CardioSentinel-AI: An Explainable, Quality-Aware Edge Framework for Personalised Multimodal Cardiorespiratory Abnormality Detection

[1]Pravat Satpathy, [2]Sonia Mohapatra
[1]Researcher, [2]Researcher
[1,2]Research and Development Division
[1,2]Tekstein Solutions Private Limited, Bhubaneswar, Odisha, India
[1]analytics.pravat@gmail.com, [2]smctc98@gmail.com

***Abstract*—Low-cost monitors usually compare isolated vital signs with fixed population limits, which raises false alarms in people whose normal physiology sits near a limit, misses change in people whose baseline sits far from it, and fails silently when a sensor loses contact. This paper presents CardioSentinel-AI, an edge framework built around a Raspberry Pi gateway that combines electrocardiogram, photoplethysmogram, oxygen saturation, respiration and temperature channels. Each modality has its own gradient-boosted expert whose log-odds are weighted by a signal-quality index and summed, so every alert decomposes exactly into per-modality contributions. Features are also expressed as deviations from a personal baseline learned during a short enrolment, shrunk toward population statistics. The framework was evaluated on a physiologically grounded synthetic cohort of 80 virtual subjects (6,400 windows of 30 s) with atrial-fibrillation-like rhythm, desaturation, tachycardia and ectopy episodes, motion artefact and sensor dropout, under subject-independent cross-validation. Personalisation dominated: at 90% sensitivity it cut false alarms from 17.5 per hour for an equivalent population model to between 1.6 and 3.2 per hour, against 28.9 per hour for a fixed-threshold rule of the NEWS2 type. The quality-aware additive fusion did not outperform a personalised early-fusion model under nominal conditions (AUROC 0.989 against 0.993), but it degraded more slowly when each of the ECG, PPG and oximeter channels was dropped with probability 0.5 (0.917 against 0.890) and was better calibrated. Continuous quality weighting added little beyond masking absent channels. All results are simulated and must next be confirmed on MIT-BIH, BIDMC and prospective recordings.**

***Index Terms*—Edge computing, electrocardiography, explainable artificial intelligence, multimodal sensor fusion, personalised anomaly detection, photoplethysmography, Raspberry Pi, signal quality index, wearable health monitoring.**

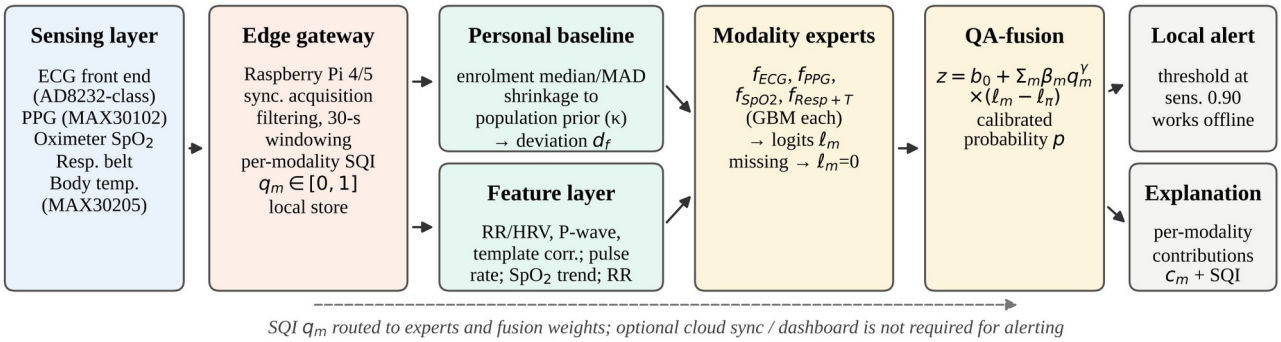


Fig. 1. Layered architecture of CardioSentinel-AI. Signal-quality indices q_m computed on the gateway feed both the modality experts and the fusion weights. Alerting runs entirely on the edge device; cloud synchronisation and the research dashboard are optional.

I. Introduction

Cardiovascular disease causes an estimated 17.9 million deaths each year [1]. Much of the deterioration that precedes serious events happens outside intensive care, where monitoring is intermittent. Single-board computers, low-cost photoplethysmography (PPG) modules and analogue electrocardiogram (ECG) front ends have made continuous multiparameter monitoring cheap to build. Building it well is a different matter.

Most low-cost monitors still reduce each signal to one number and compare it with a fixed band. Early-warning scores such as NEWS2 formalise this logic for periodic observation [2]; applied continuously, the same bands misfire. A resting heart rate of 95 min^{-1} is unremarkable for one person and a 40% rise for another; an oxygen saturation of 94% is normal for some people with chronic lung disease and alarming for a healthy adult. Consumer devices show the consequence at scale: in the Apple Heart Study, atrial fibrillation was confirmed by a later ECG patch in only 34% of notified participants who returned one [3]. PPG is also vulnerable to motion and contact artefact [4], [5], and pulse oximetry has documented accuracy limits across skin pigmentation [6].

Machine learning has shown cardiologist-level rhythm classification on curated ECG collections [7], [8]. Three deployment problems remain open. First, performance measured on windows from the same patients used in training

overstates what happens on a new person [9]; patient-specific models close part of the gap but need labelled beats from each patient [10]. Second, quality-aware fusion of ECG and PPG has mostly been studied for heart-rate estimation [11]–[13], not for abnormality detection with an explicit alarm budget. Third, compact models now run on Raspberry Pi class hardware in under a millisecond [14], yet edge studies rarely report how decisions are explained or how they behave when a channel disappears.

This paper develops a doctoral concept note on an explainable edge platform for multimodal cardiovascular monitoring and tests three of its hypotheses: that multimodal fusion improves on a single sensor (H1), that personal baselines reduce false alarms relative to population models (H2), and that quality-aware fusion degrades more gracefully when signals become unreliable (H3). Edge feasibility (H4) is profiled but not yet measured on target hardware. The contributions are:

1) A quality-aware additive logit fusion whose sum gives an exact per-modality explanation of every alert and Bayes-consistent handling of absent channels under conditional independence (Proposition 2).
2) A shrinkage personal-baseline layer and a closed-form expression for the false-alarm inflation caused by short enrolment (Proposition 1), checked by Monte Carlo.
3) A seeded synthetic cohort generator with pathology, benign activity, artefact and dropout, and a subject-independent benchmark of eleven detectors with stress tests.
4) Ablations showing that personalisation carries most of the gain, continuous quality weighting adds little beyond masking, and frozen operating points lose sensitivity under shift.

II. Related Work

Table I places the work against representative studies. De Chazal *et al.* [9] established the inter-patient protocol on the MIT-BIH Arrhythmia Database [15] and showed how strongly intra-patient splits inflate results. Kiranyaz *et al.* [10] trained a 1-D convolutional network per patient from a few minutes of expert-labelled beats; CardioSentinel-AI instead uses unlabelled enrolment data, which is easier to obtain at the bedside. Farag [14] fitted a 15-kB matched-filter CNN on a Raspberry Pi with sub-millisecond inference.

For multimodal fusion, Li *et al.* [11] combined ECG, blood-pressure and PPG heart-rate estimates in a Kalman filter whose noise covariance was driven by signal-quality indices (SQIs), and Orphanidou *et al.* [12] derived practical SQIs for ambulatory ECG and PPG. A recent preprint, CardioFusion-AI [13], compared eight ECG–PPG fusion strategies for heart-rate regression on synthetic windows and found that an SQI-conditioned gate helped mainly when a modality was missing, not under graded degradation. The present study reaches a consistent conclusion for a different task. Among the works reviewed, none evaluates personal baselines, missing-modality-aware fusion, built-in explanation and operating-point behaviour together under subject-wise splits; whether this gap holds across the wider literature still needs a systematic review.

TABLE I
Representative Prior Work and the Present Study

Ref.	Signals	Task	SQI	Pers.	XAI	Edge
[9] 2004	ECG	Beat class.	No	No	No	No
[10] 2016	ECG	Beat class.	No	Labelled	No	Real-time
[7] 2019	ECG	Rhythm	No	No	No	No
[14] 2023	ECG	Beat class.	No	No	Part	RPi
[11] 2008	ECG, ABP, PPG	HR est.	Yes	No	No	No
[13] 2026	ECG, PPG	HR reg.	Gate	No	No	No
[3] 2019	PPG (watch)	AF screen	Part	No	No	Watch
Ours	**ECG, PPG, SpO_2, R, T**	**Abnormal**	**Yes**	**Unlab.**	**Exact**	**RPi**

Unlab.: unlabelled enrolment; *target platform, not yet measured.*

III. Problem Formulation

Let subject s be observed in non-overlapping windows t of $\Delta = 30$ s. Each window carries modality observations $x_{s,t}^{(m)}$ for $m \in \mathcal{M} = \{\text{ECG, PPG, SpO}_2\text{, Resp+T}\}$, a quality index $q_m \in [0, 1]$ with $q_m = 0$ when the channel is absent, and a label $y \in \{0, 1\}$ that marks a window as abnormal when an episode covers at least half of it. The first T_b minutes of every recording form an enrolment set $\mathcal{E}_s$ assumed free of pathological episodes. The detector must output a probability $p_{s,t}$ and an alarm decision $p \geq \tau$, where τ is fixed on training subjects to give 90% sensitivity. At that operating point the quantity to minimise is the false-alarm rate per hour of normal monitoring,

$$\text{FA/h} = N_{\text{FP}} / (N_{\text{neg}}\, \Delta / 3600), \tag{1}$$

in which N_{FP} is the number of false-positive windows and N_{neg} the number of normal windows. Three constraints apply: alerting must run offline on the edge device, training and test subjects must be disjoint, and every alarm must be traceable to the modalities that caused it.

IV. Proposed Framework

A. Architecture

Fig. 1 shows the four layers. The sensing layer pairs an AD8232-class single-lead ECG front end and a MAX30102-class PPG module for prototyping with a validated pulse oximeter for SpO_2, a respiratory belt and a MAX30205-class body-temperature sensor. The Pi has no analogue input, so the ECG passes through an external converter; other sensors use I^2C or USB. The gateway aligns all channels, filters them, cuts 30-s windows and computes per-modality SQIs. Features, personal deviations, expert inference and fusion run locally, so an alert does not depend on network access.

B. Signal Quality Indices

For ECG, two R-peak detectors are run on the 5–15 Hz band: an energy-envelope detector in the Pan–Tompkins style [16] and a prominence detector on the rectified signal. Following [11], quality is their agreement,

$$q_{\text{ECG}} = 2|\mathcal{A} \cap \mathcal{B}| / (|\mathcal{A}| + |\mathcal{B}|), \tag{2}$$

where $\mathcal{A}$ and $\mathcal{B}$ are the two beat sets and a match requires peaks within 100 ms. For PPG, q_{PPG} is the square of the mean correlation between each pulse and the window template [12]. Oximeter quality is the fraction of valid readings, penalised for

physiologically implausible jumps of 3 points or more per second. Respiratory quality is the share of 0.08–0.8 Hz power within ±0.06 Hz of the dominant peak. A flat channel receives zero.

C. Personal Baseline with Shrinkage

For each of eleven features f (ECG and PPG rate, RMSSD, RR irregularity, P-wave energy ratio, pulse amplitude, mean and minimum SpO_2, respiratory rate and temperature) the median $m_{s,f}$ and scaled median absolute deviation $r_{s,f}$ are computed over the n enrolment windows whose SQI exceeds 0.5. Short enrolments give noisy estimates, so both are pulled toward the population values μ_{pop} and σ_w (pooled within-subject spread) of the training subjects:

$$\mu_{s,f} = (n\, m_{s,f} + \kappa\, \mu_{pop,f}) / (n + \kappa), \tag{3}$$

$$\sigma^2_{s,f} = (n\, r^2_{s,f} + \kappa\, \sigma^2_{w,f}) / (n + \kappa), \tag{4}$$

$$d_{s,t,f} = (x_{s,t,f} - \mu_{s,f}) / \sigma_{s,f}. \tag{5}$$

The prior strength κ behaves like a count of pseudo-windows; $\kappa = 4$ was used. With little enrolment data the deviation d stays close to a population z-score, and as n grows it becomes personal.

D. Quality-Aware Additive Logit Fusion

Each modality expert f_m is a histogram gradient-boosting classifier (200 trees, 15 leaves) trained only on windows in which that modality is present. Its inputs are the modality features, the corresponding deviations and its own SQI; its output is a logit ℓ_m. With ℓ_π the prior log-odds, the fused score is

$$z = b_0 + \Sigma_m\, \beta_m\, q_m^{\gamma}\, (\ell_m - \ell_\pi), \quad p = 1 / (1 + e^{-z}). \tag{6}$$

The weights b_0 and β_m are fitted by logistic regression on out-of-fold expert logits, obtained with a four-fold split grouped by subject inside the training set, so the stacker never sees in-sample expert outputs. The exponent $\gamma \in \{0.5, 1, 2, 4\}$ is chosen by inner log-loss. A missing modality has $q_m = 0$ and drops out of the sum.

E. Built-in Explanation

Defining the contribution of modality m as

$$c_m = \beta_m\, q_m^{\gamma}\, (\ell_m - \ell_\pi), \tag{7}$$

one obtains $z = b_0 + \Sigma_m\, c_m$ exactly. For every alarm the dashboard can state how many log-odds units came from each channel and at what quality it was read. No post-hoc approximation such as SHAP [17] is needed at this level; within-modality attribution can still be added with tree explainers.

F. Algorithm and Complexity

Algorithm 1 lists the per-window inference path executed on the gateway.

Algorithm 1: Online inference for one 30-s window

Input: window; (μ, σ); $\{f_m\}$; b_0, β, γ; τ
for each m **do** compute q_m; **if** $q_m = 0$: $c_m \leftarrow 0$, **next**
  extract features x_m; $d_m \leftarrow (x_m - \mu_m) / \sigma_m$
  $\ell_m \leftarrow f_m(x_m, d_m, q_m)$; $c_m \leftarrow \beta_m\, q_m^{\gamma}(\ell_m - \ell_\pi)$
$z \leftarrow b_0 + \Sigma\, c_m$; $p \leftarrow \sigma(z)$
if $p \geq \tau$ **then** raise local alert
log (p, c, q, model version) for optional sync

Feature extraction is linear in the window length W (3,750 samples per channel), dominated by filtering. Expert inference costs $O(M\, T\, D)$ for $M = 4$ experts of $T = 200$ trees of depth D, and the stacker adds $O(M)$. Model memory is $O(M\, T\, L)$ with $L = 15$ leaves per tree, and the baseline needs two numbers per feature per subject.

V. Analytical Properties

Proposition 1 (false-alarm inflation). Let a feature be Gaussian and stationary for a healthy subject, with mean and standard deviation estimated from n independent enrolment samples (sample mean μ_n and unbiased standard deviation s_n, $\kappa = 0$). The probability that a new healthy sample triggers the rule $|x - \mu_n| / s_n > k$ is

$$\alpha(n, k) = 2[1 - F_{t,n-1}(k / \sqrt{(1 + 1/n)})], \tag{8}$$

where $F_{t,n-1}$ is the Student-t distribution function with $n - 1$ degrees of freedom. *Proof sketch:* $x - \mu_n$ is Gaussian with variance $\sigma^2(1 + 1/n)$ and independent of s_n, so the ratio scaled by $\sqrt{(1 + 1/n)}$ follows a Student-t law with $n - 1$ degrees of freedom. As $n \to \infty$, α tends to $2[1 - \Phi(k)]$, the known-parameter rate. ∎ Short enrolments therefore inflate false alarms, which motivates (3)–(4).

Proposition 2 (missing modalities). If the modalities are conditionally independent given y and each expert returns the calibrated posterior logit ℓ_m under the common prior ℓ_π, the Bayes posterior log-odds given any available subset $\mathcal{S} \subseteq \mathcal{M}$ is

$$\text{logit}\, P(y = 1 \mid x_{\mathcal{S}}) = \ell_\pi + \Sigma_{m \in \mathcal{S}}\, (\ell_m - \ell_\pi). \tag{9}$$

Proof: Bayes' rule gives ℓ_π plus the sum of per-modality log-likelihood ratios, and each ratio equals $\ell_m - \ell_\pi$. ∎ Setting $q_m = 0$ in (6) is therefore the correct marginalisation of an absent channel, and $\beta_m = q_m = 1$ recovers (9). Learned $\beta_m \neq 1$ compensate for dependence between modalities. Early fusion with imputation has no such guarantee.

VI. Simulation Setup

A. Synthetic Multimodal Cohort

PhysioNet could not be reached from the computing environment used in this phase, so evaluation relies on a purpose-built generator. ECG beats are sums of Gaussian P, Q, R, S and T waves in the spirit of McSharry *et al.* [18], with QT scaled by √RR, respiratory sinus arrhythmia and subject-specific morphology. Each beat launches a PPG pulse (systolic and diastolic Gaussian components) after a subject-specific pulse transit time; pulse amplitude depends on the preceding RR interval, so premature beats produce weak pulses or a pulse deficit. Respiration drives a belt signal and modulates ECG and PPG amplitude; SpO_2 is reported once per second as an integer. Fig. 2(a) shows example excerpts and Table II the main settings.

Four pathological episode types are injected after enrolment: atrial-fibrillation-like rhythm (no P wave, 6-Hz fibrillatory activity, log-normal RR dispersion, rate +10 to +25 min^{-1}), desaturation (−4 to −9 points with modest rises in heart and respiratory rate), tachycardia (+25 to +45 min^{-1} at rest) and ventricular ectopy (30% premature beats with compensatory pauses). Benign activity bouts raise heart rate by 10–28 min^{-1}, usually with motion. Motion artefact (low-pass noise plus impulsive spikes that fool R-peak detection) is several times stronger on the PPG, and contact loss flattens a channel for 10–

90 s. Separate random streams for physiology and noise allow each subject to be re-rendered at other artefact severities.

TABLE II
SIMULATION AND MODEL PARAMETERS

Parameter	Value
Subjects; recording	80; 40 min (10 enrolment + 30 test)
Sampling; window	125 Hz waveforms, 1 Hz SpO_2/T; 30 s
Evaluated windows	4,800; 21.0% abnormal; 17 subjects event-free
Episodes (windows)	AF 264, desat. 231, tachy. 246, ectopy 266
Resting HR; SpO_2 baseline	$\mathcal{N}(72, 11)$ min^{-1}; $\mathcal{N}(97, 1.4)$% (clipped)
Resp. rate; temperature	$\mathcal{N}(15, 2.8)$ min^{-1}; $\mathcal{N}(36.7, 0.3)$ °C
Motion time; dropout	8.9%; 0.7–1.5 episodes/h per channel
Experts	HistGB: 200 trees, 15 leaves, lr 0.06
Baseline	T_b = 10 min, κ = 4, SQI ≥ 0.5
Operating point; seed	Sensitivity 0.90 on training subjects; 2026

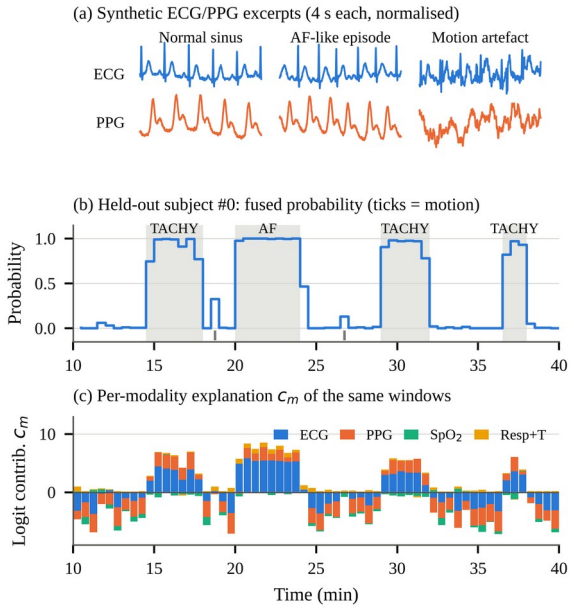


Fig. 2. (a) Synthetic ECG and PPG during sinus rhythm, an AF-like episode and motion artefact. (b) Fused probability for a held-out subject; shaded spans are ground-truth episodes and ticks mark motion. (c) Per-modality contributions c_m from (7); positive values push toward an alarm.

B. Comparators and Protocol

Eleven detectors were compared: a NEWS2-style band aggregate plus an RR-irregularity flag [2]; a personal $|z|$-max rule on (5); an Isolation Forest [19] fitted on enrolment deviations; ECG-only logistic regression and gradient boosting; early fusion by gradient boosting on all features, with and without personal deviations; naive-Bayes late fusion; stacked late fusion with q fixed to 1; and the proposed fusion with and without personalisation. Evaluation used five-fold cross-validation grouped by subject. Each operating threshold was set on out-of-fold training predictions and then frozen. Metrics were AUROC, AUPRC, sensitivity, specificity, FA/h, Brier score and expected calibration error (ECE, 10 bins); 95% intervals come from 1,000 subject-level bootstrap resamples. Stress tests removed each of ECG, PPG and SpO_2 independently with probability p at test time (three repeats), or re-rendered the test subjects at 0.5–3 times the nominal artefact rate and amplitude, without retraining. Code used Python, NumPy, SciPy and scikit-learn [20].

VII. RESULTS AND DISCUSSION

A. Overall Comparison

Table III summarises the nominal condition. Fusion helped (H1): adding PPG, SpO_2 and respiration to an ECG-only gradient-boosting model raised AUROC from 0.859 to 0.963 and cut false alarms from 73.5 to 17.5 per hour. The ECG-only models were almost blind to desaturation, detecting only 61–68% of those windows. Personalisation produced the largest single gain (H2). Early fusion with deviation features reached AUROC 0.993 and 1.6 false alarms per hour, and the proposed fusion reached 0.989 and 3.2 per hour. The NEWS2-style rule, at the same 90% sensitivity, produced 28.9 false alarms per hour; it caught AF-like and ectopic windows well through its irregularity flag but missed 24% of tachycardia windows, mostly in subjects with low resting rates.

TABLE III
SUBJECT-INDEPENDENT RESULTS AT 90% TARGET SENSITIVITY

Detector	AUROC (95% CI)	AUPRC	Sens.	Spec.	FA/h
Fixed thresholds	0.879 (.855–.899)	0.602	0.90	0.76	28.9
Personal $\|z\|$ rule	0.873 (.854–.890)	0.572	0.90	0.78	26.4
Isolation Forest	0.910 (.896–.923)	0.584	0.90	0.85	17.4
ECG-only LR	0.859 (.814–.902)	0.775	0.90	0.46	64.7
ECG-only GBM	0.859 (.823–.894)	0.783	0.92	0.39	73.5
Early GBM	0.963 (.945–.976)	0.917	0.91	0.85	17.5
Early GBM + pers.	0.993 (.990–.996)	0.980	0.90	0.99	1.6
Naive-Bayes late	0.945 (.927–.962)	0.882	0.91	0.78	26.4
QA, no pers.	0.956 (.939–.971)	0.909	0.91	0.86	17.2
Late + pers., q=1	0.988 (.984–.992)	0.969	0.90	0.97	3.1
Proposed	**0.989 (.984–.992)**	**0.969**	**0.90**	**0.97**	**3.2**

The proposed model did not beat personalised early fusion under nominal conditions. A paired subject bootstrap gave ΔAUROC = −0.0046 (95% CI −0.0074 to −0.0021), and its false-alarm rate was about twice as high. Boosting over all features can learn cross-modal interactions, such as an irregular RR series beside a normal pulse train, that an additive model cannot represent. This is a negative result for the fusion architecture; the case for it rests on Sections VII-C to VII-F.

B. Personal Baselines and Enrolment Length

Resting heart rate ranged from 51.6 to 98.8 min^{-1} and baseline SpO_2 from 94.4% to 99.2% across the cohort. Fig. 3(c) shows how the benefit depends on enrolment length. With one minute of enrolment CardioSentinel-AI produced 9.8 false alarms per hour; with ten minutes it produced 3.2. Proposition 1 explains the trend, and Fig. 3(d) confirms (8) against Monte Carlo simulation to within sampling error. With k = 3 and n = 20 windows the false-alarm probability is 0.0086, 3.2 times the known-parameter value of 0.0027; with n = 5 it is 19 times higher. Shrinkage was not uniformly helpful. For the $|z|$ rule it lowered false alarms from 63.5 to 26.4 per hour at ten minutes but raised them at one and two minutes, when the prior dominates the estimate. For the learned model it gave a modest gain at short enrolment (11.3 to 9.8 per hour at one minute) and

a slight loss at five and ten minutes (4.7 to 5.2 and 3.1 to 3.2 per hour).

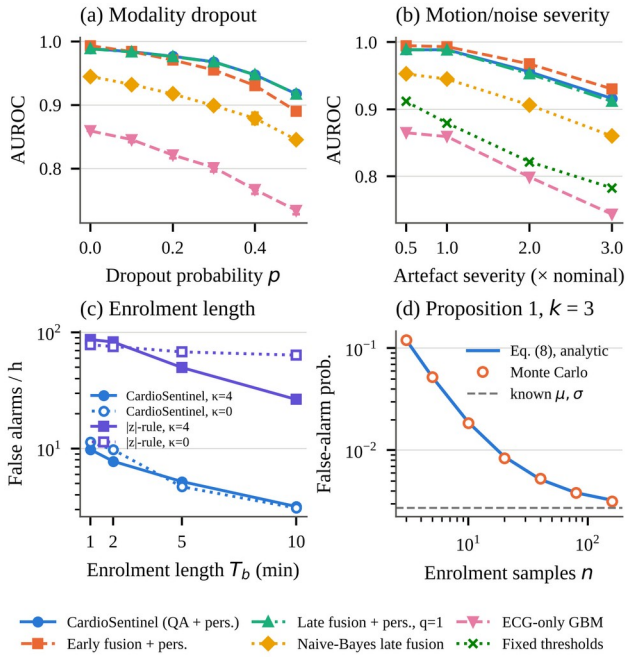


Fig. 3. Stress tests on held-out subjects without retraining. (a) AUROC versus test-time dropout of ECG, PPG and SpO_2 (mean ± s.d. of three repeats). (b) AUROC versus artefact severity. (c) False alarms per hour at the frozen operating point versus enrolment length, with (κ = 4) and without (κ = 0) shrinkage. (d) Analytic false-alarm probability (8) against Monte Carlo estimates.

C. Robustness to Dropout and Artefact

Fig. 3(a) gives partial support to H3. The personalised models were indistinguishable up to $p = 0.1$; beyond that the additive fusion pulled ahead, and at $p = 0.5$ it held AUROC 0.917 against 0.890 for personalised early fusion and 0.845 for naive-Bayes late fusion; the ECG-only model fell to 0.734. Ranking is only half of the picture: because thresholds were frozen on complete training data, sensitivity at $p = 0.5$ dropped to 0.65 for CardioSentinel-AI and 0.51 for early fusion. For a monitor this matters more than AUROC, and it argues for thresholds that depend on which channels are available. Under heavier artefact, Fig. 3(b), personalised early fusion stayed ahead (AUROC 0.930 against 0.916 at three times nominal severity), and both personalised models kept only 0.57–0.58 sensitivity. Less expectedly, at half the nominal artefact severity CardioSentinel-AI produced 11.1 false alarms per hour against 3.2 nominally: cleaner enrolment gave tighter personal spreads, so ordinary variation appeared as larger deviations. Enrolment should match routine-use conditions.

D. Where Should Signal Quality Enter?

Table IV isolates the role of the SQI. Removing quality information from both the experts and the fusion weights, while keeping the mask for absent channels, changed nominal and dropout AUROC by no more than 0.001. The inner search selected the smallest exponent, γ = 0.5, in every fold, which points the same way. Continuous quality helped only under heavy artefact, where false alarms fell from 3.7 to 2.0 per hour and AUROC rose from 0.908 to 0.916. Robustness therefore comes mostly from treating an absent channel as absent, as Proposition 2 prescribes, not from grading a channel that is present. CardioFusion-AI reported the same pattern for heart-rate regression [13].

TABLE IV
ABLATION OF SIGNAL-QUALITY PLACEMENT (PERSONALISED MODELS)

Quality used in	AUROC nominal	AUROC $p = 0.5$	AUROC art. ×3	FA/h art. ×3
Experts + fusion (proposed)	0.989	0.917	0.916	2.0
Experts only	0.988	0.917	0.912	2.5
Fusion weights only	0.988	0.918	0.913	2.6
Neither (mask only)	0.988	0.918	0.908	3.7

E. Data Leakage Check

The concept note warned against splitting adjacent windows of one person between training and test sets. When the population early-fusion model was evaluated with window-level random folds, AUROC rose from 0.963 to 0.989 and AUPRC from 0.917 to 0.971. The inflation is nearly as large as the whole personalisation gain, so a random split would have hidden the effect under study.

F. Explanation and Calibration

Fig. 2(b)–(c) traces one held-out subject. Each tachycardia and AF-like episode drives the probability close to 1, and the stacked contributions show which channel carried each decision. Two short excursions coincide with motion, where the ECG term is small because its quality is low. Across test subjects the ECG contribution to normal windows was −0.58 log-odds during motion against −2.32 at rest: a corrupted ECG loses its reassurance instead of becoming an alarm. In Fig. 4(a), desaturation is attributed mainly to SpO_2 (+4.6) and PPG (+2.4), and AF-like and ectopic windows to ECG (+4.3, +3.7). The respiration and temperature expert contributed almost nothing (stacking weight −0.24) and is a candidate for removal. Calibration (Fig. 4(b)) favoured the additive model: ECE was 0.006, against 0.010 for personalised early fusion and 0.035 for early fusion without personalisation, although early fusion had the lower Brier score (0.024 against 0.029).

G. Edge Cost

On one 2.8 GHz Xeon core, feature extraction for all channels of a 30-s window took 4.0 ms (median; 95th percentile 5.5 ms), and fusion inference took 14.8 ms, most of it Python data-frame overhead. The serialised model occupied 1.4 MB. These figures are not Raspberry Pi measurements. Under an assumed, unverified twentyfold slowdown on a Pi 4, processing would take about 0.38 s per 30-s window, under 2% of one core. H4 stays open until the prototype is profiled, including power and uptime.

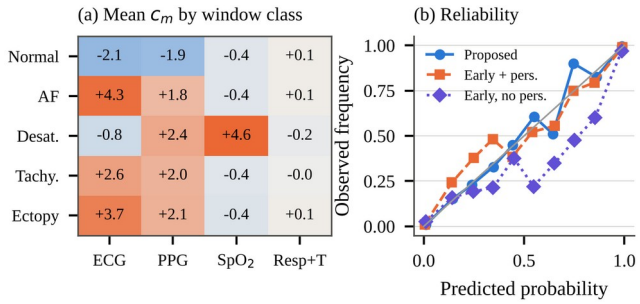


Fig. 4. (a) Mean per-modality contribution c_m for each window class across all held-out subjects. (b) Reliability diagram of pooled out-of-fold probabilities (bins with at least 15 windows).

H. Practical Significance and Limitations

For a ward or home service the decisive quantity is the false-alarm budget. At equal sensitivity the full personalised pipeline reduced it about ninefold relative to fixed bands, and within each fusion architecture personalisation alone accounted for a five- to elevenfold reduction, with no new hardware. The additive fusion is worth keeping for its explanations, calibration and dropout behaviour, not for nominal accuracy.

The limitations are substantial. All data are synthetic, and one team wrote both the generator and the feature extractors, so the features may fit simulated physiology better than real patients; the results describe design behaviour, not clinical accuracy. Enrolment was assumed to be free of events. Deep waveform models were not benchmarked. SpO_2 and PPG were simulated without skin-tone bias [6] or sensor-specific error. Nothing here supports diagnostic use. Any human-subject phase requires ethics approval, informed consent, data handling under India's Digital Personal Data Protection Act, 2023 [21], and a regulatory position under the Medical Devices Rules, 2017 [22].

VIII. Conclusion

CardioSentinel-AI combines quality-indexed modality experts, shrinkage personal baselines and an additive fusion that explains itself channel by channel. In a subject-independent simulation of 80 virtual subjects, personalisation produced most of the improvement, cutting false alarms at 90% sensitivity from 28.9 per hour for a fixed-threshold rule and 17.5 per hour for population fusion to 1.6–3.2 per hour. The additive fusion was slightly less accurate than personalised early fusion on clean data but better calibrated and more tolerant of dropout; graded quality weighting added little beyond masking. Next steps are subject-wise benchmarks on MIT-BIH [15], the 12-lead arrhythmia database [23] and BIDMC PPG and respiration recordings [24], with MIMIC-IV [25] reserved for clearly labelled deterioration outcomes; profiling on Raspberry Pi hardware; operating thresholds that adapt to the available channels; and an ethics-approved prospective study with paired reference devices.

References

[1] World Health Organization, “Cardiovascular diseases (CVDs),” Fact sheet, Geneva, Switzerland, Jun. 2021.

[2] Royal College of Physicians, *National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS*. London, U.K.: RCP, 2017.

[3] M. V. Perez *et al.*, “Large-scale assessment of a smartwatch to identify atrial fibrillation,” *N. Engl. J. Med.*, vol. 381, no. 20, pp. 1909–1917, 2019.

[4] P. H. Charlton *et al.*, “Wearable photoplethysmography for cardiovascular monitoring,” *Proc. IEEE*, vol. 110, no. 3, pp. 355–381, Mar. 2022.

[5] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, “Investigating sources of inaccuracy in wearable optical heart rate sensors,” *npj Digit. Med.*, vol. 3, Art. no. 18, 2020.

[6] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, “Racial bias in pulse oximetry measurement,” *N. Engl. J. Med.*, vol. 383, no. 25, pp. 2477–2478, 2020.

[7] A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[8] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, “Deep learning for ECG analysis: Benchmarks and insights from PTB-XL,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1519–1528, May 2021.

[9] P. de Chazal, M. O’Dwyer, and R. B. Reilly, “Automatic classification of heartbeats using ECG morphology and heartbeat interval features,” *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[10] S. Kiranyaz, T. Ince, and M. Gabbouj, “Real-time patient-specific ECG classification by 1-D convolutional neural networks,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.

[11] Q. Li, R. G. Mark, and G. D. Clifford, “Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter,” *Physiol. Meas.*, vol. 29, no. 1, pp. 15–32, 2008.

[12] C. Orphanidou *et al.*, “Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring,” *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 832–838, May 2015.

[13] N. Krishnan K. and Janakiraman K., “CardioFusion-AI: Robust ECG–PPG fusion for multimodal physiological monitoring under signal degradation,” arXiv:2608.26000, Aug. 2026.

[14] M. M. Farag, “A tiny matched filter-based CNN for inter-patient ECG classification and arrhythmia detection at the edge,” *Sensors*, vol. 23, no. 3, Art. no. 1365, 2023.

[15] G. B. Moody and R. G. Mark, “The impact of the MIT-BIH arrhythmia database,” *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001.

[16] J. Pan and W. J. Tompkins, “A real-time QRS detection algorithm,” *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[17] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

[18] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, “A dynamical model for generating synthetic electrocardiogram signals,” *IEEE Trans. Biomed. Eng.*, vol. 50, no. 3, pp. 289–294, Mar. 2003.

[19] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.

[20] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[21] Government of India, *The Digital Personal Data Protection Act, 2023* (Act No. 22 of 2023), Aug. 2023.

[22] MoHFW, Government of India, *Medical Devices Rules, 2017*, G.S.R. 78(E), Jan. 2017.

[23] J. Zheng, H. Guo, and H. Chu, “A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0),” PhysioNet, 2022.

[24] M. A. F. Pimentel *et al.*, “Toward a robust estimation of respiratory rate from pulse oximeters,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1914–1923, Aug. 2017.

[25] A. E. W. Johnson *et al.*, “MIMIC-IV, a freely accessible electronic health record dataset,” *Sci. Data*, vol. 10, Art. no. 1, 2023.