventricular tachycardia alarms from icu monitors. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*. Red Hook, NY, USA: Curran Associates Inc., 2023; .
- [10] Gu X, et al. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nat Mach Intell* 2026; 8(2).
- [11] Farayola GF, et al. Reducing False Ventricular Tachycardia Alarms in ICU Settings: A Machine Learning Approach. In *2025 10th International Conference on Machine Learning Technologies (ICMLT)*. Helsinki, Finland: IEEE, 2025; .
- [12] He K, et al. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016; .
- [13] Hu J, et al. Squeeze-and-Excitation Networks. *IEEE Trans Pattern Anal Mach Intell* 2020;42(8).
- [14] Akiba T, et al. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: ACM, 2019; .
- [15] Iwana BK, et al. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* 2021;16(7).
- [16] Lin TY, et al. Focal Loss for Dense Object Detection. *IEEE Trans Pattern Anal Mach Intell* 2020;42(2).
- [17] Raissi M, et al. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 2019;378.
- [18] Selvaraju RR, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis* 2020;128(2).

Address for correspondence:

Athanasios Papastathopoulos-Katsaros
Department of Pediatrics, Baylor College of Medicine, Houston, TX, 77030, United States of America
athanasios.papastathopoulos-katsaros@bcm.edu