Forecasting Shockable In-Hospital Cardiac Arrest from Continuous Telemetry

Mayank Keoliya* MKEOLIYA@ENGINEERING.UPENN.EDU
Seewon Choi* SEEWON@ENGINEERING.UPENN.EDU
Neelay Velingker* NEELAY@ENGINEERING.UPENN.EDU
Alaia Solko-Breslin ALAIA@ENGINEERING.UPENN.EDU
Alireza Oraii ALIREZA.ORAII@PENNMEDICINE.UPENN.EDU
Sameed M. Khatana SAMEED.KHATANA@PENNMEDICINE.UPENN.EDU
Mayur Naik MHNAIK@ENGINEERING.UPENN.EDU
Gary E. Weissman GARY.WEISSMAN@PENNMEDICINE.UPENN.EDU
Eric Wong EXWONG@ENGINEERING.UPENN.EDU
Rajeev Alur† ALUR@ENGINEERING.UPENN.EDU
Rajat Deo† RAJAT.DEO@PENNMEDICINE.UPENN.EDU
University of Pennsylvania, USA

Abstract

Recent advances in artificial intelligence (AI) for cardiology have shown promise in detecting existing cardiac conditions and stratifying long-term risk from electrocardiograms (ECGs). We introduce, and answer positively, a new question: *can AI forecast the onset of impending shockable cardiac arrest rhythms—ventricular tachycardia (VT) and ventricular fibrillation (VF)—from continuous telemetry?* A key challenge is the lack of telemetry datasets that combine hours-long recordings with precise annotations of event onsets to support modeling of evolving cardiac instability and timely prediction. We address this gap by collecting `PennVT`, a novel in-hospital telemetry dataset comprising 9,368 hours of continuous telemetry from 911 hospitalized patients, including 80 VT/VF events. We develop CAMEL, an ECG-language model that directly tokenizes raw waveforms, and evaluate it in a rolling forecasting setting. For a two-hour forecasting horizon, CAMEL provides at least 10 minutes of advance warning for 57.5% of VT/VF events, while 91.8% of control patients receive no alarm. CAMEL achieves 0.832 AUROC, outperforming ECG measurement models and vital-sign-based risk scores by 0.083 and 0.209, respectively. Furthermore, CAMEL generates 3.07 false alarms per patient-day, compared with 5.64–9.78 for prior ECG foundation models, highlighting the importance of learned waveform representations for VT/VF forecasting.

Keywords: continuous cardiac telemetry, early warning system, shockable cardiac arrest

Data and Code Availability The cardiac telemetry data used in this study were collected under an approved research protocol. Due to institutional privacy and data-use requirements, they cannot be made public. Code for reproducing the experiments is available at https://github.com/CAMEL-ECG/CAMEL-inference.

Institutional Review Board (IRB) This retrospective study was approved by the Institutional Review Board (IRB) of the University of Pennsylvania. The requirement for individual informed consent was waived given the retrospective nature of the study.

1. Introduction

In-hospital cardiac arrest (IHCA) occurs in approximately 290,000 patients annually in the United States, with 25% of cases involving shockable rhythms: ventricular tachycardia (VT) and ventricular fibrillation (VF) (Palaniappan et al., 2026; Holmberg et al., 2019). For these rhythms, a 1–2 minute delay in initial defibrillation is associated with 52% lower odds of survival to discharge, yet approximately 30% of cases experience such delays (Chan et al., 2008). Forecasting models that provide warning minutes to hours before onset could reduce treatment de-

* Equal Contribution
† Joint Supervision

© M. Keoliya *et al.*

lays and enable preventive interventions (Valenzuela et al., 1997; Marijon et al., 2022).

Continuous cardiac telemetry is well suited to real-time forecasting of shockable IHCA, as it is routinely collected in hospital settings and provides a real-time, high-frequency view of cardiac electrophysiology. However, publicly available continuous telemetry datasets remain scarce. Existing datasets typically provide a limited number of leads, lack precise annotations of VT/VF onset, or include relatively few events (Lehman et al., 2024; Moody et al., 2020). Consequently, prior ECG modeling has largely focused on widely available 10-second ECG snapshots for classifying existing abnormalities or stratifying long-term risk (Lan et al., 2025; Obermeyer et al., 2026), while shorter-term cardiac arrest prediction has relied primarily on vital signs or summary measurements derived from telemetry (Kenet et al., 2023; Sachdev et al., 2010). Forecasting imminent VT/VF onset from raw continuous telemetry therefore remains largely unexplored.

To address this gap, we collect 9,368 hours of continuous cardiac telemetry from 911 patients, including 80 patients who experience cardiac arrest due to sustained VT or VF. `PennVT` is the first dataset to combine hours-long, multi-lead telemetry recordings with expert-adjudicated VT/VF onset times. To model telemetry efficiently, we develop CAMEL (Cardiac Autoregressive Model for ECG Language Modeling), an ECG-language model that directly tokenizes each single-lead, one-second waveform segment. We evaluate CAMEL as a dynamic early warning system that periodically generates risk scores and triggers alarms throughout each recording.

In five-fold cross-validation, CAMEL provides advance warning for 57.5% of VT/VF events, between 10 minutes and two hours before onset, with false alarms concentrated in 8.2% of control patients. Closer to onset, CAMEL's risk estimates increase, and sustained alarms become more common. We compare CAMEL against the vital-sign-based Cardiac Arrest Risk Triage (CART) score, ECG measurement-based models, and ECG foundation models: OpenTSLM, xECG, and ECG-CPC. CAMEL achieves the highest discrimination and the lowest false alarm burden at matched sensitivity, suggesting that its learned waveform representations are more effective for forecasting. Overall, our results demonstrate that impending VT/VF event can be forecast from continuous telemetry sufficiently early to enable intervention.

2. Study Design

2.1. `PennVT`

Most public ECG datasets consist of 12-lead, 10-second recordings, providing limited temporal context for forecasting over minutes to hours (Perez Alday et al., 2020). Public continuous telemetry datasets are available, but typically contain short recordings, include few VT/VF patients, provide only one or two leads, or lack precise event-onset annotations, limiting their suitability for imminent VT/VF forecasting (Lehman et al., 2024; Tan et al., 2022; Moody et al., 2020; Kansal et al., 2025).

We collected eight-lead telemetry from an academic health system at the University of Pennsylvania, with VT/VF onset times adjudicated by two expert cardiologists. The `PennVT` dataset comprises 80 VT/VF cases and 831 control patients sampled from the same clinical units and study period as the cases. For VT/VF patients, we use up to three hours of telemetry immediately preceding onset; for control patients, we use a randomly selected three-hour interval. Additional cohort details are provided in Appendix A.

2.2. Rolling Prediction

We formulate real-time VT/VF forecasting as a rolling-window prediction task. At intervals of w seconds, the model uses the preceding l seconds of telemetry to estimate the risk of VT/VF onset within the next h seconds. For VT/VF patients, we exclude prediction times less than 10 minutes before onset to allow time for potential intervention (Boussina et al., 2024). An alarm is triggered when predicted risk meets or exceeds a threshold τ, selected to achieve at least 80% event sensitivity on the validation set.

An alarm is considered true if event onset occurs within its forecasting horizon h, and false otherwise. To reduce repeated alerts, subsequent alarms are suppressed for s minutes following each issued alarm (Lee et al., 2023; Shashikumar et al., 2025). Our primary configuration uses a lookback of $l = 60$ seconds, a forecasting horizon of $h = 2$ hours, a prediction interval of $w = 60$ seconds, and an alarm suppression duration of $s = 30$ minutes (Figure 1).

Rolling prediction produces repeated risk estimates, with the number of predictions varying across patients. We therefore prioritize patient- and alarm-level metrics that summarize performance over each patient's prediction trajectory. Our primary outcomes are event sensitivity, patient specificity, aver-

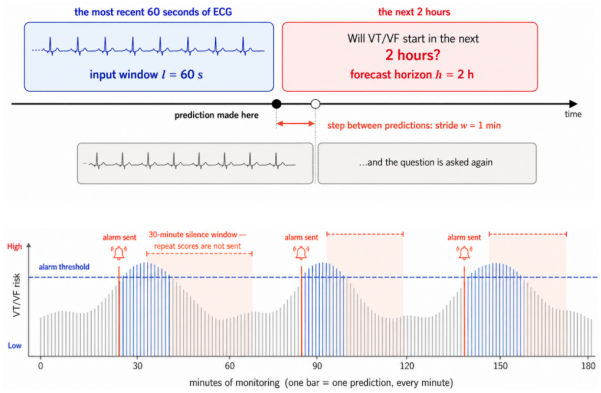


Figure 1: The primary configuration for rolling prediction: lookback window $l = 60$ seconds, forecasting horizon $h = 2$ hours, prediction interval $w = 1$ minute, and alarm suppression duration $s = 30$ minutes.

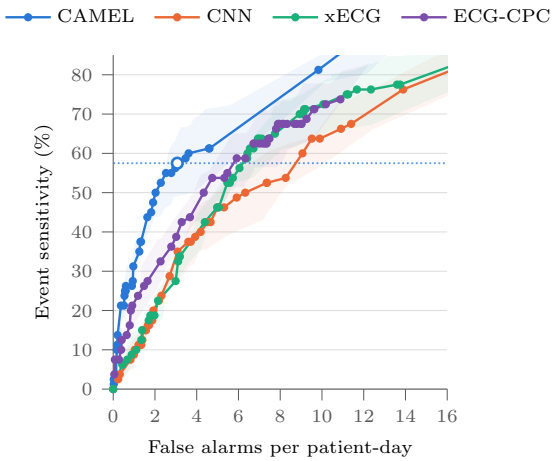


Figure 2: Event sensitivity and false alarm burden trade-offs for CAMEL, CNN, xECG, and ECG-CPC under the primary configuration. Shaded regions show 95% confidence intervals.

age number of false alarms per patient-day, and alarm positive predictive value (PPV), which together capture key aspects of clinical usability during continuous monitoring. We also report AUROC and AUPRC as secondary measures of discrimination at individual prediction times.

2.3. CAMEL

General-purpose large language models (LLMs) lack native support for raw waveform inputs, requiring ECG tracings to be presented as images. Their performance remains limited even on short ECG classification tasks (Liu et al., 2026; Zhu et al., 2024). Specialized ECG-language models improve performance by fine-tuning LLMs to jointly process ECG images and text, achieving state-of-the-art classification results (Liu et al., 2026; Lan et al., 2025). However, these image-based approaches are token-inefficient and have largely been developed for 10-second snapshots, making them difficult to scale to continuous telemetry.

CAMEL directly processes raw waveforms through a CNN-based encoder that maps each one-second, single-lead ECG segment to a single token, enabling efficient modeling of longer multi-lead recordings (Figure 5). We pretrain the ECG encoder and LoRA-adapted MedGemma (Sellergren et al., 2025) through a multi-stage ECG–text alignment curriculum to progressively learn beat- and rhythm-level representations. For VT/VF forecasting, we freeze the pretrained model and train only a lightweight two-layer MLP prediction head. Additional architecture and training details are provided in Appendix B.

3. Findings

3.1. Continuous telemetry enables clinically meaningful early warning of VT/VF

We evaluate CAMEL for real-time VT/VF forecasting on `PennVT` using five-fold cross-validation. Under the primary configuration, CAMEL provides at least 10 minutes of advance warning for 46 of 80 events (57.5%, 95% CI: 45.6–68.9) while generating 3.07 (2.46–3.81) false alarms per patient-day (Figure 2). False alarms are concentrated in 68 of 831 control patients (8.2%, 6.6–10.1). At this operating point, CAMEL achieves an alarm PPV of 34.9% (27.3–42.6) with a median lead time of 97 minutes (73–111) from the first true alarm to VT/VF onset.

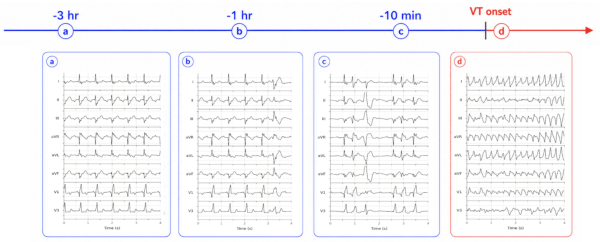
We further examine how performance varies with the forecasting configuration (Figure 6). Extending the input lookback beyond 60 seconds provides no consistent benefit, potentially reflecting CAMEL's pretraining on 10-second ECG tracings. Compared with no suppression, a 30-minute suppression duration reduces false alarms from 44.13 to 3.07 per patient-day while preserving event sensitivity. These findings highlight that early warning performance de-

pends on both the underlying risk model and the policy used to translate its predictions into alarms.

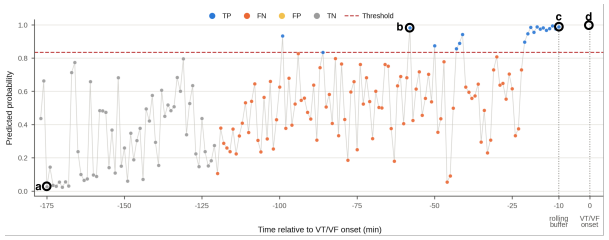
3.2. CAMEL risk increases as VT/VF onset approaches

To assess whether CAMEL captures evolving pre-event abnormalities, we divide the period from 130 to 10 minutes before onset into four 30-minute intervals and compute mean predicted risk and the longest run of consecutive above-threshold predictions for each interval. As shown in Table 5, sustained warning becomes more common as the onset approaches: the proportion of patients with at least 10 consecutive above-threshold predictions rises from 15.2% to 24.2%. The proportion of patients with any above-threshold prediction also increases by 10.6 percentage points. Accordingly, a static baseline that reuses the risk score from each patient's first available 60-second window at every prediction time generates 2.7× as many false alarms as rolling CAMEL predictions (8.17 versus 3.07 per patient-day) under the same forecasting setup, indicating that CAMEL flags deterioration beyond baseline risk. Among controls, mean risk remains essentially unchanged across the four intervals (0.285–0.287) prior to the sampled endpoint, while both sustained warning (1.6–2.8%) and any above-threshold predictions (5.2–5.8%) are less common. Together, these results suggest that CAMEL's predictions reflect proximity to VT/VF onset as well as patient-level susceptibility.

The case shown in Figure 3 illustrates this pattern. Risk estimates show an overall upward trend, with a first alarm 100 minutes before VT onset and sustained above-threshold predictions emerging 20 minutes before onset. Importantly, despite the low estimated risk, the early tracing at time point (a) already shows right bundle branch block and ST-segment changes, together with prolonged QTc and PR intervals. At time point (b), average ECG interval measurements remain broadly similar, but isolated premature wide-complex ventricular beats emerge. Time point (c) shows marked beat-to-beat irregularity, with extra and dropped beats along with increased morphological variability. Time point (d) captures the subsequent VT event. This example illustrates how forecasting requires modeling acute deviations that emerge before a VT event, beyond detecting abnormalities. Such temporal evolutions cannot be learned from isolated ECG snapshots, which are typical of existing ECG datasets.



(*a*) Telemetry at approximately (a) 3 hours, (b) 1 hour, and (c) 10 minutes before VT onset. (d) shows the subsequent VT episode. (a)–(c) show progressively more pleomorphic ventricular ectopy before onset.



(*b*) Risk trajectory generated by CAMEL. (a)–(d) mark the corresponding time points in the telemetry above. The dashed line shows the alarm threshold. Colors classify individual predictions as true positives, false negatives, false positives, or true negatives.

Figure 3: Example telemetry and corresponding CAMEL risk trajectory.

3.3. Effective waveform representation is critical for VT/VF forecasting

A natural question is whether CAMEL simply identifies the patients that conventional ECG analysis labels abnormal. We therefore train a random forest using ECG measurements as features. As shown in Figure 4 and Tables 6 and 7, the random forest exhibits substantially reduced discrimination, achieving an AUROC of 0.749, compared with 0.832 (0.787–0.869) for CAMEL. Random forest models trained on individual measurements and temporal variants also show limited discrimination, with raw QRS duration achieving the highest AUROC of 0.661. These comparisons suggest that raw telemetry contains predictive information beyond interval and heart-rate measurements. Similarly, the vital-sign-based early-warning score CART (Churpek et al., 2012) shows limited discrimination and raises approx-

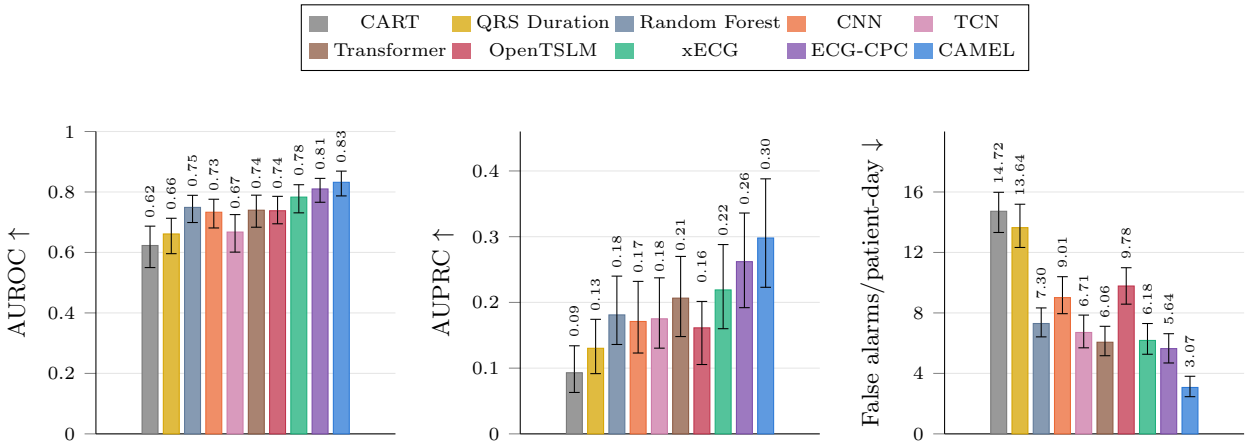


Figure 4: Forecasting performance of CART; random forests using QRS duration alone (QRS Duration) or ten ECG measurements (Random Forest); CNN, TCN, and Transformer models trained from scratch; and ECG foundation models (OpenTSLM, xECG, and ECG-CPC) and CAMEL. Error bars show 95% confidence intervals. QRS results use 909 patients (80 VT/VF, 829 controls) with QRS durations extractable; CART results use 880 patients (57 VT/VF, 823 controls) with available vital-sign measurements.

imately 4.8 times as many false alarms per patient-day as CAMEL.

We further compare CAMEL against three architectures trained from scratch—CNN, TCN, and Transformer—and three waveform-based ECG foundation models: OpenTSLM (Langer et al., 2026), xECG (Lunelli et al., 2025), and ECG-CPC (Al-Masud et al., 2026). OpenTSLM is trained with language supervision, whereas xECG and ECG-CPC are self-supervised models designed to process variable-length ECG recordings. Notably, all three architectures trained from scratch perform comparably to or worse than the measurement-based random forest, highlighting that access to the waveform alone is insufficient and that pretrained waveform representations are important for VT/VF forecasting. At matched sensitivity, ECG foundation models generate 1.8–3.2 times as many false alarms per patient-day, triggering false alarms in 13.7–30.8% of control patients, compared with 8.2% for CAMEL (Figure 4). The advantage of CAMEL persists across various forecasting horizons and lookback configurations (Figure 6 and Table 8).

4. Conclusion

In this study, we asked whether shockable cardiac arrest can be forecast from cardiac telemetry collected as part of routine hospital monitoring. On `PennVT`, CAMEL identifies 46 of 80 VT/VF events with meaningful advance warning. Furthermore, CAMEL achieves a more favorable sensitivity–false alarm burden trade-off than alternative ECG foundation models as well as vital-sign-based scores and summary measurement-based models, supporting the value of CAMEL's learned representations for real-time VT/VF forecasting.

Several limitations motivate future work. The `PennVT` cohort contains a modest number of VT/VF cases relative to the heterogeneity of patients and arrhythmia mechanisms, and all data were collected within a single health system. While our work establishes a retrospective benchmark showing feasibility of real-time VT/VF forecasting, whether the achieved sensitivity–alarm trade-off is clinically actionable requires external and prospective validation. Finally, the association between rising predicted risk and approaching VT/VF onset does not establish which underlying electrophysiological changes drive the predictions.

Acknowledgments

This research was supported by the ARPA-H program on Safe and Explainable AI under Award D24AC00253-00, NIH Awards R01-EB036016, R01-

NS141408, and R01-HL161303, the Google TPU Builders Program, and a gift from AWS AI to ASSET (Penn Engineering Center on Trustworthy AI). Partial support was provided by the Winkelman Family Fund in Cardiovascular Innovation. We also thank the Penn Advanced Research Computing Center (PARCC) for providing high-performance computing infrastructure for training.

References

M A Al-Masud, Juan Lopez Alcaraz, and Nils Strodthoff. Benchmarking ECG FMs: A reality check across clinical tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xXRqWpt3Xr.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. *Biomedizinische Technik/Biomedical Engineering*, 40(s1):317–318, 1995. ISSN 1862-278X. doi: 10.1515/bmte.1995.40.s1.317. URL http://dx.doi.org/10.1515/bmte.1995.40.s1.317.

Aaron Boussina, Supreeth P. Shashikumar, Atul Malhotra, Robert L. Owens, Robert El-Kareh, Christopher A. Longhurst, Kimberly Quintero, Allison Donahue, Theodore C. Chan, Shamim Nemati, and Gabriel Wardi. Impact of a deep learning sepsis prediction model on quality of care and survival. *npj Digital Medicine*, 7(1), January 2024. ISSN 2398-6352. doi: 10.1038/s41746-023-00986-6. URL http://dx.doi.org/10.1038/s41746-023-00986-6.

Paul S. Chan, Harlan M. Krumholz, Graham Nichol, Brahmajee K. Nallamothu, and American Heart Association National Registry of Cardiopulmonary Resuscitation Investigators. Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine*, 358(1):9–17, 2008. doi: 10.1056/NEJMoa0706467. URL https://doi.org/10.1056/NEJMoa0706467.

Matthew M. Churpek, Trevor C. Yuen, Seo Young Park, David O. Meltzer, Jesse B. Hall, and Dana P. Edelson. Derivation of a cardiac arrest prediction model using ward vital signs*. *Critical Care Medicine*, 40(7):2102–2108, July 2012. ISSN 0090-3493. doi: 10.1097/ccm.0b013e318250aa5a. URL http://dx.doi.org/10.1097/CCM.0b013e318250aa5a.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset, 2023. URL https://physionet.org/content/mimic-iv-ecg/1.0/.

Mathias J. Holmberg, Catherine E. Ross, Garrett M. Fitzmaurice, Paul S. Chan, Jordan Duval-Arnould, Anne V. Grossestreuer, Tuyen Yankama, Michael W. Donnino, Lars W. Andersen, Paul Chan, Anne V. Grossestreuer, Ari Moskowitz, Dana Edelson, Joseph Ornato, Katherine Berg, Mary Ann Peberdy, Matthew Churpek, Michael Kurz, Monique Anderson Starks, Saket Girotra, Sarah Perman, Zachary Goldberger, Anne-Marie Guerguerian, Dianne Atkins, Elizabeth Foglia, Ericka Fink, Javier J. Lasa, Joan Roberts, Melanie Bembea, Michael Gaies, Monica Kleinman, Punkaj Gupta, Robert Sutton, and Taylor Sawyer. Annual incidence of adult and pediatric in-hospital cardiac arrest in the united states. *Circulation: Cardiovascular Quality and Outcomes*, 12 (7), July 2019. ISSN 1941-7705. doi: 10.1161/circoutcomes.119.005580. URL http://dx.doi.org/10.1161/CIRCOUTCOMES.119.005580.

Chien Hua Huang, Randy Su, Hui Chun Huang, Katie Lin, Nick Foster, Nathan Juergens, Josh Risley, Katy Haynes, and Ziad Obermeyer. Subtyping cardiac arrest with ECG waveforms: A nightingale open science dataset, 2021. URL https://doi.org/10.48815/N5WC7D.

A. Kansal, E. Chen, T. Jin, P. Rajpurkar, and D. Kim. Multimodal clinical monitoring in the emergency department (MC-MED), 2025. URL https://doi.org/10.13026/WVYW-G663.

Adam L. Kenet, Rahul Pemmaraju, Sejal Ghate, Shreeya Raghunath, Yifan Zhang, Mordred Yuan, Tony Y. Wei, Jacob M. Desman, Joseph L. Greenstein, Casey O. Taylor, Timothy Ruchti, James Fackler, and Jules Bergmann. A pilot study to predict cardiac arrest in the pediatric intensive care unit. *Resuscitation*, 185:109740, April 2023. ISSN 0300-9572. doi: 10.1016/j.resuscitation.2023.109740. URL http://dx.doi.org/10.1016/j.resuscitation.2023.109740.

Zuzana Koscova, Qiao Li, Chad Robichaux, Valdery Moura Junior, Manohar Ghanta, Aditya Gupta, Jonathan Rosand, Aaron D. Aguirre, Erik Reinertsen, Steven Song, Shenda Hong, David E. Albert, Joel Xue, Aarya Parekh, Reza Sameni, Matthew A. Reyna, M. Brandon Westover, and Gari D. Clifford. The Harvard-Emory ECG database. *Scientific Data*, 13(1), February 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-06861-9. URL http://dx.doi.org/10.1038/s41597-026-06861-9.

P. Laguna, R.G. Mark, A. Goldberg, and G.B. Moody. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In *Computers in Cardiology 1997*, pages 673–676, 1997. doi: 10.1109/CIC.1997.648140. URL http://dx.doi.org/10.1109/CIC.1997.648140.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. GEM: Empowering MLLM for grounded ECG understanding with time series and images. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=idtZwmjakN.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott R. Ceresnak, Aydin Zahedivash, Alexander T Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Oppers Aalami, Filipe Barata, and Paul Schmiedmayer. OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FHDRzhKm0f.

Hyeonhoon Lee, Hyun-Lim Yang, Ho Geol Ryu, Chul-Woo Jung, Youn Joung Cho, Soo Bin Yoon, Hyun-Kyu Yoon, and Hyung-Chul Lee. Real-time machine learning model to predict in-hospital cardiac arrest using heart rate variability in icu. *npj Digital Medicine*, 6(1), November 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00960-2. URL http://dx.doi.org/10.1038/s41746-023-00960-2.

L. Lehman et al. VTaC: A benchmark dataset of ventricular tachycardia alarms from ICU monitors, 2024. URL https://doi.org/10.13026/8TD2-G363.

Hui Liu, Dan Chen, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Yinglong Wang. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific Data*, 9(1):272, June 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01403-5. URL http://dx.doi.org/10.1038/s41597-022-01403-5.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teaching multimodal llms to comprehend 12-lead electrocardiographic images. *npj Digital Medicine*, 9(1), March 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02551-3. URL http://dx.doi.org/10.1038/s41746-026-02551-3.

R. Lunelli et al. BenchECG and xECG: A benchmark and baseline for ECG foundation models. *arXiv preprint arXiv:2509.10151*, 2025. doi: 10.48550/arXiv.2509.10151. URL https://doi.org/10.48550/arXiv.2509.10151.

Eloi Marijon, Rodrigue Garcia, Kumar Narayanan, Nicole Karam, and Xavier Jouven. Fighting against sudden cardiac death: need for a paradigm shift—adding near-term prevention and pre-emptive action to long-term prevention. *European Heart Journal*, 43(15):1457–1464, February 2022. ISSN 1522-9645. doi: 10.1093/eurheartj/ehab903. URL http://dx.doi.org/10.1093/eurheartj/ehab903.

B. Moody, G. Moody, M. Villarroel, G. D. Clifford, and I. Silva. MIMIC-III waveform database matched subset, 2020. URL https://doi.org/10.13026/C2294B.

G.B. Moody and R.G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. ISSN 0739-5175. doi: 10.1109/51.932724. URL http://dx.doi.org/10.1109/51.932724.

Ziad Obermeyer, Alexander Schubert, James Ross, Sendhil Mullainathan, and Markus Lingman. An ecg biomarker for sudden cardiac death discovered with deep learning. *Nature*, 655(8121):210–218, June 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10674-6. URL http://dx.doi.org/10.1038/s41586-026-10674-6.

Latha P. Palaniappan, Norrina B. Allen, Zaid I. Almarzooq, Cheryl A.M. Anderson, Pankaj Arora, Christy L. Avery, Carissa M. Baker-Smith, Nisha Bansal, Maria E. Currie, Rebecca S. Earlie, Wenjun Fan, Jessica L. Fetterman, Bethany Barone Gibbs, Debra G. Heard, Swapnil Hiremath, Haoyun Hong, Hyacinth I. Hyacinth, Chinwe Ibeh, Tian Jiang, Michelle C. Johansen, Dhruv S. Kazi, Darae Ko, Tak W. Kwan, Michelle H. Leppert, Yilun Li, Jared W. Magnani, Karlyn A. Martin, Seth S. Martin, Erin D. Michos, Michael E. Mussolino, Oluwabunmi Ogungbe, Nisha I. Parikh, Marco V. Perez, Sarah M. Perman, Ashish Sarraju, Nilay S. Shah, Mellanie V. Springer, Marie-Pierre St-Onge, Evan L. Thacker, Seda Tierney, Sarah M. Urbut, Harriette G.C. Van Spall, Jenifer H. Voeks, Seamus P. Whelton, Sally S. Wong, Juan Zhao, and Sadiya S. Khan. 2026 heart disease and stroke statistics: A report of us and global data from the american heart association. *Circulation*, 153 (9), March 2026. ISSN 1524-4539. doi: 10.1161/cir.0000000000001412. URL http://dx.doi.org/10.1161/CIR.0000000000001412.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D Clifford, and Matthew A Reyna. Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020. *Physiological Measurement*, 41(12):124003, December 2020. ISSN 1361-6579. doi: 10.1088/1361-6579/abc960. URL http://dx.doi.org/10.1088/1361-6579/abc960.

Nicolas Pilia, Claudia Nagel, Gustavo Lenis, Silvia Becker, Olaf Dössel, and Axel Loewe. ECGdeli - an open source ECG delineation toolbox for MATLAB. *SoftwareX*, 13:100639, January 2021. ISSN 2352-7110. doi: 10.1016/j.softx.2020.100639. URL http://dx.doi.org/10.1016/j.softx.2020.100639.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. CODE-15%: a large scale annotated dataset of 12-lead ECGs, 2021. URL https://zenodo.org/doi/10.5281/zenodo.4916206.

Molly Sachdev, Barry J. Fetics, Shenghan Lai, Darshan Dalal, Jerald Insel, and Ronald D. Berger. Failure in short-term prediction of ventricular tachycardia and ventricular fibrillation from continuous electrocardiogram in intensive care unit patients. *Journal of Electrocardiology*, 43(5):400–407, September 2010. ISSN 0022-0736. doi: 10.1016/j.jelectrocard.2010.02.005. URL http://dx.doi.org/10.1016/j.jelectrocard.2010.02.005.

A. Sellergren et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. doi: 10.48550/arXiv.2507.05201. URL https://doi.org/10.48550/arXiv.2507.05201.

Supreeth P. Shashikumar, Sina Mohammadi, Rishivardhan Krishnamoorthy, Avi Patel, Gabriel Wardi, Joseph C. Ahn, Karandeep Singh, Eliah Aronoff-Spencer, and Shamim Nemati. Development and prospective implementation of a large language model based system for early sepsis prediction. *npj Digital Medicine*, 8(1), May 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01689-w. URL http://dx.doi.org/10.1038/s41746-025-01689-w.

S. Tan et al. Icentia11k single lead continuous raw electrocardiogram dataset, 2022. URL https://doi.org/10.13026/kk0v-r952.

Terence D. Valenzuela, Denise J. Roe, Shan Cretin, Daniel W. Spaite, and Mary P. Larsen. Estimating effectiveness of cardiac arrest interventions: A logistic regression survival model. *Circulation*, 96 (10):3308–3313, November 1997. ISSN 1524-4539. doi: 10.1161/01.cir.96.10.3308. URL http://dx.doi.org/10.1161/01.cir.96.10.3308.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), May 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6. URL http://dx.doi.org/10.1038/s41597-020-0495-6.

Lingxuan Zhu, Weiming Mou, Keren Wu, Yancheng Lai, Anqi Lin, Tao Yang, Jian Zhang, and Peng Luo. Multimodal chatgpt-4v for electrocardiogram interpretation: Promise and limitations. *Journal of Medical Internet Research*, 26:e54607, June 2024. ISSN 1438-8871. doi: 10.2196/54607. URL http://dx.doi.org/10.2196/54607.

Appendix A. PennVT Cohort

We collected in-hospital telemetry recordings from patients admitted to ICUs and cardiac care units at an academic health system at the University of Pennsylvania between October 2022 and January 2026. All clinically documented sustained VT and VF events during this period, identified through code calls and ICD-10 diagnostic codes, were linked to the corresponding telemetry archive. Two board-certified cardiologists adjudicated VT and VF onset times by consensus. Only the first adjudicated onset per patient was included, yielding 50 VT and 32 VF events. A total of 857 control recordings were sampled from patients in the same clinical units during the same study period and were confirmed to contain no VT or VF events.

Among these 939 patients, 21 patients were excluded because no eligible windows remained after signal-quality filtering , and 7 additional patients (1 VT, 1 VF, and 5 controls) were excluded because their recordings were shorter than 10 minutes and therefore were insufficient for evaluation. The final PennVT dataset comprised eight-lead (I, II, III, aVR, aVL, aVF, V1, and V3) recordings from 911 patients, including 80 VT/VF cases. Cohort characteristics are summarized in Table 1.

Appendix B. CAMEL

B.1. Architecture

CAMEL consists of two main components: (1) an ECG encoder that maps telemetry segments to vector representations, and (2) a LoRA-adapted pretrained LLM that jointly processes the ECG embeddings and text (Figure 5). CAMEL is built on MedGemma-4B-Instruct (Sellergren et al., 2025), leveraging its medical knowledge and language capabilities.

We first apply a standard 50/60 Hz notch filter and a 0.3 Hz high-pass filter and resample each telemetry recording to 256 Hz. To encode raw waveforms, we use a four-layer CNN, $f_\theta : \mathbb{R}^{256} \to \mathbb{R}^{64}$, that maps each single-lead, one-second ECG segment to a latent representation. A single-layer perceptron, $\mathrm{SLP}_\Phi : \mathbb{R}^{64} \to \mathbb{R}^{2560}$, then projects these representations into MedGemma's token embedding space.

CAMEL represents multi-lead telemetry as a structured sequence of ECG tokens. For an L-second recording with M leads, the projected one-second embeddings are serialized one lead at a time and passed to the LLM. Special lead-boundary tokens mark the beginning and end of each lead, preserving lead identity and delineating the corresponding ECG token sequence. ECG tokens can be interleaved with tokenized text, and multiple telemetry recordings and text segments can be included, enabling comparisons across recordings.

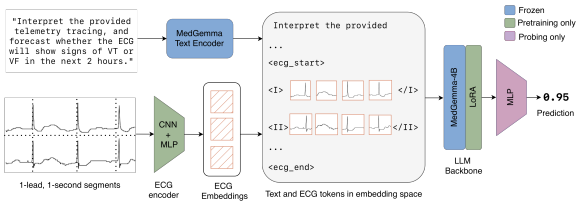


Figure 5: CAMEL architecture.

As leads are serialized sequentially, standard causal attention prevents tokens from earlier leads from attending to concurrent tokens in later leads. CAMEL therefore uses a custom attention mask that enables bidirectional attention among time-aligned ECG tokens across leads. ECG tokens may additionally attend to their corresponding lead-boundary tokens and preceding special tokens, while both start and end lead-boundary tokens may attend to the ECG tokens within their associated lead. This attention scheme enables cross-lead integration of temporally aligned information while preserving the causal structure required for autoregressive text generation.

B.2. Model Training

We employ a four-stage pretraining pipeline to learn ECG representations that support general ECG reasoning using recordings from ten public datasets. The ECG encoder is trained from scratch, while MedGemma is finetuned using LoRA adapters with rank $r = 128$.

In Stage 1, we use self-supervised learning to pretrain the CNN encoder. Given a one-second, single-lead ECG segment, the encoder–decoder pair is optimized to reconstruct the segment using MSE loss. Training uses approximately one billion segments drawn from 12.7 million ECG recordings (Moody and Mark, 2001; Huang et al., 2021; Bousseljot et al., 1995; Laguna et al., 1997; Liu et al., 2022; Perez Alday et al., 2020).

Stages 2–4 involve ECG–text instruction tuning with progressively complex and diverse tasks. In these stages, we connect the CNN encoder from Stage 1 to the LLM through an SLP layer and optimize the model using a masked autoregressive cross-entropy

Table 1: PennVT cohort characteristics.

By outcome	All	VT/VF	Control
Patients	911	80 (49 VT, 31 VF)	831
Total telemetry (h)	9,368	964	8,404
Recording duration per patient, median h [IQR]	9.2 [3.0, 15.8]	11.3 [3.0, 20.2]	8.9 [3.1, 15.3]
Time to VT/VF onset, median h [IQR]	—	5.6 [3.0, 10.7]	—

loss. We additionally apply lead-order shuffling and random lead dropping to improve robustness to variations in input lead configurations. For these stages, we use the AdamW optimizer with a linear learning-rate scheduler. Stage-specific hyperparameters are reported in Table 2.

Stage 2 focuses on ECG classification using 20 million short-answer and multiple-choice questions generated from the HEEDB dataset (Koscova et al., 2026). Stage 3 extends training to fine-grained ECG measurement and reasoning using 30 million questions involving extraction of amplitudes, intervals, and heart rate, as well as comparisons across multiple ECG tracings. These questions are generated from ECGDeli (Pilia et al., 2021)-derived measurements from HEEDB, CODE-15% (Ribeiro et al., 2021), PTB-XL (Wagner et al., 2020), and MIMIC-IV (Gow et al., 2023). Stage 4 further broadens instruction tuning to multi-turn clinical interactions using 420,000 dialogues containing one or two ECG tracings. These dialogues are generated using Gemini-3-Flash-Preview and span 28 task types and a diverse range of clinical interaction formats.

Following pretraining, we freeze all parameters and train a two-layer MLP on the last hidden states. We use the AdamW optimizer with a learning rate of 10^{-4}, weight decay of 0.01, and dropout of 0.05 with a weighted binary cross-entropy loss for up to 100 epochs with early stopping after 10 epochs without improvement.

Appendix C. Evaluation Setup

We use five-fold cross-validation with patient-level partitioning. Each fold serves once as the held-out test set (20% of patients), with 70% of patients used for training and 10% for validation. For each fold, we select the highest threshold τ that achieves at least 80% event sensitivity on the pooled validation predictions, excluding that fold's test patients. Test predictions from all five folds are then pooled for final evaluation. We compute 95% confidence intervals using a patient-level bias-corrected and accelerated (BCa) bootstrap with 1,000 resamples, preserving within-patient correlation.

Table 2: Hyperparameters for training stages 2–4.

Stage	Learning Rate	Epochs	Batch Size	Warmup Steps
Stage 2	1×10^{-4}	3	8	500
Stage 3	1×10^{-4}	1	8	500
Stage 4	1×10^{-4}	9	4	256

Prediction time points are sampled backward from 10 minutes before VT/VF onset for positive patients and from the end of the randomly selected interval for control patients. Only time points with a complete l-second lookback window are included. Windows containing persistent (> 5 s) zero-valued or missing data are excluded, and remaining missing values are zero-imputed. We further exclude windows in which all available leads are classified as noisy for more than 40% of the duration.

We train random forest, XGBoost, and logistic regression models using median imputation, with missingness indicators included as additional features. The random forest uses 400 trees, unrestricted maximum depth, and a minimum of two samples per leaf. Logistic regression uses ℓ_2 regularization with $C = 1$ and at most 1,000 iterations. XGBoost uses 300 trees, a maximum depth of 5, and a learning rate of 0.1. For each of five individual measurements, we additionally train separate random forest models using absolute values, deviations from the patient's earliest recording, percentiles among the patient's previous windows, or trailing slopes as features.

The CART score (Churpek et al., 2012) is an early warning score for cardiac arrest, death, and ICU transfer. It is computed from age and measurements of diastolic blood pressure, heart rate, and respiratory rate, which are recorded approximately hourly in PennVT.

For architectures trained from scratch, we use a three-layer CNN, a TCN, and a four-layer Trans-

Table 3: CAMEL performance for VT and VF separately, with 95% confidence intervals.

	Event count	AUROC	AUPRC	Event sensitivity
VT	49	0.870 [0.824–0.904]	0.245 [0.168–0.345]	65.3% [50.0–76.9]
VF	31	0.788 [0.681–0.855]	0.140 [0.070–0.275]	45.2% [27.7–62.1]

Table 4: Baseline ECG measurements (median [IQR]) and signal quality (mean) in controls with and without CAMEL false alarms. P values are from two-sided Mann–Whitney U tests.

Measurement	Controls with ≥ 1 false alarm	Controls without false alarms	P value
Baseline QRS duration (ms)	117.2 [101.6–138.7]	93.8 [85.9–117.2]	1.3×10^{-8}
Baseline QTc interval (ms)	477.7 [420.4–519.0]	440.9 [397.1–476.5]	5.1×10^{-4}
Baseline heart rate (bpm)	87.3 [70.7–108.0]	82.1 [71.4–93.7]	0.049
Any noise flag (%)	0.860	1.870	0.617
Flatline (%)	0.844	1.774	0.305
Powerline artifact (%)	0.001	0.001	0.551
High-frequency artifact (%)	0.016	0.095	0.865

former with four attention heads per layer. TCN and Transformer are each jointly trained with a four-layer CNN tokenizer for encoding 10-second ECG segments. They are trained using AdamW with batch size 16 and weight decay of 0.01. We use a learning rate of 1×10^{-4} for CNN and 3×10^{-4} for TCN and Transformer. For OpenTSLM, we use the same probing configuration as CAMEL. For xECG and ECG-CPC, we follow their original implementations.

For matched-sensitivity comparisons, we select alarm thresholds post-hoc on held-out predictions. We choose the highest threshold that attains event sensitivity at least as high as CAMEL's. The achieved sensitivities are 57.9% for CART, 60.0% for QRS duration, and 57.5% for the remaining models.

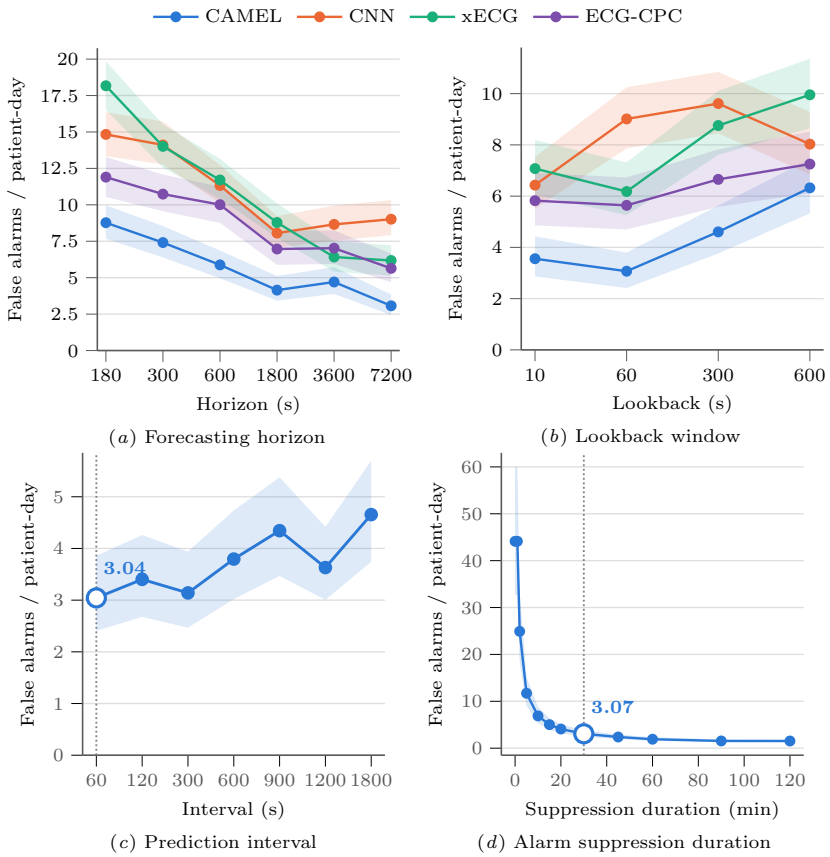
Appendix D. Additional Experimental Results

We report performance for VT and VF separately in Table 3. The false alarms generated by CAMEL on control patients are associated with wider baseline QRS durations and longer QTc intervals and modestly with higher heart rate, but not with noise levels (Table 4). We compare the risk predicted by CAMEL over time for VT/VF and control patients in Table 5.

We vary each parameter (l, h, w, s) while holding the other three at their primary values. For forecasting horizons shorter than 1 hour, we exclude predictions within the last 1 minute, instead of 10 minutes. False alarm burden generally decreases with longer horizons h (Figure 6(a)). Longer horizons relax the requirement for precise temporal forecasting and allow a broader future window of event onset, progressively shifting the task toward risk stratification.

A lookback of $l = 60$ seconds yields the lowest false alarm burden, while 10 seconds performs similarly (Figure 6(b)). Performance declines beyond 60 seconds. Shorter intervals generally produce better alarm-level performance, with $w = 60$ seconds yielding the lowest false alarm burden (Figure 6(c)) on a matched subset of patients. A 300-second interval performs similarly, whereas longer intervals are less favorable. Finally, increasing the alarm-suppression duration s sharply reduces alarm burden, with most of the reduction occurring within the first 30 minutes and only modest gains thereafter (Figure 6(d)). At $s = 30$ minutes, false alarms decrease from 44.13 to 3.07 per patient-day compared with no suppression.

Tables 6 and 7 report the performance of ECG measurement-based models. Figure 6 and Table 8 compare alternative waveform-based models across lookback windows and forecasting horizons. OpenTSLM is omitted from these sweeps because it is pretrained only on 10-second ECG recordings.



(a) Forecasting horizon (b) Lookback window

(c) Prediction interval (d) Alarm suppression duration

Figure 6: False alarm burden across forecasting configurations. Parameters not varied are fixed at their primary values: $l = 60$ seconds, $h = 2$ hours, $w = 1$ minute, and $s = 30$ minutes. Shading shows 95% confidence intervals. Prediction interval (w) comparisons include a subset of patients with at least 30 minutes of continuous telemetry.

Table 5: CAMEL risk trajectories 130–10 minutes before VT/VF onset for positive patients and before the sampled endpoint for control patients Max columns report the percentage of patients with at least the indicated number of consecutive above-threshold predictions. Results use 565 patients (66 VT/VF, 499 controls) with at least 130 minutes of telemetry.

(a) VT/VF patients

Minutes before event	Max ≥ 1	Max ≥ 2	Max ≥ 5	Max ≥ 10	Mean risk	Median risk
100–130	37.9%	34.8%	25.8%	15.2%	0.628	0.680
70–100	34.8%	31.8%	27.3%	18.2%	0.632	0.691
40–70	39.4%	36.4%	25.8%	19.7%	0.644	0.687
10–40	48.5%	39.4%	30.3%	24.2%	0.666	0.710

(b) Control patients

Minutes before interval end	Max ≥ 1	Max ≥ 2	Max ≥ 5	Max ≥ 10	Mean risk	Median risk
100–130	5.2%	3.8%	2.4%	1.8%	0.285	0.219
70–100	5.8%	4.0%	2.8%	1.6%	0.285	0.219
40–70	5.8%	4.2%	2.8%	2.4%	0.287	0.222
10–40	5.2%	4.2%	3.4%	2.8%	0.287	0.223

Table 6: Performance of ECG measurement-based models with 95% confidence intervals. Each model uses ten features: QRS duration, QT interval, QTc, PR interval, T-peak-to-T-end (Tp–Te) interval, Tp–Te/QT ratio, root mean square of successive differences (RMSSD), standard deviation of normal-to-normal intervals (SDNN), RR-interval interquartile range (IQR), and heart rate.

	Random Forest	**XGBoost**	**Logistic Regression**
AUROC	0.749 [0.699–0.789]	0.703 [0.653–0.749]	0.699 [0.643–0.753]
AUPRC	0.181 [0.136–0.240]	0.159 [0.118–0.213]	0.153 [0.112–0.210]

Table 7: AUROC and AUPRC with 95% confidence intervals for random forest models trained on individual ECG measurements and temporal features. Trailing slopes are computed across the five most recent prediction windows.

Measurement	**Metric**	**Absolute value**	**Deviation from own baseline**	**Percentile in patient history**	**Trailing slope**
QRS duration	AUROC	0.661 [0.596–0.713]	0.530 [0.463–0.588]	0.592 [0.547–0.629]	0.509 [0.490–0.526]
	AUPRC	0.130 [0.092–0.174]	0.095 [0.069–0.136]	0.090 [0.070–0.114]	0.085 [0.066–0.106]
QTc (Bazett)	AUROC	0.641 [0.584–0.698]	0.520 [0.457–0.579]	0.500 [0.469–0.532]	0.500 [0.494–0.507]
	AUPRC	0.122 [0.088–0.164]	0.091 [0.066–0.126]	0.071 [0.056–0.088]	0.083 [0.064–0.104]
Heart rate	AUROC	0.605 [0.535–0.664]	0.520 [0.455–0.580]	0.512 [0.465–0.552]	0.505 [0.498–0.513]
	AUPRC	0.116 [0.082–0.165]	0.084 [0.059–0.111]	0.075 [0.057–0.096]	0.074 [0.059–0.092]
RMSSD	AUROC	0.570 [0.515–0.627]	0.524 [0.461–0.584]	0.520 [0.495–0.546]	0.506 [0.501–0.510]
	AUPRC	0.092 [0.067–0.122]	0.104 [0.073–0.156]	0.076 [0.060–0.094]	0.082 [0.065–0.102]
PR interval	AUROC	0.535 [0.481–0.598]	0.505 [0.441–0.562]	0.552 [0.510–0.586]	0.502 [0.495–0.510]
	AUPRC	0.079 [0.060–0.105]	0.084 [0.061–0.115]	0.080 [0.063–0.100]	0.078 [0.061–0.097]

Table 8: Performance with 95% confidence intervals for CAMEL and for TCN and Transformer models trained from scratch across varying lookback windows l.

Model	Metric	10 s	1 min	5 min	10 min
TCN	AUROC	0.736 [0.677–0.788]	0.667 [0.601–0.725]	0.658 [0.590–0.731]	0.643 [0.571–0.701]
	AUPRC	0.209 [0.152–0.278]	0.175 [0.130–0.237]	0.195 [0.138–0.263]	0.170 [0.118–0.242]
Transformer	AUROC	0.729 [0.662–0.781]	0.740 [0.683–0.789]	0.664 [0.592–0.727]	0.722 [0.668–0.776]
	AUPRC	0.200 [0.151–0.261]	0.207 [0.148–0.270]	0.176 [0.125–0.235]	0.200 [0.140–0.272]
CAMEL	AUROC	0.840 [0.797–0.875]	0.832 [0.787–0.869]	0.815 [0.766–0.851]	0.799 [0.755–0.838]
	AUPRC	0.306 [0.231–0.385]	0.298 [0.223–0.388]	0.254 [0.188–0.325]	0.237 [0.171–0.317]