

Physiology-augmented deep representations improve cross-database electrocardiogram beat classification under unseen-domain shift

Mujahid Shahid

mshahid1@st.knust.edu.gh

Kwame Nkrumah University of Science and Technology

Oliver Konyo

Kwame Nkrumah University of Science and Technology

Article

Keywords: electrocardiogram classification, deep learning, domain generalization, physiological features, cross-database evaluation, arrhythmia detection

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11024765/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Physiology-augmented deep representations improve cross-database electrocardiogram beat classification under unseen-domain shift

Mujahid Shahid^{1,*}, Dr Oliver Konyo¹

¹ *Department of Computer Science, College of Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.*

** Corresponding author. Correspondence and requests for materials should be addressed to M.S.*

Abstract

Cardiovascular disease is the leading cause of death worldwide, and automated electrocardiogram analysis increasingly relies on deep neural networks trained within a single database. Such models can perform well within their development data yet degrade when applied to recordings from a different institution, device or population, a form of cross-database distribution shift. We asked whether physiological descriptors, obtained from a validated waveform delineation procedure, could complement a frozen deep representation and improve five-class AAMI heartbeat classification when a source-trained model is transferred, without target-domain adaptation, to a completely unseen external database. Using the MIT-BIH Arrhythmia Database for source development, the Lobachevsky University Database for delineation validation, and the INCART Database as an unseen external target, we combined a 256-dimensional deep embedding with a 40-dimensional physiological representation in a capacity-matched fusion classifier. On the external database, the hybrid representation achieved a macro-F1 of 0.433 versus 0.392 for the deep representation alone, a difference of +0.041 supported by a 95% bootstrap confidence interval of 0.016–0.057 and a randomization P value of approximately 0.0001; the advantage persisted after excluding a rare class and across most training seeds. These findings indicate physiological information provides complementary, partially domain-stable signal beyond deep features for cross-database classification.

Keywords: electrocardiogram classification; deep learning; domain generalization; physiological features; cross-database evaluation; arrhythmia detection

Introduction

Cardiovascular disease remains the leading cause of death worldwide¹, and the electrocardiogram (ECG) is the most widely used non-invasive tool for detecting abnormalities of cardiac rhythm, conduction and repolarization. The growing volume of ambulatory, wearable and Holter ECG data has made automated beat classification essential for timely clinical interpretation^{2,3}. Early systems relied on manually engineered morphological and interval features^{4,5}; later work refined feature selection and inter-patient evaluation^{6,8}. Deep convolutional networks then replaced hand-crafted descriptors with representations learned directly from the waveform^{9,11}, scaling to cardiologist-level and multi-million-recording studies^{12,15}.

A strength within one database, however, does not guarantee performance on another. ECG databases differ in patient population, instrumentation, sampling rate, lead configuration, noise and annotation convention, producing a distribution shift between the data used for model development and the data encountered later¹⁶. This is consistent with learning theory showing that target-domain error depends on both source-domain

error and the source–target discrepancy¹⁶, and it has motivated a family of domain-adaptation and domain-generalization methods^{17,21}. Because a deployed ECG classifier may eventually meet a hospital, device or population never seen during development, domain generalization — improving performance without access to the target domain during development — is the more clinically relevant setting than domain adaptation, which assumes some, even unlabelled, access to target data^{19,20}.

Cross-database ECG classification has begun to receive direct attention. Pyramid-style networks²², explicit-feature approaches²³ and domain-generalization benchmarks²⁴ have all reported substantial degradation when a classifier is transferred across databases; an unsupervised domain-adaptation approach combining deep and temporal features has shown that the two feature types can be complementary, but it uses target-domain information during adaptation²⁵, leaving open whether a strictly source-trained model can generalize with no such access at all. Most recently, hybrid architectures that pair a morphology-sensitive convolutional branch with a context-sensitive attention branch have themselves been described as physiology-guided²⁶, and cross-domain feature-alignment methods have reported substantially higher performance on INCART than earlier approaches²⁷; however, these results were obtained either under a stratified, non-patient-independent, single-database split rather than a genuine cross-database transfer²⁶, or with some form of access to the target domain during development^{25,27}, so neither establishes whether an explicit, source-only physiological representation improves generalization under a true zero-shot cross-database transfer.

A second body of work is a prerequisite for any explicit physiological representation: automated waveform delineation. Clinically meaningful PR, QRS, QT, ST and related intervals cannot be measured unless P, QRS and T wave boundaries are localized accurately^{28,29}; the availability of an expert-annotated waveform-landmark database now permits independent validation of a delineation procedure before its output is used for classification³⁰.

This study asks whether a learned deep ECG representation and an explicit, independently validated physiological representation of rhythm, timing, morphology and delineation confidence describe complementary aspects of the same heartbeat, such that combining them improves classification after a genuine, unseen cross-database transfer. We use the MIT-BIH Arrhythmia Database³¹ as the source database for model development and patient-independent internal testing, the Lobachevsky University Electrocardiography Database (LUDB)³⁰ exclusively to develop and validate a continuous-signal P/QRS/T delineation procedure, and the St Petersburg INCART 12-lead Arrhythmia Database³² as a completely unseen external target, evaluated only once, with no retraining, threshold tuning or normalization fitted on INCART data. Classification follows the five AAMI heartbeat superclasses N, S, V, F and Q^{4,5}. We compare a capacity-matched deep-control representation, a physiology-only representation and a hybrid representation combining both, and we report patient-level statistical inference, robustness across training seeds, a rare-class sensitivity analysis and a semantic-group occlusion analysis of the physiological features actually used by the frozen classifier.

Methods

Study design and database roles

Three publicly available ECG databases were assigned non-overlapping methodological roles to prevent information leakage. The MIT-BIH Arrhythmia Database³¹ served as the source database for representation development, model selection and patient-independent internal testing. The Lobachevsky University Electrocardiography Database (LUDB)³⁰ was used exclusively, and independently of the classification task, to develop and validate the continuous-signal P/QRS/T delineation procedure. The St Petersburg INCART 12-lead Arrhythmia Database³² served as the unseen external target domain and was not used, at any stage, to fit encoder weights, physiological preprocessing statistics, classifier weights, decision thresholds or the training seed. Figure 1 summarizes the overall framework, including the two parallel representation pathways, the three representation conditions, and the separation of development, delineation-validation and external-evaluation roles across the three databases.

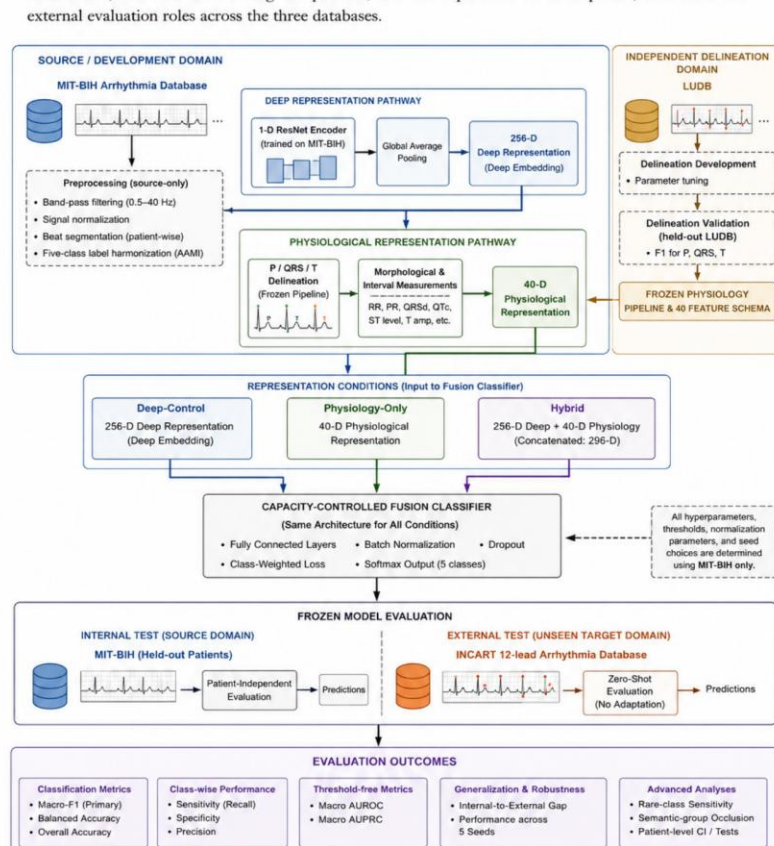


Figure 1. Proposed physiology-augmented cross-database ECG classification framework. Overview of the two parallel representation pathways (deep and physiological), the independent delineation-development and delineation-validation process on LUDB, the three representation conditions supplied to a capacity-controlled fusion classifier, and the separation of internal (MIT-BIH, patient-independent) and external (INCART, zero-shot) frozen-model evaluation. All hyperparameters, thresholds, normalization parameters and seed choices were fixed using MIT-BIH and LUDB only, before any INCART data were scored.

ECG databases

MIT-BIH contains 48 approximately 30-minute two-channel recordings from 47 patients, digitized at 360 Hz³¹. The MLII channel was used where available (V5 for records 102 and 104). After applying the frozen AAMI label mapping^{4,5} and excluding beats that could not yield a complete one-second window, 109,436 beats remained and were split at patient level into training (33 patients, 75,132 beats), validation (7 patients,

17,131 beats) and internal-test (7 patients, 17,173 beats) partitions; records 201/202, from the same patient, were kept together in the internal-test partition (Table 1). INCART contains 75 twelve-lead Holter recordings from 32 patients sampled at 257 Hz³²; lead II was used, giving 175,816 eligible external beats ($N = 153,629$; $S = 1,958$; $V = 20,004$; $F = 219$; $Q = 6$) across all 32 patients. LUDB comprises 200 ten-second, 12-lead recordings sampled at 500 Hz with expert P/QRS/T annotations³⁰; using lead II, 140 records were used to develop the delineation procedure and 60 were held out for a single, final, independent validation, after which no delineation parameter was modified.

Table 1. ECG databases and their methodological roles.

Database	Records	Patients	Sampling rate	Role in this study	Eligible beats used
MIT-BIH	48	47	360 Hz	Source training, validation and patient-independent internal testing	109,436
INCART	75	32	257 Hz	Unseen external classification evaluation	175,816
LUDB	200	200	500 Hz	Independent P/QRS/T delineation development (140 records) and held-out validation (60 records)	Not used for classification

Deep representation pathway

For the deep pathway, the selected lead was band-pass filtered (fourth-order Butterworth, zero-phase, 0.5–40 Hz), resampled to a common 250 Hz, robustly normalized per record, and segmented into one-second beat-centred windows (250 samples). A one-dimensional residual convolutional network (a convolutional stem followed by four residual stages of 32, 32, 64, 128 and 256 channels, then global average pooling; 2,185,637 parameters) was trained on the MIT-BIH training partition with a linear source classification head, selected on the validation partition, and then frozen; its global-average-pooling output, a 256-dimensional vector z_d , was extracted for every beat in all subsequent experiments, including INCART, without further adaptation.

Physiological representation pathway

The physiological pathway operated on each ECG at its native sampling frequency to preserve temporal resolution. A continuous-signal delineation procedure localized P, QRS and T wave boundaries, developed on the 140-record LUDB development subset and evaluated once, without retuning, on the 60-record held-out subset. From the delineated waveform, 29 continuous descriptors (6 rhythm: previous/next/mean RR, RR difference, RR ratio, heart rate; 9 timing/interval: P/T peak timing, P/QRS/T durations, PR, QT, P–QRS gap, ST; 14 morphology/confidence: signed and peak-to-peak amplitude, area, slope, P/T candidate scores) and 11 binary structural-validity indicators (RR validity, waveform validity, non-overlap and PR/QT/complete-complex validity) were computed, giving a 40-dimensional physiological vector. Morphology was measured in physical millivolt units from the filtered, non-amplitude-normalized signal. The 29 continuous features were median-imputed and robustly scaled (median/interquartile range) using

statistics fitted exclusively on the MIT-BIH training partition and then frozen; the 11 binary indicators were retained unscaled. No preprocessing statistic was refitted on validation, internal-test or external data.

Fusion classifier and representation conditions

A common downstream network compared three representation conditions using an identical architecture, optimization procedure and validation criterion, differing only in the input supplied: deep-control (real 256-D deep embedding, zeroed 40-D physiological input), physiology-only (zeroed 256-D input, real 40-D physiological vector), and hybrid (both real). The classifier projected each branch independently (256→128 and 40→64, each with layer normalization, ReLU and dropout), concatenated the two hidden representations (192-D), and passed them through a fusion layer (192→128) and a linear output layer (128→5). This equal-capacity design ensured that any performance difference reflected the information content of the physiological input rather than additional classifier parameters (61,509 parameters in all three conditions).

For heartbeat i , with $z_{d,i} \in \mathbb{R}^{256}$ the frozen deep embedding and $\tilde{z}_{p,i} \in \mathbb{R}^{40}$ the processed physiological vector, the three input conditions were defined as

$$x_{i,\text{Deep}} = [z_{d,i}, 0_{40}], \quad x_{i,\text{Phys}} = [0_{256}, \tilde{z}_{p,i}], \quad x_{i,\text{Hybrid}} = [z_{d,i}, \tilde{z}_{p,i}], \quad (1)$$

All models were trained with AdamW (learning rate 1×10^{-3} , weight decay 1×10^{-4}), batch size 512, up to 50 epochs with early stopping (patience 8 epochs), gradient-norm clipping at 5, dropout 0.2, and a class-weighted cross-entropy loss (weights 0.2126, 1.2083, 0.7659, 1.9530 and 0.8602 for N, S, V, F and Q respectively, fixed from the source training distribution and shared across conditions). The epoch with the highest validation macro-F1 was selected, with ties broken first by balanced accuracy and then by the earliest epoch. Seed 42 was pre-specified as the primary training seed; seeds 137, 271, 314 and 911 were used only for the robustness analysis (Fig. 5). No component of the pipeline was retrained, re-normalized or re-thresholded using INCART data at any stage; INCART was scored once, after every parameter had been frozen on MIT-BIH and LUDB.

Evaluation metrics and class definitions

The classification task used the five AAMI heartbeat superclasses $Y = \{N, S, V, F, Q\}^{4,5}$. Macro-F1, the unweighted mean of the five per-class F1 scores,

$$F1_{\text{macro}} = (1/5)(F1_N + F1_S + F1_V + F1_F + F1_Q), \quad (2)$$

was the pre-specified primary outcome because of the severe class imbalance in both MIT-BIH and INCART; balanced accuracy, overall accuracy, class-specific F1, macro-AUROC and macro-AUPRC were evaluated as secondary measures.

Statistical analysis

Uncertainty in the primary external hybrid-minus-deep macro-F1 difference was estimated with a non-parametric, patient-level bootstrap (10,000 replicates; random seed 20260707): in each replicate, INCART patients were resampled with replacement, all beats belonging to a resampled patient were included together, and the 95% confidence interval was obtained by the percentile method. A paired, patient-level randomization test (10,000 randomizations for INCART; an exact test over all $2^7 = 128$ patient-swap

configurations for the seven-patient MIT-BIH internal-test set) assessed the null hypothesis that the deep-control and hybrid models were exchangeable at the patient level. The same bootstrap procedure was applied to the difference in internal-to-external generalization gap, $G_{\text{Deep}} - G_{\text{Hybrid}}$, where $G_M = F1_{M,\text{internal}} - F1_{M,\text{INCART}}$ for model M. A supportive four-class (N, S, V, F) sensitivity analysis excluded the six true external Q beats. A semantic-group occlusion analysis replaced each of four physiological feature groups (rhythm; timing/intervals; morphology/confidence; structural validity) in turn with its neutral value, holding the deep representation, the remaining physiological features and all classifier weights fixed, and recorded the resulting change in macro-F1.

Software and reproducibility

Experiments used Python 3.12.13 with WFDB 4.3.1³⁷ for waveform and annotation handling, NumPy 2.0.2³⁴, SciPy 1.16.3³⁵ and scikit-learn 1.6.1³⁶ for numerical and classical machine-learning routines, and PyTorch 2.10.0+cu128³³ for all neural-network components, executed on a CUDA-enabled NVIDIA Tesla T4 GPU. All preprocessing policies, network architectures, training configurations, evaluation metrics and statistical procedures were fixed before internal and external evaluation.

Results

Source-validation and patient-independent internal-test performance

Using the pre-specified primary training seed, the three representation conditions were first compared on the MIT-BIH source-validation partition (7 patients, 17,131 beats). The hybrid representation achieved the highest validation macro-F1 (0.564), exceeding the deep-control (0.524) and physiology-only (0.451) conditions (Fig. 2; Table 2). On the held-out, patient-independent MIT-BIH internal-test set (7 different patients, 17,173 beats), physiology-only achieved the highest macro-F1 (0.507), with the hybrid (0.499) again exceeding deep-control (0.479); the hybrid-minus-deep internal difference of +0.020 was not statistically significant (exact paired patient-level randomization test over all $2^7 = 128$ possible patient-level swaps, $P = 0.578$).

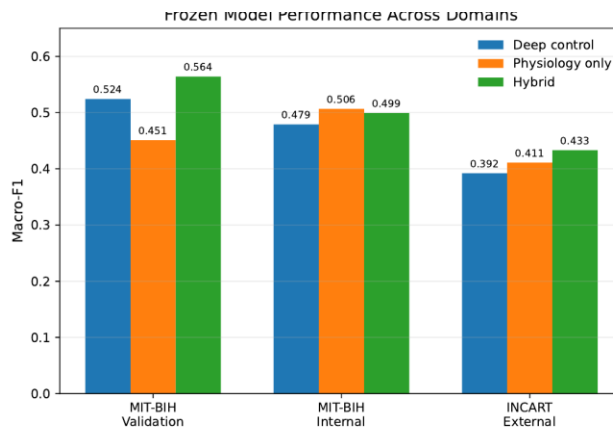


Figure 2. Frozen representation performance across evaluation domains. Macro-F1 of the deep-control, physiology-only and hybrid representation conditions on the MIT-BIH source-validation partition, the held-out patient-independent MIT-BIH internal-test partition, and the unseen external INCART database, using the pre-specified primary training seed (42). Bars show point macro-F1 for $n = 17,131$ (validation), 17,173 (internal-test) and 175,816 (external) beats.

Table 2. Summary of primary classification and statistical results.

Evaluation domain / analysis	Deep-control	Physiology-only	Hybrid	Hybrid – deep-control	95% CI	P value
MIT-BIH validation (macro-F1)	0.524	0.451	0.564	+0.040	—	—
MIT-BIH internal-test (macro-F1)	0.479	0.507	0.499	+0.020	—	0.578 (exact permutation)
INCART external, primary outcome (macro-F1)	0.392	0.411	0.433	+0.041	[0.016, 0.057]	≈ 0.0001 (patient randomization)
INCART four-class sensitivity, N/S/V/F (macro-F1)	0.490	—	0.541	+0.051	—	—
Internal→external generalization gap	0.087	—	0.066	−0.021 (reduction)	[−0.025, 0.062]	not significant

Physiological augmentation improves classification on the unseen external database

The critical test of the study was zero-shot transfer to the INCART database (175,816 beats, 32 patients), which was never used to fit any component of the source-developed pipeline. Without retraining, threshold adjustment or target-domain normalization, deep-control achieved a macro-F1 of 0.392, physiology-only achieved 0.411, and the hybrid representation achieved the highest external macro-F1 of 0.433 (Fig. 2; Table 2). The primary hybrid-minus-deep effect (Δ macro-F1 = +0.041) was supported by a 95% patient-level bootstrap confidence interval of [0.016, 0.057] (10,000 replicates resampled by patient; Fig. 3) and a paired patient-level randomization test (10,000 randomizations) that rejected the null hypothesis of no systematic advantage ($P \approx 0.0001$; Table 2). Because both models used an identical downstream architecture that differed only in whether the true physiological vector or a zero vector was supplied, this difference reflects the information content of the physiological representation rather than added classifier capacity.

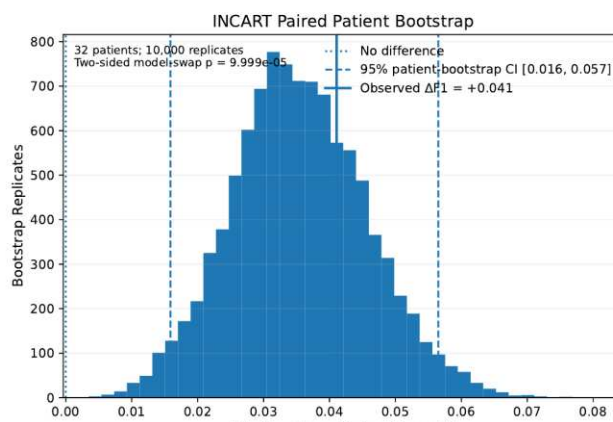


Figure 3. Patient-level bootstrap distribution of the primary external effect. Distribution of the hybrid-minus-deep-control macro-F1 difference on INCART across 10,000 patient-level bootstrap replicates (32 patients)

resampled with replacement). Dashed lines show the 95% percentile confidence interval [0.016, 0.057]; the solid line shows the observed difference (+0.041). A two-sided paired patient-level randomization (model-swap) test gave $P \approx 0.0001$.

Class-wise analysis showed that the external improvement was distributed across several AAMI categories rather than confined to the majority N class (Fig. 4): the hybrid model outperformed deep-control on N (0.944 vs 0.933), S (0.431 vs 0.317) and F (0.113 vs 0.076), while physiology-only alone achieved the highest F1 for the V class (0.870 vs 0.634 for deep-control). All three conditions failed on the six-beat Q class.

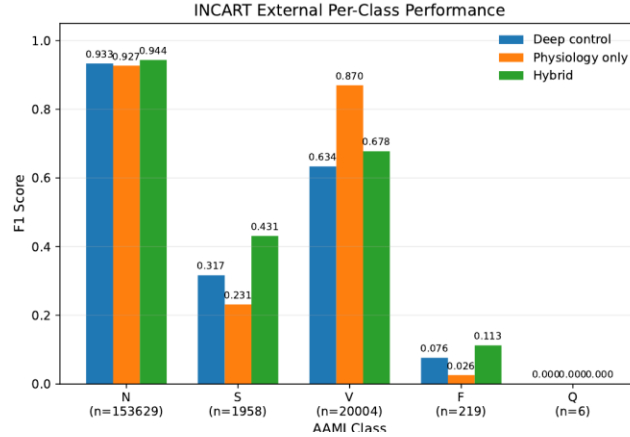


Figure 4. Class-wise external classification performance. F1 score for each AAMI heartbeat class (N, S, V, F, Q) on the INCART external evaluation set ($n = 175,816$ beats from 32 patients) for the three representation conditions. Beat counts per class are shown in parentheses beneath each class label.

The external advantage does not depend on the rarest class

Because only six true Q beats remained in the external evaluation set, a supportive four-class sensitivity analysis (N, S, V, F only) was performed. The hybrid-minus-deep difference remained positive and was, if anything, larger in this analysis: 0.541 versus 0.490 ($\Delta = +0.051$; Table 2), confirming that the five-class result was not an artefact of the extremely rare Q class.

The advantage is robust across training seeds

Across five predeclared training seeds {42, 137, 271, 314, 911}, the mean source-validation hybrid-minus-deep difference was +0.016 (s.d. 0.017), positive for four of the five seeds (Fig. 5). The magnitude of the effect therefore varied with training stochasticity, but its direction was not confined to the pre-specified primary seed.

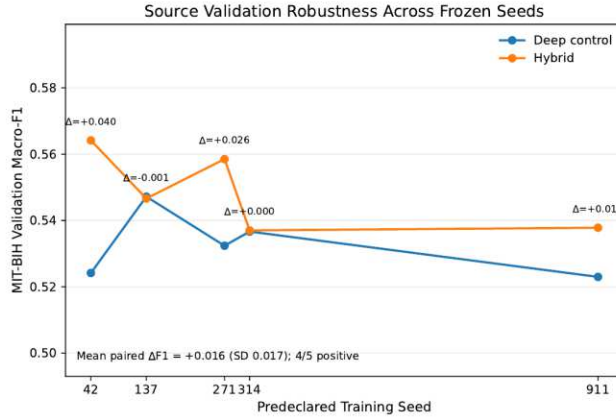


Figure 5. Robustness of the hybrid-versus-deep-control advantage across training seeds. MIT-BIH source-validation macro-F1 for the deep-control and hybrid models across five predeclared training seeds (42, 137, 271, 314, 911). Δ denotes the hybrid-minus-deep-control difference at each seed; the mean paired difference and its standard deviation across seeds are annotated.

The frozen hybrid classifier relies most on rhythm-related physiological features

To characterize how the trained hybrid classifier used the physiological representation, its 40 features were partitioned into four semantic groups (rhythm; timing/intervals; morphology/confidence; structural validity) and each was occluded in turn by neutralizing its values while holding all other inputs and weights fixed. On the external database, neutralizing the rhythm group produced the largest reduction in macro-F1 (-0.021), followed by morphology/confidence (-0.008); timing/intervals and structural validity produced smaller reductions (Fig. 6). Because the deep encoder operated on an isolated one-second beat window, whereas the physiological pathway explicitly encoded neighbouring RR intervals, this pattern is consistent with the two representations capturing different temporal scopes of the same heartbeat.

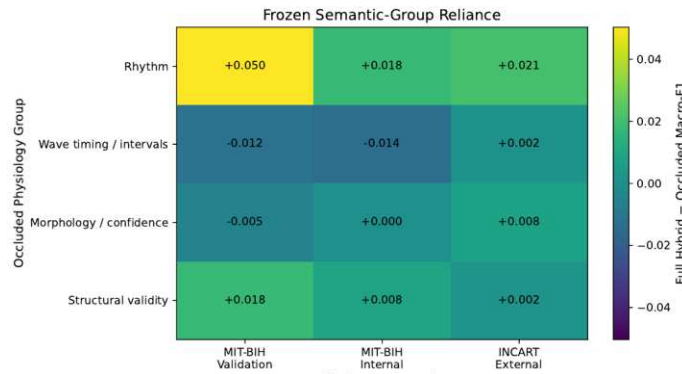


Figure 6. Semantic-group occlusion analysis of the frozen hybrid classifier. Reduction in macro-F1 (full hybrid model minus the model with one physiological feature group neutralized) for each of four semantic feature groups (rhythm; wave timing/intervals; morphology/confidence; structural validity), evaluated on the MIT-BIH validation, MIT-BIH internal-test and INCART external domains. Positive values indicate reliance on that feature group.

Internal-to-external generalization gap

The internal-to-external macro-F1 reduction was numerically smaller for the hybrid model (0.066) than for deep-control (0.087), a gap reduction of 0.021. However, the 95% confidence interval for this difference ($[-0.025, 0.062]$) included zero (Table 2), so the study does not statistically establish that physiological augmentation reduced the generalization gap, despite the improvement in absolute external performance.

Discussion

The principal finding is that explicitly encoding cardiac rhythm, waveform timing, morphology and delineation confidence — derived from an independently validated P/QRS/T delineation procedure — provided information beyond a frozen, capacity-matched deep representation when a classifier trained on MIT-BIH was applied, without any adaptation, to the unseen INCART database. The hybrid representation was superior to either single representation on the external test (0.433 vs 0.392 for deep-control and 0.411 for physiology-only), indicating that the two pathways were not simply redundant descriptions of the same signal.

The semantic-group occlusion analysis suggests a specific mechanism for this complementarity. The deep encoder classified an isolated one-second beat-centred window, whereas the physiological pathway explicitly carried previous- and next-beat RR-interval information and derived rhythm relationships. Occluding the rhythm group produced by far the largest reduction in the frozen hybrid model's external performance, consistent with the interpretation that inter-beat temporal context — information lying outside the deep encoder's input window — was the primary source of added value, although morphology and delineation confidence also contributed measurably.

Three findings temper this conclusion. First, the corresponding improvement on the internal, patient-independent MIT-BIH test set was numerically positive but not statistically significant (exact permutation $P = 0.578$); the strongest evidence for complementarity was therefore obtained specifically under database shift and does not establish a general within-database advantage. Second, although the internal-to-external generalization gap was numerically smaller for the hybrid model, its confidence interval included zero, so a formal reduction in generalization gap is not established: the data support higher external accuracy, not proven gap-closure. Third, the magnitude, though not the direction, of the effect varied across training seeds, indicating some sensitivity to optimization stochasticity that any replication should anticipate.

These effect sizes are more modest than the accuracy or F1 figures reported by some recent cross-database electrocardiogram studies, and this difference reflects evaluation protocol rather than a weaker underlying effect. Studies reporting substantially higher external performance on INCART have generally done so either by allowing some form of access to the target domain during development, for example through unsupervised domain adaptation on unlabelled target data²⁵ or joint multi-database training for domain generalization²⁷, or by evaluating physiology-guided architectures on a stratified, non-patient-independent, single-database split rather than a genuine cross-database transfer²⁶. Overall accuracy is also a more forgiving metric than macro-F1 under the roughly 700:1 class imbalance present in INCART, since a model that predicts the majority N class well can achieve high accuracy while still performing poorly on rarer classes. The present zero-shot, no-target-access, macro-F1-based protocol is consequently a stricter test, and the reported effect sizes should be read in that context rather than compared directly against accuracy or F1 figures obtained under more permissive settings.

The physiology-only representation performed competitively throughout — matching or exceeding deep-control on both the internal and external test sets — showing that explicit descriptors retained discriminative value on their own rather than acting as merely auxiliary inputs. At the same time, the best external result required both representations together, arguing against either replacing deep learning with handcrafted features or dismissing physiological structure as redundant once deep features are available.

Several limitations bound the generality of these results. Only one independent external database was available for confirmatory evaluation, so robustness across a wider range of hospitals, devices and populations remains untested. The AAMI class distribution was heavily imbalanced, and the Q class, with only six beats in the external set, could not be evaluated reliably despite the reassuring four-class sensitivity analysis. The physiological features depend on automated delineation that, although validated independently on a held-out subset of LUDB, showed larger tail errors for some T-wave measurements; delineation-derived descriptors are consequently not equivalent to expert manual annotation. The final physiological representation used predominantly a single ECG lead rather than the full multi-lead information available in INCART and LUDB, and the deep encoder was frozen before fusion, so the results speak to the value of adding physiological information to an already-trained deep model rather than to joint end-to-end optimization. The occlusion analysis measures the reliance of the trained hybrid classifier under perturbation of already-learned weights, not the causal or additive importance of each feature group in isolation.

Within these bounds, the results support a specific and testable claim: physiological descriptors derived from validated waveform delineation can provide information that is partially stable across a genuine, unseen database shift and complementary to what a frozen deep encoder extracts from an isolated beat window. This motivates external validation across additional databases and, eventually, joint training of deep and physiological pathways, before any translation toward clinical decision support.

References

1. World Health Organization. Cardiovascular diseases (CVDs). WHO Fact Sheet [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)) (2025).
2. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, e215–e220 (2000).
3. Luz, E. J. S., Schwartz, W. R., Cámara-Chávez, G. & Menotti, D. ECG-based heartbeat classification for arrhythmia detection: a survey. *Comput. Methods Programs Biomed.* 127, 144–164 (2016).
4. de Chazal, P., O’Dwyer, M. & Reilly, R. B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Trans. Biomed. Eng.* 51, 1196–1206 (2004).
5. Association for the Advancement of Medical Instrumentation. ANSI/AAMI EC57:1998/(R)2008: Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms (AAMI, 2008).
6. Ince, T., Kiranyaz, S. & Gabbouj, M. A generic and robust system for automated patient-specific classification of ECG signals. *IEEE Trans. Biomed. Eng.* 56, 1415–1426 (2009).
7. Llamedo, M. & Martínez, J. P. Heartbeat classification using feature selection driven by database generalization criteria. *IEEE Trans. Biomed. Eng.* 58, 616–625 (2011).

8. Ye, C., Vijaya Kumar, B. V. K. & Coimbra, M. T. Heartbeat classification using morphological and dynamic features of ECG signals. *IEEE Trans. Biomed. Eng.* 59, 2930–2941 (2012).
9. Kiranyaz, S., Ince, T. & Gabbouj, M. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Trans. Biomed. Eng.* 63, 664–675 (2016).
10. Acharya, U. R. et al. A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* 89, 389–396 (2017).
11. Kachuee, M., Fazeli, S. & Sarrafzadeh, M. ECG heartbeat classification: a deep transferable representation. In *Proc. 2018 IEEE Int. Conf. Healthcare Informatics* 443–444 (2018).
12. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 25, 65–69 (2019).
13. Ribeiro, A. H. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat. Commun.* 11, 1760 (2020).
14. Strodthoff, N., Wagner, P., Schaeffter, T. & Samek, W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. *IEEE J. Biomed. Health Inform.* 25, 1519–1528 (2021).
15. Wagner, P. et al. PTB-XL, a large publicly available electrocardiography dataset. *Sci. Data* 7, 154 (2020).
16. Ben-David, S. et al. A theory of learning from different domains. *Mach. Learn.* 79, 151–175 (2010).
17. Ganin, Y. et al. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 1–35 (2016).
18. Sun, B. & Saenko, K. Deep CORAL: correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops* 443–450 (2016).
19. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. Preprint at <https://arxiv.org/abs/1907.02893> (2019).
20. Gulrajani, I. & Lopez-Paz, D. In search of lost domain generalization. In *Int. Conf. Learning Representations* (2021).
21. Koh, P. W. et al. WILDS: a benchmark of in-the-wild distribution shifts. In *Proc. 38th Int. Conf. Machine Learning* 139, 5637–5664 (2021).
22. He, J., Sun, L., Rong, J., Wang, H. & Zhang, Y. A pyramid-like model for heartbeat classification from ECG recordings. *PLoS ONE* 13, e0206593 (2018).
23. Aziz, S., Ahmed, S. & Alouini, M.-S. ECG-based machine-learning algorithms for heartbeat classification. *Sci. Rep.* 11, 18738 (2021).
24. Ballas, A. & Diou, C. Towards domain generalization for ECG and EEG classification: algorithms and benchmarks. Preprint at <https://arxiv.org/abs/2303.11338> (2023).
25. Imtiaz, M. N. & Khan, N. Cross-database and cross-channel electrocardiogram arrhythmia heartbeat classification based on unsupervised domain adaptation. *Expert Syst. Appl.* 244, 122960 (2024).
26. Li, G., Zhang, Z., Qiao, G., Chen, X., Zhou, G. & Chipusu, K. Physiology-guided beat-level arrhythmia classification from ECG using a CNN-transformer hybrid neural network. *Front. Cardiovasc. Med.* 13, 1802210 (2026).

27. Li, R., Aierken, Y., Xu, Y., Liu, J. & Tang, Y. Research on cross-dataset cardiac signal domain generalization and feature interpretability. *Sci. Rep.* 16, 3138 (2025).
28. Laguna, P., Jané, R. & Caminal, P. Automatic detection of wave boundaries in multilead ECG signals: validation with the CSE database. *Comput. Biomed. Res.* 27, 45–60 (1994).
29. Martínez, J. P., Almeida, R., Olmos, S., Rocha, A. P. & Laguna, P. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Trans. Biomed. Eng.* 51, 570–581 (2004).
30. Kalyakulina, A. I. et al. LUDB: a new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access* 8, 186181–186190 <https://doi.org/10.13026/eegm-h675> (2020).
31. Moody, G. B. & Mark, R. G. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng. Med. Biol. Mag.* 20, 45–50 <https://doi.org/10.13026/C2F305> (2001).
32. Yakushenko, E. St Petersburg INCART 12-lead Arrhythmia Database. *PhysioNet* <https://doi.org/10.13026/C2V88N> (2008).
33. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32, 8024–8035 (2019).
34. Harris, C. R. et al. Array programming with NumPy. *Nature* 585, 357–362 (2020).
35. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272 (2020).
36. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
37. Xie, C. et al. Waveform Database Software Package (WFDB) for Python. *PhysioNet* <https://doi.org/10.13026/egpf-2788> (2021).

Author contributions

M.S. conceived the study, developed the computational pipeline, performed the experiments and statistical analyses, and drafted the manuscript. O.K. supervised the research and reviewed the manuscript. Both authors read and approved the final manuscript.

Data availability

All ECG data used in this study are publicly available from PhysioNet. The MIT-BIH Arrhythmia Database³¹ (version 1.0.0) is available at <https://doi.org/10.13026/C2F305>; the St Petersburg INCART 12-lead Arrhythmia Database³² (version 1.0.0) is available at <https://doi.org/10.13026/C2V88N>; and the Lobachevsky University Electrocardiography Database³⁰ (LUDB; version 1.0.1) is available at <https://doi.org/10.13026/eegm-h675>.

Code availability

The experimental notebooks used to generate the reported results are available through Kaggle at <https://www.kaggle.com/work/collections/19034579>.

Ethics declarations

This study used only publicly available, de-identified, retrospective ECG databases (MIT-BIH, INCART and LUDB). No new data were collected from human participants, and no additional ethical approval was required.

Competing interests

The authors declare no competing interests.