Early Warning of In-Hospital Cardiac Arrest from Photoplethysmography Using Deep Residual Networks

Somin Mindy Lee
Seoul National University Hospital
Seoul, South Korea
somindy.lee@mail.utoronto.ca

Hyeonhoon Lee
Seoul National University Hospital
Seoul, South Korea
hhoon@snu.ac.kr

Hyung-Chul Lee
Seoul National University Hospital
Seoul, South Korea
vital@snu.ac.kr

Abstract

Early detection of in-hospital cardiac arrest remains a critical challenge for improving patient outcomes. We propose a deep learning framework that leverages continuous photoplethysmography (PPG) signals to predict cardiac arrest within a 24-hour window. We used the SCOPE dataset, a recently released collection that includes PPG waveforms from 4,517 ICU admissions across 3,785 patients at Seoul National University Hospital. A residual 1D convolutional neural network was trained on 5-minute PPG segments sampled at 125 Hz and evaluated using patient-level stratified 5-fold cross-validation. We benchmarked the model against a logistic regression baseline and a deeper ResNet variant, and performed ablations over class imbalance, windowing strategies, and model capacity. The model achieved strong discrimination with both AUROC and AUPRC, demonstrating that PPG signals contain predictive signatures of impending deterioration. These findings highlight the feasibility of non-invasive waveform-based risk prediction, and position PPG monitoring as a promising biomarker in critical care.

1 Introduction

Cardiac arrest is among the most severe clinical emergencies, with survival rates below 25% despite advances in monitoring and resuscitation (1; 2; 3). Early identification of patients at risk is critical, as interventions before arrest significantly improve outcomes (4). Conventional early warning systems such as the Modified (MEWS) and the National Early Warning Score (NEWS) rely on manual and sparse vital signs and struggle to detect subtle changes that precede cardiac arrest (5). Although NEWS has stratified risk up to 24 hours prior to in-hospital cardiac arrest, its discriminative ability is limited (AUC of 0.58–0.64) (6). These limitations show the need for continuous, data-driven monitoring to detect physiologic deterioration earlier.

PPG, obtained through pulse oximetry, represents a promising signal for this task. PPG is non-invasive, low-cost, and already collected continuously in many hospital settings and wearable devices. Beyond heart rate and oxygen saturation, PPG captures beat-to-beat variability and morphological changes reflecting circulatory and autonomic status (7)(8). Prior studies have demonstrated associations between PPG features and cardiovascular instability (9), suggesting that PPG may contain early signs of clinical deterioration that are invisible to current early warning scores. Despite this, relatively little work has examined the use of continuous PPG time series for predicting in-hospital cardiac arrest.

Deep learning has transformed time-series analysis in healthcare, with CNNs and residual architectures showing strong performance on ECG, EEG, and other biosignals (10; 11; 12). Residual networks in particular allow efficient training of deep architectures while preserving temporal detail, making them well-suited for noisy physiological data. In this study, we investigate whether deep

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

residual networks applied directly to raw PPG signals can provide early warning of in-hospital cardiac arrest.

2 Methods

We constructed an experimental cohort by linking clinical outcomes with the full set of cached continuous PPG recordings sampled at 125 Hz. The dataset contained 71 cardiac arrest patients and more than 3,500 non-arrest patients. For model development, we retained all eligible non-arrest patients for the main class-weighted analysis, and additionally created balanced and semi-balanced subsets (1:1, 1:2) for controlled ablation experiments. Class imbalance was addressed using class-weighted loss and patient-level stratified sampling, ensuring that no waveform segments from the same patient appeared in both training and evaluation (Figure 1).

Continuous PPG recordings were segmented into five-minute windows, with or without 50% overlap. For cardiac arrest patients, only the last 24 h before the event were used; for non-arrest patients, only the first 24 h of admission. Segments were normalized (zero mean, unit variance; near-zero variance handled by mean subtraction) and missing values were replaced with zero. Each window received a binary label indicating whether it fell within the predictive horizon.

We implemented a 1D residual neural network in PyTorch with three residual blocks (16, 32, 64 channels) including convolution, batch normalization, and ReLU, using projection shortcuts when dimensions changed. The network ended with adaptive average pooling, a fully connected layer, and sigmoid activation to output the probability of cardiac arrest.

To contextualize the performance of our 1D ResNet, we implemented a simple logistic regression model using handcrafted PPG morphology features extracted from each prediction window. For every 5-minute segment, we computed mean amplitude, standard deviation (beat-to-beat variability), peak-to-peak range, and an approximate heart rate estimated from peak counts. These features were z-scored and used to train a logistic regression classifier with balanced class weights. This baseline allows us to quantify the value of learning directly from raw waveforms compared to using simple interpretable summary statistics.

We trained a deeper variant of the 1D ResNet with additional residual blocks to assess whether increasing model capacity yields improved performance. to test the architectural simplicity of the baseline model.

Models were trained for 10-20 epochs using Adam (lr = 1e-3), batch size 64, and binary cross-entropy loss. Evaluation followed patient-level five-fold stratified cross-validation, ensuring windows from the same patient remained in a single fold. Performance was assessed using AUROC and AUPRC on held-out patients.

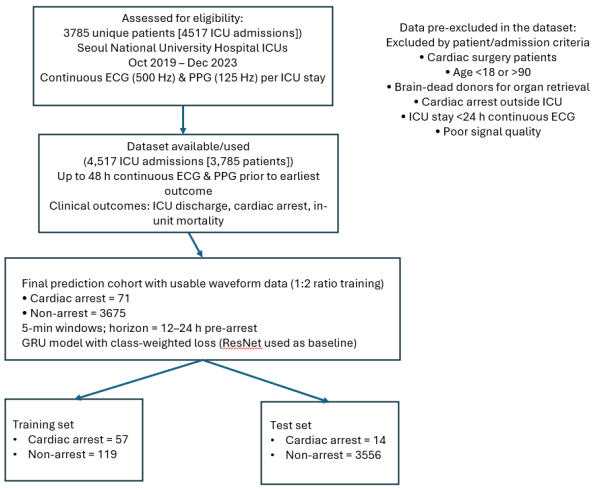

Figure 1: Flowchart of Study Cohort.

3 Results

Table 1: 24 Hr Horizon, 50% windows overlap, 1:2 dataset training

Stage	AUROC	AUPRC
Fold_1	0.892	0.890
Fold_2	0.892	0.893
Fold_3	0,890	0.889
Fold_4	0,865	0.870
Fold_5	0.894	0.883
Mean	0.888 ± 0.011	0.885 ± 0.008

Table 2: Comparing Different Models

Model	AUROC	AUPRC
Logistical Regression	0.616	0.386
Shallow ResNet (3 blocks)	0.888 ± 0.011	0.885 ± 0.008
Deeper ResNet (4 blocks)	0.841 ± 0.071	0.716 ± 0.105

Fold-wise AUROC and AUPRC values are reported in Table 1. In this setting, the model was trained on a 1:2 balanced subset but evaluated on the more realistic imbalanced patient distribution. The similar performance across all five folds indicates that the model generalizes well despite being trained on a balanced cohort and tested on a distribution that more closely reflects real-world imbalance.

Figure 2 presents the ROC and precision–recall curves for a 12-hour horizon, no overlap, and 1:2 cardiac arrest to non prediction. The model achieves strong discriminative performance (AUROC = 0.833) and maintains high precision (AUPRC = 0.831), despite the class imbalance. Figure 3 shows the corresponding curves for a 24-hour horizon, where ROC performance remains similar, but precision drops more notably at lower recall levels reflecting, as expected, increased uncertainty for longer-range predictions. Figure 4 displays the ROC curves stratified by fold for the 24-hour horizon, illustrating trends across cross-validation splits.

In additional (non-CV) sensitivity analyses, we varied the fixed train–test split (80/20 vs. 70/30), windowing strategy (300s/300s non-overlapping vs. 300s/150s overlapping), prediction horizon (24h vs. 48h), and class ratio (1:1 vs. 1:2, via down-sampling of non-arrest windows). These ablations yielded results that were qualitatively consistent with those from cross-validation. Among the variations tested, overlapping windows, a 1:1 class ratio, and a shorter prediction horizon (24h) led to modest improvements in both AUROC and AUPRC, indicating that these choices may enhance early warning performance.

Table 2 compares the performance of the logistic regression baseline with two neural architectures. The logistic regression model, which uses only four handcrafted morphology features, achieved modest discrimination (AUROC 0.616; AUPRC 0.386). In contrast, the shallow 1D ResNet substantially outperformed the baseline, yielding the highest and most consistent results across folds (AUROC 0.888 ± 0.011; AUPRC 0.885 ± 0.008). Surprisingly, increasing model depth to a four-block ResNet reduced both average performance (AUROC 0.841; AUPRC 0.716) and stability (std = 0.071 AUROC), suggesting that the predictive signal in PPG waveforms may be captured primarily by lower-level temporal features rather than deeper hierarchical representations.

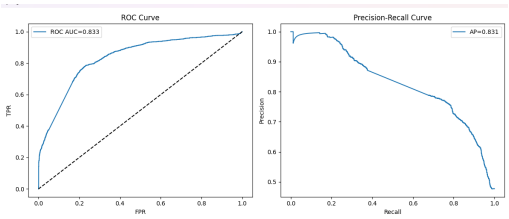

Figure 2: 12 Hr Horizon, No windows overlap, 1:2 dataset

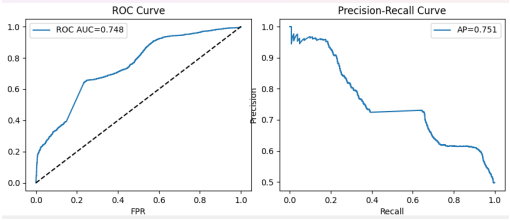

Figure 3: 24 Hr Horizon, No windows overlap, 1:2 dataset.

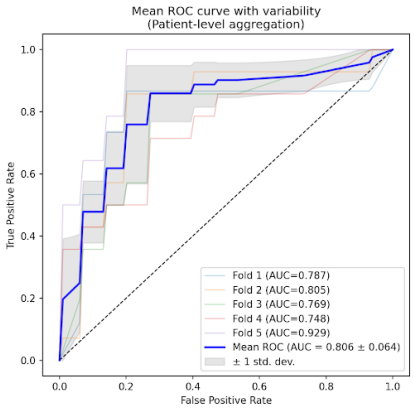

Figure 4: 24 Hr Horizon, 50% windows overlap, 1:2 dataset ROC curve full.

4 Discussion

Our findings demonstrate that deep residual networks trained on PPG segments can extract signals for early warning of in-hospital cardiac arrest. Importantly, evaluation was performed using patient-level cross-validation, which mitigates the optimistic bias that occurs when windows from the same patient appear in both training and validation. The stability of AUROC and AUPRC across folds suggests that the model is learning features that generalize beyond individual patients.

Our dataset contained 74 cardiac arrest cases within a much larger cohort of non-arrest patients. For computational efficiency and to perform controlled imbalance ablations, we created balanced and semi-balanced training subsets (e.g., 1:1 or 1:2). While this yields a more balanced cohort and the subsampled training improves stability under limited positive cases, it does restrict exposure to the full range of control-patient variability. We also constrained the prediction horizon to 24 hours before arrest and limited control segments to the first 24 hours of admission. While this simplification aids model learning, it may not capture the full diversity of clinical trajectories, including gradual deterioration or recurrent instability.

The current model also relies on a relatively shallow architecture and short fixed-length windows. More expressive architectures such as temporal transformers, diffusion models, or self-supervised pretraining approaches could capture longer-term dependencies and subtle temporal trends in the data. Finally, while AUROC and AUPRC provide useful measures of discrimination, clinical deployment would require additional evaluation of calibration, interpretability, and fairness across subgroups.

We evaluated two additional baselines: a logistic-regression classifier and a deeper convolutional network variant. The logistic-regression model represents a simple physiologic-feature baseline, while the deeper network tests whether additional depth meaningfully improves representation learning on short PPG segments. Across experiments, the logistic-regression baseline achieved substantially lower AUROC/AUPRC, underscoring that simple morphology trends alone are insufficient for early cardiac-arrest prediction. The deeper CNN improved window-level fitting but did not consistently outperform the original 1D-ResNet, suggesting that extremely deep temporal models may overfit the limited number of positive arrests available in SCOPE. These results highlight that (1) PPG indeed

contains physiologically meaningful early-warning structure, but (2) large, highly expressive models do not necessarily translate to better patient-level generalization under extreme class imbalance.

Future work should validate these methods in larger multi-institutional cohorts, explore multimodal integration with ECG or electronic health records, and investigate model interpretability techniques like saliency maps to identify physiologic correlates of risk. By grounding early warning in high-frequency physiological signals, this approach may complement existing early warning scores.

4.1 Morphology Analysis

To better understand which waveform segments influenced the model's decisions, we analyzed simple beat-to-beat morphology features extracted from each validation window and examined how they correlated with the model's predicted risk. (15)

Our morphology analysis showed that the deep model's predicted risk was weakly correlated with basic waveform descriptors, suggesting that the CNN was not relying on any single handcrafted feature, but rather on more complex combinations of morphology and temporal patterns. Correlation coefficients between predicted risk and morphology features ranged from –0.18 to –0.01, indicating small but interpretable trends. High-risk windows tended to show lower amplitude variability (stdamp: 0.63 vs. 1.00), reduced peak-to-peak amplitude (6.29 vs. 10.61), and lower estimated heart rate (59 vs. 82 bpm). These patterns are consistent with physiologic deterioration preceding cardiac arrest, where PPG signals become damped, low-variability, or noisy due to peripheral vasoconstriction, hypoperfusion, or irregular circulation. Importantly, the model learned to associate these subtle reductions in pulsatility and slope irregularity with increased risk, even though these features were not explicitly engineered.

4.2 Related Works

Kataria et al. developed PPG-GPT (13), a large pre-trained foundation model trained on over 200 million 30-second PPG segments. Their Feature Extractor–Aggregator Network leverages representations from PPG-GPT variants up to 1 billion parameters, achieving an AUROC of 0.79 over a 24-hour horizon peaking at 0.82 one hour before cardiac arrest. These results highlight the potential of large-scale pretraining to extract rich PPG representations for downstream prediction tasks. In comparison, our lightweight 1D ResNet achieved an AUROC of 0.888 over 24 hours using only supervised training on the smaller SCOPE dataset, showing that carefully designed task-specific models can still outperform foundation-model-based pipelines on domain-specific prediction tasks.

Park et al. incorporated multimodal EMR features to achieve strong discriminative performance for in-hospital cardiac arrest prediction (14). Lee et al. combined baseline features with LSTM-modeled vital signs, obtaining AUROC of 0.91 up to 13 hours in advance using MIMIC-IV, with 89% detection one hour prior (16). Lu et al. proposed PedCA-FT, a transformer architecture that fuses tabular and textual pediatric EHR features, outperforming ten baselines (18). These multimodal approaches demonstrate that integrating diverse clinical data sources can improve accuracy but require broad data availability and substantial integration infrastructure. Notably, our unimodal PPG-based approach obtained comparable AUROC using waveform data alone, underscoring the predictive richness of continuous physiological signals where such monitoring is already routine.

Unlike multimodal or foundation-model-based systems, our method uses a compact 3-block 1D ResNet trained directly on short sliding-window segments without handcrafted features or auxiliary clinical inputs. We performed controlled experiments across window sizes, prediction horizons, and class imbalance strategies, and enforced strict patient-level stratified 5-fold cross-validation to avoid data leakage and optimistic bias. This design enables efficient deployment with minimal data dependencies while providing realistic performance estimates under the severe class imbalance characteristic of cardiac arrest datasets (17).

5 Conclusion

We presented a residual neural network for early warning of in-hospital cardiac arrest using only PPG time series. By segmenting continuous PPG recordings into fixed windows and restricting positive samples to the 24-hour pre-arrest horizon, we framed the task as a patient-level binary pre-

diction problem. Training with five-fold stratified cross-validation demonstrated that the model could distinguish arrest from non-arrest cases using waveform-derived features alone. These results indicate that routinely collected PPG signals contain predictive information about clinical deterioration, and that residual networks can effectively extract these patterns. Future work will involve scaling to larger and more diverse datasets, incorporating additional modalities such as ECG or EHR-derived features, and extending to self-supervised representation learning for improved generalization.

6 Acknowledgments

This work was supported by the Digital Medical Products Development Based on Medical Data Synthesis and AI Technologies Program(RS-2025-02304834,Development of On-Device AI DIgital Medical Products using Structured Synthetic Medical Data Technology) funded by the Ministry of Trade, Industry and Resources (MOTIR) of Korea

References

[1] Abella, B.S. Not all cardiac arrests are the same. CMAJ. 2011;183(14):1572-1573. doi:10.1503/cmaj.110982

[2] Andersen, L.W., Holmberg, M.J., Berg, K.M., Donnino, M.W. & Granfeldt, A. (2019) In-Hospital Cardiac Arrest: A Review. *JAMA – Journal of the American Medical Association* **321**(12):1200-1210.

[3] Ricceri, S., Salazar, J.W., Vu, A., Vittinghoff, E., Moffatt, E. & Tseng, Z.H. (2021) Factors Predisposing to Survival after Resuscitation for Sudden Cardiac Arrest. *Journal of the American College of Cardiology* **77**(19):2353-2362.

[4] Chugh, S.S. (2010) Early identification of risk factors for sudden cardiac death. *Nature Reviews Cardiology* **7**(6):318-326.

[5] Mathukia, C., Fan, W.Q., Vadyak, K., Biege, C. & Krishnamurthy, M. (2015) Modified Early Warning System improves patient safety and clinical outcomes in an academic community hospital. *Journal of Community Hospital Internal Medicine Perspectives* **5**(2):26716. doi:10.3402/jchimp.v5.26716.

[6] Spångfors M, Molt M, Samuelson K. In-hospital cardiac arrest and preceding National Early Warning Score (NEWS): A retrospective case-control study. Clin Med (Lond). 2020;20(1):55-60. doi:10.7861/clinmed.2019-0137

[7] Tamura, T. (2019) Current progress of photoplethysmography and SPO for monitoring and screening: a review. *Journal of Medical and Biological Engineering* **39**(6):845-851. doi:10.1007/s40846-019-00488-3.

[8] Almarshad, M.A., Islam, M.S., Al-Ahmadi, S. & BaHammam, A.S. (2022) Diagnostic Features and Potential Applications of PPG Signal in Healthcare: A Systematic Review. *Healthcare (Basel)* **10**(3):547. doi:10.3390/healthcare10030547.

[9] Karimpour, P., May, J.M. & Kyriacou, P.A. (2023) Photoplethysmography for the assessment of arterial stiffness. *Sensors (Basel)* **23**(24):9882. doi:10.3390/s23249882.

[10] Petmezas, G., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J.A., Katsaggelos, A.K. & Maglaveras, N. (2022) State-of-the-art deep learning methods on electrocardiogram data: systematic review. *JMIR Medical Informatics* **10**(8):e38454. doi:10.2196/38454.

[11] Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H. & Faubert, J. (2019) Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering* **16**(5):051001. doi:10.1088/1741-2552/ab260c.

[12] Ganapathy, N., Swaminathan, R. & Deserno, T.M. (2018) Deep learning on 1-D biosignals: a taxonomy-based survey. *Yearbook of Medical Informatics* **27**(1):98. doi:10.1055/s-0038-1667083.

[13] Kataria, S., Xiao, R., Ruchti, T., Clark, M.T., Lu, J., Lee, R.J., Grunwell, J. & Hu, X. (2025) Continuous cardiac arrest prediction in ICU using PPG foundation model. *arXiv preprint* arXiv:2502.08612.

[14] Park, H. & Park, C.S. (2025) A machine learning approach for predicting in-hospital cardiac arrest using single-day vital signs, laboratory test results, and International Classification of Disease-10 block for diagnosis. *Annals of Laboratory Medicine* **45**(2):209-217. doi:10.3343/alm.2024.0315.

[15] Elgendi M. On the analysis of fingertip photoplethysmogram signals. Curr Cardiol Rev. 2012 Feb;8(1):14-25. doi: 10.2174/157340312801215782. PMID: 22845812; PMCID: PMC3394104.

[16] Lee, H., Yang, HL., Ryu, H.G. et al. Real-time machine learning model to predict in-hospital cardiac arrest using heart rate variability in ICU. npj Digit. Med. 6, 215 (2023). https://doi.org/10.1038/s41746-023-00960-2

[17] Penketh, J., Nolan, J.P. In-hospital cardiac arrest: the state of the art. Crit Care 26, 376 (2022). https://doi.org/10.1186/s13054-022-04247-y

[18] Lu, J., Brown, S.R., Wu, Y., Dong, K., Bold, D., Liu, D., Grunwell, J.R. Hu, X. (2025) Early Risk Prediction of Pediatric Cardiac Arrest from Electronic Health Records via Multimodal Fused Transformer. *arXiv preprint* arXiv:2502.07158.Retry