Computers in Biology and Medicine 214 (2026) 111890

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

HRV-anchored autonomic-morphological structured factorization for interpretable multimodal cardiovascular risk prediction

Jiantao Xu [a], Dexing Zhou [b], Jianbo Xu [c,*], Tianruo Wang [d]

[a] The University of Aizu, Aizuwakamatsu, 965-8580, Fukushima, Japan
[b] Hainan Medical University, Xueyuan Road, Haikou, 570-331, Hainan, China
[c] Hainan University, Renmin Road, Haikou, 570-228, Hainan, China
[d] Southwest Forestry University, Panlong Road, Kunming, 650-224, Yunnan, China

HIGHLIGHTS

- HRV anchors autonomic structure in multimodal ICU risk modeling.
- Quality-aware fusion improves tolerance to missing and noisy inputs.
- PCG is assessed only as an optional component-level morphology branch.
- The target is a retrospective composite under observed care practices.
- Latent analyses support internal consistency rather than causality.

ARTICLE INFO

Keywords:
Cardiovascular risk prediction
Heart rate variability
Autonomic–morphological structured factorization
Multimodal physiological signals
Quality-aware fusion
Interpretable deep learning

ABSTRACT

Cardiovascular risk prediction from heterogeneous physiological signals supports early warning in bedside and wearable monitoring, where ECG, PPG, HRV, signal-quality indicators, activity context, and clinical metadata jointly carry evidence of short-term deterioration. However, existing approaches face two limitations: HRV-based models are physiologically interpretable yet cannot represent waveform morphology, whereas conventional multimodal deep models entangle autonomic regulation, cardiovascular morphology, activity-related interference, patient-specific baseline, and acquisition uncertainty in one latent space. We propose AM-DiMNet, an HRV-anchored autonomic–morphological structured factorization framework predicting clinically recorded cardiovascular deterioration within 24 h. In the primary end-to-end cohort, it integrates ECG, PPG, explicit HRV features, movement- and waveform-instability proxies, clinical metadata, modality-availability masks, and signal-quality scores, then factorizes the fused representation into autonomic, morphological, activity, baseline, and noise components via HRV anchoring and quality-aware fusion. Because no public dataset jointly provides all modalities with longitudinal outcomes, PCG is treated as an architecturally compatible optional branch and is assessed only through task-specific component-level morphology experiments, not as an empirically validated contributor to the primary 24-h endpoint. AM-DiMNet attained an AUROC of 0.895, AUPRC of 0.684, Macro-F1 of 0.804, MCC of 0.641, and ECE of 0.036, raising AUROC from 0.786 and AUPRC from 0.503 over HRV-based XGBoost and surpassing attention and modality-dropout fusion in discrimination, calibration, and incomplete-input robustness. The prediction target was a retrospective composite of time-stamped physiological adverse events and care-process-mediated interventions recorded under the clinical practices represented in the source cohort, and should not be interpreted as a policy-invariant estimate of untreated biological deterioration. These results show HRV can act as an interpretable autonomic anchor for calibrated, physiologically structured risk modeling; the latent-factor analyses offer internal, mechanism-consistent evidence rather than causal or clinician-validated explanations.

* Corresponding author.
Email address: jianbo_xu@hainanu.edu.cn (J. Xu).

https://doi.org/10.1016/j.compbiomed.2026.111890
Received 21 May 2026; Received in revised form 25 July 2026; Accepted 5 August 2026

1. Introduction

Cardiovascular diseases remain a leading cause of mortality, disability, and health-system burden worldwide [1]. Intensive-care monitoring, wearable sensors, digital stethoscopes, and large-scale electrocardiographic repositories now make continuous risk assessment feasible at scale. Within this setting, ECG, PPG, PCG, HRV, accelerometry, and clinical metadata each report on a different facet of the cardiovascular system: electrical activation, peripheral hemodynamics, acoustic cardiac events, autonomic regulation, activity context, and baseline risk. The bottleneck, however, is no longer signal availability. The harder problem is to turn heterogeneous, noisy, and often incomplete measurements into risk evidence that is reliable, well calibrated, and clinically interpretable.

Deep learning has driven much of the recent progress. Large-scale ECG models perform arrhythmia classification [2], 12-lead diagnosis [3], left ventricular systolic dysfunction screening [4], mortality prediction [5], and AI-enabled cardiovascular risk estimation [6]. The same data-driven philosophy has reached unobtrusive vital-sign monitoring. Wang et al. combined empirical mode decomposition with convolutional feature extraction and temporal modeling in an EMD–LSTM–CNN model, recovering heart-rate and respiratory-rate signals from a microbend optical-fiber sensor in perioperative infants and handling nonstationary noise better than frequency- or wavelet-domain baselines [7]. Such work shows the payoff of jointly considering sensing characteristics, signal decomposition, and temporal modeling. Its target, however, is sensor-specific parameter estimation, not prediction of future clinical deterioration from several modalities at once.

Closer to risk prediction, wearable heart-rate and activity streams have supported semi-supervised cardiovascular risk modeling [8], ECG-derived HRV features have predicted in-hospital cardiac arrest [9], and PCG/ECG foundation models have enabled digital-stethoscope-based assessment [10]. These results confirm the prognostic value of physiological signals, yet most pipelines pursue only one of signal reconstruction, parameter estimation, single-modality diagnosis, or general-purpose representation learning. None separates, inside a single risk model, the autonomic, morphological, activity-related, baseline, and acquisition-quality sources that jointly shape a prediction. This gap motivates our HRV-anchored structured factorization approach.

Two modeling traditions make the gap concrete. HRV-based models are interpretable and physiologically grounded, but HRV alone cannot encode ECG morphology, PPG pulse dynamics, PCG acoustic patterns, or structural cardiovascular abnormality [11,12]. Conversely, black-box multimodal networks raise discrimination yet rarely reveal whether an elevated score reflects autonomic dysfunction, morphological abnormality, activity-related confounding, baseline vulnerability, or unreliable acquisition. Neither tradition, on its own, meets what bedside deployment also demands: tolerance to missing modalities, robustness to low signal quality, probability calibration, external transfer, subgroup stability, and patient-level explanation.

We hypothesize that risk prediction becomes more reliable and interpretable when autonomic regulation and cardiovascular morphology are modeled as separate physiologically structured factors rather than as one mixed latent vector. In this view, HRV anchors autonomic regulation, while ECG and PPG supply morphology evidence from electrical activation and peripheral pulse dynamics. PCG adds acoustic morphology when available, but we do not assume it is present in the primary risk-prediction cohort. Activity/motion proxy information together with signal-quality indicators models confounding and acquisition uncertainty rather than serving as direct disease evidence.

To test this hypothesis, we propose AM-DiMNet, an HRV-anchored autonomic–morphological structured factorization framework for interpretable cardiovascular risk prediction. In the primary end-to-end cohort it consumes ECG, PPG, activity/motion proxy information, explicit HRV features, available clinical metadata, modality masks, and signal-quality scores; PCG-related morphology is examined only through complementary component-level validation datasets rather than claimed as simultaneous all-modality evidence. The framework couples modality-specific encoders, quality-aware gated fusion, HRV and morphology anchoring, adversarial activity deconfounding, orthogonality regularization, and calibration-aware prediction, so that each design element targets one of the shortcomings identified above. Throughout, the prediction target is a retrospective 24-h composite of observed physiological adverse events and care-process-mediated interventions recorded under the source cohort's clinical practices, rather than a policy-invariant biological deterioration endpoint.

Our contributions are fourfold. First, we cast cardiovascular risk prediction as HRV-anchored autonomic–morphological structured representation learning rather than conventional black-box multimodal fusion. Second, we design a flexible multimodal architecture that jointly models ECG/PPG morphology, HRV-based autonomic regulation, activity context, signal quality, missing modalities, and clinical metadata in the primary cohort; PCG is included as an architecturally compatible optional branch and is assessed only through task-specific component-level morphology experiments, not as an empirically validated contributor to the primary endpoint. Third, we introduce a structured training objective that combines physiological anchoring, adversarial deconfounding, latent-factor regularization, and calibration-aware prediction. Fourth, we develop a dataset-aware evaluation protocol that separates primary end-to-end risk prediction from modality-specific validation, mechanism-oriented analysis, robustness testing, calibration assessment, subgroup analysis, and patient-level interpretation.

2. Related work

2.1. ECG-based deep learning for cardiovascular diagnosis and risk prediction

ECG-based deep learning has matured quickly in computational cardiology. Convolutional and residual networks reach strong accuracy in ambulatory arrhythmia classification [2], 12-lead diagnosis [3], left ventricular dysfunction screening [4], and mortality prediction from raw voltage traces [5], and AI-enabled ECG scoring has recently moved from diagnosis toward long-term risk stratification [6]. The recurring weakness is interpretability: the learned codes are compact but physiologically opaque. Latent-factorization models such as FactorECG improve morphology interpretation [13,14], yet they stay ECG-centric and leave HRV anchoring, PPG hemodynamics, activity confounding, missing modalities, signal-quality uncertainty, and calibration unaddressed—precisely the axes AM-DiMNet sets out to organize.

2.2. HRV-based physiological modeling and its limitations

HRV offers an interpretable window onto autonomic regulation and is widely used for cardiovascular assessment, stress monitoring, critical-care prediction, and medical decision support [11,12]; ECG-derived HRV features have even enabled real-time in-hospital cardiac arrest prediction in the ICU [9]. Its reliability, however, is sensitive to age, respiration, medication, posture, physical activity, arrhythmia, signal quality, and window length. More fundamentally, HRV describes beat-to-beat variability without encoding ECG morphology, PPG pulse shape, PCG acoustic patterns, or structural abnormality. We therefore use HRV as a physiological anchor for the autonomic factor rather than as a standalone predictor.

2.3. Multimodal physiological signal learning

Multimodal learning fuses heterogeneous streams such as waveforms, activity signals, acoustic recordings, and clinical metadata [15]. In cardiology these channels are complementary—ECG for electrical activation, PPG for peripheral blood-volume dynamics, PCG for mechanical-acoustic events, ACC for activity context, and metadata for baseline risk—and prior work has exploited wearable heart-rate and activity data [8], PPG-based atrial fibrillation monitoring [16], PCG

heart-sound classification [17,18], and PCG/ECG foundation models [10]. Most methods stop at feature fusion and leave the physiological roles of the modalities entangled; attention weights need not correspond to mechanisms, and noisy or motion-contaminated channels can inflate false alarms. AM-DiMNet counters this with modality masks, signal-quality scores, and activity-aware structured representation learning.

2.4. Self-supervised and foundation models for cardiac signals

Because labeled cardiac data are costly and task-specific, self-supervised pretraining has become central to cardiac signal analysis. CLOCS contrasts cardiac signals across space, time, and patients [19], ECG-MAE reconstructs masked multi-lead ECG [20], and digital-stethoscope foundation models extend the idea to synchronous PCG and ECG [10]. These approaches improve representation quality and transferability, but they optimize general-purpose embeddings rather than clinically structured factors. AM-DiMNet is complementary: pretrained encoders can be plugged in, while its supervised objective forces the final representation to organize around autonomic regulation, waveform morphology, activity confounding, baseline variation, and signal quality.

Recent time-series foundation models provide stronger pretrained representation-learning baselines. MOMENT is a general-purpose time-series foundation model pretrained across heterogeneous time-series domains [21], whereas NormWear is designed for multivariate wearable physiological signals, including ECG and PPG [22]. We therefore evaluate both models in the primary cardiovascular deterioration cohort to examine whether general-purpose or physiology-oriented pretraining can match the proposed physiologically structured factorization. Unlike AM-DiMNet, these baselines do not explicitly separate autonomic, morphological, activity-related, baseline-related, and signal-quality-associated representations; they instead provide pretrained waveform embeddings that are then combined with the same downstream features for a fair comparison.

2.5. Structured latent representation and explainability

Disentangled representation learning aims to isolate independent factors of variation, as in beta-VAE, FactorVAE, and beta-TCVAE [23–25]. Without inductive bias or supervision, however, such disentanglement is underdetermined [26]—a problem that is acute in medicine, where the recovered factors must map to physiologically meaningful sources. Interpretable ECG modeling has begun to address this through VAE-based factorization and knowledge-guided diagnosis [13,14,27,28]. We carry this line from ECG-only interpretation to multimodal cardiovascular risk prediction, supplying the missing inductive bias through HRV anchoring, morphology anchoring, activity deconfounding, signal-quality modeling, and calibration-aware learning.

2.6. Datasets, evaluation protocols, and remaining gaps

Open datasets underpin reproducible cardiovascular AI: PTB-XL for large-scale 12-lead labels [29], MIMIC-IV and MIMIC-IV Waveform for clinical records and bedside waveforms [30,31], MIMIC-IV-ECG for ECG–clinical linkage [32], Icentia11k for long-duration ECG representation learning [33], and the PhysioNet/CinC 2016 Challenge and EPHNOGRAM for heart-sound and ECG–PCG analysis [17,34]. None of them, though, simultaneously contains ECG, PPG, PCG, ACC, HRV, clinical metadata, signal-quality indicators, and time-resolved cardiovascular outcomes. This absence forces a dataset-aware evaluation protocol in which complementary datasets must not be read as simultaneous all-modality validation. Accordingly, we train and test end-to-end deterioration prediction on a primary clinical waveform cohort and reserve complementary ECG, PCG, ECG–PCG, and activity datasets for component-level or mechanism-oriented validation, so that ECG/PPG risk modeling, optional PCG morphology, activity deconfounding, quality-aware fusion, missing-modality robustness, calibration, and latent-factor interpretability are each assessed without overclaiming all-modality external validity. Consistent with current guidance for clinical prediction models, the study is reported with reference to TRIPOD + AI [35], and the PROBAST + AI framework [36] is used to structure the assessment of potential bias and applicability.

3. Methodology

3.1. Problem definition

This study aims to predict future cardiovascular deterioration from heterogeneous physiological signals by learning physiologically structured latent representations of autonomic regulation, waveform morphology, activity-related confounding, patient-specific baseline variation, and signal-quality uncertainty. For each patient i, physiological recordings are segmented into a series of chronological observation windows indexed by w. Each window contains synchronized or partially synchronized physiological measurements collected before the future prediction interval.

The proposed framework is defined for a flexible set of physiological modalities, including ECG, PPG, PCG, ACC, explicit HRV features, signal-quality indicators, modality-availability masks, and optional clinical metadata. However, the actually observed modality set is dataset-dependent. In the primary end-to-end cardiovascular risk-prediction cohort, ECG, PPG, activity/motion proxy information, explicit HRV features, signal-quality scores, modality masks, and available clinical metadata are used. PCG is treated as an optional morphology-related branch and is evaluated only through complementary PCG or ECG–PCG component-level validation datasets rather than being assumed to be simultaneously available in the primary risk cohort.

For the w-th observation window of patient i, the model input is defined as

$$\mathcal{X}_{i,w} = \left\{ \mathbf{X}^{r}_{i,w}, \mathbf{h}^{hrv}_{i,w}, \mathbf{c}_{i,w}, \mathbf{m}_{i,w}, \mathbf{q}_{i,w} \right\}_{r \in \mathcal{R}_{i,w}}, \tag{1}$$

where $\mathcal{R}_{i,w}$ denotes the set of modalities available in the corresponding observation window. The term $\mathbf{X}^{r}_{i,w}$ represents the raw or transformed signal of modality r, $\mathbf{h}^{hrv}_{i,w}$ denotes explicit HRV features computed from normal-to-normal intervals, $\mathbf{c}_{i,w}$ denotes available clinical metadata, $\mathbf{m}_{i,w}$ is the modality-availability mask, and $\mathbf{q}_{i,w}$ denotes modality-specific signal-quality scores.

Let the observation window be $[t^{start}_{i,w}, t^{end}_{i,w}]$. To reduce near-event leakage, the observation window and the prediction window are separated by a temporal gap Δ_g. Given a prediction horizon Δ_p, the future outcome interval is defined as

$$\mathcal{T}^{pred}_{i,w} = [t^{end}_{i,w} + \Delta_g,\ t^{end}_{i,w} + \Delta_g + \Delta_p]. \tag{2}$$

In the primary setting, $\Delta_g = 1.5$ h and $\Delta_p = 24$ h.

The binary label for each patient-window pair is defined as

$$y_{i,w} = \begin{cases} 1, & \text{if at least one predefined cardiovascular} \\ & \text{deterioration event occurs within } \mathcal{T}^{pred}_{i,w}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The model outputs a future risk probability

$$\hat{y}_{i,w} = P(y_{i,w} = 1 | \mathcal{X}_{i,w}). \tag{4}$$

For notational simplicity, the window index w is omitted in subsequent equations when no ambiguity arises.

3.2. Endpoint definition and label construction

The primary endpoint is defined as a time-resolved composite cardiovascular deterioration event occurring within the future prediction interval. The endpoint is designed to capture clinically meaningful short-term deterioration rather than static cardiovascular disease status. Therefore, discharge-level diagnosis codes are not used as direct

Table 1
Endpoint-extraction summary for the primary composite cardiovascular deterioration label. Item identifiers follow MIMIC-IV `d_items` and table conventions; the complete protocol, including the SQL queries and full identifier lists, is given in Appendix A.

Endpoint component	Source tables	Item IDs / codes	Medications / procedures	Initiation and escalation rule	Timestamp
Cardiac arrest or CPR	`procedureevents`, `chartevents`	225,466, 225,468, 224,385	CPR, defibrillation	First documented resuscitation event during the ICU stay	`starttime` `charttime`
ICU or in-hospital death	`admissions`, `icustays`, `patients`	`deathtime`, `dod`	—	Death with a resolvable timestamp within the prediction interval	`deathtime`
Vasoactive support escalation	`inputevents`	221,906, 221,289, 221,662, 221,653, 222,315	Norepinephrine, epinephrine, dopamine, dobutamine, vasopressin	New infusion, or a rate increase > 30% sustained ≥30 min after a stable baseline	`starttime`
Acute heart-failure deterioration	`inputevents`, `prescriptions`	221,794, 225,152, 228,340	Intravenous furosemide, inotropes	Escalation of intravenous diuretics or inotropes, or documented decompensation	`starttime` `charttime`
Malignant arrhythmia	`chartevents`, `procedureevents`	224,168, 225,448, 225,468	VT/VF, high-grade block, defibrillation	Documented VT/VF or severe brady/high-grade conduction block requiring treatment	`charttime`
Urgent cardiovascular intervention	`procedureevents`	225,752, 225,459, 225,802	Non-elective cardiovascular procedure	Urgent procedure related to hemodynamic or electrical instability, excluding elective cases	`starttime`

time-resolved labels. Instead, diagnosis codes are used only for baseline comorbidity characterization, cohort stratification, and sensitivity analysis when appropriate.

In the primary MIMIC-IV Waveform experiment, a positive label is assigned if at least one predefined cardiovascular deterioration event occurs within the prediction interval. The composite endpoint includes cardiac arrest or cardiopulmonary resuscitation, ICU or in-hospital death with available event time, initiation or marked escalation of vasoactive support, acute heart-failure deterioration requiring intensified treatment, clinically documented malignant arrhythmia, or urgent cardiovascular intervention. Event timestamps are obtained from ICU chart events, procedure records, medication and input events, and linked hospital records. When multiple endpoint components occur within the same prediction interval, the earliest eligible event is used as the first qualifying endpoint for label construction.

To improve reproducibility, each endpoint component is defined using an operational rule before model training. The event counts report the first qualifying endpoint per positive patient in the primary cohort. The database source tables, item identifiers, medication generic names, procedure codes, and initiation/escalation and timestamp rules used for label construction are summarized in Table 1, and the complete protocol, including exclusion and precedence rules together with the SQL queries and item-identifier lists, is given in Appendix A and released in the public repository stated in the Code Availability section.

The extraction rules are designed to reduce ambiguous label assignments. For vasoactive support, new initiation and marked escalation are distinguished from maintenance therapy. For arrhythmia-related events, clinically documented malignant rhythm deterioration is used rather than unsupervised rhythm inference from noisy bedside waveforms. For urgent cardiovascular intervention, only time-stamped procedures plausibly associated with cardiovascular instability are considered. To examine whether treatment-related components dominate the model behavior, hard-endpoint and treatment-driven endpoint sensitivity analyses are reported in Section 4.10.

Treatment-driven endpoint components require a different interpretation from hard clinical outcomes. Cardiac arrest, death, and malignant arrhythmia primarily represent observed adverse events, whereas vasoactive support escalation, intensified heart-failure treatment, and urgent cardiovascular intervention are partly determined by clinical decisions. These decisions may depend not only on the patient's physiological state but also on clinician judgment, monitoring intensity, staffing, institutional protocols, resource availability, and local intervention thresholds. Consequently, treatment-driven labels represent care-process-mediated outcomes rather than purely biological events.

No medication initiation, dose escalation, procedure record, or endpoint indicator occurring after the observation window is included in the model input. This design prevents direct leakage of the future intervention into the predictor. Nevertheless, exclusion of post-observation treatment information does not remove decision-related label bias because the recorded outcome itself remains dependent on the clinical policy under which the patient was treated. The primary prediction target should therefore be interpreted as the risk of clinically recorded cardiovascular deterioration under the care practices represented in the development cohort, rather than the risk of an untreated or policy-invariant biological event.

To reduce label leakage, all physiological signals, HRV features, signal-quality indicators, activity proxies, and clinical covariates are extracted only from the observation window. The gap window and prediction window are not used for feature extraction, normalization, quality-score estimation, activity-threshold estimation, or model input construction. Windows are excluded if any endpoint criterion is already satisfied before or during the observation window, because such cases represent ongoing clinical instability rather than prospective risk prediction. Windows are also excluded if the first eligible endpoint occurs within the gap interval, since these cases do not satisfy the predefined prediction horizon.

Patients with insufficient waveform duration, unreliable ECG-derived RR intervals, unavailable event-time information, or insufficient temporal alignment between waveform and clinical records are excluded according to predefined quality-control criteria. This design ensures that each positive label corresponds to a future deterioration event occurring after the observation period and following the leakage-prevention gap.

Because multiple overlapping windows can be generated from the same patient, window-level samples are not statistically independent. Therefore, all train-validation-test partitions are performed at the patient level. Confidence intervals and statistical comparisons are computed using patient-level resampling or patient-level aggregation rather than treating all windows as independent observations.

3.3. Overall framework

We propose an HRV-anchored autonomic–morphological structured multimodal network, named AM-DiMNet, for interpretable cardiovascular risk prediction. The framework is designed around three principles. First, HRV is used as an explicit physiological anchor for autonomic regulation rather than as a simple auxiliary feature. Second, ECG and PPG waveforms are used for morphology and hemodynamic modeling in the primary risk-prediction cohort, while PCG is incorporated as an optional morphology-related branch only when PCG is available and is evaluated through complementary PCG datasets. Third, activity/motion proxy information, signal-quality scores, and modality-availability masks are used to reduce confounding caused by physical activity, motion artifacts, and incomplete physiological recordings. When PCG is available, its encoded representation is projected into the shared embedding space

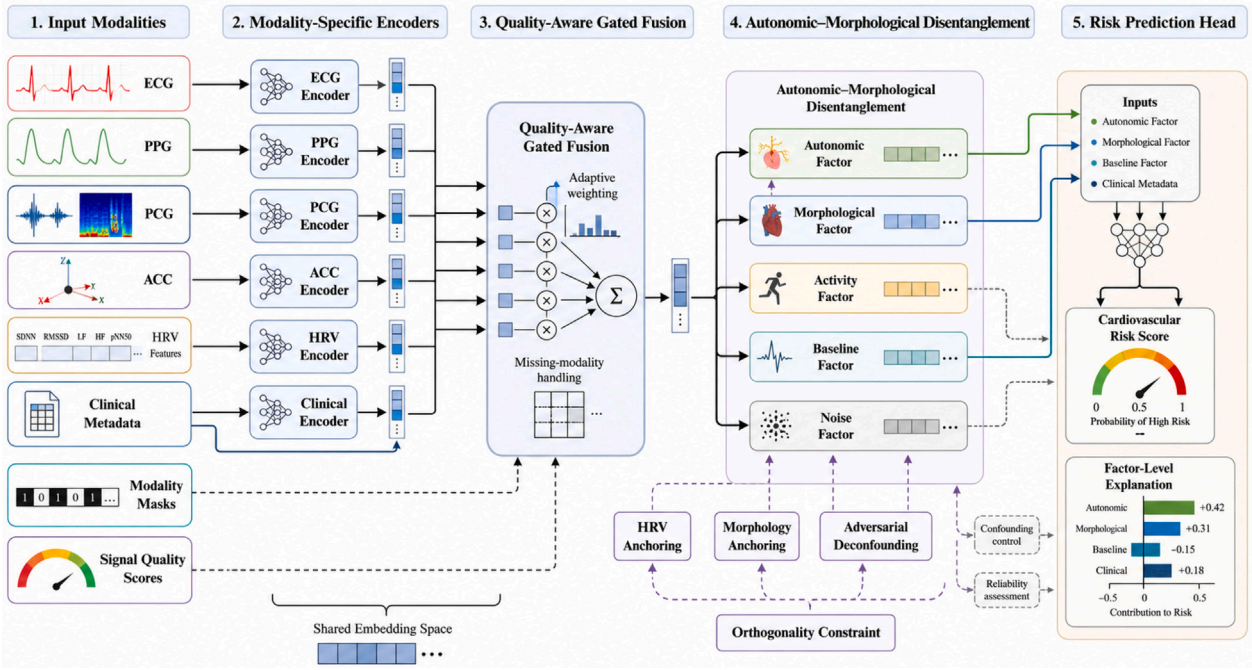

Fig. 1. Overall workflow of the proposed AM-DiMNet framework. The primary end-to-end risk-prediction pathway uses ECG, PPG, activity/motion proxy information, HRV features, available clinical metadata, modality masks, and signal-quality scores. PCG is shown as an optional, design-compatible branch evaluated only through task-specific component-level morphology experiments; it is not used in the primary 24-h endpoint model.

and incorporated through the same modality-availability mask and quality-aware fusion gate used for the other modalities. PCG-derived descriptors may also contribute to the morphology-anchoring objective. When PCG is unavailable, its modality mask is set to zero, and it contributes neither to the fused representation nor to the PCG-related anchoring term. This mechanism permits sample-dependent modality availability without imputing an artificial PCG signal. This architectural compatibility should be distinguished from task-level evidence. Because synchronized PCG recordings and future cardiovascular outcomes are unavailable in the primary cohort, the present study does not quantify the incremental contribution of PCG to the 24-h deterioration endpoint.

The proposed flexible pipeline consists of five stages. First, raw physiological signals are preprocessed, quality-controlled, and segmented into leakage-safe observation windows. Second, available modalities are encoded by modality specific encoders. Third, a quality-aware gated fusion module integrates available representations while suppressing unreliable or missing modalities. Fourth, the fused representation is factorized into autonomic, morphological, activity-related, baseline-related, and noise-related latent components. Finally, cardiovascular risk is predicted from physiologically motivated latent factors, and patient-level explanations are generated from factor-level contributions.

The proposed workflow is illustrated in Fig. 1. The primary risk-prediction pathway uses ECG, PPG, activity/motion proxy information, HRV features, signal-quality scores, modality masks, and available clinical metadata. The PCG branch is optional and is used only for component-level morphology validation when PCG recordings are available.

3.4. Signal preprocessing and feature construction

For ECG signals, baseline wander and high-frequency noise are removed by band-pass filtering. R peaks are detected from the filtered waveform, and RR intervals shorter than 300 ms or longer than 2000 ms are treated as unreliable and removed or locally interpolated according to signal quality. Normal-to-normal intervals are then used to compute the HRV vector

$$\mathbf{h}_i^{hrv} = [\text{MeanNN}, \text{SDNN}, \text{RMSSD}, \text{pNN50}, \text{LF}, \text{HF}, \text{LF/HF}, \text{SD1}, \text{SD2}, \text{SampEn}]_i. \tag{5}$$

These features serve as predictive inputs and physiological targets for autonomic anchoring.

For PPG signals, pulse peaks are detected after band-pass filtering. Pulse-to-pulse intervals and morphology descriptors are extracted, including amplitude, width, rise time, decay time, and slope-related features. In optional PCG datasets, PCG recordings are converted into log-Mel spectrograms or wavelet scalograms to characterize S1, S2, systolic, and diastolic patterns.

ACC denotes triaxial accelerometry measuring body-motion acceleration along the x, y, and z axes. When direct ACC is available, movement intensity is summarized by the signal magnitude area

$$\text{SMA}_{i,w} = \frac{1}{T}\sum_{t=1}^{T}\left(|a_{x,t}| + |a_{y,t}| + |a_{z,t}|\right). \tag{6}$$

Standard triaxial ACC is not consistently available in the primary ICU cohort. Therefore, this cohort uses an activity- and motion-related proxy vector

$$\mathbf{a}_{i,w}^{proxy} = \left[b_{i,w}^{bed}, d_{i,w}^{ecg}, d_{i,w}^{ppg}, r_{i,w}^{ppg\text{-}motion}\right], \tag{7}$$

where $b_{i,w}^{bed}$ is an available bed-motion indicator, $d_{i,w}^{ecg}$ and $d_{i,w}^{ppg}$ denote normalized waveform-envelope instability, and $r_{i,w}^{ppg\text{-}motion}$ is the proportion of PPG subsegments rejected because of abrupt amplitude variation, pulse undetectability, or motion contamination. For modality $r \in \{ecg, ppg\}$, envelope instability is computed as

$$d_{i,w}^{r} = \frac{\text{MAD}\left(\mathbf{e}_{i,w}^{r}\right)}{\text{median}\left(\mathbf{e}_{i,w}^{r}\right) + \epsilon}, \tag{8}$$

where $\mathbf{e}_{i,w}^{r}$ is the local waveform envelope and ϵ prevents numerical instability. These proxies characterize movement-related interference rather than directly measured physical activity.

Table 2
Signal-quality descriptors used for quality-aware fusion.

Modality	Main quality descriptors	Low-quality indicators	Role in the model
ECG	R-peak consistency, RR plausibility, baseline drift, high-frequency noise, missing-beat ratio	Implausible RR intervals, unstable R-peak detection, high local noise, severe baseline drift	HRV reliability estimation and ECG morphology weighting
PPG	Pulse detectability, amplitude stability, saturation ratio, abrupt amplitude change, motion contamination	Missing pulse peaks, saturation, abrupt amplitude fluctuation, severe motion artifact	Peripheral hemodynamic reliability estimation
ACC / activity-motion proxy	Movement-intensity stability, bed-motion indicator, ECG/PPG envelope instability, rejected PPG-segment ratio	High movement intensity, abrupt waveform instability, motion-contaminated physiological signals	Activity-confounding estimation and subgroup stratification
Optional PCG branch	Energy distribution, clipping ratio, time-frequency noise, S1/S2 detectability	Clipping, broadband noise, weak heart-sound components, unstable time-frequency patterns	Component-level acoustic morphology validation
Clinical metadata	Completeness, value plausibility, missingness pattern	Missing core covariates, implausible values, inconsistent metadata	Baseline-risk representation and missingness control

For stratification in the primary ICU cohort, the four proxy components are min–max normalized to $[0, 1]$ using training-set percentiles and averaged into a scalar activity/motion proxy score

$$s_{i,w}^{proxy} = \frac{1}{4}\left(\tilde{b}_{i,w}^{bed} + \tilde{d}_{i,w}^{ecg} + \tilde{d}_{i,w}^{ppg} + \tilde{r}_{i,w}^{ppg\text{-}motion}\right), \tag{9}$$

where $\tilde{(\cdot)}$ denotes the training-set normalized component, so that $s_{i,w}^{proxy} \in [0, 1]$ and larger values indicate stronger movement-related interference.

In datasets with direct triaxial accelerometry, movement is stratified into low-, moderate-, and high-activity groups using tertiles of the SMA in Eq. (6); in the primary ICU cohort, the same three strata are defined using training-set tertiles of the activity/motion proxy score $s_{i,w}^{proxy}$. All thresholds are estimated from the training set and then fixed for validation and testing. The proxy score is used only to represent movement-related interference and is not equivalent to directly measured physical activity.

Each modality is assigned a signal-quality score $q_i^r \in [0, 1]$, where larger values indicate higher reliability:

$$q_i^r = \mathrm{clip}\left(\sum_{j=1}^{J_r} \eta_j^r s_{i,j}^r, 0, 1\right). \tag{10}$$

Here, $s_{i,j}^r$ is a normalized quality descriptor, η_j^r is a training-set-defined weight, and J_r is the number of descriptors. ECG quality is estimated from R-peak consistency, RR plausibility, missing-beat ratio, baseline drift, and local noise. PPG quality uses pulse detectability, amplitude stability, saturation, and motion contamination. Optional PCG quality is estimated from energy distribution, clipping, S1/S2 detectability, and time-frequency noise.

The modality-availability mask is defined as

$$m_i^r = \begin{cases} 1, & \text{if modality } r \text{ is available,} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The pair (m_i^r, q_i^r) distinguishes unavailable modalities from available but low-quality signals.

Table 2 summarizes the quality descriptors used in the primary and component-level validation settings.

Table 3 summarizes the main notations.

3.5. Modality-specific encoders

Each available modality is first processed by a dedicated encoder. The ECG encoder learns electrical morphology and rhythm-related features from raw waveform segments. The PPG encoder learns pulse morphology and peripheral hemodynamic patterns. The optional PCG encoder learns acoustic cardiac patterns from spectrogram-like representations when PCG recordings are available. The ACC encoder models activity intensity and motion-related variation when direct ACC is available; otherwise, a proxy encoder maps movement-related and waveform-instability descriptors into an activity-context representation. The HRV encoder processes explicit autonomic features, while the clinical encoder maps tabular clinical metadata into a compact representation.

Table 3
Notation used in the proposed framework.

Symbol	Definition
$\mathcal{R}_{i,w}$	Modalities available for patient i in window w
$\mathcal{M}$	Fixed candidate-modality set (ECG, PPG, PCG, proxy, HRV, clinical)
$\mathbf{X}^{ecg}$	ECG waveform segment
$\mathbf{X}^{ppg}$	PPG waveform segment
$\mathbf{X}^{pcg}$	Optional PCG waveform or time-frequency representation
$\mathbf{X}^{acc}$	Direct triaxial ACC sequence, when available
$\mathbf{a}^{proxy}$	Activity- and motion-related proxy vector used in the primary ICU cohort
$\mathbf{h}^{hrv}$	HRV feature vector computed from NN intervals
$\mathbf{c}$	Available clinical metadata
$\mathbf{m}$	Modality-availability mask
$\mathbf{q}$	Modality-specific signal-quality scores
$\mathbf{z}^{auto}$	Autonomic latent representation
$\mathbf{z}^{morph}$	Morphological latent representation
$\mathbf{z}^{act}$	Activity-related latent representation
$\mathbf{z}^{base}$	Patient-specific baseline representation
$\mathbf{z}^{noise}$	Signal-noise and acquisition-artifact representation
$\hat{y}$	Predicted cardiovascular deterioration risk

For modality $r \in \mathcal{R}_{i,w} \cup \{hrv, clin\}$, the encoded representation is

$$\mathbf{u}_i^r = E_r(\mathbf{X}_i^r; \theta_r), \tag{12}$$

where $E_r(\cdot)$ denotes the modality-specific encoder and θ_r represents its trainable parameters. For HRV and clinical metadata, $\mathbf{X}_i^r$ refers to the corresponding feature vector rather than a raw waveform.

Before fusion, each representation is projected into a shared latent dimension:

$$\mathbf{v}_i^r = \mathbf{W}_r \mathbf{u}_i^r + \mathbf{b}_r. \tag{13}$$

This projection reduces dimensional mismatches among waveform, spectrogram, sequence, and tabular inputs.

3.6. Quality-aware multimodal fusion

Simple early fusion may amplify noisy modalities, while late fusion may fail to capture cross-modal physiological interactions. Therefore, we introduce a quality-aware gated fusion module defined over the fixed candidate-modality set $\mathcal{M} = \{\mathrm{ECG}, \mathrm{PPG}, \mathrm{PCG}, \mathrm{proxy}, \mathrm{HRV}, \mathrm{clinical}\}$, so that the availability mask, rather than a sample-dependent index set, controls which modalities contribute. For each modality $r \in \mathcal{M}$, the fusion weight is determined by its projected representation, signal-quality score, and availability mask:

$$\alpha_i^r = \frac{m_i^r \exp\left(\mathbf{w}_g^T [\mathbf{v}_i^r; q_i^r; m_i^r]\right)}{\sum_{s \in \mathcal{M}} m_i^s \exp\left(\mathbf{w}_g^T [\mathbf{v}_i^s; q_i^s; m_i^s]\right) + \epsilon}, \tag{14}$$

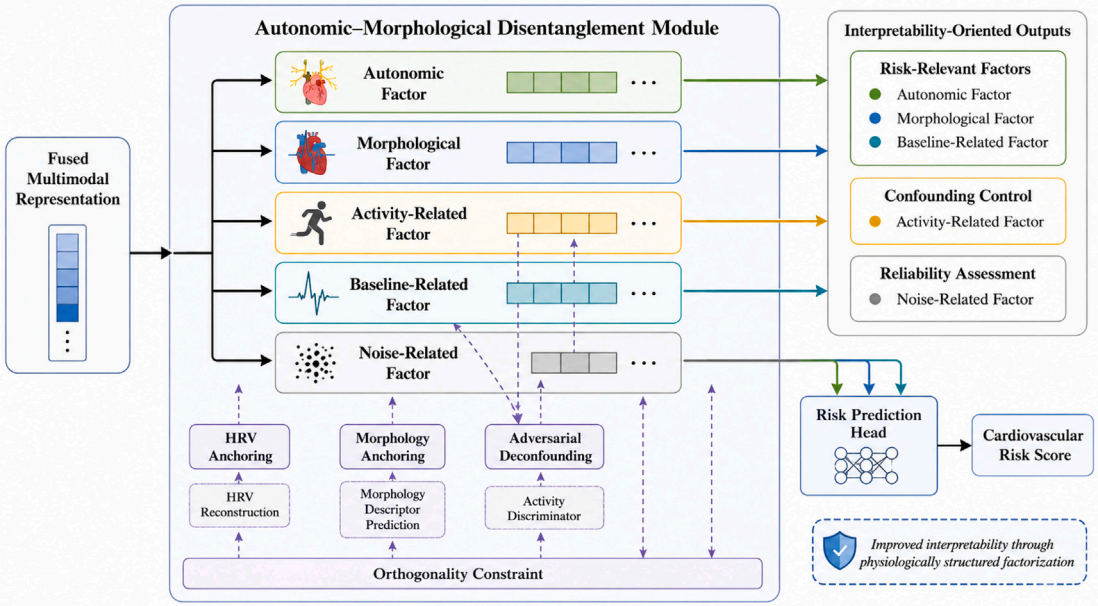

Fig. 2. Architecture of the autonomic–morphological structured factorization module. The fused representation is decomposed into autonomic, morphological, activity-related, baseline-related, and noise-related latent factors. In the primary cohort, the morphology factor is constrained by ECG and PPG descriptors; PCG-related morphology is an optional, design-compatible branch evaluated only in task-specific component-level experiments and is not part of the primary 24-h endpoint model.

where α_i^r is the normalized contribution of modality r, $\mathbf{w}_g$ is a trainable gating vector, and ϵ is a small constant for numerical stability. For an unavailable modality, $m_i^r = 0$, $\mathbf{v}_i^r = \mathbf{0}$, and $q_i^r = 0$, so it contributes to neither the numerator nor the denominator, and the mask is therefore not redundant. Signal-quality scores are defined for every modality in $\mathcal{M}$: waveform quality follows Table 2, HRV quality is derived from RR-interval reliability, the activity/motion-proxy quality from waveform-envelope stability, and clinical-metadata quality from covariate completeness and value plausibility.

The fused multimodal representation is computed as

$$\mathbf{h}_i = \sum_{r \in \mathcal{M}} \alpha_i^r \mathbf{v}_i^r. \tag{15}$$

This mechanism allows the model to down-weight unreliable signals and maintain prediction stability when one or more modalities are missing.

3.7. Autonomic–morphological structured factorization

The central design of AM-DiMNet is to factorize the fused representation into physiologically meaningful latent components. We use the term structured factorization to emphasize that these latent variables are guided by physiological anchors and probing analyses, but they should not be interpreted as proving strict causal disentanglement. The fused vector $\mathbf{h}_i$ is mapped into five latent components:

$$\begin{aligned} \mathbf{z}_i^{auto} &= G_{auto}(\mathbf{h}_i), \\ \mathbf{z}_i^{morph} &= G_{morph}(\mathbf{h}_i), \\ \mathbf{z}_i^{act} &= G_{act}(\mathbf{h}_i), \\ \mathbf{z}_i^{base} &= G_{base}(\mathbf{h}_i), \\ \mathbf{z}_i^{noise} &= G_{noise}(\mathbf{h}_i). \end{aligned} \tag{16}$$

Here, $\mathbf{z}_i^{auto}$ is designed to encode autonomic regulation patterns reflected by HRV and beat-to-beat variability. The vector $\mathbf{z}_i^{morph}$ is designed to encode cardiovascular morphology from available waveform modalities. In the primary cohort, this factor is mainly constrained by ECG and PPG morphology; in component-level validation datasets, PCG morphology is additionally evaluated when PCG is available. The vector $\mathbf{z}_i^{act}$ is intended to represent activity-related confounding, $\mathbf{z}_i^{base}$ patient-specific baseline variation, and $\mathbf{z}_i^{noise}$ signal quality and acquisition artifacts.

The risk prediction head uses physiologically motivated risk-related components:

$$\hat{y}_i = \sigma\left(f_{risk}\left([\mathbf{z}_i^{auto}; \mathbf{z}_i^{morph}; \mathbf{z}_i^{base}]\right)\right), \tag{17}$$

where $\sigma(\cdot)$ is the sigmoid function. Clinical metadata enter the model only through the clinical encoder and the resulting fused and baseline representations; they are not concatenated a second time at the risk head. Continuous clinical variables are imputed using training-set medians, categorical missingness is represented by an explicit unknown category, and every clinical variable is accompanied by a missingness indicator, while the clinical-modality mask and quality score are supplied to the fusion gate. The activity and noise factors are not used as direct risk evidence. Instead, they are used for deconfounding, uncertainty estimation, and robustness analysis.

The structured factorization module is shown in Fig. 2.

3.8. Physiological anchoring objectives

To encourage $\mathbf{z}^{auto}$ to represent autonomic regulation, an HRV reconstruction head is attached to the autonomic latent factor:

$$\hat{\mathbf{h}}_i^{hrv} = R_{hrv}(\mathbf{z}_i^{auto}). \tag{18}$$

The HRV anchoring loss is defined as

$$\mathcal{L}_{hrv} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\mathbf{h}}_i^{hrv} - \mathbf{h}_i^{hrv} \right\|_2^2. \tag{19}$$

This loss encourages the autonomic latent space to preserve explicit HRV information.

Similarly, morphology-related descriptors are used to constrain $\mathbf{z}^{morph}$. Let $\mathbf{r}_i^{morph}$ denote morphology descriptors derived from available morphology modalities. In the primary risk-prediction cohort, $\mathbf{r}_i^{morph}$ includes ECG and PPG descriptors, such as QRS-related descriptors, pulse

width, pulse amplitude, rise time, and decay time. In optional PCG validation datasets, PCG time-frequency energy descriptors and heart-sound abnormality descriptors are additionally evaluated.

The morphology prediction head is

$$\hat{\mathbf{r}}_i^{morph} = R_{morph}(\mathbf{z}_i^{morph}), \tag{20}$$

and the morphology anchoring loss is

$$\mathcal{L}_{morph} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{r}}_i^{morph} - \mathbf{r}_i^{morph}\right\|_2^2. \tag{21}$$

This constraint reduces the risk that morphological abnormalities are absorbed into the autonomic latent factor.

3.9. Adversarial deconfounding and orthogonality constraint

Physical activity and motion artifacts can alter HRV, PPG morphology, and signal quality, leading to false risk elevation. To reduce this confounding effect, an activity classifier is trained on $\mathbf{z}^{act}$, while adversarial classifiers are attached to $\mathbf{z}^{auto}$ and $\mathbf{z}^{morph}$ through gradient reversal layers.

The activity prediction loss from the activity latent factor is

$$\mathcal{L}_{act} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} a_{i,k}\log \hat{a}_{i,k}, \tag{22}$$

where $a_{i,k}$ is the activity-intensity label and $\hat{a}_{i,k}$ is the predicted probability for class k.

The adversarial deconfounding loss is defined as

$$\mathcal{L}_{adv} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} a_{i,k}\left[\log D_{auto}^{k}(\mathbf{z}_i^{auto}) + \log D_{morph}^{k}(\mathbf{z}_i^{morph})\right], \tag{23}$$

where D_{auto} and D_{morph} are adversarial activity discriminators. During optimization, gradient reversal encourages $\mathbf{z}^{auto}$ and $\mathbf{z}^{morph}$ to become less predictive of activity intensity.

To further reduce redundancy among latent factors, we first center and row-normalize each batch-level latent matrix,

$$\tilde{\mathbf{Z}}^{p} = \mathrm{RowNorm}\left(\mathbf{H}_B\mathbf{Z}^{p}\right), \qquad \mathbf{H}_B = \mathbf{I}_B - \tfrac{1}{B}\mathbf{1}\mathbf{1}^{T}, \tag{24}$$

where $\mathbf{H}_B$ removes the batch mean and $\mathrm{RowNorm}(\cdot)$ scales each row to unit norm, so that the penalty reflects cross-factor redundancy rather than latent magnitude or batch mean. The orthogonality constraint is then

$$\mathcal{L}_{orth} = \frac{1}{|\mathcal{P}|(|\mathcal{P}|-1)}\sum_{p\neq q}\left\|\frac{(\tilde{\mathbf{Z}}^{p})^{T}\tilde{\mathbf{Z}}^{q}}{B}\right\|_F^2, \tag{25}$$
$$p, q \in \mathcal{P} = \{auto, morph, act, base, noise\},$$

where B is the batch size, $\mathbf{Z}^{p}$ is the batch-level latent matrix for factor p, and $\|\cdot\|_F$ denotes the Frobenius norm.

3.10. Risk prediction, calibration, and threshold selection

The primary risk prediction loss is the weighted binary cross-entropy:

$$\mathcal{L}_{risk} = -\frac{1}{N}\sum_{i=1}^{N}\left[\omega_1 y_i\log(\hat{y}_i) + \omega_0(1-y_i)\log(1-\hat{y}_i)\right], \tag{26}$$

where ω_1 and ω_0 are class weights used to address class imbalance.

To improve clinical usability, probability calibration is encouraged using a Brier-type regularization term:

$$\mathcal{L}_{cal} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2. \tag{27}$$

The total training objective is

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{risk} + \lambda_{hrv}\mathcal{L}_{hrv} + \lambda_{morph}\mathcal{L}_{morph} + \lambda_{act}\mathcal{L}_{act} \\ &+ \lambda_{adv}\mathcal{L}_{adv} + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{cal}\mathcal{L}_{cal}.\end{aligned} \tag{28}$$

Each loss term corresponds to a specific methodological purpose: risk discrimination, autonomic anchoring, morphology anchoring, activity modeling, deconfounding, latent factor separation, and probability calibration.

For binary classification metrics, the operating threshold is selected on the validation set and then fixed for the test set. Unless otherwise specified, the threshold is selected by maximizing validation Macro-F1 under the constraint that validation sensitivity is not lower than 0.75. For the early-warning operating point, consecutive above-threshold windows are consolidated into alarm episodes using a refractory interval, and the resulting alarm burden and lead time are reported in Section 4.16. This constraint reflects the early-warning purpose of the model and prevents a high-specificity but low-sensitivity threshold from being selected. AUROC, AUPRC, Brier score, and ECE are threshold-independent and are computed from predicted probabilities.

3.11. Temporal alignment and leakage-safe prediction

Temporal alignment is designed to ensure that the model predicts future deterioration rather than detecting ongoing or imminent events. For each patient, candidate windows are generated chronologically using a fixed observation length and sliding step. In the primary configuration, the observation window is 5 min, the sliding step is 60 s, the leakage-prevention gap is 1.5 h, and the prediction horizon is 24 h.

For each candidate window, the model receives only information available within the observation window. ECG, PPG, activity/motion proxy information, HRV features, modality masks, signal-quality scores, and available clinical metadata are all constructed from the observation interval. No physiological samples, clinical measurements, medication changes, procedure records, or event indicators from the gap window or prediction window are used as inputs. The outcome label is assigned only after the input window has been fixed.

The leakage-safe temporal structure is defined as

$$[t_{i,w}^{start}, t_{i,w}^{end}] \rightarrow (t_{i,w}^{end}, t_{i,w}^{end} + \Delta_g) \rightarrow [t_{i,w}^{end} + \Delta_g, t_{i,w}^{end} + \Delta_g + \Delta_p], \tag{29}$$

where the first interval is used for feature extraction, the second interval is discarded as a leakage-prevention gap, and the third interval is used only for outcome assessment.

For patients with multiple valid windows, all windows from the same patient are assigned exclusively to the training, validation, or test split. Standardization parameters, HRV normalization statistics, activity-intensity thresholds, signal-quality thresholds, and calibration parameters are estimated using only the training or validation sets and are then fixed for testing. This prevents information from test patients from influencing preprocessing, threshold selection, or calibration.

For patient-level evaluation, the pre-specified primary patient-level risk score $\hat{Y}_i$ is the mean of the highest decile of valid window-level probabilities:

$$\hat{Y}_i = \frac{1}{|\Omega_i^{top}|}\sum_{w\in\Omega_i^{top}}\hat{y}_{i,w}, \tag{30}$$

where Ω_i denotes the set of valid windows for patient i and Ω_i^{top} is the subset corresponding to the highest decile of predicted risk scores. Maximum-risk aggregation,

$$\hat{Y}_i^{max} = \max_{w\in\Omega_i}\hat{y}_{i,w}, \tag{31}$$

is reported only as a sensitivity analysis.

Unless otherwise stated, the main prediction, calibration, ablation, hyperparameter-sensitivity, endpoint-sensitivity, subgroup, decision-curve, and statistical-significance analyses report patient-level results

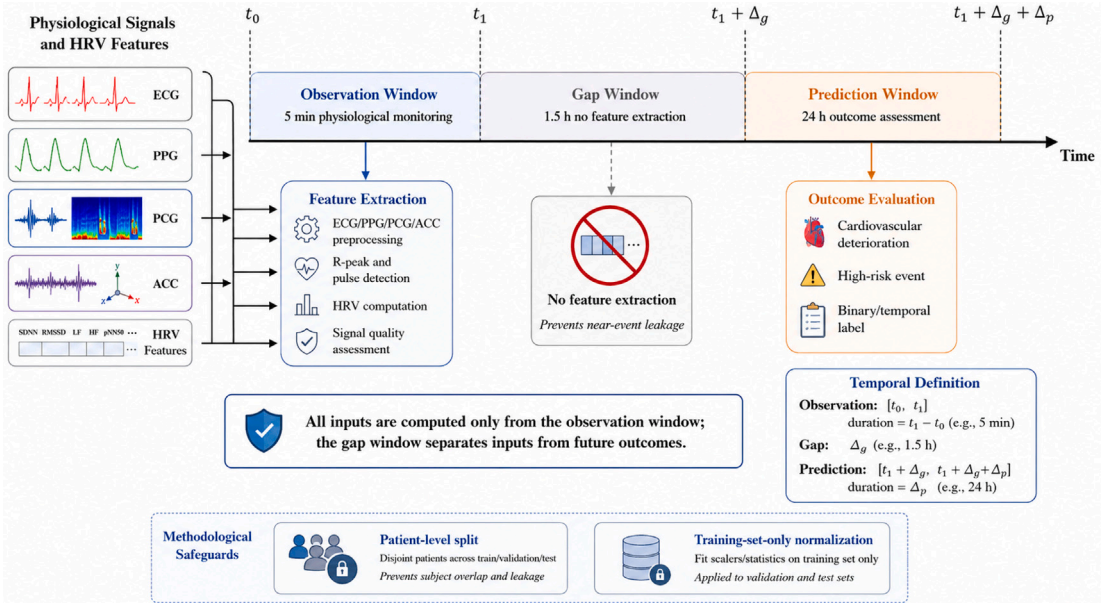

Fig. 3. Leakage-safe temporal alignment for cardiovascular deterioration prediction. Physiological signals, HRV features, signal-quality indicators, modality masks, and available clinical metadata are extracted only from the observation window. The gap interval is discarded to reduce near-event leakage, and the composite endpoint is evaluated only within the future prediction window.

based on $\hat{Y}_i$, with all confidence intervals obtained from 1000 patient-level bootstrap samples. The signal-quality and activity-stratified analysis is reported at the window level, with confidence intervals obtained by patient-cluster bootstrap resampling so that overlapping windows from the same patient are never treated as independent; window-level predictions are also used for the time-resolved case study and early-warning lead-time illustration. For missing-modality stress tests, modalities are masked at the window level and predictions are then aggregated to the patient level using $\hat{Y}_i$ before reporting. Every table and figure caption states its evaluation unit.

The proposed temporal design is illustrated in Fig. 3.

3.12. Training algorithm

Algorithm 1 summarizes the training procedure of AM-DiMNet. The algorithm is written in a modality-flexible form so that the primary cohort and component-level validation datasets can be processed under the same framework without assuming simultaneous availability of all modalities.

3.13. Implementation details

The ECG and PPG encoders use one-dimensional residual blocks, while the optional PCG encoder uses two-dimensional convolutional blocks on log-Mel spectrograms and is activated only for datasets where PCG is available. The ACC encoder uses temporal convolutional layers or feature-level activity proxies depending on the dataset, and HRV and clinical metadata are processed by multilayer perceptrons. All modality embeddings are projected to a shared latent dimension before fusion, and modality dropout is applied during training to improve robustness to missing modalities. Parameters are optimized with Adam, and early stopping follows validation AUPRC and calibration performance.

Table 4 provides the implementation configurations used in the experiments.

The default loss weights are initialized as

$$\lambda_{hrv} = 0.35, \quad \lambda_{morph} = 0.28, \quad \lambda_{act} = 0.20,$$
$$\lambda_{adv} = 0.15, \quad \lambda_{orth} = 0.06, \quad \lambda_{cal} = 0.08. \tag{32}$$

Algorithm 1 Training procedure of AM-DiMNet.

Require: Training set $\mathcal{D}_{\text{train}}$ and validation set $\mathcal{D}_{\text{val}}$
Require: Available modality set $\mathcal{R}_{i,w}$ for each patient-window pair
Require: Encoders $\{E_r\}$, factorization heads $\{G_p\}$, and risk head f_{risk}
Ensure: Trained model parameters Θ
Initialize all trainable parameters Θ
Estimate normalization statistics from $\mathcal{D}_{\text{train}}$
Estimate activity and signal-quality thresholds from $\mathcal{D}_{\text{train}}$
for each training epoch **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{train}}$ **do**
    for each available modality $r \in \mathcal{R}_{i,w}$ **do**
      Preprocess modality r within the observation window
      Extract modality-specific waveform or feature representation
      Compute modality mask m_i^r and quality score q_i^r
      Encode modality representation using E_r as in Eq. (12)
    end for
    Extract HRV features from ECG-derived NN intervals
    Construct morphology descriptors from available waveform modalities
    Compute quality-aware fusion weights using Eq. (14)
    Obtain fused representation using Eq. (15)
    Generate structured latent factors using Eq. (16)
    Predict future cardiovascular risk using Eq. (17)
    Compute the total training loss using Eq. (28)
    Update Θ by minimizing the total loss
  end for
  Evaluate AUROC, AUPRC, Macro-F1, Brier score, and ECE on $\mathcal{D}_{\text{val}}$
  Save the model if validation AUPRC and calibration performance improve
end for
return Trained AM-DiMNet model

These weights are selected on the validation set and are not tuned on the test set.

The hyperparameter search space is summarized in Table 5, and the final configuration is chosen by validation AUPRC and calibration performance. Beyond reporting the candidate search space, we performed a one-factor-at-a-time sensitivity analysis for the five principal objective weights, including λ_{hrv}, λ_{morph}, λ_{adv}, λ_{orth}, and λ_{cal}, as well as the shared

Table 4
Preprocessing and implementation settings used in the experimental configuration.

Item	Setting
ECG sampling rate	125–500 Hz, dataset dependent
PPG sampling rate	62.5–125 Hz, dataset dependent
PCG sampling rate	1000–2000 Hz after resampling
ACC sampling rate	32–100 Hz, dataset dependent
Observation window	5.0 min
Sliding step	60 s
Prediction gap	1.5 h
Prediction horizon	24 h
ECG band-pass filter	0.5–40 Hz
PPG band-pass filter	0.5–8 Hz
PCG band-pass filter	20–400 Hz
ECG heartbeat length	512 samples
PCG spectrogram bins	64 Mel bins
Shared embedding dimension	128
Latent factor dimension	32 per factor
Batch size	96
Learning rate	7.5×10^{-4}
Weight decay	1.2×10^{-4}
Dropout rate	0.27
Modality dropout rate	0.18
Maximum epochs	120
Early stopping patience	16 epochs
Random seeds	5 independent runs

Table 5
Hyperparameter search space used for validation-set model selection.

Hyperparameter	Candidate values
Shared embedding dimension	64, 128, 192
Latent factor dimension	16, 32, 48
Batch size	64, 96, 128
Learning rate	3×10^{-4}, 7.5×10^{-4}, 10^{-3}
Weight decay	5×10^{-5}, 1.2×10^{-4}, 3×10^{-4}
Dropout rate	0.15, 0.27, 0.35
Modality dropout rate	0.10, 0.18, 0.30
λ_{hrv}	0.20, 0.35, 0.50
λ_{morph}	0.15, 0.28, 0.40
λ_{adv}	0.05, 0.15, 0.25
λ_{orth}	0.03, 0.06, 0.10
λ_{cal}	0.04, 0.08, 0.12

Table 6
Computational profile of AM-DiMNet in the primary risk-prediction configuration.

Item	Value
Trainable parameters	2.84M
Model size	11.6 MB
Mean training time per epoch	74.3 s
Total training time	2.7 h
Inference time per window	3.8 ms
Peak GPU memory	3.2 GB
Hardware	NVIDIA RTX 3090

embedding dimension and the latent factor dimension. For each setting, only the parameter under examination was changed, while all remaining parameters were fixed at their validation-selected values. Each configuration was evaluated over the same five independent random seeds used for the main experiments (Table 4), and we report the mean and standard deviation across these seeds, so that the sensitivity table reflects run-to-run variability rather than a single run. The test set was used neither for hyperparameter selection nor for sensitivity assessment.

To clarify computational cost, Table 6 reports the model complexity and runtime profile measured in the primary risk-prediction configuration.

The values in Table 6 characterize server-side execution on an NVIDIA RTX 3090 and should not be read as evidence of wearable-device deployment. The reported inference time covers neural-network execution for one preprocessed observation window and excludes signal acquisition, filtering, R-peak detection, HRV computation, data transfer, and device-specific scheduling overhead. Because risk estimates are updated every 60 s, this latency is compatible with workstation- or bedside-server-assisted monitoring, whereas direct execution on a mobile processor, smartwatch, or microcontroller has not been evaluated.

On-device deployment would therefore require model compression and hardware-specific evaluation. Candidate strategies include post-training quantization, structured pruning, knowledge distillation, shared lightweight waveform encoders, and selective activation of available modality branches. End-to-end measurements of latency, memory, energy consumption, and thermal behavior remain necessary before claiming wearable or fully on-device feasibility.

3.14. Model outputs and interpretability

AM-DiMNet produces three types of outputs. The first output is the future cardiovascular deterioration risk probability $\hat{y}_i$. The second output is a set of structured latent factors, including $\mathbf{z}^{auto}$, $\mathbf{z}^{morph}$, $\mathbf{z}^{act}$, $\mathbf{z}^{base}$, and $\mathbf{z}^{noise}$. The third output is an explanation vector that summarizes the contributions of each physiological factor to the final prediction.

For patient-level explanation, the contribution of each risk-head factor $p \in \{\text{auto}, \text{morph}, \text{base}\}$ is estimated by the change in predicted risk after replacing that factor with its fixed training-set reference vector $\bar{\mathbf{z}}^p = \frac{1}{N_{\text{train}}} \sum_{i \in \mathcal{D}_{\text{train}}} \mathbf{z}_i^p$:

$$\Delta_i^p = \left| \hat{y}_i - \sigma\left(f_{risk}\left([\mathbf{z}_i^{auto}; \mathbf{z}_i^{morph}; \mathbf{z}_i^{base}]_{p \leftarrow \bar{\mathbf{z}}^p} \right) \right) \right|, \tag{33}$$

where $[\cdot]_{p \leftarrow \bar{\mathbf{z}}^p}$ denotes the representation with factor p replaced by its training-set reference vector. Only the autonomic, morphological, and baseline factors are analyzed this way, because these are the factors that directly enter the risk-prediction head; the activity and noise factors, which do not enter the risk head, are instead assessed through association, leakage probing, and robustness analyses. This replacement is a sensitivity analysis on the learned representation, not a physiological or clinical intervention.

To verify whether the learned factors are physiologically meaningful, we compute associations between latent factors and interpretable physiological descriptors. The autonomic factor is expected to correlate with SDNN, RMSSD, LF/HF, and sample entropy. The morphological factor is expected to correlate with ECG and PPG morphology descriptors in the primary cohort and with PCG acoustic descriptors in optional component-level validation datasets. The activity factor is expected to encode movement intensity from the activity/motion proxy (or direct ACC intensity in datasets with accelerometry), whereas the noise factor is expected to reflect signal-quality scores and missing-modality patterns.

In addition to correlation analysis, factor-level leakage probing is used to assess whether activity information remains recoverable from $\mathbf{z}^{auto}$ or $\mathbf{z}^{morph}$ and whether risk labels are spuriously predictable from $\mathbf{z}^{noise}$. Latent-factor replacement is also used to estimate the change in predicted risk after replacing one risk-head factor with its training-set reference vector. These analyses provide empirical evidence of physiological structure, but they are interpreted as mechanism-consistent validation rather than causal proof.

The factor-level outputs are interpreted as post hoc internal model explanations. Associations with physiological descriptors and changes in predicted risk after factor replacement indicate model consistency with the intended factor definitions, but they do not establish causal physiological effects, clinical correctness, or agreement with independent clinicians. In particular, latent-factor replacement is a sensitivity analysis on the learned latent representation rather than an intervention on the underlying biological process.

Table 7
Principal sources of cross-dataset heterogeneity and their implications for interpretation.

Dataset	Evaluation context	Principal heterogeneity	Harmonization used	Interpretation boundary
MIMIC-IV Waveform	ICU bedside risk prediction	Bedside devices, variable waveform quality, incomplete modalities, composite future outcome	Dataset-specific filtering, training-set normalization, modality masks, and quality-aware fusion	Primary end-to-end evidence within the ICU waveform cohort
MIMIC-IV-ECG	Clinical 12-lead ECG analysis	Different lead configuration, recording duration, acquisition workflow, and diagnostic distribution	ECG-specific preprocessing, common embedding dimension, and dataset-specific prediction head	ECG morphology adaptability, not full multimodal risk transfer
PTB-XL	External diagnostic ECG analysis	Population, device, diagnostic label, prevalence, and acquisition-setting shift	Lead-aware preprocessing, resampling, normalization, and ECG encoder evaluation	External ECG representation assessment only
CirCor DigiScope	PCG abnormality analysis	Acoustic sensor, recording site, age distribution, noise profile, and heart-sound label shift	PCG-specific time–frequency transformation and dataset-specific task head	Optional PCG branch validation, not contribution to the primary endpoint
EPHNOGRAM	Synchronous ECG–PCG analysis	Electrical–acoustic synchronization, sensor configuration, and task-definition shift	Temporal alignment and shared morphology representation	Mechanism-oriented ECG–PCG consistency evidence
WESAD	Wearable activity-confounding analysis	Wearable sensors, controlled activity/stress protocol, and non-ICU population	Activity-stratified preprocessing and activity-aware representation analysis	Activity-deconfounding evidence rather than cardiovascular outcome validation

4. Experimental results

4.1. Experimental design

This section evaluates whether the proposed AM-DiMNet improves future cardiovascular deterioration prediction by learning physiologically structured representations of autonomic regulation, waveform morphology, activity-related confounding, patient-specific baseline variation, and signal-quality uncertainty. The experiments follow the leakage-safe prediction setting described in Section 3. For each patient-window sample, all physiological features are extracted only from the observation window, a temporal gap is inserted to reduce near-event leakage, and the outcome is evaluated only within the future prediction interval.

Because no single public dataset simultaneously provides ECG, PPG, PCG, ACC, HRV, clinical metadata, signal-quality annotations, and time-resolved cardiovascular outcomes, the experiments are organized according to a dataset-aware evaluation protocol.

Cross-dataset evaluation in this study is affected by several forms of heterogeneity. First, the datasets represent different clinical and experimental environments, ranging from intensive-care bedside monitoring to diagnostic ECG acquisition, digital auscultation, synchronous ECG–PCG recording, and controlled wearable sensing. Second, the waveform modalities differ in sensor hardware, lead configuration, sampling frequency, recording duration, noise characteristics, and signal-quality distribution. Third, the corresponding tasks and labels are not identical: the primary cohort uses a future composite cardiovascular deterioration endpoint, whereas the complementary datasets provide diagnostic, morphology-related, heart-sound, cross-modal consistency, or activity-related labels. Differences in population characteristics and class prevalence further introduce covariate and label-distribution shifts.

To reduce avoidable technical discrepancies, each modality is processed using dataset-appropriate filtering and resampling, with normalization parameters estimated only from the corresponding training data. Modality-specific encoders map heterogeneous input formats into a common embedding dimension, while dataset-specific task heads are used when the label spaces differ. These procedures provide architectural and representation-level compatibility, but they do not eliminate domain shifts caused by population, device, acquisition protocol, clinical environment, or endpoint definition. Accordingly, results from the complementary datasets are interpreted as evidence of component-level adaptability and physiological consistency rather than as direct estimates of full-system external generalization.

The primary evidence comes from end-to-end cardiovascular deterioration prediction on the MIMIC-IV Waveform cohort, where ECG, PPG, activity/motion proxy information, explicit HRV features, signal-quality scores, modality masks, and available clinical metadata are used. PCG is not assumed to be simultaneously available in the primary risk-prediction cohort. Instead, PCG-related morphology is evaluated through complementary component-level validation datasets. Therefore, these complementary datasets are used to assess modality-specific transferability and mechanism consistency rather than to claim external validation of the full model in a single all-modality cohort.

The experiments are organized around seven questions. First, we evaluate whether AM-DiMNet improves discrimination and calibration compared with HRV-only, single-modality, and conventional fusion baselines. Second, we examine whether each methodological component contributes to performance. Third, we test robustness under missing modalities and low-quality signals. Fourth, we evaluate whether activity-related confounding can be reduced. Fifth, we analyze whether the composite endpoint definition affects model performance. Sixth, we assess whether the learned latent factors are physiologically structured and interpretable. Seventh, we analyze whether performance remains stable across clinically relevant subgroups.

Table 7 summarizes these sources of heterogeneity and their interpretation boundaries, Table 8 summarizes the dataset roles and evidence levels used in the evaluation protocol.

4.2. Cohort construction and endpoint statistics

In the primary MIMIC-IV Waveform experiment, each sample corresponds to a leakage-safe patient-window pair. The observation window is 5 min, the sliding step is 60 s, the leakage-prevention gap is 1.5 h, and the prediction horizon is 24 h. A positive label is assigned if at least one predefined cardiovascular deterioration event occurs within the future prediction interval. The endpoint is a time-resolved composite outcome and is not defined by discharge-level ICD codes alone. ICD codes are used only for baseline comorbidity characterization, cohort stratification, or sensitivity analysis.

The composite endpoint includes cardiac arrest or cardiopulmonary resuscitation, ICU or in-hospital death with available event time, initiation or marked escalation of vasoactive support, acute heart-failure deterioration requiring intensified treatment, clinically documented malignant arrhythmia, and urgent cardiovascular intervention. Event timestamps are obtained from ICU chart events, procedure records, medication and input events, and linked hospital records. When multiple endpoint components occur in the same prediction interval, the earliest eligible event is used as the first qualifying event.

Windows are excluded if the patient already satisfies an endpoint criterion before or during the observation window. Windows are also excluded if the first eligible endpoint occurs within the gap interval,

Table 8
Dataset roles and evidence levels in the dataset-aware evaluation protocol.

Dataset	Available modalities used	Task	Evidence level	Role in this study
MIMIC-IV Waveform	ECG, PPG, activity/motion proxy scores, HRV, SQI, modality masks, available clinical metadata	24-h future cardiovascular deterioration prediction	Primary end-to-end evaluation	Training, validation, and testing of the main AM-DiMNet risk-prediction model
MIMIC-IV-ECG	12-lead ECG and clinical metadata	ECG morphology and CVD phenotype validation	Component-level validation	Evaluation of ECG morphology-related representation and clinical ECG transferability
PTB-XL	12-lead ECG and diagnostic labels	External ECG diagnostic validation	Component-level validation	External validation of ECG morphology encoding rather than full multimodal risk prediction
CirCor DigiScope	PCG and metadata	Abnormal heart sound detection	PCG branch validation	Evaluation of PCG morphology encoding in an independent heart-sound dataset
EPHNOGRAM	Synchronous ECG and PCG	ECG–PCG consistency analysis	Mechanism-oriented validation	Assessment of electrical–acoustic morphology consistency
WESAD	ECG, BVP, ACC	Activity-confounding analysis	Mechanism-oriented validation	Validation of activity-related deconfounding under physiological activity variation

Table 9
Primary cohort statistics for leakage-safe cardiovascular deterioration prediction.

Item	Value
Patients after quality control	4826
Valid observation windows	118,342
Median windows per patient	22 [14, 37]
Positive patients	1055 (21.9%)
Positive windows	22,148 (18.7%)
Observation window	5.0 min
Sliding step	60 s
Leakage-prevention gap	1.5 h
Prediction horizon	24 h
Median age	66.8 [56.1, 76.9] years
Male patients	2643 (54.8%)
Median ECG SQI	0.83 [0.71, 0.92]
Median PPG SQI	0.78 [0.64, 0.89]
Low activity/motion proxy threshold	$s^{proxy} < 0.34$
High activity/motion proxy threshold	$s^{proxy} > 0.61$
Low SQI threshold	Combined SQI < 0.62
High SQI threshold	Combined SQI > 0.84

Table 10
Patient-level data partition with the corresponding window and endpoint counts for each split. Splits are performed at the patient level, so all windows from a patient belong to a single split.

Item	Training	Validation	Test	Total
Patients	3379	723	724	4826
Positive patients	741	157	157	1055
Negative patients	2638	566	567	3771
Observation windows	82,951	17,612	17,779	118,342
Positive windows	15,522	3287	3339	22,148
First qualifying endpoints	741	157	157	1055

Table 11
Patient-level first-qualifying-endpoint counts by component for each data split. Each positive patient contributes one first qualifying endpoint, so column totals match the positive-patient counts in Table 10 and row totals match the component counts in Table 13.

Endpoint component	Training	Validation	Test	Total
Cardiac arrest or CPR	64	14	13	91
ICU or in-hospital death	153	33	32	218
Vasoactive support escalation	231	49	49	329
Acute heart-failure deterioration	122	26	26	174
Malignant arrhythmia	106	22	23	151
Urgent cardiovascular intervention	65	13	14	92
Composite (first qualifying endpoints)	741	157	157	1055

Table 12
Cohort-construction flow for the primary MIMIC-IV waveform experiment. Each row reports the patients (and, once windowing begins, the observation windows) remaining after the corresponding eligibility or exclusion step, ending with the patient-level train–validation–test split.

Cohort-construction step	Patients	Windows
Initial linked ICU stays	18,342	—
Bedside waveform records available	14,905	—
ECG available	12,760	—
PPG available	10,133	—
Sufficient observation duration	8472	—
Valid ECG-derived RR intervals	7015	—
Successful waveform–clinical alignment	5912	152,004
After pre-existing-endpoint window removal	5410	138,470
After gap-window endpoint removal	4826	118,342
Final eligible cohort	4826	118,342
Training split	3379	82,951
Validation split	723	17,612
Test split	724	17,779

because these cases do not satisfy the predefined prediction horizon. Patients with insufficient waveform duration, unreliable ECG-derived RR intervals, unavailable event-time information, or poor temporal alignment between waveform and clinical records are excluded before model training.

Table 9 reports the primary cohort statistics.

The patient-level train–validation–test partition, together with the corresponding window and endpoint counts, is summarized in Table 10. Splits are performed at the patient level, so all windows from a given patient belong to a single split.

Table 11 further breaks down the first-qualifying-endpoint counts by component for each split. Because each positive patient contributes exactly one first qualifying endpoint, the column totals equal the positive-patient counts in Table 10, and the row totals equal the component counts in Table 13.

The cohort-construction flow, reporting the patients and windows remaining after each eligibility and exclusion step and ending with the final patient-level split, is shown in Table 12. This funnel makes the selection process auditable and clarifies how the initially linked ICU stays are reduced to the final analysis cohort.

Table 13 summarizes the endpoint components and timestamp sources. The reported counts correspond to the first qualifying endpoint per positive patient. The prevalence values are computed among all patients after quality control.

The manuscript follows the TRIPOD + AI recommendations for prediction-model studies [35], with attention to the data source, participant flow (Table 12), predictor and outcome definitions, missing-data handling, data partitioning (Table 10), calibration, model specification, and reproducibility. A completed reporting checklist and a

Table 13
Endpoint components, extraction rules, timestamp sources, and prevalence for the primary composite cardiovascular deterioration label.

Endpoint component	Operational definition	Timestamp source	First-event patients	Prevalence	Median time to event
Cardiac arrest or CPR	Documented cardiac arrest, cardiopulmonary resuscitation, defibrillation during resuscitation, or resuscitation procedure during ICU stay	ICU chart events and procedure records	91 (8.6%)	1.9%	7.4 h
ICU or in-hospital death	Death with sufficient temporal resolution occurring within the future prediction interval	Linked hospital records, ICU discharge records, and death-time fields	218 (20.7%)	4.5%	13.2 h
Vasoactive support escalation	New initiation of vasopressor or inotropic infusion, or dose escalation greater than 30% sustained for at least 30 min after a stable baseline period	Medication events, input events, and vasoactive infusion records	329 (31.2%)	6.8%	9.3 h
Acute heart-failure deterioration	Treatment intensification consistent with acute cardiac decompensation, including escalation of intravenous diuretics, inotropes, or documented decompensated heart-failure management	Chart events, medication records, and linked clinical records	174 (16.5%)	3.6%	11.7 h
Malignant arrhythmia	Documented ventricular tachyarrhythmia, severe bradyarrhythmia, high-grade conduction deterioration, or rhythm deterioration requiring treatment or clinical documentation	Rhythm-related chart events, procedure records, and linked diagnostic records	151 (14.3%)	3.1%	8.9 h
Urgent cardiovascular intervention	Urgent cardiovascular procedure or intervention related to hemodynamic or electrical instability, excluding elective or scheduled procedures	Procedure records and linked hospital records	92 (8.7%)	1.9%	15.4 h
Composite endpoint	Earliest occurrence of any eligible endpoint component within the future prediction interval	Combined time-stamped records	1055 (100.0%)	21.9%	10.5 h

PROBAST + AI-informed [36] risk-of-bias and applicability assessment are provided in Appendix B (Tables S1 and S2). This is a retrospective model-development study rather than a prospective clinical evaluation.

4.3. Evaluation metrics and statistical analysis

The primary discrimination metrics include AUROC, AUPRC, accuracy, precision, recall, Macro-F1, and Matthews correlation coefficient (MCC). Because cardiovascular deterioration prediction is affected by class imbalance, AUPRC and Macro-F1 are considered more informative than accuracy alone. Calibration is evaluated using the Brier score and expected calibration error (ECE).

The Brier score is computed as

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{34}$$

where $\hat{y}_i$ is the predicted risk probability and y_i is the ground-truth label.

For binary deterioration-risk prediction, calibration is assessed by comparing the mean predicted risk with the observed event frequency within each probability bin, rather than by comparing thresholded classification accuracy with mean predicted confidence. Predictions are divided into M equal-width probability bins over $[0, 1]$. For the m-th bin B_m, the observed event frequency is $\text{obs}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} y_i$ and the mean predicted risk is $\text{pred}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} \hat{y}_i$. The expected calibration error is

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{obs}(B_m) - \text{pred}(B_m)\right|, \tag{35}$$

where y_i is the ground-truth deterioration label, $\hat{y}_i$ is the predicted risk probability, and N is the number of evaluated predictions. We use $M = 15$ equal-width probability bins in the main experiments. In the primary analysis, i indexes patient-level aggregated predictions, so that observed event frequency and mean predicted risk are compared at the patient level; window-level calibration is treated as a secondary analysis and reported separately.

Because multiple overlapping windows can be generated from the same patient, window-level predictions are not treated as statistically independent. All train-validation-test partitions are performed at the patient level. For patient-level evaluation, window-level probabilities are aggregated within each patient using the maximum risk score or the mean of the highest decile of risk scores. Confidence intervals are estimated using patient-level bootstrap resampling with 1000 repetitions. Pairwise comparisons between AM-DiMNet and baseline models are performed using the Wilcoxon signed-rank test on patient-level prediction summaries, followed by Holm–Bonferroni correction for multiple comparisons.

For calibration fairness, all models are calibrated on the validation set before reporting final calibration metrics. Neural models are calibrated using temperature scaling, whereas logistic regression, random forest, and XGBoost are calibrated using validation-set Platt scaling. The test set is not used for calibration parameter selection. For threshold-dependent metrics, the operating threshold is selected on the validation set by maximizing Macro-F1 under the constraint that sensitivity is not lower than 0.75, and the selected threshold is then fixed for testing.

4.4. Baseline methods

The proposed AM-DiMNet is compared with six groups of baselines in the primary risk-prediction cohort. The first group includes traditional HRV-based machine learning models, including logistic regression, random forest, and XGBoost. The second group includes single-modality deep models using ECG or PPG alone. The third group includes conventional multimodal fusion methods, including early fusion, late fusion, and attention fusion using the modalities available in the primary cohort. The fourth group includes missing-modality-aware fusion using modality dropout. The fifth group includes a structured-representation baseline using adversarial deconfounding without physiological anchoring. The sixth group contains two pretrained representation-learning baselines, MOMENT and NormWear: MOMENT is a general-purpose time-series foundation model [21], whereas NormWear is a foundation model pretrained specifically on heterogeneous wearable physiological signals [22].

For a fair comparison, both pretrained encoders received the same leakage-safe ECG and PPG observation windows used by AM-DiMNet. The standardized ECG and PPG channels were resampled to a common 100 Hz rate, segmented to each model's native input length, and processed with the official pretrained checkpoints (MOMENT-1-large for MOMENT and NormWear-base for NormWear); the resulting token

Table 14
Implementation Details of the pretrained representation-learning baselines. Both models use the same downstream inputs, prediction head, and calibration as AM-DiMNet, so that differences reflect the pretrained waveform representation rather than access to additional covariates.

Model	Pretrained checkpoint	Waveform inputs	Additional downstream inputs	Adaptation	Calibration
MOMENT	MOMENT-1-large	ECG, PPG (100 Hz, 5-min windows)	HRV, motion proxy, clinical metadata, SQI, modality masks	Frozen backbone + 2-layer MLP head	Temperature scaling
NormWear	NormWear-base	ECG, PPG (100 Hz, 5-min windows)	HRV, motion proxy, clinical metadata, SQI, modality masks	Frozen backbone + 2-layer MLP head	Temperature scaling

Table 15
Baseline implementation summary. ✓ indicates that the baseline received the corresponding input and – indicates it did not (HRV, explicit HRV features; Clin., clinical metadata; SQI, signal-quality scores; Mask, modality-availability masks). All neural models use a two-layer MLP head, Adam, and validation-set temperature scaling, whereas HRV machine-learning models use validation-set Platt scaling; hyperparameters are selected by grid search on validation AUPRC over five seeds. * frozen backbone; only the MLP head (≈0.1M) is trained. Pretrained-model specifics are in Table 14.

Baseline	Architecture	Waveforms	HRV	Clin.	SQI	Mask	Params	Calib.
Logistic regression (HRV)	Linear, L2-regularized	None	✓	–	–	–	0.001M	Platt
Random forest (HRV)	500 trees	None	✓	–	–	–	–	Platt
XGBoost (HRV)	600 trees, depth 6	None	✓	–	–	–	–	Platt
ECG-CNN	1D ResNet, 5 blocks	ECG	–	–	–	–	1.9M	Temp.
PPG-CNN	1D ResNet, 4 blocks	PPG	–	–	–	–	1.4M	Temp.
ECG–PPG early fusion	1D ResNet, input concat	ECG, PPG	–	–	–	–	2.1M	Temp.
ECG–PPG late fusion	Two 1D ResNets, late concat	ECG, PPG	–	–	–	–	3.3M	Temp.
Attention fusion	Modality encoders + attention	ECG, PPG	✓	✓	✓	✓	3.6M	Temp.
Modality-dropout fusion	Attention + modality dropout	ECG, PPG	✓	✓	✓	✓	3.6M	Temp.
Adversarial deconfounding	Encoders + GRL activity head	ECG, PPG	✓	✓	✓	✓	3.9M	Temp.
MOMENT (matched inputs)	MOMENT-1-large (frozen) + MLP	ECG, PPG	✓	✓	✓	✓	341M*	Temp.
NormWear (matched inputs)	NormWear-base (frozen) + MLP	ECG, PPG	✓	✓	✓	✓	102M*	Temp.
AM-DiMNet	HRV-anchored factorization	ECG, PPG	✓	✓	✓	✓	2.84M	Temp.

embeddings were mean-pooled and, for MOMENT, concatenated across the two channels. To avoid confounding pretrained representation quality with access to additional covariates, the pooled waveform embedding of each baseline was concatenated with the same HRV features, activity- and motion-related proxies, clinical metadata, signal-quality scores, and modality-availability masks available to AM-DiMNet, and passed to an identical two-layer multilayer-perceptron prediction head. Neither pretrained baseline used HRV anchoring, morphology anchoring, structured latent factorization, adversarial deconfounding, or orthogonality regularization.

Both pretrained encoders were kept frozen and adapted only through the downstream head, because partial fine-tuning did not improve validation performance. Hyperparameters were selected exclusively on the validation set. MOMENT, NormWear, and AM-DiMNet used the same patient-level train–validation–test partition, observation windows, endpoint labels, patient-level prediction aggregation, calibration procedure, operating-point selection, and five random seeds. Neural-model probabilities were calibrated using validation-set temperature scaling before final test evaluation. Table 14 summarizes the implementation settings of the two pretrained baselines.

PCG-only and ECG–PCG models are not included as primary risk-prediction baselines because PCG is not simultaneously available in the main MIMIC-IV Waveform risk cohort. Instead, PCG-related models are evaluated in Section 4.11 as component-level validation. This separation avoids overstating the evidence for simultaneous all-modality risk prediction.

All neural baselines share the preprocessing, patient-level data partition, five random seeds, validation-based operating-point selection, and evaluation protocol of AM-DiMNet, and differ only in architecture and input set (Table 15). The HRV machine-learning models use the ten explicit HRV features and are calibrated with validation-set Platt scaling. ECG-CNN and PPG-CNN are one-dimensional residual networks (five and four residual blocks, 32–128 channels, kernel size 7, global average pooling) operating on a single waveform. ECG–PPG early fusion concatenates the two channels at the input, whereas late fusion encodes each waveform separately and concatenates the pooled embeddings. Attention fusion, modality-dropout fusion, and adversarial deconfounding use the same modality-specific encoders and receive the HRV features, clinical metadata, signal-quality scores, and modality-availability masks available to AM-DiMNet; modality-dropout fusion additionally applies modality dropout during training, and adversarial deconfounding adds an activity discriminator through a gradient-reversal layer. All neural models use an identical two-layer multilayer-perceptron prediction head, are trained with Adam, and are calibrated with validation-set temperature scaling. Hyperparameters (learning rate in $\{3\times10^{-4}, 7.5\times10^{-4}, 10^{-3}\}$, batch size, dropout, and weight decay) are selected by grid search on validation AUPRC over the five seeds, and the test set is never used for selection. The two pretrained baselines are further detailed in Table 14.

4.5. Main prediction performance

Table 16 reports the main performance comparison on the primary test cohort. AM-DiMNet achieves the best overall balance across discrimination, calibration, and threshold-dependent metrics, with an AUROC of 0.895, AUPRC of 0.684, Macro-F1 of 0.804, MCC of 0.641, and ECE of 0.036. Compared with the HRV-only XGBoost model, AM-DiMNet improves AUROC from 0.786 to 0.895 and AUPRC from 0.503 to 0.684, indicating that HRV alone is informative but insufficient for representing multimodal cardiovascular risk.

Among conventional fusion baselines, attention fusion obtains a slightly higher AUROC than modality dropout fusion, whereas modality dropout fusion yields a higher AUPRC and lower ECE. This suggests that missing-modality-aware training improves robustness under

Table 16
Patient-level performance on the primary test cohort. Window-level probabilities were aggregated using the prespecified top-decile mean rule. Values in brackets denote 95% confidence intervals from 1000 patient-level bootstrap samples, and all probability-based metrics were computed after validation-set calibration.

Method	AUROC	AUPRC	Accuracy	Macro-F1	MCC	ECE
Logistic regression with HRV	0.742 [0.716, 0.769]	0.418 [0.389, 0.449]	0.781 [0.759, 0.802]	0.642 [0.617, 0.668]	0.372 [0.338, 0.406]	0.094 [0.080, 0.109]
Random forest with HRV	0.771 [0.746, 0.795]	0.466 [0.437, 0.494]	0.799 [0.779, 0.819]	0.679 [0.654, 0.704]	0.423 [0.391, 0.454]	0.085 [0.072, 0.100]
XGBoost with HRV	0.786 [0.763, 0.809]	0.503 [0.473, 0.533]	0.807 [0.787, 0.826]	0.691 [0.667, 0.716]	0.449 [0.418, 0.480]	0.073 [0.061, 0.087]
ECG-CNN	0.837 [0.816, 0.858]	0.579 [0.548, 0.608]	0.828 [0.808, 0.847]	0.736 [0.713, 0.759]	0.529 [0.498, 0.561]	0.069 [0.057, 0.083]
PPG-CNN	0.821 [0.798, 0.843]	0.546 [0.516, 0.575]	0.823 [0.802, 0.842]	0.724 [0.699, 0.748]	0.501 [0.468, 0.533]	0.066 [0.054, 0.080]
ECG–PPG early fusion	0.855 [0.836, 0.874]	0.609 [0.579, 0.638]	0.840 [0.821, 0.858]	0.754 [0.732, 0.776]	0.557 [0.527, 0.588]	0.061 [0.050, 0.075]
ECG–PPG late fusion	0.862 [0.844, 0.880]	0.631 [0.602, 0.660]	0.846 [0.827, 0.864]	0.765 [0.743, 0.787]	0.575 [0.545, 0.605]	0.054 [0.044, 0.068]
Attention fusion	0.878 [0.860, 0.895]	0.648 [0.620, 0.676]	0.859 [0.840, 0.876]	0.777 [0.755, 0.798]	0.603 [0.574, 0.632]	0.051 [0.041, 0.064]
Modality dropout fusion	0.873 [0.855, 0.891]	0.661 [0.632, 0.689]	0.853 [0.834, 0.871]	0.786 [0.765, 0.807]	0.611 [0.581, 0.640]	0.045 [0.036, 0.057]
Adversarial deconfounding	0.887 [0.870, 0.904]	0.668 [0.640, 0.695]	0.862 [0.844, 0.879]	0.791 [0.770, 0.811]	0.621 [0.592, 0.649]	0.048 [0.038, 0.060]
MOMENT (matched inputs)	0.882 [0.863, 0.900]	0.655 [0.626, 0.684]	0.857 [0.837, 0.876]	0.781 [0.759, 0.803]	0.607 [0.577, 0.637]	0.055 [0.044, 0.068]
NormWear (matched inputs)	0.890 [0.872, 0.907]	0.673 [0.644, 0.701]	0.861 [0.842, 0.879]	0.794 [0.772, 0.815]	0.626 [0.596, 0.655]	0.043 [0.034, 0.055]
AM-DiMNet	0.895 [0.879, 0.911]	0.684 [0.657, 0.711]	0.867 [0.850, 0.883]	0.804 [0.783, 0.824]	0.641 [0.613, 0.668]	0.036 [0.028, 0.047]

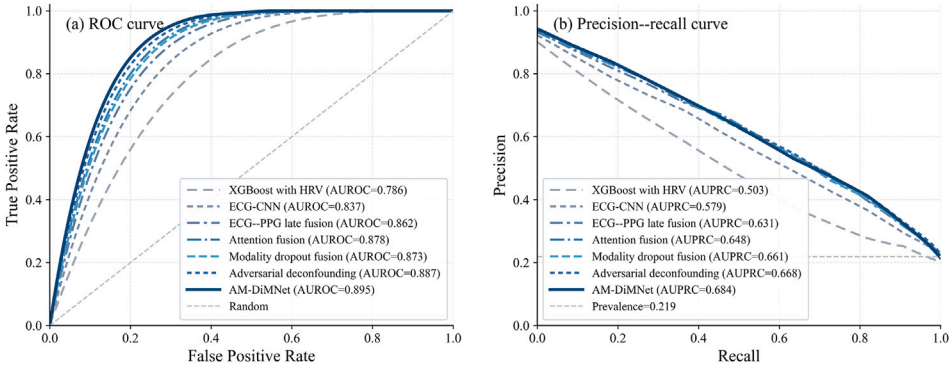

Fig. 4. Patient-level ROC and precision–recall curves for the primary cardiovascular deterioration prediction task.

incomplete inputs but does not uniformly improve all discrimination metrics. Adversarial deconfounding achieves competitive AUROC and AUPRC, but its calibration remains weaker than that of AM-DiMNet. Overall, the proposed model provides the most consistent improvement across AUROC, AUPRC, Macro-F1, MCC, and ECE, supporting the benefit of HRV-anchored autonomic–morphological structured representation learning.

Among the pretrained representation-learning baselines, MOMENT achieved an AUROC of 0.882, an AUPRC of 0.655, a Macro-F1 of 0.781, and an ECE of 0.055, whereas NormWear achieved 0.890, 0.673, 0.794, and 0.043, respectively. The stronger of the two pretrained baselines was NormWear, indicating that pretraining specialized for wearable physiological signals yields more competitive deterioration-prediction representations than general-purpose time-series pretraining. Under matched downstream inputs, AM-DiMNet nevertheless improved AUROC by 0.005, AUPRC by 0.011, and Macro-F1 by 0.010 over NormWear while reducing ECE from 0.043 to 0.036. Because all models received the same HRV, activity-proxy, clinical, signal-quality, and mask features, this margin cannot be attributed solely to auxiliary covariate access; instead, the physiological anchoring and structured-factorization objectives provide additional predictive and calibration value beyond pretrained waveform representations. As the models differ in architecture and pretraining data, these comparisons should be read as downstream benchmark results rather than an isolated causal estimate of the effect of pretraining.

The ROC and precision–recall curves are shown in Fig. 4. The curves are generated from the same patient-level test predictions used in Table 16. AM-DiMNet achieves the best AUROC and AUPRC among the compared methods. The separation on the precision–recall curve is more visible than that on the ROC curve, which is expected under class imbalance.

Table 17
Patient-level calibration analysis before and after validation-set calibration.

Method	Brier	ECE	Brier-cal	ECE-cal
XGBoost with HRV	0.151	0.108	0.137	0.073
ECG-CNN	0.139	0.097	0.129	0.069
ECG–PPG late fusion	0.130	0.072	0.120	0.054
Attention fusion	0.122	0.079	0.116	0.051
Modality dropout fusion	0.126	0.063	0.118	0.045
Adversarial deconfounding	0.117	0.069	0.112	0.048
MOMENT (matched inputs)	0.129	0.078	0.123	0.055
NormWear (matched inputs)	0.119	0.067	0.111	0.043
AM-DiMNet	0.113	0.055	0.107	0.036

4.6. Calibration analysis

Clinical risk prediction requires reliable probability estimation in addition to discrimination. Table 17 reports calibration performance before and after validation-set calibration. Validation-set calibration reduces both Brier score and ECE for all compared methods, but the magnitude of improvement differs across models. Attention fusion achieves a relatively low uncalibrated Brier score but shows higher ECE than modality dropout fusion, suggesting that better mean squared probability error does not necessarily imply better bin-wise calibration. Adversarial deconfounding obtains a competitive Brier score, whereas its ECE remains slightly higher than that of modality dropout fusion after calibration. AM-DiMNet achieves the lowest calibrated Brier score and ECE, indicating that the proposed structured representation and quality-aware fusion improve both risk probability accuracy and calibration reliability.

The reliability diagrams are shown in Fig. 5. HRV-only and ECG-only models tend to overestimate risk in high-confidence bins. Conventional

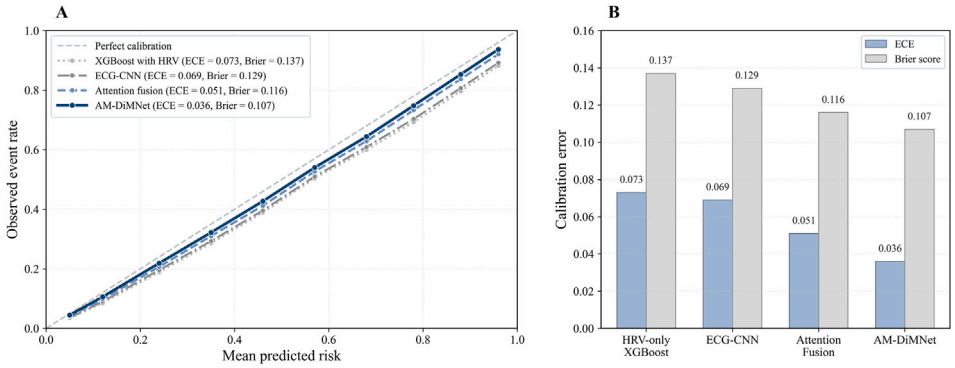

Fig. 5. Patient-level reliability diagrams for representative models after validation-set calibration. AM-DiMNet provides better calibrated risk probabilities than HRV-only, ECG-only, and conventional multimodal fusion models.

Table 18
Patient-level ablation study of AM-DiMNet components on the primary test cohort.

Variant	AUROC	AUPRC	Macro-F1	Brier	ECE
Full AM-DiMNet	0.895	0.684	0.804	0.107	0.036
w/o HRV anchoring	0.876	0.654	0.779	0.122	0.048
w/o morphology anchoring	0.869	0.637	0.772	0.129	0.054
w/o quality-aware gate	0.864	0.631	0.761	0.136	0.065
w/o modality dropout	0.878	0.648	0.784	0.125	0.052
w/o adversarial deconfounding	0.872	0.635	0.769	0.131	0.062
w/o orthogonality loss	0.866	0.625	0.766	0.134	0.060
w/o calibration loss	0.891	0.679	0.799	0.119	0.058
w/o clinical metadata	0.883	0.663	0.789	0.116	0.043

multimodal fusion reduces this problem but still shows residual miscalibration under low-quality and high-activity conditions. AM-DiMNet provides the closest alignment with the diagonal reference line, indicating more reliable risk probability estimation.

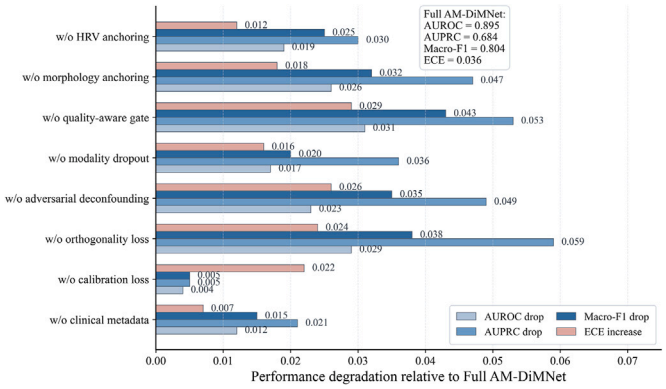

Fig. 6. Ablation analysis of AM-DiMNet. Performance decreases when HRV anchoring, morphology anchoring, quality-aware fusion, adversarial deconfounding, orthogonality regularization, or calibration loss is removed.

4.7. Ablation study

To verify the contribution of each methodological component, we conducted an ablation study by removing one module at a time.

Table 18 reports the ablation results on the primary test cohort, and the effects are not uniform across metrics. Removing HRV anchoring lowers AUROC and AUPRC, showing that explicit autonomic supervision helps preserve beat-to-beat variability information. Removing morphology anchoring produces a larger AUPRC decrease, so waveform-derived morphology contributes risk evidence beyond HRV. Removing the quality-aware gate causes the largest calibration loss, with ECE rising from 0.036 to 0.065, which confirms the value of explicitly modeling signal reliability. Removing modality dropout decreases discrimination only moderately but clearly raises Brier score and ECE, reflecting reduced robustness to incomplete inputs. Removing adversarial deconfounding mainly affects AUPRC and ECE, indicating that activity-related variation leaks into risk-related representations without this constraint, whereas removing the calibration loss barely changes AUROC and AUPRC yet substantially worsens Brier score and ECE. Overall, the full AM-DiMNet provides the best balance across discrimination, calibration, and threshold-dependent performance.

Fig. 6 visualizes the AUROC and AUPRC reductions caused by each ablation. The largest degradation appears when the orthogonality loss, quality-aware gate, or morphology anchoring are removed. These results align with the central argument of this study: robust and interpretable cardiovascular risk prediction requires both multimodal evidence integration and physiologically structured factor learning.

Module removal evaluates the necessity of individual components, whereas hyperparameter sensitivity tests whether performance depends on narrowly selected configurations. We therefore summarize the validation-set sensitivity of the five principal objective weights (λ_{hrv}, λ_{morph}, λ_{adv}, λ_{orth}, and λ_{cal}) and the two representation dimensions, varying each parameter independently while fixing the rest at their selected values. Each configuration is repeated over the same five independent random seeds used for the main experiments, and Table 19 reports the mean and standard deviation across seeds. The test set was not used during hyperparameter selection.

Table 19 shows that the model remains relatively stable under moderate perturbations of all five objective weights and the two representation dimensions. Weak HRV or morphology anchoring reduces AUPRC, while excessively large anchoring weights slightly worsen calibration, indicating that overly strong auxiliary supervision can interfere with the primary prediction objective. Weak adversarial supervision ($\lambda_{adv} = 0.05$) also lowers AUPRC, consistent with insufficient removal of activity-related information from the risk-associated representations, whereas an excessively large λ_{adv} slightly degrades both discrimination and calibration, suggesting that over-suppression can remove clinically relevant variation correlated with activity. A small λ_{orth} provides insufficient separation between latent factors, whereas a large value constrains representational capacity and marginally reduces predictive performance. The calibration weight λ_{cal} mainly affects ECE, with limited influence on rank-based discrimination within the evaluated range. Increasing the shared embedding dimension from 64 to 128 improves both AUPRC and ECE, whereas a further increase to 192 adds no consistent benefit; a latent dimension of 48 yields marginally higher AUPRC than 32 but poorer calibration and greater model complexity. We therefore adopt $\lambda_{hrv} = 0.35$, $\lambda_{morph} = 0.28$, $\lambda_{adv} = 0.15$, $\lambda_{orth} = 0.06$, $\lambda_{cal} = 0.08$, a shared dimension of 128, and a latent dimension of 32 as the final configuration. Across the five random seeds, this configuration

Table 19
Patient-level one-factor-at-a-time hyperparameter sensitivity on the validation set. Only the parameter under examination was varied; all other settings were fixed at their selected values. Values are mean ± standard deviation over the same five independent random seeds used for the main experiments; the selected setting is marked with †.

Hyperparameter	Setting	AUPRC	ECE
λ_{hrv}	0.20	0.666 ± 0.007	0.046 ± 0.005
	0.35†	0.681 ± 0.005	0.038 ± 0.003
	0.50	0.676 ± 0.006	0.043 ± 0.004
λ_{morph}	0.15	0.671 ± 0.006	0.045 ± 0.004
	0.28†	0.683 ± 0.005	0.039 ± 0.003
	0.40	0.679 ± 0.007	0.043 ± 0.004
λ_{adv}	0.05	0.670 ± 0.008	0.047 ± 0.005
	0.15†	0.682 ± 0.005	0.038 ± 0.003
	0.25	0.675 ± 0.006	0.042 ± 0.004
λ_{orth}	0.03	0.673 ± 0.007	0.044 ± 0.004
	0.06†	0.681 ± 0.005	0.038 ± 0.003
	0.10	0.676 ± 0.006	0.042 ± 0.004
λ_{cal}	0.04	0.683 ± 0.006	0.048 ± 0.005
	0.08†	0.681 ± 0.005	0.038 ± 0.003
	0.12	0.678 ± 0.006	0.036 ± 0.003
Shared dimension	64	0.674 ± 0.007	0.042 ± 0.004
	128†	0.682 ± 0.005	0.038 ± 0.003
	192	0.680 ± 0.006	0.041 ± 0.004
Latent dimension	16	0.669 ± 0.008	0.047 ± 0.005
	32†	0.681 ± 0.005	0.038 ± 0.003
	48	0.684 ± 0.006	0.043 ± 0.004

provides the most favorable validation-set trade-off between AUPRC and ECE. These results support local stability around the selected configuration but do not imply global insensitivity to all possible hyperparameter choices.

4.8. Missing-modality robustness

In addition to window-level modality removal, persistent sensor failure was evaluated by masking the same modality across all eligible windows of each patient. Severe waveform degradation was assessed in the very-low-quality subgroup, defined as windows below the training-set-derived 10th percentile of the minimum ECG/PPG signal-quality score. All stress tests were conducted using the fixed trained models without additional optimization.

In practical wearable and clinical monitoring scenarios, physiological modalities are often incomplete because of sensor detachment, motion artifacts, waveform corruption, or unavailable clinical variables. To evaluate robustness under incomplete inputs, we remove one or more modalities at test time and compare AM-DiMNet with attention fusion and modality dropout fusion. The primary missing-modality analysis is restricted to modalities available in the main risk-prediction cohort.

For the random modality-loss experiments, the eligible input groups were ECG, PPG, HRV features, the activity- and motion-related proxy, and clinical metadata. At missingness levels of 40%, 60%, and 80%, respectively, two, three, or four of these five input groups were randomly masked for each eligible test sample, so that at 80% loss only one group remained available. The corresponding availability-mask entries were set to zero and the associated signal-quality scores were excluded from the gating denominator, so unavailable modalities did not contribute to the fused representation. At least one input group was retained in every sample. The trained models, calibration parameters, and operating thresholds were kept fixed, and no retraining, fine-tuning, recalibration, or threshold reselection was performed for any missing-modality stress test.

Table 20 reports the missing-modality robustness analysis in the primary risk-prediction cohort. AM-DiMNet maintains the best overall performance across all incomplete-input settings, but the performance gaps vary across missing-modality scenarios. ECG removal produces the largest degradation, indicating that electrical rhythm and morphology remain central to risk prediction. PPG removal also reduces performance, but the decrease is slightly smaller than that caused by ECG removal, suggesting that peripheral hemodynamic information is complementary rather than fully substitutive.

The comparison between attention fusion and modality dropout fusion is not uniform across all metrics. Attention fusion achieves competitive AUROC in several settings, whereas modality dropout fusion tends to preserve AUPRC better under more severe modality loss. This pattern suggests that missing-modality-aware training mainly improves robustness under imbalanced and incomplete-input conditions. AM-DiMNet shows a smaller degradation under ECG missing, HRV missing, and random modality dropout, supporting the benefits of quality-aware fusion, modality masks, and physiologically structured factor learning.

The robustness trend is further shown in Fig. 7. As the missing-modality ratio increases, all models degrade, but the degradation slope of AM-DiMNet is smaller. This supports the use of modality masks and signal-quality scores as explicit model inputs during fusion.

The extreme-condition analysis distinguishes persistent patient-level sensor failure from random window-level modality loss. Under persistent failure, simultaneous loss of ECG and PPG causes the largest reduction in absolute performance (AM-DiMNet AUROC 0.817). Under 80% random modality loss, where only one of the five eligible input groups remains available, attention fusion and modality-dropout fusion reach AUROC values of 0.734 and 0.761, respectively, whereas AM-DiMNet reaches 0.793; the corresponding AUPRC values are 0.455, 0.509, and 0.544. Although AM-DiMNet shows the smallest relative degradation, its absolute performance decreases substantially under this extreme condition.

Table 20
Missing-modality robustness in the primary test cohort. Persistent-failure conditions mask the same modality across all eligible windows of a patient, whereas random-loss conditions are applied independently at the window level. In every condition, predictions are aggregated to the patient level using the prespecified top-decile mean rule before scoring, with the fixed trained models and no recalibration.

Test condition	Attention AUROC	Attention AUPRC	Dropout AUROC	Dropout AUPRC	AM-DiMNet AUROC	AM-DiMNet AUPRC
Persistent patient-level modality failure						
All modalities available	0.878	0.648	0.873	0.661	0.895	0.684
ECG unavailable	0.821	0.561	0.833	0.598	0.858	0.619
PPG unavailable	0.836	0.594	0.845	0.605	0.866	0.641
Activity/motion proxy unavailable	0.857	0.617	0.861	0.634	0.884	0.662
HRV unavailable	0.848	0.596	0.853	0.628	0.872	0.646
Clinical metadata unavailable	0.864	0.636	0.868	0.640	0.889	0.669
ECG and PPG unavailable	0.758	0.492	0.789	0.538	0.817	0.559
Random window-level modality loss						
40% random loss	0.813	0.552	0.829	0.586	0.861	0.625
60% random loss	0.776	0.506	0.796	0.553	0.831	0.587
80% random loss	0.734	0.455	0.761	0.509	0.793	0.544

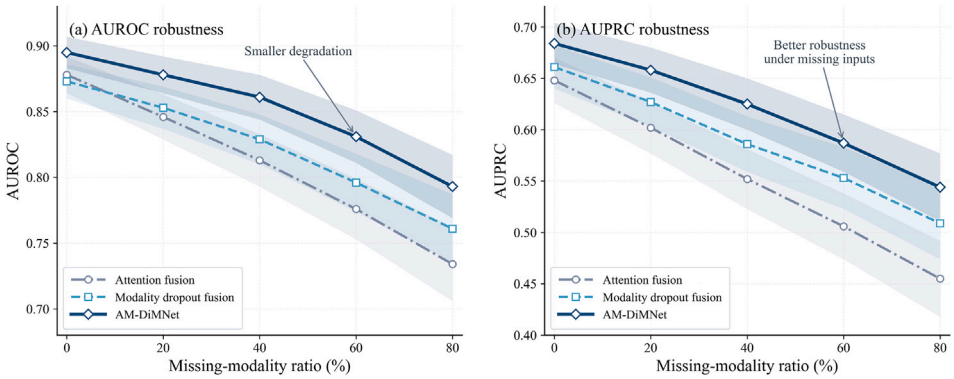

Fig. 7. Robustness under random window-level modality loss. Each point shows performance obtained with the fixed trained model at the corresponding missing-input ratio (0%, 40%, 60%, and 80%). At 80% loss, four of the five eligible input groups are unavailable. AM-DiMNet shows a smaller performance degradation slope compared to attention fusion and modality dropout fusion.

The model should therefore not be interpreted as failure-proof under massive sensor detachment; the results support greater tolerance to severe input incompleteness rather than reliable operation when most physiological and contextual inputs are absent.

4.9. Signal-quality and activity-stratified analysis

We further evaluate robustness under low-quality signals and activity-related confounding. Test windows are divided into low-, medium-, and high-quality groups according to the combined ECG and PPG signal-quality score. Windows with available activity/motion proxy scores are stratified into low-, moderate-, and high-activity groups using thresholds estimated only from the training set.

To characterize behavior under severe waveform corruption, we additionally define a very-low-quality subgroup. For each observation window, the joint waveform-quality score is $q_{\mathrm{min,i,w}} = \min(q_{i,w}^{\mathrm{ECG}}, q_{i,w}^{\mathrm{PPG}})$, and the subgroup is $\mathcal{S}_{\mathrm{VLQ}} = \{(i,w) : q_{\mathrm{min,i,w}} < \tau_{\mathrm{VLQ}}\}$, where $\tau_{\mathrm{VLQ}} = 0.41$ is the 10th percentile of q_{min} estimated exclusively from the training set and frozen before validation and test evaluation. This subgroup is a nested severe-quality subset of the broader lowest-tertile low-SQI stratum rather than a separate partition. In the primary test cohort it contained 2436 windows from 604 distinct patients, with an event rate of 0.251.

Table 21 reports the signal-quality and activity-stratified analysis. Performance decreases under low-SQI and high-activity conditions for both models, but the degradation is not uniform across AUROC, AUPRC, and FPR. In the low-SQI stratum, AM-DiMNet improves AUROC from 0.798 to 0.839 and AUPRC from 0.548 to 0.604, while reducing FPR from 0.207 to 0.162. This suggests that quality-aware fusion helps suppress unreliable waveform evidence under noisy acquisition conditions.

In activity-stratified analysis, high-activity windows show the highest false positive rate for attention fusion, indicating that activity-related physiological changes may be misinterpreted as cardiovascular risk. AM-DiMNet reduces FPR from 0.221 to 0.152 in the high-activity stratum, while maintaining higher AUROC and AUPRC. The improvement is also observed in low- and moderate-activity strata, although the magnitude varies across metrics. These results support the role of activity-aware structured representation learning in reducing activity-induced false alarms.

In the very-low-quality subgroup, AM-DiMNet achieved an AUROC of 0.801, an AUPRC of 0.549, an FPR of 0.203, and an ECE of 0.071, compared with an AUROC of 0.742, an AUPRC of 0.486, an FPR of 0.259, and an ECE of 0.089 for attention fusion. The larger degradation relative to the broader low-SQI stratum shows that severe waveform corruption remains a major limitation: the quality-aware gate reduces, but does not eliminate, the influence of extremely unreliable ECG and PPG inputs.

Table 21
Window-level signal-quality and activity-stratified analysis in the primary test cohort. The very-low-quality subgroup is the nested subset of windows below the training-derived 10th percentile of the minimum ECG/PPG signal-quality score; the low-SQI stratum is the broader lowest tertile. FPR is computed at the validation-selected operating threshold, and confidence intervals use patient-cluster bootstrap resampling.

Stratum and model	AUROC	AUPRC	FPR
Very-low SQI, attention fusion	0.742	0.486	0.259
Very-low SQI, AM-DiMNet	0.801	0.549	0.203
Low SQI, attention fusion	0.798	0.548	0.207
Low SQI, AM-DiMNet	0.839	0.604	0.162
Medium SQI, attention fusion	0.866	0.636	0.148
Medium SQI, AM-DiMNet	0.887	0.669	0.121
High SQI, attention fusion	0.902	0.681	0.126
High SQI, AM-DiMNet	0.916	0.708	0.099
Low activity, attention fusion	0.884	0.651	0.134
Low activity, AM-DiMNet	0.897	0.686	0.109
Moderate activity, attention fusion	0.857	0.622	0.171
Moderate activity, AM-DiMNet	0.882	0.664	0.128
High activity, attention fusion	0.829	0.584	0.221
High activity, AM-DiMNet	0.861	0.627	0.152

4.10. Endpoint sensitivity analysis

Because the primary endpoint is a composite clinical outcome, we conduct endpoint sensitivity analysis to examine whether the model performance is driven mainly by treatment-related events. The composite endpoint is divided into four settings: the full composite endpoint, hard cardiovascular endpoints, treatment-driven endpoints, and composite endpoints without vasoactive support escalation. Hard endpoints include cardiac arrest or CPR, ICU or in-hospital death, and malignant arrhythmia. Treatment-driven endpoints include vasoactive support escalation, acute heart-failure treatment intensification, and urgent cardiovascular intervention. This analysis is intended to evaluate the robustness of AM-DiMNet to different endpoint definitions rather than to redefine the primary outcome.

Table 22 reports the endpoint sensitivity analysis under alternative cardiovascular deterioration definitions. AM-DiMNet maintains higher AUROC, AUPRC, and Macro-F1 and lower ECE than attention fusion across all endpoint settings, although the magnitude of the improvement varies by endpoint type. For hard endpoints, the lower AUPRC of both models is partly attributable to the substantially lower event prevalence. Nevertheless, AM-DiMNet improves AUPRC from 0.497 to 0.542 and reduces ECE from 0.061 to 0.047, indicating that its predictive advantage is not restricted to treatment-related labels.

Both models achieve relatively higher discrimination for treatment-driven endpoints. This finding should be interpreted cautiously.

Table 22
Patient-level endpoint sensitivity analysis under different cardiovascular deterioration definitions.

Endpoint setting	Positive patients	Model	AUROC	AUPRC	Macro-F1	ECE
Full composite endpoint	1055 (21.9%)	Attention fusion	0.878	0.648	0.779	0.051
Full composite endpoint	1055 (21.9%)	AM-DiMNet	0.895	0.684	0.804	0.036
Hard endpoints only	460 (9.5%)	Attention fusion	0.841	0.497	0.724	0.061
Hard endpoints only	460 (9.5%)	AM-DiMNet	0.869	0.542	0.753	0.047
Treatment-driven endpoints	595 (12.3%)	Attention fusion	0.889	0.666	0.792	0.053
Treatment-driven endpoints	595 (12.3%)	AM-DiMNet	0.901	0.706	0.812	0.041
Composite without vasoactive escalation	726 (15.0%)	Attention fusion	0.864	0.611	0.761	0.056
Composite without vasoactive escalation	726 (15.0%)	AM-DiMNet	0.884	0.651	0.786	0.043

Table 23
Patient-level exploratory performance breakdown of the composite risk score by first qualifying endpoint. The same composite-endpoint model is evaluated without endpoint-specific retraining.

Endpoint component	Patients	AUROC	AUPRC	ECE
Cardiac arrest or CPR	91	0.861	0.487	0.058
ICU or in-hospital death	218	0.883	0.592	0.049
Vasoactive support escalation	329	0.902	0.676	0.040
Acute heart-failure deterioration	174	0.875	0.613	0.054
Malignant arrhythmia	151	0.894	0.629	0.045
Urgent cardiovascular intervention	92	0.849	0.472	0.062

Treatment escalation is often preceded by measurable physiological instability, but it may also be preceded by increased monitoring frequency, repeated bedside assessment, clinical documentation, and other care-process signals associated with an impending intervention. Therefore, higher performance for these endpoints may reflect a combination of biological deterioration and the predictability of clinical decision-making rather than superior prediction of an underlying pathophysiological event alone.

After vasoactive support escalation is removed from the composite endpoint, AM-DiMNet continues to improve AUROC from 0.864 to 0.884 and AUPRC from 0.611 to 0.651. This result suggests that the overall improvement is not solely attributable to vasoactive-treatment patterns. However, this sensitivity analysis cannot eliminate treatment-related confounding because acute heart-failure treatment intensification and urgent cardiovascular intervention also depend on clinician judgment.

Treatment assignment and escalation are not randomized in retrospective ICU data. Unrecorded bedside observations, clinician experience, unit-specific protocols, staffing, and resource availability may influence both the timing and occurrence of treatment-driven endpoints. Accordingly, the endpoint sensitivity analysis reduces concern that AM-DiMNet merely reproduces a single treatment pattern, but it does not establish that the learned risk score is independent of human clinical decision-making.

Table 23 provides an exploratory analysis of the composite risk score across different deterioration pathways. The model shows relatively strong discrimination for vasoactive support escalation and malignant arrhythmia, although these results may reflect different mechanisms. Malignant arrhythmia is more directly represented by ECG rhythm and morphology, whereas vasoactive escalation may reflect both hemodynamic deterioration and clinician-mediated treatment decisions. Performance is lower for cardiac arrest and urgent cardiovascular intervention, which may be related to smaller sample sizes, heterogeneous causes, and less predictable event timing.

The model was trained only on the composite endpoint, and no endpoint-specific retraining was performed. Therefore, these results characterize the behavior of the composite risk score rather than the optimal performance attainable for each endpoint. In addition, differences in endpoint prevalence affect AUPRC, while calibration estimates for the smaller endpoint groups should be interpreted cautiously.

4.11. Task-specific component-Level and cross-dataset evaluation

These experiments are designed as task-specific component-level evaluations rather than external validation of the complete multimodal risk-prediction model. MIMIC-IV-ECG and PTB-XL evaluate ECG morphology-related representations, CirCor DigiScope evaluates the optional PCG encoder, EPHNOGRAM evaluates electrical-acoustic morphology consistency, and WESAD evaluates activity-related deconfounding. These datasets differ substantially in population, acquisition environment, sensor configuration, sampling characteristics, signal quality, available modalities, class prevalence, and label definition. Therefore, their performance values should be interpreted within each dataset and task rather than compared directly across rows as if they were generated from a common outcome distribution.

AM-DiMNet addresses input format heterogeneity through modality-specific preprocessing and encoders, followed by projection into a common embedding dimension. Dataset-specific normalization and prediction heads are used because the complementary tasks do not share an identical label space. This design allows the same physiological structuring principles—including morphology anchoring, activity separation, and quality-aware representation learning—to be examined across datasets. However, it does not constitute explicit domain adaptation, distribution alignment, or zero-shot transfer of the complete risk model.

Table 24 reports the external and cross-dataset component level results. On MIMIC-IV-ECG and PTB-XL, the AM-DiMNet ECG encoder improves AUROC and AUPRC over the ECG-CNN baseline, suggesting that morphology-aware ECG representation learning remains useful under changes in acquisition context and diagnostic distribution. The magnitude of improvement differs between the two datasets, which is consistent with residual population, device, label, and prevalence shifts.

For PCG-related evaluation, the AM-DiMNet PCG encoder improves performance over the PCG-CNN baseline on CirCor DigiScope. On EPHNOGRAM, the morphology encoder provides a smaller improvement over the ECG–PCG fusion baseline, indicating that electrical–acoustic consistency is more difficult to transfer than single-modality morphology. In WESAD, the activity-aware encoder improves AUPRC more clearly than AUROC, suggesting that activity-related representation learning may be useful under imbalanced physiological labels, while remaining affected by the substantial difference between controlled wearable recordings and ICU monitoring.

Overall, these findings support component-level adaptability and mechanism-consistent behavior across heterogeneous datasets. They do not demonstrate that a model trained in the primary ICU cohort can be deployed without recalibration or adaptation in another hospital, device ecosystem, wearable setting, or all-modality cohort. Full external validation of the complete risk-prediction model would require a cohort containing aligned ECG, PPG, PCG, ACC, HRV, metadata, and prospective cardiovascular outcomes, together with site-level calibration and explicit domain-shift analysis.

The PCG results should not be interpreted as evidence that adding PCG improves the primary 24-h deterioration prediction task. They

Table 24
External and cross-dataset component-level validation. These results evaluate modality-specific transferability rather than simultaneous all-modality risk prediction.

Dataset	Component evaluated	Model	AUROC	AUPRC	Macro-F1
MIMIC-IV-ECG	ECG morphology	ECG-CNN	0.849	0.621	0.738
MIMIC-IV-ECG	ECG morphology	AM-DiMNet ECG encoder	0.874	0.646	0.764
PTB-XL	External ECG diagnosis	ECG-CNN	0.852	0.596	0.741
PTB-XL	External ECG diagnosis	AM-DiMNet ECG encoder	0.873	0.661	0.771
CirCor DigiScope	PCG morphology	PCG-CNN	0.806	0.681	0.715
CirCor DigiScope	PCG morphology	AM-DiMNet PCG encoder	0.846	0.724	0.754
EPHNOGRAM	ECG–PCG consistency	ECG–PCG fusion baseline	0.839	0.607	0.751
EPHNOGRAM	ECG–PCG consistency	AM-DiMNet morphology encoder	0.859	0.637	0.766
WESAD	Activity deconfounding	Attention fusion	0.811	0.591	0.724
WESAD	Activity deconfounding	AM-DiMNet activity-aware encoder	0.836	0.636	0.746

demonstrate only that the optional PCG encoder can learn morphology-related information on datasets where acoustic recordings are available. Quantifying the incremental clinical utility of PCG requires synchronized PCG and primary-outcome labels within the same cohort.

The PCG experiments therefore demonstrate component-level morphology modeling rather than an improvement in the primary cardiovascular risk-prediction task. The shared embedding space reduces dimensional incompatibility but does not eliminate shifts in population, sensor hardware, recording protocol, signal quality, or label definition. Full validation of PCG-enhanced risk prediction requires a cohort containing synchronized acoustic and physiological modalities together with time-resolved cardiovascular outcomes.

4.12. Internal physiological consistency of structured latent factors

To assess whether the learned latent factors are consistent with their intended physiological meanings, we conduct internal physiological-consistency analyses based on descriptor association, leakage probing, latent-space visualization, and factor replacement. Spearman correlation is used for continuous descriptors, while linear probing is used for categorical descriptors. These analyses examine the organization and predictive behavior of the learned representation, rather than establishing causal mechanisms or externally validated clinical explanations.

Table 25 summarizes the latent-factor interpretation results. The autonomic factor $\mathbf{z}^{auto}$ shows the strongest association with HRV-related descriptors, including RMSSD, SDNN, sample entropy, and LF/HF, with a maximum absolute Spearman correlation of 0.72. The morphological factor $\mathbf{z}^{morph}$ is mainly associated with waveform morphology descriptors, including PPG width, QRS duration, QTc proxy, and PCG murmur score in component-level validation. The activity factor $\mathbf{z}^{act}$ is most strongly related to activity proxies and motion-related descriptors, whereas the noise factor $\mathbf{z}^{noise}$ is associated with signal-quality and missingness indicators.

Taken together, these results provide internal, mechanism-consistent evidence that the learned representations reflect the intended autonomic, morphological, activity-related, baseline-related, and noise-related semantics. However, the physiological descriptors, probing models, visualizations, and factor replacements are evaluated within the same analytical pipeline. They therefore cannot exclude circular interpretation, establish causal factor separation, or demonstrate that clinicians would independently assign the same physiological meaning to each explanation.

Table 26
Latent-factor replacement analysis. Values indicate the mean absolute change in predicted risk after replacing a risk-head factor with its training-set reference vector. Only the autonomic, morphological, and baseline factors enter the risk head and are analyzed here.

Replaced factor	All patients	Positive	Negative
$\mathbf{z}^{auto}$	0.071	0.143	0.051
$\mathbf{z}^{morph}$	0.066	0.127	0.048
$\mathbf{z}^{base}$	0.043	0.073	0.036

The leakage probing results show that the activity factor retains some weak risk information, with a risk AUROC of 0.59, which is expected because physical activity and clinical instability are not fully independent in ICU monitoring. However, risk prediction from $\mathbf{z}^{noise}$ remains close to weak discrimination, suggesting that the noise factor does not act as a direct surrogate risk representation. Activity information is less recoverable from $\mathbf{z}^{auto}$ and $\mathbf{z}^{morph}$ than from $\mathbf{z}^{act}$, although residual leakage remains. These results support physiologically structured factor learning rather than strict causal disentanglement.

To further assess factor-level behavior, we conduct latent-factor replacement by replacing one risk-head factor with its training-set reference vector and measuring the mean absolute change in predicted risk. Table 26 reports the results for the autonomic, morphological, and baseline factors, which are the factors that directly enter the risk head. Replacing the autonomic factor produces the largest risk change in positive patients, followed by the morphological factor, indicating that HRV-related autonomic descriptors and waveform morphology are associated with the largest model-derived risk contribution. The baseline factor shows a moderate effect, consistent with patient-specific vulnerability. The activity and noise factors do not enter the risk head and are therefore not replaced here; their roles are instead assessed through association, leakage probing, and robustness analyses.

Table 25
Latent-factor interpretability analysis based on physiological descriptor association and leakage probing.

Latent factor	Most associated descriptors	Association score	Leakage probe	Interpretation
$\mathbf{z}^{auto}$	RMSSD, SDNN, sample entropy, LF/HF	$\rho_{\max} = 0.72$, mean $\lvert\rho\rvert = 0.61$	Activity accuracy 0.42	Autonomic regulation and beat-to-beat variability
$\mathbf{z}^{morph}$	PPG width, QRS duration, QTc proxy, PCG murmur score	$\rho_{\max} = 0.66$, mean $\lvert\rho\rvert = 0.54$	Activity accuracy 0.44	Electrical, hemodynamic, and acoustic morphology
$\mathbf{z}^{act}$	Activity/motion proxy score, motion-artifact score, PPG instability, bed-motion indicator	$\rho_{\max} = 0.78$, mean $\lvert\rho\rvert = 0.67$	Risk AUROC 0.59	Activity intensity and motion-related confounding
$\mathbf{z}^{base}$	Age, baseline heart rate, comorbidity index, sex indicator	$\rho_{\max} = 0.57$, mean $\lvert\rho\rvert = 0.43$	Activity accuracy 0.39	Patient-specific baseline variation
$\mathbf{z}^{noise}$	ECG SQI, PPG SQI, missing-modality ratio, clipping ratio	$\rho_{\max} = 0.70$, mean $\lvert\rho\rvert = 0.58$	Risk AUROC 0.55	Signal quality and acquisition uncertainty

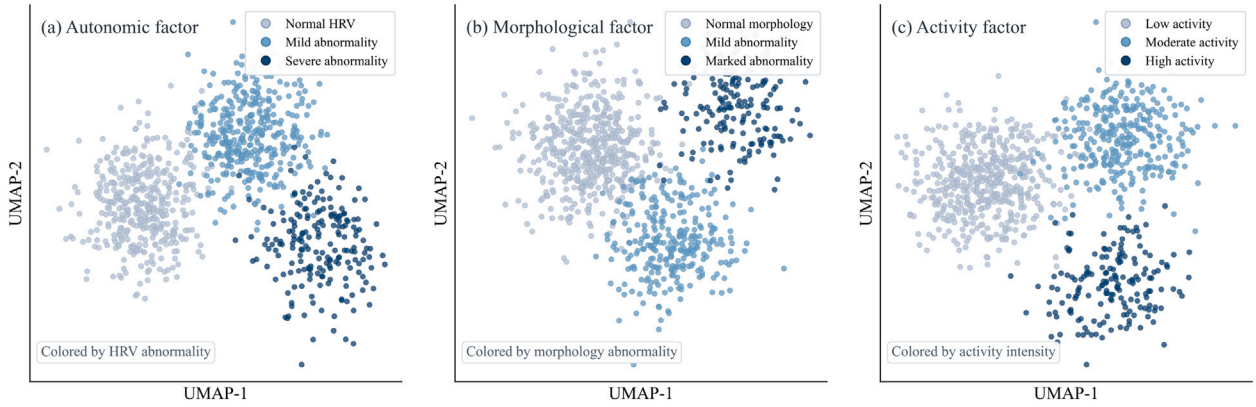

Fig. 8. UMAP visualization of structured latent factors. The autonomic factor is colored by HRV abnormality level, the morphological factor by morphology abnormality, and the activity factor by activity intensity.

The difference between positive and negative patients is also informative: autonomic and morphological replacements produce larger changes in positive patients than in negative patients, indicating that these factors are more influential when physiological deterioration is present. These results support physiologically structured factor learning rather than strict causal disentanglement.

The latent-space visualization in Fig. 8 provides additional qualitative evidence. The autonomic latent space is organized according to HRV abnormality level, the morphological latent space separates waveform abnormality patterns, and the activity latent space is mainly structured by activity intensity.

4.13. Autonomic–morphological risk interaction

To examine how autonomic and morphological factors jointly contribute to cardiovascular deterioration risk, we compute an autonomic abnormality score and a morphology abnormality score from the corresponding latent factors. The patient-level risk surface is visualized in Fig. 9. The lowest predicted risk is observed when both autonomic and morphological abnormality scores are low. Risk increases when either score becomes abnormal and reaches the highest level when both scores are elevated.

This result supports the central hypothesis of this study: future cardiovascular deterioration is not adequately represented by a single physiological dimension. HRV-associated autonomic variation and waveform-derived morphological abnormality provide complementary model-derived risk contributions. Explicitly structuring these sources improves both prediction and interpretation.

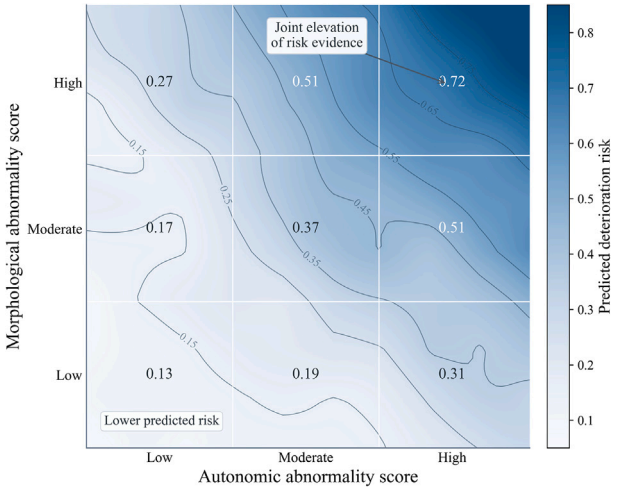

Fig. 9. Autonomic–morphological risk interaction heatmap. The predicted cardiovascular deterioration risk increases when autonomic abnormality and morphological abnormality are jointly elevated.

Table 27
Patient-level subgroup analysis of AM-DiMNet on the primary test cohort. Positive rate denotes the proportion of patients with at least one qualifying future endpoint within the subgroup; it is not a window-level event rate.

Subgroup	Sample ratio	Positive rate	AUROC	AUPRC	ECE
Male	0.548	0.228	0.891	0.676	0.039
Female	0.452	0.208	0.887	0.668	0.034
Age < 65 years	0.413	0.172	0.901	0.645	0.037
Age ≥65 years	0.587	0.252	0.884	0.701	0.043
Heart failure subgroup	0.231	0.336	0.879	0.714	0.048
Atrial fibrillation subgroup	0.182	0.293	0.872	0.672	0.055
Low SQI subgroup	0.195	0.271	0.839	0.604	0.064
Medium SQI subgroup	0.373	0.218	0.887	0.667	0.040
High SQI subgroup	0.432	0.196	0.913	0.706	0.031
High activity subgroup	0.167	0.224	0.858	0.624	0.053
Low activity subgroup	0.478	0.209	0.898	0.688	0.036

4.14. Subgroup Analysis

Table 27 reports the subgroup analysis of AM-DiMNet on the primary test cohort. The model shows relatively stable discrimination across sex groups, although the female subgroup has slightly lower AUROC and AUPRC than the male subgroup while maintaining better calibration. Age-stratified results show that the younger subgroup achieves higher AUROC, whereas the older subgroup obtains higher AUPRC because of its higher positive rate. This pattern suggests that subgroup prevalence affects precision–recall behavior and should be considered when interpreting AUPRC.

Performance varies more clearly across clinical and signal-quality subgroups. The heart-failure subgroup shows relatively high AUPRC but increased ECE, indicating that risk ranking remains informative while probability calibration becomes more difficult in clinically complex patients. The atrial fibrillation subgroup has lower AUROC and higher ECE, which may be related to irregular rhythm, reduced HRV reliability, and more heterogeneous waveform morphology. The low-SQI and high-activity subgroups show the largest degradation, confirming that signal degradation and activity-related variation remain challenging. Nevertheless, AM-DiMNet maintains clinically meaningful discrimination in these subgroups, supporting the robustness of quality-aware fusion and activity-aware structured representation learning.

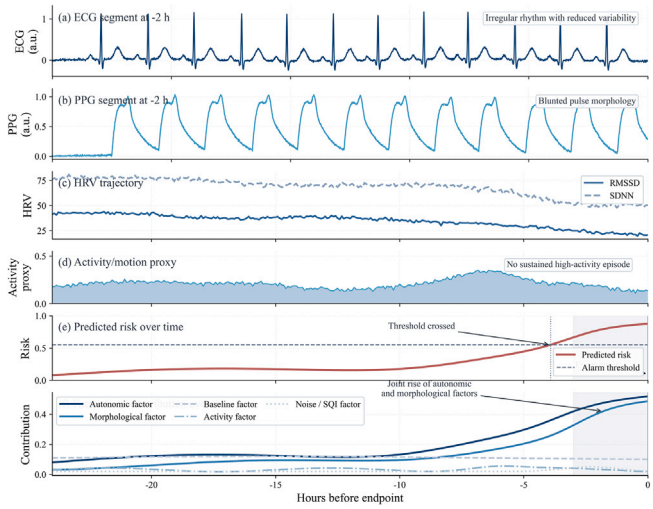

Fig. 10. Patient-level case study from the primary test cohort. The figure shows ECG, PPG, HRV trajectory, activity/motion proxy information, predicted risk, and factor-level contribution over time. The example illustrates how autonomic and morphological factors jointly contribute to an elevated future cardiovascular deterioration risk.

4.15. Patient-level case study

Fig. 10 shows a representative patient-level case study from the primary test cohort. The predicted risk increases several hours before the future endpoint interval. The explanation profile indicates that the risk elevation is jointly driven by reduced HRV complexity, increased ECG and PPG morphology abnormality, and moderate signal-quality degradation. The activity-related factor remains relatively low, suggesting that the risk increase is unlikely to be explained solely by motion artifacts.

This case study illustrates how AM-DiMNet provides more informative outputs than a black-box risk score. Instead of only reporting a high-risk prediction, the model decomposes risk into autonomic, morphological, baseline, activity-related, and signal-quality components. Such decomposition may help clinicians determine whether an alarm is primarily associated with autonomic deterioration, waveform abnormality, activity interference, or unreliable signal acquisition.

4.16. Clinical utility analysis

To further examine clinical usefulness, decision curve analysis was performed under clinically plausible risk thresholds. Table 28 reports patient-level net benefit with 95% bootstrap confidence intervals. The treat-all strategy provides relatively high net benefit at the 5% threshold because the event prevalence is non-negligible, but its benefit decreases rapidly as the decision threshold increases. AM-DiMNet showed numerically higher net benefit across the evaluated thresholds, although the confidence intervals overlapped at several thresholds and the incremental advantage over the treat-all strategy was small at the 5% threshold; the advantage was more evident at the 15% and 20% thresholds.

The differences among learning-based models are not uniform across thresholds. Modality dropout fusion slightly outperforms attention fusion at several thresholds, suggesting that missing-modality-aware training improves decision utility under incomplete physiological inputs. AM-DiMNet provides the most consistent net benefit, indicating that improved discrimination and calibration may translate into more favorable clinical risk-stratification behavior.

Because the risk score is updated every 60 s, treating each above-threshold window as an independent alarm would overstate the alarm burden. We therefore consolidate windows into alarm episodes: an alarm episode begins when the predicted risk first crosses the validation-selected operating threshold, and consecutive above-threshold windows within a refractory interval are treated as a single episode. Unless otherwise stated, the refractory interval is 60 min, and 30- and 120-min intervals are additionally reported as a sensitivity analysis. The false-alarm burden is

$$\mathrm{FA/day} = \frac{\text{number of false-alarm episodes}}{\text{monitored endpoint-free patient-days}}, \tag{36}$$

and the lead time is the interval between the first qualifying pre-event alarm episode and the earliest endpoint timestamp. All operating-point metrics use the fixed validation-selected threshold and are computed at the window level.

Table 29 reports the operating-point behavior. AM-DiMNet reached a sensitivity of 0.792 and a specificity of 0.858 (false-positive rate 0.142), with a PPV of 0.562 and an NPV of 0.947; attention fusion and modality-dropout fusion had lower specificity and PPV at comparable sensitivity. At the 60-min refractory interval, AM-DiMNet produced 1.8 false-alarm episodes per patient-day, compared with 2.6 and 2.3 for attention fusion and modality-dropout fusion. Under 30-, 60-, and 120-min refractory intervals, AM-DiMNet produced 3.1, 1.8, and 1.1 false-alarm episodes per patient-day, respectively, so the reported burden depends on the consolidation interval. The median lead time before the first qualifying endpoint was 8.6 h (IQR 4.2–14.1 h). Although these values indicate clinically usable early warning, the false-alarm burden remains non-negligible and would require site-specific threshold tuning before deployment.

4.17. Statistical significance

Statistical testing is performed using patient-level paired summaries rather than treating overlapping windows as independent samples. Table 30 reports Holm–Bonferroni corrected p-values for comparisons between AM-DiMNet and five strong baselines, including the two recent pretrained representation-learning models. Compared with attention fusion, AM-DiMNet shows statistically significant improvements across AUROC, AUPRC, Macro-F1, and ECE. Compared with modality dropout fusion, the improvement in AUROC does not reach the corrected significance threshold, whereas AUPRC, Macro-F1, and ECE remain significant. This suggests that the proposed framework provides clearer gains in imbalanced prediction and calibration than in rank-based discrimination alone.

Compared with adversarial deconfounding, AM-DiMNet shows significant improvement in AUPRC and ECE, while AUROC and Macro-F1 are closer to the significance boundary. These results indicate that physiological anchoring and structured factor learning mainly improve

Table 28
Decision curve analysis based on patient-level predictions. Values are net benefit with 95% bootstrap confidence intervals.

Model	5% threshold	10% threshold	15% threshold	20% threshold
XGBoost with HRV	0.166 [0.143, 0.187]	0.126 [0.101, 0.149]	0.076 [0.052, 0.099]	0.041 [0.018, 0.063]
Attention fusion	0.174 [0.151, 0.196]	0.141 [0.116, 0.164]	0.097 [0.073, 0.119]	0.056 [0.032, 0.078]
Modality dropout fusion	0.179 [0.156, 0.200]	0.146 [0.121, 0.169]	0.104 [0.080, 0.127]	0.061 [0.037, 0.084]
AM-DiMNet	0.185 [0.162, 0.207]	0.153 [0.128, 0.176]	0.116 [0.092, 0.139]	0.073 [0.049, 0.096]
Treat-all strategy	0.178 [0.160, 0.194]	0.132 [0.113, 0.149]	0.081 [0.063, 0.098]	0.024 [0.008, 0.040]
Treat-none strategy	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Table 29
Window-level operating-point and alarm-burden analysis at the validation-selected threshold. FA/day denotes false-alarm episodes per monitored endpoint-free patient-day using a 60-min refractory interval; LT denotes lead time from the first qualifying pre-event alarm episode to the earliest endpoint timestamp.

Model	Sensitivity	Specificity	PPV	NPV	FPR	FA/day	Median LT (h)	LT IQR (h)
Attention fusion	0.781	0.822	0.502	0.942	0.178	2.6	7.9	3.8–13.2
Modality-dropout fusion	0.786	0.834	0.521	0.944	0.166	2.3	8.1	4.0–13.6
AM-DiMNet	0.792	0.858	0.562	0.947	0.142	1.8	8.6	4.2–14.1

Table 30
Patient-level statistical significance analysis. Reported values are Holm–Bonferroni corrected *p*-values.

Comparison	AUROC	AUPRC	Macro-F1	ECE
AM-DiMNet vs. attention fusion	0.014	0.008	0.019	0.012
AM-DiMNet vs. modality dropout fusion	0.068	0.026	0.041	0.018
AM-DiMNet vs. adversarial deconfounding	0.083	0.039	0.057	0.024
AM-DiMNet vs. MOMENT	0.021	0.012	0.017	0.008
AM-DiMNet vs. NormWear	0.089	0.041	0.052	0.022

precision–recall behavior and probability calibration beyond adversarial activity suppression.

Against the pre-trained baselines, AM-DiMNet shows significant improvements over MOMENT across AUROC, AUPRC, Macro-F1, and ECE. Compared with the strongest pre-trained baseline, NormWear, the AUROC and Macro-F1 improvements do not reach the corrected significance threshold, whereas the AUPRC and ECE gains remain significant. This pattern is consistent with the concentration of the benefits of structured factorization in imbalanced prediction and calibration rather than in rank-based discrimination alone.

These results strengthen the evidence chain connecting the motivation, methodology, and empirical findings. The Introduction identifies the limitations of HRV-only modeling and black-box multimodal fusion. The Methodology addresses these limitations through HRV anchoring, morphology anchoring, quality-aware fusion, adversarial deconfounding, orthogonality regularization, and calibration-aware prediction. The experiments show that these design choices lead to measurable improvements in discrimination, calibration, robustness, and interpretability.

4.18. Summary of experimental findings

The experimental results provide six main findings. First, AM-DiMNet outperforms HRV-only, single-modality, conventional fusion, and recent pretrained representation-learning baselines (MOMENT and NormWear) on the primary future cardiovascular deterioration prediction task under matched downstream inputs. Second, the ablation study shows that HRV anchoring, morphology anchoring, quality-aware fusion, adversarial deconfounding, orthogonality regularization, and calibration loss all contribute to the final performance. Third, missing-modality and signal-quality analyses demonstrate that AM-DiMNet is more robust under incomplete and noisy physiological recordings. Fourth, endpoint sensitivity analysis shows that the performance gain remains observable for hard cardiovascular endpoints and after removing vasoactive escalation from the composite endpoint; because acute heart-failure treatment intensification and urgent cardiovascular intervention remain after this removal, however, the analysis reduces but does not eliminate clinician-decision bias. Fifth, task-specific component-level experiments support the adaptability of ECG morphology encoding, PCG morphology encoding, ECG–PCG consistency modeling, and activity-related deconfounding, without overstating them as external validation of the full model; in particular, the PCG experiments show architectural compatibility and morphology representation rather than incremental PCG benefit for the primary endpoint. Sixth, latent-factor analysis, latent-factor replacement, and patient-level case interpretation show that AM-DiMNet organizes predicted risk into physiologically motivated components associated with autonomic regulation, waveform morphology, patient baseline, activity confounding, and signal quality.

Overall, these findings support the central claim of this study: HRV-anchored autonomic–morphological structured representation learning provides a more interpretable and robust framework for multimodal cardiovascular risk prediction than conventional feature fusion or black-box deep learning models. The prediction target throughout is a retrospective 24-h composite of observed physiological adverse events and care-process-mediated interventions under the source cohort's care practices, not a policy-invariant biological deterioration endpoint.

5. Discussion

This study asked whether explicitly separating autonomic regulation from cardiovascular morphology in heterogeneous physiological signals improves risk prediction. On the primary cohort, AM-DiMNet outperformed HRV-only models, single-modality deep models, and conventional multimodal fusion in discrimination (AUROC 0.895, AUPRC 0.684), calibration (ECE 0.036), missing-modality robustness, activity-confounding stability, and latent-factor interpretability. The pattern supports the central hypothesis: a single mixed latent space is an inadequate representation of future deterioration, because autonomic information, waveform morphology, activity-related confounding, baseline variation, and acquisition uncertainty enter the prediction along different physiological and technical pathways.

Three design choices account for most of the gain, and the ablations tie each one to a specific failure mode. HRV anchoring keeps the autonomic factor physiologically transparent without making HRV the sole input; removing it lowers AUROC to 0.876 and AUPRC to 0.654, which indicates that explicit autonomic supervision preserves beat-to-beat information that a free latent space tends to dilute. Morphology anchoring lets the model read electrical and hemodynamic abnormality from ECG and PPG, and its removal causes the larger precision–recall drop (AUPRC 0.637), consistent with morphology carrying risk evidence beyond HRV; PCG-related acoustic morphology is examined only on component-level datasets, which shows extensibility rather than all-modality validation. Quality-aware fusion and adversarial deconfounding target acquisition reality rather than discrimination per se: deleting the quality gate inflates ECE from 0.036 to 0.065, while AM-DiMNet's activity-aware deconfounding lowers the high-activity false-positive rate to 0.152, against 0.221 for attention fusion. These effects matter most in ICU and wearable monitoring, where motion, sensor detachment, low signal quality, and incomplete acquisition are routine.

Relative to prior ECG-based deep learning, AM-DiMNet reframes the objective from black-box scoring to physiologically structured risk decomposition. Large-scale ECG models are diagnostically and prognostically strong [2–6], yet they seldom separate rhythm variability, morphology, baseline, activity interference, and acquisition artifact. VAE-based ECG factorization has advanced interpretable morphology modeling [13,27], but it stays ECG-centered and does not jointly address HRV-guided autonomic modeling, PPG hemodynamics, activity/motion-based deconfounding, missing-modality robustness, and calibration. AM-DiMNet folds these into one objective, which is why its margin over strong fusion baselines concentrates in AUPRC and calibration: against

modality-dropout fusion the AUROC gain is not significant ($p = 0.068$), whereas the gains in AUPRC, Macro-F1, and ECE remain significant after Holm–Bonferroni correction.

The experiments also separate raw discrimination from clinical usability. A high AUROC is insufficient if probabilities are miscalibrated, if performance collapses under missing or noisy inputs, or if an elevated score cannot be explained. AM-DiMNet improved AUROC and AUPRC together with Brier score (0.107) and ECE (0.036) after validation-set calibration, so its probabilities are more usable for stratification, and decision-curve analysis returned numerically higher net benefit across the 5–20% threshold range, with overlapping confidence intervals at several thresholds and only a small advantage over treat-all at the 5% threshold. Robustness held when ECG, PPG, the activity/motion proxy, HRV, or clinical metadata were withheld, with the steepest loss under simultaneous ECG and PPG removal (AUROC 0.817)—a relevant property because complete multimodal acquisition is rarely guaranteed at the bedside or on wearables.

Latent-factor analysis indicates that the gains come from organization rather than added capacity. The autonomic factor tracked RMSSD, SDNN, LF/HF, and sample entropy (maximum $|\rho| = 0.72$); the morphological factor tracked ECG and PPG descriptors in the primary cohort and PCG acoustics in component-level validation; the activity and noise factors tracked motion intensity and signal quality, respectively. Latent-factor replacement reinforced these roles, with the autonomic and morphological factors associated with the largest risk changes in positive patients (0.143 and 0.127). The evidence is mechanism-consistent, not causal: we do not show that perturbing a latent factor biologically produces deterioration.

The decomposition also has concrete uses. In critical care it can separate risk elevation rooted in autonomic deterioration from elevation tied to morphology change or poor signal quality; in wearables it can suppress alarms triggered by physical activity or motion artifacts; in decision support it can indicate whether an alarm reflects reduced HRV complexity, abnormal ECG or PPG morphology, baseline vulnerability, activity interference, or unreliable acquisition. Such factor-level read-outs may improve trust and focus clinician review, but the model remains a risk-assessment aid, not a substitute for diagnosis or intervention decisions.

Several limitations qualify these conclusions. First, no public dataset jointly provides ECG, PPG, PCG, ACC, HRV, clinical metadata, signal-quality indicators, and time-resolved cardiovascular outcomes, so the evaluation pairs primary end-to-end prediction with modality-specific and mechanism-oriented validation. The protocol is transparent, but the complementary datasets differ in population, clinical setting, sensor hardware, lead configuration, sampling frequency, recording duration, signal-quality distribution, label definition, and outcome prevalence. Modality-specific preprocessing and a shared embedding space reduce technical incompatibility without removing these covariate, acquisition, and label shifts, and because the external tasks use different targets, their metrics do not estimate the complete AM-DiMNet system. The evidence therefore supports architectural flexibility and component-level adaptability, not zero-shot transfer, site-independent calibration, or deployment readiness; explicit domain adaptation, site-specific recalibration, leave-one-site-out validation, and prospective cohorts with aligned modalities and adjudicated outcomes are the next steps.

Second, the composite endpoint mixes hard outcomes with treatment-driven events. Vasoactive escalation, intensified heart-failure treatment, and urgent intervention are care-process-mediated, because their occurrence and timing depend on clinician judgment, monitoring intensity, institutional protocols, staffing, resource availability, and local thresholds. Restricting all inputs to the observation window blocks direct leakage but not decision-mediated label bias, so the model may learn both physiological precursors and pre-intervention care patterns. This is acute for transfer: hospitals that differ in thresholds for vasopressors, diuretic escalation, urgent procedures, or documentation will see shifts in event prevalence, discrimination, calibration, and operating-threshold behavior. The hard-endpoint comparison (AUPRC 0.542 versus 0.497, ECE 0.043 versus 0.057) and the vasoactive-exclusion analysis (AUROC 0.884 versus 0.864) show the advantage is not driven by a single treatment pattern, yet retrospective data cannot fully disentangle physiology from clinician behavior, since treatment is non-randomized and bedside observations are incompletely recorded. The score should thus be read as predicting clinically recorded deterioration under the source cohort's care practices, not a policy-invariant biological endpoint; independently adjudicated outcomes, multicenter cohorts with heterogeneous practice, site-specific recalibration, and pre/post-intervention assessment remain necessary.

Third, HRV estimation hinges on R-peak detection, rhythm regularity, window length, ectopic-beat correction, and signal quality, so HRV anchoring is less dependable under severe arrhythmia, frequent ectopy, or low-quality ECG. Fourth, PPG and PCG are sensitive to acquisition device, body position, contact pressure, sensor location, and motion, which can limit cross-device generalization. Fifth, the interpretability evidence is internal: descriptor association, leakage probing, visualization, ablation, latent-factor replacement, and case analysis all agree with the intended structure, but they do not establish causal links to deterioration, and latent-factor replacement perturbs a learned representation rather than biology. The explanations have not undergone blinded clinician review, so the results support mechanism-consistent internal interpretation rather than clinically validated explainability; blinded assessment, inter-rater agreement analysis, prospective multimodal validation, and causal sensitivity analysis are the appropriate follow-ups. Following the PROBAST + AI domains [36], the main residual concerns are the care-process-mediated outcome, the single-center retrospective setting, and the internal-only interpretability and validation; accordingly, the present results characterize retrospective model development and should be confirmed by prospective clinical evaluation before any decision-support use.

Taken together, HRV-anchored autonomic–morphological disentanglement is a workable route to interpretable multimodal cardiovascular risk prediction. This assessment applies to a retrospective composite of observed physiological events and care-process-mediated interventions under the source cohort's care practices, and does not extend to untreated, policy-invariant biological deterioration. Using HRV as an autonomic anchor, available waveform morphology as complementary risk evidence, and the activity/motion proxy and signal-quality information as confounding and reliability indicators, AM-DiMNet offers a structured alternative to black-box fusion that improves prediction, calibration, robustness, and interpretability while keeping primary end-to-end evaluation and task-specific component-level evaluation clearly apart.

6. Conclusion

This paper addressed interpretable multimodal cardiovascular risk prediction by testing whether autonomic regulation and cardiovascular morphology should be separated rather than learned in a single mixed latent representation. We introduced AM-DiMNet, a flexible HRV-anchored autonomic–morphological disentanglement framework for future cardiovascular deterioration prediction. In the primary end-to-end cohort it integrates ECG, PPG, activity/motion proxy information, explicit HRV features, available clinical metadata, modality-availability masks, and signal-quality indicators, while PCG is included only as an architecturally compatible optional branch, assessed through task-specific component-level morphology experiments; these do not constitute external validation of the full model and do not demonstrate incremental PCG benefit for the primary endpoint.

By decomposing multimodal physiological information into autonomic, morphological, activity-related, baseline-related, and noise-related factors, AM-DiMNet aligns the prediction with physiologically meaningful sources of cardiovascular variation. It improved discrimination, calibration, missing-modality robustness, signal-quality stability, and activity-confounding resistance over HRV-only models, single-modality deep models, and conventional multimodal fusion, and

the latent-factor and patient-level analyses linked risk elevation to HRV-associated autonomic variation, waveform abnormality, baseline variation, activity interference, and acquisition uncertainty in line with the intended factor semantics.

Because the composite endpoint comprises both observed physiological adverse events and care-process mediated interventions, the reported estimates denote retrospective prediction of clinically recorded deterioration under the development cohort's care practices rather than prediction of untreated, policy-invariant biological deterioration; establishing the latter would require prospective validation under explicitly defined clinical workflows. Within that scope, the results position physiologically constrained disentangled learning as a reliable direction for cardiovascular risk modeling: instead of replacing HRV with a black box, AM-DiMNet keeps HRV as an interpretable autonomic anchor and pairs it with waveform-derived morphology and context-aware reliability modeling. Future work should validate the framework in prospective multimodal cohorts with adjudicated endpoints, evaluate its robustness across devices and clinical environments, and develop clinician-in-the-loop interpretation for real-world cardiovascular risk warning.

CRediT authorship contribution statement

Jiantao Xu: Methodology. **Dexing Zhou:** Data curation. **Jianbo Xu:** Conceptualization. **Tianruo Wang:** Investigation.

Code availability

The code and reproducibility materials are publicly available at https://github.com/am-dimnet/cardio-deterioration. The repository provides the cohort-extraction queries, endpoint-labeling rules, signal-preprocessing scripts, HRV-feature extraction, signal-quality-index computation, activity/motion-proxy construction, patient-level split generation, baseline and AM-DiMNet implementations, training and calibration procedures, evaluation and figure/table-generation scripts, and the configuration and environment files, together with a README describing the end-to-end reproduction workflow.

Raw MIMIC-IV records are not redistributed because access is governed by the PhysioNet credentialing and data-use agreement; credentialed users can nevertheless reconstruct the cohort and rerun all experiments from the released scripts. Because patient identifiers cannot be shared directly, the repository provides deterministic, fixed-seed split-generation code and cryptographic hashes of the patient-level training, validation, and test partitions, so that an independent user can verify that a reconstructed cohort reproduces the exact splits reported here.

Ethics statement

This paper used publicly available datasets and did not involve new data collection from human participants. Ethical approval and informed consent were handled by the original dataset providers.

Funding

This work was supported in part by the Joint Program on Health Science & Technology Innovation of Hainan Province (WSJK2024MS182) and by the Hainan Provincial Natural Science Foundation of China under Grant 823QN229.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the contributors and maintainers of the public physiological and clinical datasets used in this study, including MIMIC-IV Waveform, MIMIC-IV-ECG, PTB-XL, CirCor DigiScope, EPHNOGRAM, and WESAD. These resources have provided valuable support for reproducible research in cardiovascular signal processing and multimodal health intelligence.

Appendix A. Endpoint extraction protocol

This appendix specifies the complete protocol used to construct the composite cardiovascular deterioration label so that the endpoint can be reconstructed independently. The source tables, item identifiers, and rules are summarized in Table 1, and the full SQL queries, item-identifier lists, and configuration files are released in the public repository stated in the Code Availability section.

S1. Database versions and linkage keys. Labels are built from MIMIC-IV v3.1 and the MIMIC-IV Waveform Database, linked to MIMIC-IV-ECG. Records are joined on `subject_id`, `hadm_id`, and `stay_id`, and all event timestamps are aligned to the ICU-stay clock.

S2. Observation-window eligibility. Candidate windows use a 5-min observation length, a 60-s sliding step, a 1.5-h leakage-prevention gap, and a 24-h prediction horizon. A window is discarded if any endpoint criterion is already satisfied before or during the observation window, or if the first eligible endpoint falls within the gap interval.

S3. Endpoint component definitions. The composite label is the earliest of the six components listed below within the prediction interval; the earliest eligible event defines the first qualifying endpoint used for the patient-level count.

S4. Cardiac arrest and CPR. `procedureevents` and `chartevents` identifiers for cardiac arrest, cardiopulmonary resuscitation, and defibrillation during resuscitation; the earliest resuscitation event during the ICU stay is used.

S5. Death-time extraction. `deathtime` from `admissions` and `dod` from `patients`; only deaths with a resolvable timestamp within the prediction interval are included.

S6. Vasoactive initiation and escalation. `inputevents` infusion rates for norepinephrine, epinephrine, dopamine, dobutamine, and vasopressin are converted to weight normalized units. Initiation is the first non-zero infusion after a $\geq$30-min zero-rate baseline; escalation is a sustained rate increase $> 30\%$ maintained for $\geq$30 min after a stable baseline.

S7. Heart-failure treatment escalation. Escalation of intravenous furosemide or inotropes, or documented decompensated heart-failure management, from `inputevents`, `prescriptions`, and `chartevents`.

S8. Malignant arrhythmia. Documented ventricular tachyarrhythmia, severe bradyarrhythmia, or high-grade conduction block requiring treatment or clinical documentation, from rhythm-related `chartevents` and `procedureevents`.

S9. Urgent cardiovascular interventions. Non-elective cardiovascular procedures related to hemodynamic or electrical instability in `procedureevents`; elective or scheduled procedures are excluded.

S10. Exclusion and precedence rules. Duplicate records within a 5-min window are collapsed to the earliest timestamp. When several components occur in the same prediction interval, the earliest eligible event is retained. Windows with a pre-existing or gap-window endpoint are removed, as quantified in the cohort-construction flow (Table 12).

S11. SQL queries and item-identifier lists. The exact SQL for each component, together with the complete item-identifier and medication lists, is provided in the public repository; representative identifiers are listed in Table 1.

S12. Patient-level split generation. Patients are partitioned at the patient level into training, validation, and test sets (Table 10) with a fixed random seed. All windows from a patient are assigned to a single split, and normalization, threshold, and calibration parameters are estimated from the training or validation sets only.

Appendix B. Reporting standards and risk-of-bias assessment

This study is reported with reference to the TRIPOD and AI statement for prediction-model studies [35], and the PROBAST and AI

Table S1
TRIPOD + AI reporting items and their locations in this manuscript. Items marked as Partial or component-level only indicate limitations discussed in the main text; the study is not claimed to comply fully with every item.

Reporting item	Addressed	Location
Study type and objective (model development)	Yes	Title, Abstract, Introduction
Data source and setting	Yes	Cohort construction; Appendix A
Eligibility and participant flow	Yes	Cohort construction; Table 12
Outcome definition and timing	Yes	Endpoint definition; Table 13
Predictor definition and timing	Yes	Problem definition; signal preprocessing
Missing-data handling	Yes	Signal preprocessing; structured factorization
Sample size	Partial	Cohort construction (no a priori calculation)
Model specification	Yes	Methodology (framework to training objective)
Discrimination	Yes	Main prediction performance; Table 16
Calibration	Yes	Evaluation metrics; calibration analysis; Table 17
Data partition and internal validation	Yes	Leakage-safe prediction; Table 10
External validation	Component-level only	Task-specific component-level evaluation
Subgroup and fairness	Yes	Subgroup analysis; Table 27
Interpretability	Internal only	Internal physiological-consistency analysis
Reproducibility (code and data)	Yes	Code and Data Availability; Appendix A
Intended use and deployment claims	Moderated	Discussion; Conclusion

Table S2
PROBAST + AI-informed risk-of-bias and applicability considerations. Ratings are the authors' assessment intended to guide cautious interpretation rather than to certify low risk of bias.

PROBAST + AI domain	Rating	Basis and mitigation
Participants	Some concerns	Single-center retrospective ICU cohort selected by waveform, ECG, and PPG availability; participant flow is reported in Table 12.
Predictors	Low to some concerns	Predictors are extracted only from the observation window with a leakage-prevention gap; activity is represented by a proxy score rather than direct accelerometry.
Outcome	High concern	The composite endpoint includes care-process-mediated components whose timing depends on clinical decisions and is not assessed independently of care.
Analysis	Some concerns	Patient-level partitioning, class weighting, and validation-set calibration are used, but validation is internal and PPV depends on event prevalence.
Applicability	High concern	Developed in a single MIMIC-IV setting without prospective or all-modality external validation; deployment claims are limited to workstation- or bedside-server assistance.

framework [36] is used to structure the assessment of potential bias and applicability. Table S1 maps the principal reporting items to their manuscript locations, and Table S2 summarizes the domain-level risk-of-bias and applicability considerations. We do not claim full compliance with every checklist item; instead, we report where each item is addressed and flag the items that remain partial. This is a retrospective model-development study and should not be interpreted as a prospective clinical evaluation.

Data availability

The datasets used in this study are publicly available from their original repositories under the corresponding access conditions. The MIMIC-IV Waveform Database and MIMIC-IV-ECG Database are available through PhysioNet and require credentialed access, completion of the required training, and compliance with the PhysioNet data-use agreement. PTB-XL and the CirCor DigiScope dataset are also available from PhysioNet under their respective usage terms. EPHNOGRAM and WESAD are available from their original dataset repositories according to the access policies specified by the dataset providers. No new human-subject data were collected in this study. Because patient-level identifiers are governed by the PhysioNet data-use agreement and cannot be redistributed, the public repository described in the Code Availability section provides deterministic, fixed-seed split-generation code together with cryptographic hashes of the training, validation, and test partitions, so that credentialed users can regenerate and verify the exact patient-level splits used here; the derived feature-matrix format and preprocessing configuration files are included in the same repository. The complete endpoint-extraction protocol—database versions, source tables, item identifiers, medication and procedure lists, and timestamp and precedence rules—is provided in Appendix A and Table 1, and the corresponding SQL scripts and item-identifier lists are released in the public repository stated in the Code Availability section.

References

[1] G.A. Mensah, V. Fuster, C.J.L. Murray, G.A. Roth, G.B. of Cardiovascular Diseases, R. Collaborators, Global burden of cardiovascular diseases and risks, 1990–2022, J. Am. Coll. Cardiol. 82 (25) (2023) 2350–2473.
[2] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, A.Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nat. Med. 25 (1) (2019) 65–69.
[3] A.H. Ribeiro, M.H. Ribeiro, G.M.M. Paixão, D.M. Oliveira, P.R. Gomes, J.A. Canazart, M.P.S. Ferreira, C.R. Andersson, P.W. Macfarlane, W. Meira, T.B. Schön, A.L.P. Ribeiro, Automatic diagnosis of the 12-lead ECG using a deep neural network, Nat. Commun. 11 (2020) 1760.
[4] Z.I. Attia, S. Kapa, F. Lopez-Jimenez, P.M. McKie, D.J. Ladewig, G. Satam, P.A. Pellikka, M. Enriquez-Sarano, P.A. Noseworthy, T.M. Munger, et al., Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram, Nat. Med. 25 (1) (2019) 70–74.
[5] S. Raghunath, A.E. Ulloa Cerna, L. Jing, D.P. vanMaanen, J. Stough, D.N. Hartzel, J.B. Leader, H.L. Kirchner, M.C. Stumpe, A. Hafez, et al., Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network, Nat. Med. 26 (6) (2020) 886–891.
[6] A. Sau, et al., Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study, Lancet Digit. Health 6 (11) (2024) e791–e802.
[7] Q. Wang, Y. Zhang, G. Chen, Z. Chen, H.I. Hee, Assessment of heart rate and respiratory rate for perioperative infants based on ELC model, IEEE Sens. J. 21 (12) (2021) 13685–13694.
[8] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G.H. Tison, G.M. Marcus, J.M. Sanchez, C. Maguire, J.E. Olgin, M.J. Pletcher, DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018.
[9] H. Lee, et al., Real-time machine learning model to predict in-hospital cardiac arrest using heart rate variability in ICU, NPJ Digit. Med. 6 (2023) 215.
[10] G. Mathew, D. Barbosa, J. Prince, et al., Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes, npj Cardiovasc. Health 1 (2024) 25.
[11] F. Shaffer, J.P. Ginsberg, An overview of heart rate variability metrics and norms, Front. Public Health 5 (2017) 258.

[12] O. Faust, W. Hong, H. Loh, S. Xu, R.S. Tan, S. Chakraborty, P. Barua, F. Molinari, U.R. Acharya, Heart rate variability for medical decision support systems: a review, Comput. Biol. Med. 145 (2022) 105407.
[13] R.R. van de Leur, M.N. Bos, K. Taha, A. Sammani, M.W. Yeung, S. van Duijvenboden, et al., Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders, Eur. Heart J. - Digit. Health 3 (3) (2022) 390–404.
[14] P.C. Wouters, R.R. van de Leur, et al., Electrocardiogram-based deep learning improves outcome prediction after cardiac resynchronization therapy, Eur. Heart J. 44 (8) (2023) 680–692.
[15] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.
[16] P. Antiperovitch, D. Mortara, J. Barrios, R. Avram, K. Yee, A.N. Khaless, A. Cristal, G.H. Tison, J.E. Olgin, Continuous atrial fibrillation monitoring from photoplethysmography: comparison between supervised deep learning and heuristic signal processing, JACC Clin. Electrophysiol. 10 (2) (2024) 334–345.
[17] G.D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, R.G. Mark, Classification of Normal/Abnormal heart sound recordings: the PhysioNet/Computing in cardiology challenge 2016, Comput. Cardiol. 43 (2016) 609–612.
[18] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, F. Castells, J.M. Roig, I. Silva, A.E.W. Johnson, et al., An open access database for the evaluation of heart sound algorithms, Physiol. Meas. 37 (12) (2016) 2181–2213.
[19] D. Kiyasseh, T. Zhu, D.A. Clifton, CLOCS: contrastive learning of cardiac signals across space, time, and patients, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 5606–5615.
[20] R. Hu, J. Chen, L. Zhou, et al., Spatiotemporal self-supervised representation learning from multi-lead ECG signals, Biomed. Signal Process. Control 84 (2023) 104772.
[21] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, A. Dubrawski, MOMENT: a family of open time-series foundation models, in: Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.
[22] Y. Luo, Z. Zhang, et al., Toward a foundation model for multivariate wearable physiological signals, arXiv preprint arXiv:2412.09758, 2024.
[23] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: learning basic visual concepts with a constrained variational framework, in: International Conference on Learning Representations, 2017.
[24] H. Kim, A. Mnih, Disentangling by factorising, in: Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018, pp. 2649–2658.
[25] R.T.Q. Chen, X. Li, R.B. Grosse, D.K. Duvenaud, Isolating sources of disentanglement in variational autoencoders, in: Advances in Neural Information Processing Systems, 31, 2018.
[26] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 4114–4124.
[27] M. Kapsecker, M.C. Möller, S.M. Jonas, Disentangled representational learning for anomaly detection in single-lead electrocardiogram signals using variational autoencoder, Comput. Biol. Med. 184 (2025) 109422.
[28] Y. Jin, Z. Li, M. Wang, J. Liu, Y. Tian, Y. Liu, X. Wei, L. Zhao, C. Liu, Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis, Commun. Med. 4 (2024) 49.
[29] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F.I. Lunze, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, Sci. Data 7 (2020) 154.
[30] A.E.W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, L.-W.H. Lehman, L.A. Celi, R.G. Mark, MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (2023) 1.
[31] B. Moody, S. Hao, B. Gow, T. Pollard, W. Zong, R. Mark, MIMIC-IV waveform database, PhysioNet 2022. Version 0.1.0.
[32] B. Gow, T. Pollard, L.A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J.W. Waks, P. Eslami, T. Carbonati, A. Chaudhari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, S. Horng, MIMIC-IV-ECG: diagnostic electrocardiogram matched subset, PhysioNet 2023. Version 1.0.
[33] S. Tan, G. Androz, A. Chamseddine, P. Fecteau, A. Courville, Y. Bengio, J.P. Cohen, Icentia11K: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery, 2019.
[34] A. Kazemnejad, P. Gordany, R. Sameni, EPHNOGRAM: a simultaneous electrocardiogram and phonocardiogram database, PhysioNet, 2021. Version 1.0.0.
[35] G.S. Collins, K.G.M. Moons, P. Dhiman, et al., TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378.
[36] K.G.M. Moons, J.A.A. Damen, T. Kaul, et al., PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods, BMJ 388 (2025) e082505.