Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Modality-agnostic late fusion of electrocardiogram, phonocardiogram and photoplethysmogram signals for cardiovascular disease detection

MD Tasnim Alam

mdtasnim.alam@uq.edu.au

University of Queensland https://orcid.org/0009-0005-7798-2035

Md Zakir Hossain

University of New South Wales

Mohammad Ali Moni

University of Queensland

Research Article

Keywords: Cardiovascular Disease, AI, ML, Late fusion, Missing modalities, Electrocardiogram, Phonocardiogram, Photoplethysmogram

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11551830/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Modality-agnostic late fusion of electrocardiogram, phonocardiogram and photoplethysmogram signals for cardiovascular disease detection

1 Introduction

Cardiovascular diseases remain the leading cause of death worldwide (Roth et al. 2020), and much of their early detection relies on three inexpensive, non-invasive signals. The electrocardiogram (ECG) records the heart's electrical activation, the phonocardiogram (PCG) the sounds produced by valve closure and turbulent flow, and the photoplethysmogram (PPG) the arrival of each pulse in the peripheral circulation (Allen 2007; Charlton et al. 2023). Each has supported deep-learning detectors of cardiologist-level skill for specific tasks: arrhythmia from single-lead ECG (Hannun et al. 2019), abnormal heart sounds from PCG (Clifford et al. 2016; Liu et al. 2016) and atrial fibrillation (AF) from PPG (Pereira et al. 2020; Torres-Soto and Ashley 2020).

The three signals increasingly arrive together, but rarely all at once and rarely all usable. Digital stethoscopes record a PCG together with a single-lead ECG (Bachtiger et al. 2022); smartwatches combine continuous PPG with an on-demand ECG (Tison et al. 2018; Perez et al. 2019); intensive-care monitors record ECG and PPG side by side (Johnson et al. 2016; Moody et al. 2020). In each setting a signal can be absent by design, or recorded but rejected by a quality check (Orphanidou et al. 2015; Elgendi 2016). A deployable model therefore has to give a decision from whatever subset of signals is present, and the clinically useful question is not only how accurate the full combination is, but what each additional signal adds and how the model behaves when one is missing (Acosta et al. 2022; Huang et al. 2020; Baltrušaitis et al. 2019). We call a model modality-agnostic when, trained once, it returns a calibrated decision from any combination of available signals, without imputing the missing ones or keeping a separate model for each combination.

Multimodal ECG–PCG studies have reported high accuracies on the only public dataset with synchronised recordings and a diagnostic label, the MIT heart-sound database included in the PhysioNet/CinC 2016 Challenge as training-a (Li et al. 2019; H. Li et al. 2021; P. Li et al. 2021; Hettiarachchi et al. 2021; Li et al. 2022; Zhang et al. 2024; Li et al. 2025; Calzoni et al. 2025; Bargarai et al. 2026), and ECG–PPG models have been proposed for AF

(Aldughayfiq et al. 2023; Malone et al. 2025; John et al. 2025; Guo et al. 2025). Five aspects of this literature limit what can be concluded about fusion itself. First, protocols differ widely and subject identity is not released for training-a, whose 409 recordings come from 121 patients (Clifford et al. 2016); splits by recording, or by beat-length segments of the same recording, let information leak between training and test data (Saeb et al. 2017; Kapoor and Narayanan 2023). Second, gains of a fused model over its best single-signal component are usually reported as point differences, without a paired test on the same records. Third, why fusion helps is rarely examined: classifier-combination theory predicts the largest gains when the experts' errors are independent (Kittler et al. 1998; Kuncheva 2014), yet the dependence between signal-specific experts is seldom measured. Fourth, fused models are usually evaluated only on the records that have every signal, so how they behave on the records that lack one, which may differ systematically from the rest, is rarely reported. Fifth, explanations are presented as heat maps without tests of whether they reflect the evidence the model uses (Adebayo et al. 2018; Hooker et al. 2019; Rong et al. 2022; Ghassemi et al. 2021).

Late fusion, in which one expert per signal produces a score and a combiner merges the scores, is modality-agnostic by construction: the combiner can be defined on every subset of signals, experts can be retrained or replaced independently, and the contribution of each signal can be computed exactly, because the model's output without a signal is well defined. Intermediate-fusion networks need additional machinery for the same purpose, such as modality dropout during training (Neverova et al. 2016), architectures designed for missing modalities (Ma et al. 2021, 2022), or reading out a single-signal branch with the missing input set to zero (Zhang et al. 2024). Late fusion is also the yardstick for more elaborate architectures. In one recent study of ECG–PCG fusion, for example, a late-fusion baseline raised the area under the receiver operating characteristic curve (AUC) of its ECG component on training-a by 0.005, whereas a progressive intermediate-fusion network raised it by 0.022 (Zhang et al. 2024). Before such comparisons can be interpreted, the behaviour of late fusion itself has to be characterised: which combiner, whether calibration matters, and what limits the gain.

This study provides that characterisation. Its contributions are:

1. A modality-agnostic late-fusion framework: experts are trained in nested, subject-grouped cross-validation and calibrated per expert, and seven combination rules are each fitted once and applied to whichever signals a record has (Fig. 1). We evaluate it on every combination of signals, compared on the same records, and, as it would be deployed, on every record, with paired, cluster-aware statistics.

2. An evaluation on three studies whose labels are never pooled: abnormal heart sounds from ECG and PCG (CinC 2016), AF from ECG and PPG in an intensive-care cohort (MIMIC-III-Ext-PPG (Moulaeifard, Charlton, et al. 2026; Moulaeifard, Kutscher, et al. 2026)), and a synthetic tri-modal benchmark with planted ground truth.

3. Evidence that modality agnosticism matters in practice: quality control removed PPG preferentially during AF, so a model requiring both signals could not have assessed almost half of the AF segments; the modality-agnostic model assessed every segment and discriminated better than ECG alone.

4. A redundancy analysis that explains the size of the fusion gain, combining within-class expert correlation, binormal fusion ceilings and an empirical oracle linear combination. It shows that well-chosen combiners are already close to optimal, that correlated expert errors, not the combiner, bound the gain, and how much stronger experts could add.

5. Practical guidance on combination rules: per-expert calibration acts as implicit expert weighting, the poor performance of majority voting is largely an artefact of its few output levels, soft voting nearly matches the best rule, and either-positive or both-positive votes serve as screening or confirmation operating points.

6. A tested explanation layer: exact modality Shapley values and interactions at the decision level, and signal-level attributions checked for deletion faithfulness, dependence on learned parameters, beat-aligned physiology and, on synthetic data, localisation of the planted abnormality.

7. A comparison with published ECG–PCG and ECG–PPG fusion results, including methods re-implemented under one protocol.

2 Related work

2.1 Multimodal ECG–PCG analysis

Synchronised ECG and PCG provide the electrical and mechanical views of each cardiac cycle. Early work extracted hand-crafted features from both signals (Chakir et al. 2020). Deep models followed, mostly for coronary artery disease or the CinC 2016 abnormal-heart-sound label: dual-input networks combining features and raw signals (Li et al. 2019), multi-domain deep features with feature selection (H. Li et al. 2021), multi-modal features with classical learners (P. Li et al. 2021), transfer-learned scalogram networks fused at the classification stage (Hettiarachchi et al. 2021) and coupling signals derived by deconvolving PCG with ECG (Sun, C. Liu, et al. 2024; Sun, X. Liu, et al. 2024). Decision-level combination has also been explored, through Dempster–Shafer evidence theory (Li et al. 2022) and, recently, quality-aware weighting of calibrated gradient-boosting experts (Bargarai et al. 2026). Intermediate fusion now dominates: progressive dense fusion with co-learning (Zhang et al. 2024), progressive attention-based cross-modal fusion (Li et al. 2025) and multi-scale spectral–spatial networks (Reehana and Siddique Ibrahim 2025). Calzoni et al. combined a one-dimensional convolutional neural network (CNN) ECG branch with an audio-pretrained PCG branch by late fusion and quantified modality contributions with three explanation techniques (Calzoni et al. 2025). Simultaneous ECG–PCG recordings of healthy volunteers (Kazemnejad et al. 2021) have been used to separate

information common to both signals from information exclusive to each (Karimi et al. 2025), the quantity that ultimately limits what fusion can add. Large PCG-only corpora such as CirCor DigiScope (Oliveira et al. 2022) have no ECG.

2.2 PPG, ECG and atrial fibrillation

PPG-based AF detection exploits the irregular timing and varying amplitude of pulses (Pereira et al. 2020). Smartwatch studies established its feasibility at scale (Tison et al. 2018; Perez et al. 2019), multi-task networks jointly assessed signal quality and rhythm (Torres-Soto and Ashley 2020), and intensive-care data allowed training on millions of alarm-labelled segments (Ding et al. 2024). Combining ECG and PPG has been approached by stacking both signals as input channels (Aldughayfiq et al. 2023), by fusing signal-quality-weighted wavelet coefficients of ECG, PPG and arterial blood pressure (Malone et al. 2025), by convolutional fusion at one or several levels with quality indices as inputs (John 2025; John et al. 2025), and by learning a shared latent space so that one model can classify either signal (Guo et al. 2025). Foundation models pretrained on large PPG corpora offer generic embeddings (Pillai et al. 2025). The MIMIC-III-Ext-PPG benchmark (Moulaeifard, Charlton, et al. 2026; Moulaeifard, Kutscher, et al. 2026) provides millions of 30-s PPG segments with rhythm labels and companion ECG; PPG-only AF benchmarks on its predefined folds reached AUCs of 0.94–0.96 (Moulaeifard, Aston, et al. 2026). Signal-quality indices decide whether a PPG segment is used at all (Orphanidou et al. 2015; Elgendi 2016); whether those decisions depend on the rhythm, and so make missingness informative (Little and Rubin 2019; Groenwold 2020), has received little attention.

2.3 Combining classifiers and missing modalities

Fixed combination rules (sum, product, maximum, vote) have a decision-theoretic basis under conditional independence of the experts (Kittler et al. 1998); the product of calibrated experts corresponds to a product-of-experts model (Hinton 2002). Learned combiners include stacked generalisation (Wolpert 1992) and input-dependent gating, as in mixtures of experts (Jacobs et al. 1991). The benefit of any combiner depends on the diversity of its members (Kuncheva 2014). In diagnostic statistics, the best linear combination of correlated binormal markers has a closed form (Su and Liu 1993; Pepe 2003), which we use to bound the attainable gain. Multimodal networks trained end to end can be dominated by one modality (Wang et al. 2020) and degrade when a modality is missing, which motivated modality dropout (Neverova et al. 2016) and dedicated architectures for severely or partially missing modalities (Ma et al. 2021, 2022).

2.4 Explanations and their evaluation

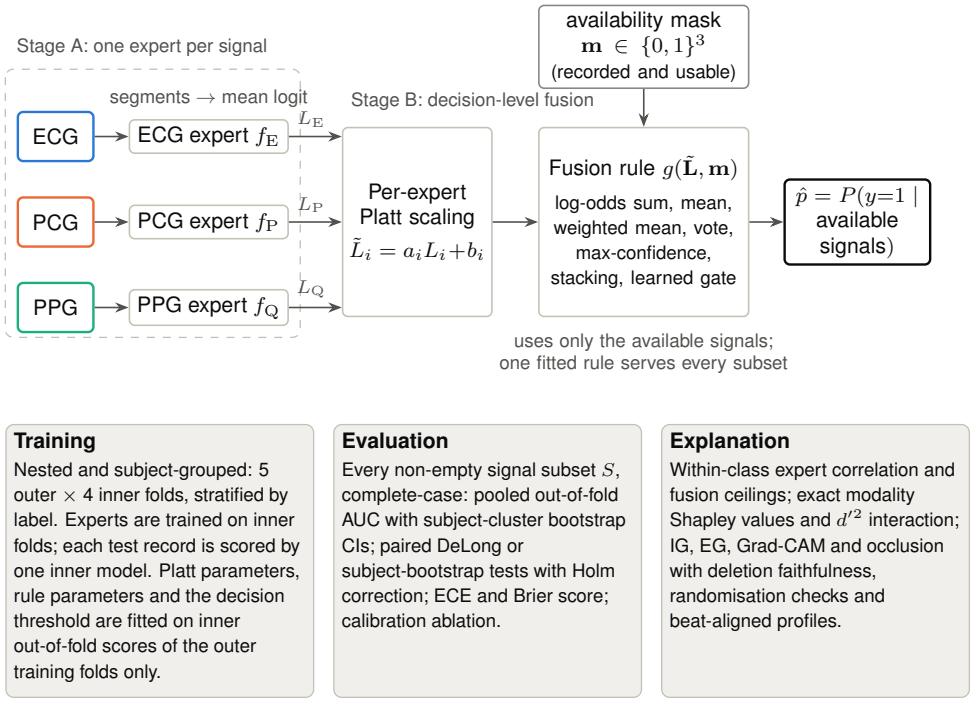
Shapley values allocate a model's output among its inputs (Shapley 1953; Lundberg and Lee 2017); at the level of modalities they quantify the contribution of each signal (Parcalabescu and Frank 2023), and the interaction index separates synergy from redundancy (Grabisch and Roubens 1999). Within a signal, gradient-based methods

(Integrated Gradients (Sundararajan et al. 2017), Expected Gradients (Erion et al. 2021), Grad-CAM (Selvaraju et al. 2017)) and perturbation methods (occlusion (Zeiler and Fergus 2014)) highlight samples. Whether such maps are faithful can be tested by deleting samples in order of relevance (Samek et al. 2017), ideally with retraining (Hooker et al. 2019) or with imputation that keeps perturbed inputs close to the data (Rong et al. 2022); whether they depend on what the model learned can be tested by randomising its parameters (Adebayo et al. 2018). Without such tests, explanations can mislead clinicians (Ghassemi et al. 2021).

3 Materials and methods

3.1 Problem formulation

Let $\mathcal{M} = \{\mathrm{E}, \mathrm{P}, \mathrm{Q}\}$ denote ECG, PCG and PPG. Record i has a binary label y_i, an availability mask $\boldsymbol{m}_i \in \{0,1\}^3$ (signal recorded and usable) and, for each available signal k, a set of fixed-length segments. An expert f_k maps a segment to a logit, and the record logit L_{ik} is the mean of its segment logits. A fusion rule g maps the calibrated logits $\tilde{\boldsymbol{L}}_i$ and the mask to a probability $\hat{p}_i = g(\tilde{\boldsymbol{L}}_i, \boldsymbol{m}_i)$ and depends only on the available entries, so one fitted rule serves every non-empty subset $S \subseteq \mathcal{M}$. A subset is evaluated on the records that have every signal in S, using only those signals (complete-case evaluation); because these record sets differ between subsets, subsets are compared with one another on the records that have all signals (Section 3.7). Scoring each record with the subset that matches its own available signals gives the modality-agnostic prediction a deployed model would return; Section 3.8 evaluates it on every record.



The late-fusion framework. Stage A trains one expert per signal; Stage B calibrates each expert and combines the available experts with a fusion rule that is defined on every subset of signals. Training, evaluation and explanation follow the nested, subject-grouped protocol summarised in the lower panels.

3.2 Datasets

Three studies were analysed with one pipeline and never pooled, because their positive classes mean different things (Table 1).

CinC 2016 training-a (ECG + PCG). The training-a folder of the PhysioNet/CinC 2016 Challenge (Clifford et al. 2016; Liu et al. 2016) contains 409 synchronised PCG and ECG recordings from 121 patients, sampled at 2 kHz, labelled normal or abnormal heart sound by the organisers. The Challenge's validation folder duplicates 80 training-a recordings; because one sensitivity analysis uses a PCG encoder whose checkpoint was selected on that folder, the primary analysis excludes the 80 duplicates and uses the remaining 329 recordings (252 abnormal, 76.6%). Three recordings have no ECG channel (ECG treated as missing). Sparse invalid samples (the invalid-sample code of the WaveForm DataBase (WFDB) format, typically clipping; at most 0.86% of a channel, runs of at most 2.5 ms) were restored by linear interpolation in 14 ECG and 19 PCG channels, and two PCG channels with more than 1% invalid samples were treated as missing, leaving ECG in 326, PCG in 327 and both in 324 recordings. Subject identifiers are not released, so cross-validation groups by recording; Section 5.7 discusses the consequences. A sensitivity analysis uses all 409 recordings.

MIMIC-III-Ext-PPG (ECG + PPG). MIMIC-III-Ext-PPG (version 1.1.0) (Moulaeifard, Charlton, et al. 2026; Moulaeifard, Kutscher, et al. 2026) provides non-overlapping 30-s PPG segments at 125 Hz from intensive-care patients of the MIMIC-III waveform database matched subset (Johnson et al. 2016; Moody et al. 2020), with companion ECG lead II, rhythm labels from charted rhythm annotations and per-window quality codes. We used a sample of the files, kept at most eight segments per patient and label, and defined AF or atrial flutter as positive and sinus rhythm (SR), sinus tachycardia (STACH) or sinus bradycardia (SBRAD) as negative; tachycardia and bradycardia were included so that heart rate alone could not separate the classes. Supraventricular tachycardia (408 segments) was excluded, and no atrial-flutter segment was present in the sample. The cohort comprised 40,792 segments from 5,642 patients: AF 7,464 segments (1,021 patients), SR 14,386 (4,863), STACH 11,940 (2,625) and SBRAD 7,002 (1,416); 597 patients contributed segments of both classes. Every segment had a usable ECG. PPG was treated as missing when it failed the dataset's quality flags, which happened in 7,254 segments, leaving 33,538 segments (5,523 patients) with both signals. Seventy-seven files were skipped because they were truncated, unreadable or lacked a data file. Folds were grouped by patient.

Synthetic tri-modal benchmark. No public dataset provides ECG, PCG and PPG with a disease label, so a synthetic benchmark tests the machinery and supplies ground truth for the explanations. Each of 250 subjects contributes two 9-s records generated from a shared beat train (60–90 beats per minute, 3% beat-to-beat jitter): an ECG with P, QRS and T waves, a PCG with first and second heart-sound (S1, S2) tone bursts, and a PPG with systolic and dicrotic waves, each with baseline wander and noise. A positive subject (target prevalence 0.4) with severity $s \sim \mathcal{U}(0.5,1.5)$ manifests its abnormality independently in each signal: T-wave flattening or inversion in the ECG (probability 0.7; window 0.20–0.42 s after the R peak), a systolic murmur in the PCG (0.7; 0.10–0.32 s) and a damped dicrotic wave in the PPG (0.5; 0.42–0.62 s). Because manifestation is independent, some positives are visible to only one signal and fusion has genuinely complementary information to exploit. PPG was removed at random from 20% of records, leaving 404 of 500 records with all three signals; the realised prevalence was 37.6%.

The three studies. Records are recordings (CinC 2016) or 30-s segments (MIMIC-III-Ext-PPG).

	CinC 2016	MIMIC-III-Ext-PPG	Synthetic
Signals	ECG, PCG	ECG, PPG	ECG, PCG, PPG
Positive class	abnormal heart sound	atrial fibrillation	planted abnormality
Records	329	40,792	500
Subjects	121[a]	5,642	250
Prevalence (%)	76.6	18.3	37.6
With all signals	324	33,538	404
Grouping of folds	recording[a]	patient	subject

3.3 Preprocessing

Every signal was band-pass filtered (zero-phase fourth-order Butterworth), resampled by polyphase filtering, cut into non-overlapping windows and z-scored per window. CinC 2016: ECG 0.5–40 Hz, 250 Hz, 5-s windows; PCG 25–400 Hz, 1 kHz, 5-s windows. MIMIC: ECG 0.5–40 Hz resampled to 100 Hz in 10-s windows (the input format of the pretrained ECG encoder; three per segment); PPG 0.5–8 Hz, 125 Hz, 10-s windows (three per segment, one per quality flag). Synthetic: ECG 0.5–40 Hz, 100 Hz, 3-s windows; PCG 10–90 Hz, 200 Hz, 3-s windows; PPG 0.3–8 Hz, 25 Hz, 4.5-s windows.

3.4 Signal experts

From-scratch experts. Unless stated otherwise, each expert is a one-dimensional convolutional network of four blocks (convolution with kernel 7, batch normalisation, ReLU, max-pooling by 4) with 16, 32, 64 and 64 channels, global average pooling, dropout 0.3 and a linear output (47,121 parameters; three blocks of 8, 16 and 16 channels, 2,841 parameters, for the synthetic data). Segments inherit the record label. Training used class-balanced binary cross-entropy (positive weight $n_{\mathrm{neg}}/n_{\mathrm{pos}}$), Adam (Kingma and Ba 2015) (learning rate 10^{-3}, weight decay 10^{-4}, batch 64) for at most 40 epochs, with early stopping (patience 8) on the record-level AUC of a subject-grouped 15% early-stopping slice. Each model's logits were standardised by their mean and standard deviation on its early-stopping slice, so that the four inner models of an outer fold produce comparable scores.

Pretrained experts. The MIMIC ECG expert is an InceptionTime-style network (Ismail Fawaz et al. 2020; Szegedy et al. 2015) pretrained on 12-lead ECGs of the PhysioNet/CinC 2021 Challenge (Reyna et al. 2021) and fine-tuned inside the cross-validation. Its input layer was rebuilt for one lead by summing the 12 pretrained lead filters, which reproduces the pretrained network for a lead copied into all inputs; batch-normalisation statistics were re-estimated on each fold's training segments (AdaBN (Li et al. 2018)); the input layer alone was trained for 10 epochs before all 29,793 parameters were unfrozen. Two sensitivity analyses replace experts: on CinC 2016, the same fine-tuned ECG encoder and a frozen PCG ResNet (He et al. 2016) (366,082 parameters) trained on CinC 2016 training-b to training-f, whose checkpoint was selected on the Challenge validation folder (hence the exclusion of the 80 duplicated recordings from the primary cohort); on MIMIC, a logistic-regression head (standardised inputs, balanced classes, $C = 1$) fitted per inner fold on frozen embeddings of the PaPaGei-S foundation model (Pillai et al. 2025) (5.8 million parameters; 10-s windows at 125 Hz). PaPaGei was pretrained on corpora that include MIMIC-III waveforms, so this variant is not independent of its test data.

3.5 Nested, subject-grouped cross-validation

Five outer folds were formed, stratified by label and grouped by subject. Within each outer training set, four inner folds were formed in the same way. Nesting keeps every data-dependent choice (early stopping, calibration, fusion parameters and thresholds) inside the outer training set, which avoids the optimistic bias of selecting models on the data used to evaluate them (Varma and Simon 2006). Each inner model was trained on three inner folds minus its early-stopping slice and scored (i) its held-out inner fold, giving inner out-of-fold scores (the *fit* rows) for every outer-training record, and (ii) the outer test fold. Each test subject was scored by exactly one inner model, assigned by a balanced seeded permutation, so that fit and test scores come from the same procedure and every calibration map, rule parameter and threshold learned on fit rows transfers to test rows; averaging the inner models would shrink the test logits relative to the fit logits. Disjointness of subjects between training, early-stopping, fit and test sets was asserted in code. All results are pooled over the five outer test folds, so every record is a test case exactly once (seed 42).

3.6 Calibration and fusion rules

In each outer fold, each expert was calibrated by Platt scaling (Platt 1999), $\tilde{L}_k = a_k L_k + b_k$, fitted on that expert's available fit rows (logistic regression, $C = 10^4$). Let $A_i = \{k : m_{ik} = 1\}$, σ the logistic function and $\lambda = \mathrm{logit}\bar{y}$ the log-odds of the fit-set prevalence. Seven rules were compared:

$$\text{Mean:} \quad \hat{p} = 1/|A| \sum_{k \in A} \sigma\,(\tilde{L}_k),$$

$$\text{Weighted mean:} \quad \hat{p} = \sum_{k \in A} w_k\, \sigma(\tilde{L}_k) / \sum_{k \in A} w_k\,, \quad w_k = \max(\mathrm{AUC}_k - 0.5,\, 10^{-3}),$$

$$\text{Log-odds sum:} \quad \mathrm{logit}\hat{p} = \lambda + \sum_{k \in A} (\,\tilde{L}_k - \lambda),$$

$$\text{Majority vote:} \quad \hat{p} = 1/|A| \sum_{k \in A} \mathbb{1}\,[\sigma(\tilde{L}_k) \geq t_k],$$

$$\text{Max-confidence:} \quad \hat{p} = \sigma(\tilde{L}_{k^*}), \quad k^* = \underset{k \in A}{\mathrm{argmax}} \left|\sigma(\tilde{L}_k) - 0.5\right|,$$

$$\text{Stacking:} \quad \mathrm{logit}\hat{p} = \beta_0 + \sum_{k} m_k\,(\beta_k \tilde{L}_k + \gamma_k),$$

$$\text{Learned gate:} \quad \mathrm{logit}\hat{p} = e^s \sum_{k \in A} \omega_k\, \tilde{L}_k + c, \quad \omega = \mathrm{softmax}_{k \in A}\left(\alpha_k + \beta_k |\tilde{L}_k|\right).$$

The weights w_k use each expert's fit-set AUC, and t_k is each expert's Youden threshold on the fit rows (Youden 1950). The log-odds sum [eq:logodds] is the Bayes posterior when the calibrated experts are conditionally independent given the label, that is, a product of experts (Kittler et al. 1998; Hinton 2002). Stacking (Wolpert 1992) is a logistic regression ($C = 1$) whose mask terms give each availability pattern its own offset. The gate, in the

spirit of mixtures of experts (Jacobs et al. 1991), has $2d + 2$ parameters, fitted with Adam (600 steps, learning rate 0.05, L_2 penalty 10^{-3} on α, β). Both learned rules were fitted on subset-augmented fit rows, in which every row is replicated once per non-empty subset of its available signals, so they learn every availability pattern. Each rule's decision threshold on each subset was the Youden threshold on fit rows. The mean rule is soft voting and the majority vote is hard voting; with two experts the vote fraction takes only the values 0, 0.5 and 1, so choosing its threshold by Youden's index amounts to declaring a record positive either when any expert is positive or only when both are. In a calibration ablation the whole of Stage B was repeated without Platt scaling.

3.7 Evaluation and statistics

Discrimination was measured by the pooled out-of-fold AUC, with 95% percentile confidence intervals (CIs) from 2,000 bootstrap resamples of subjects (Efron and Tibshirani 1993; Field and Welsh 2007), and summarised per outer fold. Sensitivity, specificity and balanced accuracy were computed at the fit-row threshold; calibration by the Brier score (Brier 1950) and the expected calibration error (ECE, ten equal-width bins) (Naeini et al. 2015; Guo et al. 2017). Each fused rule on the full signal set was compared with each single expert on the same records. To compare combinations of signals with one another, every combination was also evaluated on the records that have all of a study's signals, so that combinations differ only in the signals they use; for these comparisons we also report accuracy and the F1 score, and with three signals we compared the full combination with each pair. When each record is its own grouping unit (CinC 2016, which does not release subject identifiers), DeLong's test (DeLong et al. 1988; Sun and Xu 2014) gave the p-value and the 95% CI of the difference. When subjects contribute several records (MIMIC, synthetic), DeLong's independence assumption fails, so a paired bootstrap over subjects (2,000 resamples) gave a percentile CI and a two-sided p-value, whose resolution is about 0.001. P-values were Holm-adjusted (Holm 1979) over the comparisons within each study (14 for two signals, 21 for three; the comparisons between combinations of signals formed a separate family). The study follows the relevant items of the TRIPOD+AI reporting guideline (Collins et al. 2024).

3.8 Deployment, voting variants and decision rules

Modality-agnostic deployment. Each test record was scored by the same fitted rule restricted to the signals it had, and these predictions were evaluated on every record. Coverage was the share of records, and of positive records, that a model could score; a complete-case fusion model, which needs every signal, can score only the records that have all of them. On every record with an ECG, the modality-agnostic model was compared with the ECG expert by the paired tests described below.

Voting variants. On the records with all signals, hard voting was compared with three variants that keep the experts' confidence: hard voting with ties broken by the mean calibrated probability (records ranked first by their number of positive votes, then by the mean probability), soft voting (the mean rule) and weighted soft voting (the weighted mean).

Decision rules. At the experts' own Youden thresholds, a record was declared positive when any expert was positive (OR), when all were (AND) or, with three experts, when at least two were; these rules were compared with the single experts, the soft vote and the log-odds sum by sensitivity, specificity and positive predictive value.

Robustness to missing signals. On the records with every signal of a study, each non-ECG signal was removed independently of the label and of the other records with probability m (missing completely at random (Little and Rubin 2019)), the ECG being always kept, and each record was scored by the same fitted rule from the signals it kept. With K signals, a record keeps subset s, the ECG and k_s of the other $K-1$ signals, with probability $\pi_s(m) = (1-m)^{k_s} m^{K-1-k_s}$. Let A_{st} be the probability that a positive record scored from subset s outranks a negative record scored from subset t, ties counting one half, so that A_{ss} is the AUC of subset s. Because removals are independent between records, the expected pooled AUC over all possible removals is exactly

$$\mathrm{E}[\mathrm{AUC}(m)] = \sum_s \sum_t \pi_s(m)\, \pi_t(m)\, A_{st},$$

and no Monte Carlo sampling of removals is needed; for CIs, A was recomputed on every bootstrap resample. The cross terms A_{st}, $s \neq t$, are where the scores of different subsets must share one scale: if records scored from more signals receive systematically higher scores, positive records scored from fewer signals rank below negative records scored from more, and the pooled AUC can fall below the AUC of every subset. The pipeline was also run with calibration switched off, with every rule fitted on the raw logits, and each rule was deployed as the modality-agnostic model on every record with an ECG.

Decision curves. The net benefit of flagging the records whose predicted probability is at least the threshold probability p_t is $\mathrm{NB}(p_t) = \mathrm{TP}/n - (\mathrm{FP}/n)\, p_t/(1-p_t)$ (Vickers and Elkin 2006), where $p_t/(1-p_t)$ is the number of false positives a clinician would accept to find one more true positive; a difference in net benefit of δ is equivalent to $1000\,\delta$ more true positives per 1,000 records at the same number of false positives. Net benefit was computed on every record with an ECG for the ECG expert, the modality-agnostic model (log-odds rule), complete-case fusion, which can flag only the records that have every signal, and treating every record or no record as positive, at $p_t = 0.01, \ldots, 0.99$. ROC curves were computed on the records with every signal, with the operating points of the decision rules.

All statistics in this section used one set of 2,000 subject-level bootstrap resamples per study: percentile CIs for AUCs and rates; paired differences from the ECG expert by DeLong's test on CinC 2016 and by the paired subject bootstrap otherwise, with $p = 2\,(1 + \min(\#\{\Delta^* \leq 0\}, \#\{\Delta^* \geq 0\}))/(B+1)$ for B resamples Δ^*; and Holm adjustment within each study and family of comparisons.

3.9 Redundancy analysis

Within-class correlation. Fusion gains most when the experts' errors are independent given the class. For each pair of experts and each class we computed Spearman's ρ between the two experts' test scores after centring each score within its outer fold, so that fold-level offsets cannot create correlation, with a subject-cluster bootstrap CI (2,000 resamples). Positive ρ means that a record that looks more abnormal to one expert also tends to look more abnormal to the other within the same class, that is, redundant evidence.

Fusion ceilings. Under the equal-variance binormal model an expert with AUC a_k separates the classes by $d'_k = \sqrt{2}\,\Phi^{-1}(a_k)$. If the class-conditional correlation matrices of the experts' scores are R_0 and R_1, the best linear combination of the scores attains (Su and Liu 1993; Pepe 2003)

$$\mathrm{AUC}^{*} = \Phi\left(\sqrt{1/2\ \boldsymbol{d}'^{\top}\bar{R}^{-1}\boldsymbol{d}'}\right), \qquad \bar{R} = 1/2\,(R_0 + R_1).$$

With $\bar{R} = I$ this is the ceiling for conditionally independent experts; with $\bar{R}$ estimated from the within-class Spearman correlations (converted to Pearson-equivalent correlations $2\sin(\pi\rho/6)$), it is a correlation-aware ceiling. Because real scores are not binormal, we also computed an assumption-free reference: the in-sample AUC of the best linear combination of the two calibrated scores on the records with both signals, found by a search over directions. This oracle is optimistic by construction, as its weights are chosen on the evaluation data.

Projection. To ask what a stronger second expert could add, Eq. [eq:ceiling] was evaluated for our ECG expert and a second expert of any AUC at the measured class-averaged correlation, including the AUCs reported for PCG (Zhang et al. 2024; Li et al. 2025) and PPG (Moulaeifard, Aston, et al. 2026) experts elsewhere. The projection assumes that a stronger expert's errors would be as correlated with the ECG expert's as ours were.

Interactions. With modality Shapley values (below), the pairwise interaction index (Grabisch and Roubens 1999) was computed on the d'^{2} scale, on which conditionally independent binormal evidence is additive, so that a negative interaction indicates redundancy and a positive one synergy. On the AUC scale, interactions are negative even for independent experts, because the AUC saturates.

3.10 Explanations

Decision level. Because the fused model is defined on every subset of signals, exact Shapley values (Shapley 1953) over modalities require no baseline imputation and, with at most three signals, no sampling. For a cohort, the value of a coalition S is $v(S) = \mathrm{AUC}_S - 0.5$ on a fixed record set, and the Shapley values split $\mathrm{AUC}_{\mathcal{M}} - 0.5$ among the signals; CIs come from a subject bootstrap. For a record, $v(S)$ is the fused probability with only the signals in S and $v(\emptyset)$ the fit-set prevalence. To see what fusion does when experts disagree,

test records with all signals were grouped by which experts classified them correctly at their own thresholds, and the accuracy of the fused decision was computed per group.

Signal level. Each test segment was explained with the inner model that scored it, by four methods: Integrated Gradients (zero baseline, 32 steps) (Sundararajan et al. 2017), Expected Gradients (32 references drawn from the model's training segments) (Erion et al. 2021), one-dimensional Grad-CAM at the last ReLU (Selvaraju et al. 2017) and occlusion with a 0.2-s window replaced by a linear bridge (half-window stride) (Zeiler and Fergus 2014). Faithfulness was tested by deletion (Samek et al. 2017; Rong et al. 2022): for 200 test segments per expert, the fraction $q \in \{0, 0.05, \ldots, 0.5\}$ of samples ranked most relevant (MoRF) or least relevant (LeRF) was removed and replaced by noisy linear imputation, as in the RemOve And Debias (ROAD) benchmark (Rong et al. 2022), and the faithfulness gap is the mean over q of the expert logit after LeRF minus that after MoRF removal (logit units; larger is better, about zero is no better than chance). Dependence on learned parameters was tested by model randomisation (Adebayo et al. 2018): the Spearman correlation between the |attribution| of the trained expert and that of the same network with random weights, averaged over 60 segments (lower is better). Beat-aligned profiles placed every beat on a common time axis: R peaks were detected on the ECG (squared derivative energy, 0.3-s refractory period, refinement to the largest $|x|$ within 60 ms), each segment's attribution was normalised to a mean |attribution| of 1, and |attribution| was interpolated onto a 5-ms grid from -0.25 to 0.70 s around each R peak and averaged per class; signals recorded in synchrony with the ECG (PCG in CinC 2016, PPG in MIMIC) were aligned to the same R peaks. On synthetic data, localisation was scored as enrichment: the attribution density inside the planted window relative to the whole cycle, in abnormal records divided by the same ratio in normal records (1 means no enrichment).

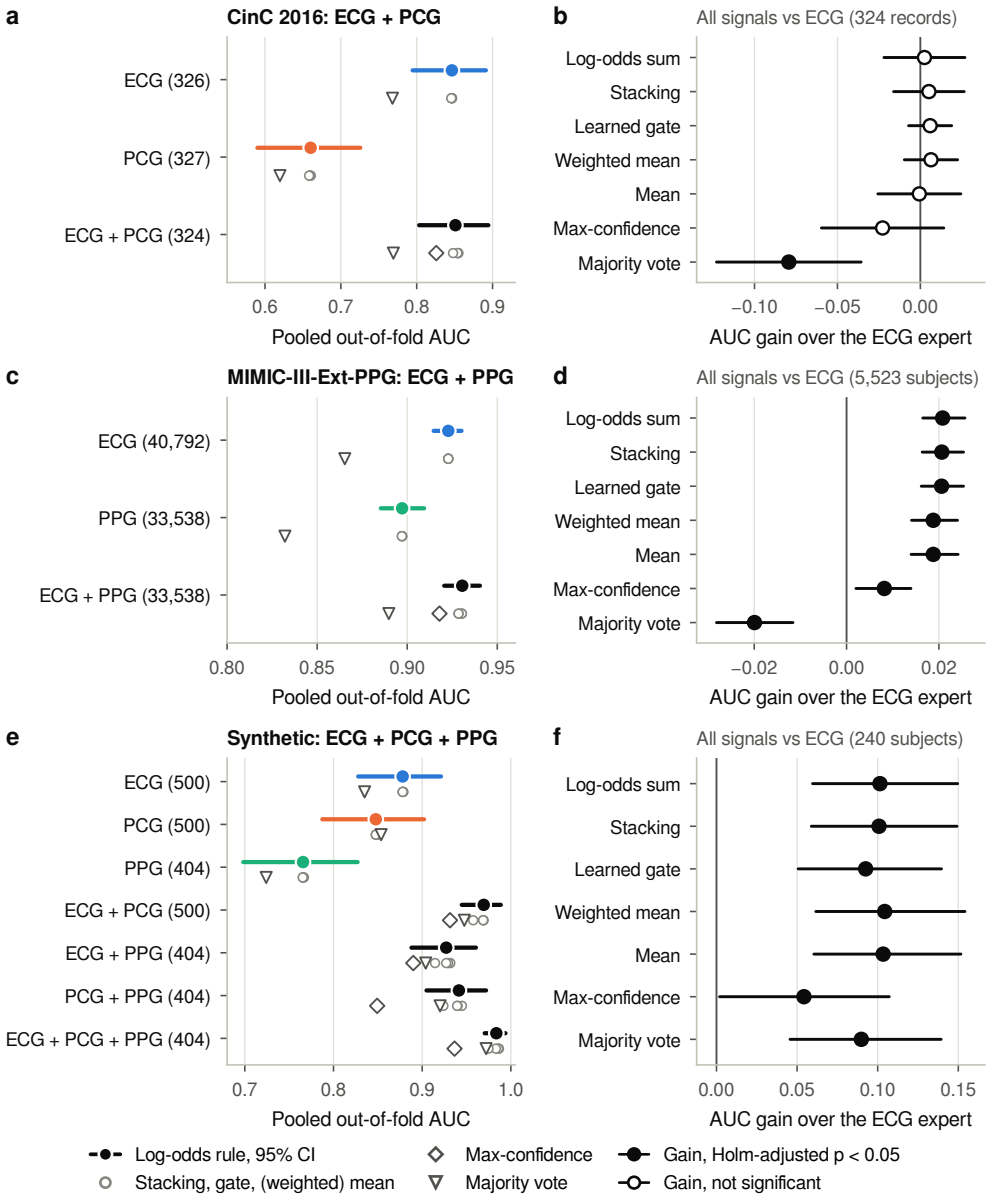
3.11 Implementation

The pipeline was implemented in Python 3.12 with PyTorch 2.10 (Paszke et al. 2019) and scikit-learn (Pedregosa et al. 2011) and run on one NVIDIA Tesla T4 graphics processing unit. On MIMIC, training the experts took 95 min, fusion and statistics 30 min and the 222,990 segment explanations 34 min; on CinC 2016 all stages, including the calibration ablation, took under 6 min. The redundancy analyses, figures and tables were computed from the pipeline outputs by scripts provided with this article. An AI assistant, Claude Opus 5.5 (Anthropic; model identifier claude-opus-5-5, accessed through the Claude application in October 2026), was used to help develop the analysis and figure scripts, to search the literature for published results and to check references. Representative prompts asked it to implement specified analyses and figures, to find published single- and dual-signal results on the same datasets and to check bibliographic details. Every number, table and figure was verified by re-running the scripts on the pipeline outputs, every reference and published value was checked against a publisher, index or dataset page, and the authors take full responsibility for the content.

4 Results

4.1 Discrimination of single experts and fused models

Table [tab:main] and Fig. 2 summarise the clinical studies. On CinC 2016, the ECG expert reached an AUC of 0.846 (95% CI 0.795–0.891) and the PCG expert 0.660 (95% CI 0.590–0.726). On MIMIC, the ECG expert reached 0.923 (95% CI 0.915–0.930) on all segments and the PPG expert 0.897 (95% CI 0.885–0.909). Per-fold AUCs varied more on CinC 2016 (ECG 0.782–0.929, standard deviation (SD) 0.057) than on MIMIC (ECG 0.913–0.930, SD 0.008; fused 0.921–0.942).



Discrimination on every subset of signals and the gain from fusion, by study (rows). ***a,c,e*** *Pooled out-of-fold AUC of the log-odds rule with 95% subject-cluster bootstrap CI (filled marker and line), and the AUCs of the other rules on the same subset (open markers on the row beneath); numbers in parentheses are the records evaluated.* ***b,d,f*** *Paired AUC difference between the fused model on all signals and the ECG expert on the same records, for each rule (95% CI; DeLong for CinC 2016, subject-level bootstrap otherwise). Filled markers: Holm-adjusted* $p < 0.05$.

ECG + PPG. On the 33,538 segments with both signals, where the ECG expert's AUC was 0.910, every rule except majority voting improved on it (Fig. 2d). The log-odds rule reached 0.931 (95% CI 0.921–0.940), a gain of +0.021 (95% CI 0.017–0.026; Holm-adjusted $p = 0.014$, the bootstrap resolution) over ECG and of +0.033 (0.025–0.042) over PPG; stacking

(0.930) and the learned gate (0.930) were indistinguishable from it. Max-confidence added only +0.008 (0.002–0.014), and majority voting lost −0.020 (−0.028 to −0.012) relative to ECG. At the fit-row threshold, fusion raised specificity from 0.895 (ECG) to 0.923 at a sensitivity of 0.808.

ECG + PCG. On CinC 2016, no rule improved on the ECG expert (Fig. 2b). For the five rules other than majority voting and max-confidence, differences from ECG ranged from −0.001 (mean) to +0.006 (weighted mean), all with Holm-adjusted $p = 1.00$; the log-odds rule gave +0.002 (−0.022 to 0.027) and max-confidence −0.023 (−0.060 to 0.014). Majority voting was significantly worse than ECG alone (−0.079, −0.123 to −0.036; $p = 0.003$); Section 4.3 shows why. Every rule except voting improved substantially on the PCG expert (log-odds +0.194, 0.134–0.254). The same pattern held on all 409 recordings and with pretrained encoders (Appendix 7, Table 5).

All three signals. On the synthetic benchmark (Table 2), where every signal carries independent planted evidence, fusion raised the AUC from 0.878 (ECG), 0.848 (PCG) and 0.766 (PPG) alone to 0.984 (95% CI 0.970–0.994) for all three with the log-odds rule, +0.101 (0.060–0.150) above ECG on the same 404 records (240 subjects; $p = 0.021$). Here max-confidence was the weakest rule (0.936), because choosing one expert per record wastes the evidence of the others.

Every combination of signals. Table [tab:combinations] compares all combinations of signals on the records that have every signal of a study, so that they differ only in the signals used. On the 404 synthetic records with all three signals, both pairs that contained ECG improved on ECG alone (ECG + PCG +0.094, 0.050–0.143; ECG + PPG +0.045, 0.017–0.074), and all three signals gave the highest AUC (0.984), accuracy (0.946) and F1 score (0.927) of any combination. The third signal added most to the weaker pairs: all three signals improved on ECG + PPG by +0.057 (0.024–0.093) and on PCG + PPG by +0.042 (0.015–0.075), but on ECG + PCG, which already reached 0.976, by only +0.008 (−0.003 to 0.021; Holm-adjusted $p = 0.33$). Because records without PPG were scored from ECG and PCG by the same model, the modality-agnostic model reached 0.977 on all 500 synthetic records. In the clinical studies, where only pairs were recorded, the second signal alone was less discriminative than ECG on the same records (PPG −0.013, −0.024 to −0.001; PCG −0.191, −0.271 to −0.112).

@llcccccc@ Signals & Model & AUC (95% CI) & ΔAUC vs ECG (95% CI) & p & Sens & Spec & ECE

ECG & Expert alone (n=326) & 0.846 (0.795–0.891) & – & – & 0.731 & 0.805 & 0.048
PCG & Expert alone (n=327) & 0.660 (0.590–0.726) & – & – & 0.668 & 0.571 & 0.020
ECG + PCG & Log-odds sum & 0.851 (0.803–0.895) & +0.002 (−0.022 to 0.027) & 1.00 & 0.741 & 0.831 & 0.035
& Stacking & 0.854 (0.808–0.897) & +0.005 (−0.016 to 0.026) & 1.00 & 0.717 & 0.831 & 0.043
& Learned gate & 0.854 (0.805–0.897) & +0.006 (−0.007 to 0.019) & 1.00 & 0.761 & 0.792 &

0.054
& Weighted mean & 0.855 (0.808–0.897) & +0.006 (−0.010 to 0.022) & 1.00 & 0.713 & 0.844 & 0.077
& Mean & 0.848 (0.801–0.893) & −0.001 (−0.026 to 0.024) & 1.00 & 0.704 & 0.857 & 0.092
& Max-confidence & 0.826 (0.774–0.873) & −0.023 (−0.060 to 0.014) & 1.00 & 0.684 & 0.779 & 0.081
& Majority vote & 0.769 (0.715–0.823) & −0.079 (−0.123 to −0.036) & 0.003 & 0.607 & 0.831 & 0.205

ECG & Expert alone (n=40,792) & 0.923 (0.915–0.930) & – & – & 0.836 & 0.895 & 0.017
PPG & Expert alone (n=33,538) & 0.897 (0.885–0.909) & – & – & 0.742 & 0.922 & 0.011
ECG + PPG & Log-odds sum & 0.931 (0.921–0.940) & +0.021 (0.017–0.026) & 0.014 & 0.808 & 0.923 & 0.030
& Stacking & 0.930 (0.920–0.940) & +0.021 (0.016–0.025) & 0.014 & 0.810 & 0.922 & 0.021
& Learned gate & 0.930 (0.920–0.940) & +0.021 (0.016–0.025) & 0.014 & 0.806 & 0.923 & 0.016
& Weighted mean & 0.929 (0.919–0.938) & +0.019 (0.014–0.024) & 0.014 & 0.801 & 0.924 & 0.033
& Mean & 0.929 (0.919–0.938) & +0.019 (0.014–0.024) & 0.014 & 0.803 & 0.923 & 0.033
& Max-confidence & 0.918 (0.907–0.928) & +0.008 (0.002–0.014) & 0.016 & 0.782 & 0.918 & 0.025
& Majority vote & 0.890 (0.877–0.903) & −0.020 (−0.028 to −0.012) & 0.014 & 0.845 & 0.855 & 0.083

Synthetic tri-modal benchmark: pooled out-of-fold AUC of every signal subset under every rule (500 records from 250 subjects; PPG present in 404). The log-odds column gives the 95% subject-cluster bootstrap CI. Single-signal rows of the mean, weighted, log-odds and max-confidence rules are the calibrated expert itself.

Signals (n)	Log-odds (95% CI)	Stacking	Gate	Weighted	Mean	Max-conf.	Vote
ECG (500)	0.878 (0.828–0.921)	0.878	0.879	0.878	0.878	0.878	0.835
PCG (500)	0.848 (0.787–0.902)	0.848	0.849	0.848	0.848	0.848	0.854
PPG (404)	0.766 (0.698–0.827)	0.765	0.767	0.766	0.766	0.766	0.724
ECG + PCG (500)	0.969 (0.945–0.989)	0.969	0.958	0.969	0.969	0.931	0.948
ECG + PPG (404)	0.927 (0.888–0.961)	0.927	0.914	0.929	0.932	0.890	0.904
PCG + PPG (404)	0.941 (0.905–	0.940	0.924	0.939	0.945	0.849	0.920

Signals (n)	Log-odds (95% CI)	Stacking	Gate	Weighted	Mean	Max-conf.	Vote
	0.972)						
ECG + PCG + PPG (404)	0.984 (0.970–0.994)	0.983	0.975	0.987	0.986	0.936	0.972

@lcccccc@ Signals & AUC (95% CI) & Acc & Sens & Spec & F1 & ΔAUC (95% CI)

ECG & 0.849 (0.798–0.892) & 0.750 & 0.733 & 0.805 & 0.817 & –
PCG & 0.657 (0.587–0.722) & 0.642 & 0.664 & 0.571 & 0.739 & −0.191 (−0.271 to −0.112)*
ECG + PCG & 0.851 (0.803–0.895) & 0.762 & 0.741 & 0.831 & 0.826 & +0.002 (−0.022 to 0.027)
Any available (n=326) & 0.849 (0.801–0.892) & 0.761 & 0.739 & 0.831 & 0.825 & +0.003 (−0.022 to 0.027)

ECG & 0.910 (0.898–0.921) & 0.894 & 0.790 & 0.909 & 0.650 & –
PPG & 0.897 (0.885–0.909) & 0.900 & 0.742 & 0.922 & 0.649 & −0.013 (−0.024 to −0.001)*
ECG + PPG & 0.931 (0.921–0.940) & 0.908 & 0.808 & 0.923 & 0.688 & +0.021 (0.017–0.026)*
Any available (n=40,792) & 0.938 (0.931–0.944) & 0.896 & 0.846 & 0.907 & 0.749 & +0.015 (0.012–0.018)*

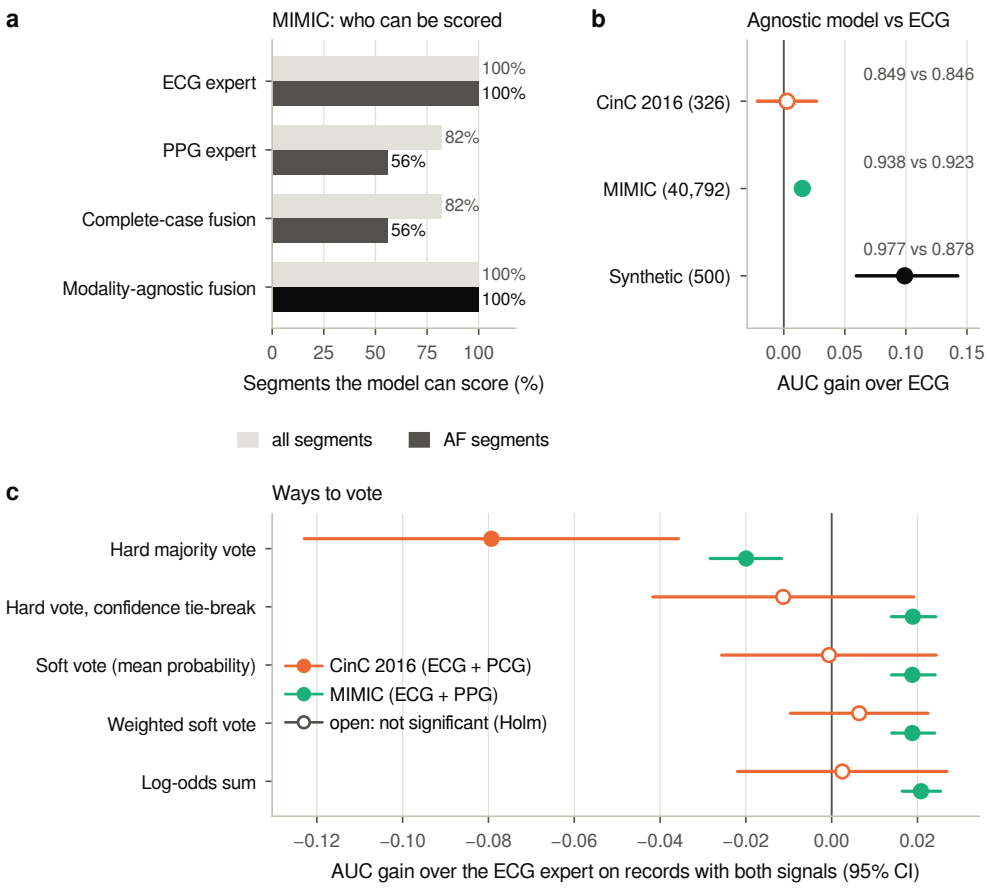
ECG & 0.882 (0.833–0.926) & 0.876 & 0.711 & 0.976 & 0.812 & –
PCG & 0.842 (0.777–0.898) & 0.881 & 0.717 & 0.980 & 0.820 & −0.040 (−0.128 to 0.046)
PPG & 0.766 (0.698–0.827) & 0.770 & 0.539 & 0.909 & 0.638 & −0.117 (−0.195 to −0.044)*
ECG + PCG & 0.976 (0.957–0.990) & 0.943 & 0.888 & 0.976 & 0.922 & +0.094 (0.050–0.143)*
ECG + PPG & 0.927 (0.888–0.961) & 0.874 & 0.783 & 0.929 & 0.824 & +0.045 (0.017–0.074)*
PCG + PPG & 0.941 (0.905–0.972) & 0.903 & 0.842 & 0.940 & 0.868 & +0.059 (−0.001 to 0.123)
ECG + PCG + PPG & 0.984 (0.970–0.994) & 0.946 & 0.914 & 0.964 & 0.927 & +0.101 (0.060–0.150)*
vs ECG + PCG & & & & & & +0.008 (−0.003 to 0.021)
vs ECG + PPG & & & & & & +0.057 (0.024–0.093)*
vs PCG + PPG & & & & & & +0.042 (0.015–0.075)*
Any available (n=500) & 0.977 (0.958–0.991) & 0.938 & 0.904 & 0.958 & 0.916 & +0.099 (0.059–0.142)*

4.2 One model for every combination of signals

Because one fitted rule serves every subset, the modality-agnostic model scored every record in all three studies (Fig. 3a). On MIMIC, a model requiring both signals could have scored only 82.2% of the segments and 56.1% of the AF segments, because PPG was missing far more often in AF (Section 4.7): it failed the quality flags in 43.9% of AF segments

but in 11.9% of non-AF segments (3.7-fold), so that 45.1% of the segments without PPG were AF, against 12.5% of those with both signals. Missingness was therefore informative, and the 3,273 AF segments without usable PPG are exactly those a complete-case model cannot assess. The modality-agnostic model scored them with the calibrated ECG expert, whose AUC was similar with and without a usable PPG (0.906 on the 7,254 segments without PPG and 0.910 on those with it), and fused both signals wherever both were present. On CinC 2016 and the synthetic data, complete-case fusion would have scored 98.5% and 80.8% of the records.

On every record with an ECG, the modality-agnostic model improved on the ECG expert on MIMIC (AUC 0.938 against 0.923 on all 40,792 segments; difference +0.015, 95% CI 0.012–0.018; Holm-adjusted $p = 0.007$) and on the synthetic benchmark (+0.099, 0.059–0.142, all 500 records), and matched it on CinC 2016 (+0.003, −0.022 to 0.027; Fig. 3b). At the fit-row thresholds it raised both sensitivity (0.846 against 0.836) and specificity (0.907 against 0.895) on MIMIC. Stacking (+0.016) and the learned gate (+0.015) behaved like the log-odds rule, and hard voting was again worse than ECG alone (−0.026). Pooled AUCs over all segments exceed the AUCs within the two availability groups because AF prevalence differs between the groups; the comparison with the ECG expert is nevertheless fair, because both are evaluated on the same segments.



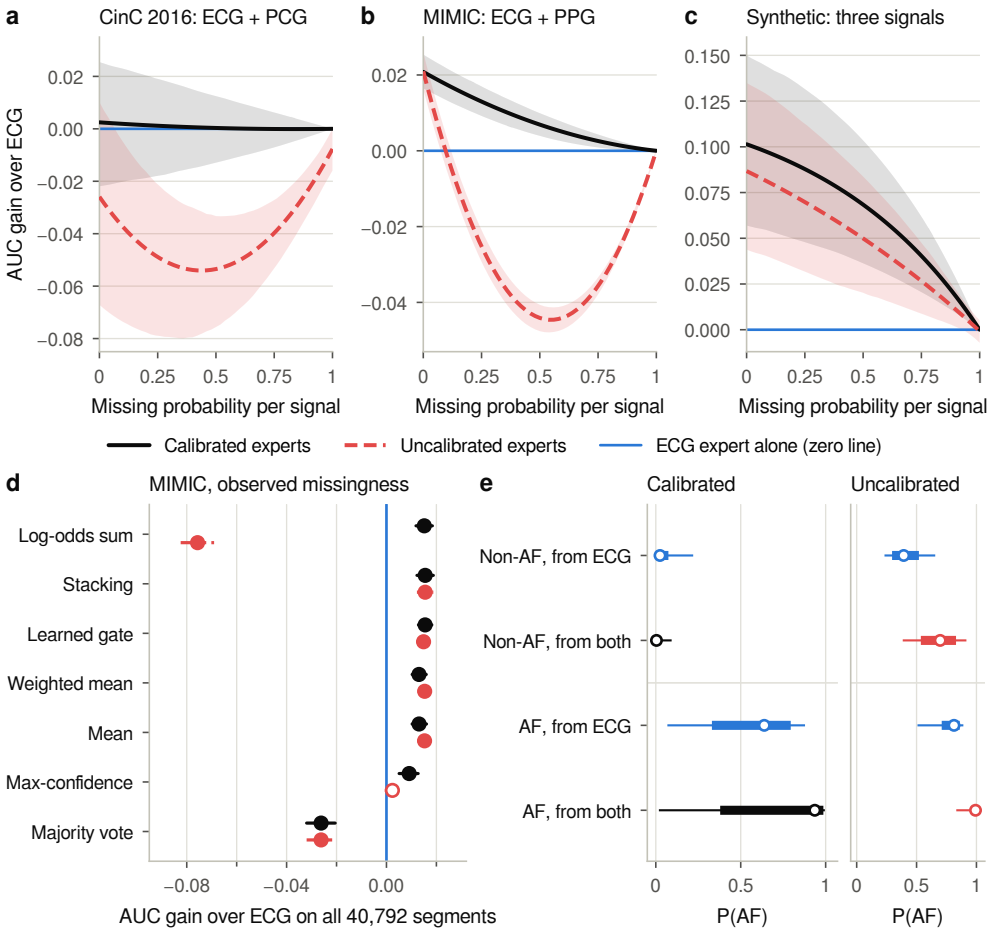
Deploying one modality-agnostic model, and ways to vote. ***a*** *Share of MIMIC segments, and of AF segments, that each model can score: a single expert needs its signal, complete-case fusion needs both, and the modality-agnostic model scores every segment with whichever signals it has.* ***b*** *AUC gain of the modality-agnostic model (log-odds rule) over the ECG expert on every record with an ECG (95% CI; DeLong for CinC 2016, subject-level bootstrap otherwise; number of records in parentheses; AUCs at right).* ***c*** *AUC gain over the ECG expert of hard voting and of voting variants that keep the experts' confidence, on the records with both signals. Filled markers: Holm-adjusted* $p < 0.05$.

For the log-odds sum, calibration is what allows one rule to serve every subset. When non-ECG signals were removed at random, the expected AUC of the calibrated sum declined smoothly from that of fusion to that of ECG alone and never dropped below ECG by more than 0.001 (Fig. 4a–c). Summed without calibration, the experts matched calibrated fusion on the MIMIC segments with both signals (AUC 0.931 against 0.931), but once segments scored from different subsets were pooled, the sum fell below ECG alone, by up to 0.045 at $m = 0.54$ (95% CI 0.041–0.048) on MIMIC and by up to 0.054 at $m = 0.44$ (0.032–0.078) on CinC 2016; on the synthetic data it stayed above ECG but below the calibrated sum (-0.018 at $m = 0.5$, -0.027 to -0.011). With PPG missing as observed, the uncalibrated sum reached an AUC of 0.847 on every MIMIC segment (0.838–0.856), against 0.923 for ECG alone and 0.938 with calibration (Fig. 4d). Without calibration the subsets did not share one scale: non-AF segments scored from both signals (median predicted probability

0.70) scored nearly as high as AF segments scored from ECG alone (0.81), so that such an AF segment outranked such a non-AF segment with probability 0.63, against 0.95 with calibration (Fig. 4e). Without calibration, the mean of the experts' probabilities changed the AUC by −0.026 on CinC 2016 (−0.045 to −0.006) and +0.002 on MIMIC, and stacking and the learned gate, which were fitted on every availability pattern, changed little in any study (Table 3).

*Every rule deployed as the modality-agnostic model on every record with an ECG (signals missing as observed), with calibrated and uncalibrated experts: AUC gain over the ECG expert. CinC 2016: 326 records, ECG AUC 0.846; MIMIC: 40,792 records, ECG AUC 0.923; Synthetic: 500 records, ECG AUC 0.878. *Holm-adjusted $p < 0.05$ (DeLong's test on CinC 2016, paired subject bootstrap otherwise). CIs are in Online Resource 2.*

	CinC 2016		MIMIC		Synthetic	
2-3 Rule	calib.	uncalib.	calib.	uncalib.	calib.	uncalib.
Log-odds sum	+0.003	−0.024	+0.015*	−0.076*	+0.099*	+0.086*
Stacking	+0.005	+0.004	+0.016*	+0.016*	+0.099*	+0.098*
Learned gate	+0.006	+0.004	+0.015*	+0.015*	+0.087*	+0.095*
Weighted mean	+0.006	−0.003	+0.013*	+0.015*	+0.099*	+0.094*
Mean	−0.001	−0.027	+0.013*	+0.015*	+0.098*	+0.089*
Max-confidence	−0.023	−0.036	+0.009*	+0.002	+0.055*	+0.066*
Majority vote	−0.079*	−0.079*	−0.026*	−0.026*	+0.087*	+0.087*



One model for any missing signals, and why it needs calibrated experts. ***a–c*** *AUC gain of the log-odds sum over the ECG expert (zero line) on the records with every signal, when each non-ECG signal is missing at random with probability m and each record is scored from the signals it keeps: expected value over the removals (Eq. [eq:missing]) with 95% CI (subject-level bootstrap), for calibrated (solid) and uncalibrated (dashed) experts.* ***d*** *Every rule deployed as the modality-agnostic model on all MIMIC segments, with PPG missing as observed, with calibrated and uncalibrated experts: AUC gain over the ECG expert (95% CI; filled markers: Holm-adjusted $p < 0.05$).* ***e*** *Predicted probability of AF from ECG alone and from both signals on the MIMIC segments with both signals, by class (dot: median; bar: interquartile range; line: 10th to 90th percentile).*

4.3 Better ways to vote

Majority voting was the only rule that was significantly worse than ECG alone in both clinical studies (Fig. 2). With two experts a hard vote takes only three values, so most of its AUC deficit is lost resolution rather than lost information (Fig. 3c). Breaking ties by the experts' mean calibrated probability raised the AUC of the hard vote from 0.890 to 0.929 on MIMIC, turning a loss of −0.020 against ECG into a gain of +0.019 (0.014–0.024), and from 0.769 to 0.837 on CinC 2016. Soft voting, which averages the calibrated probabilities

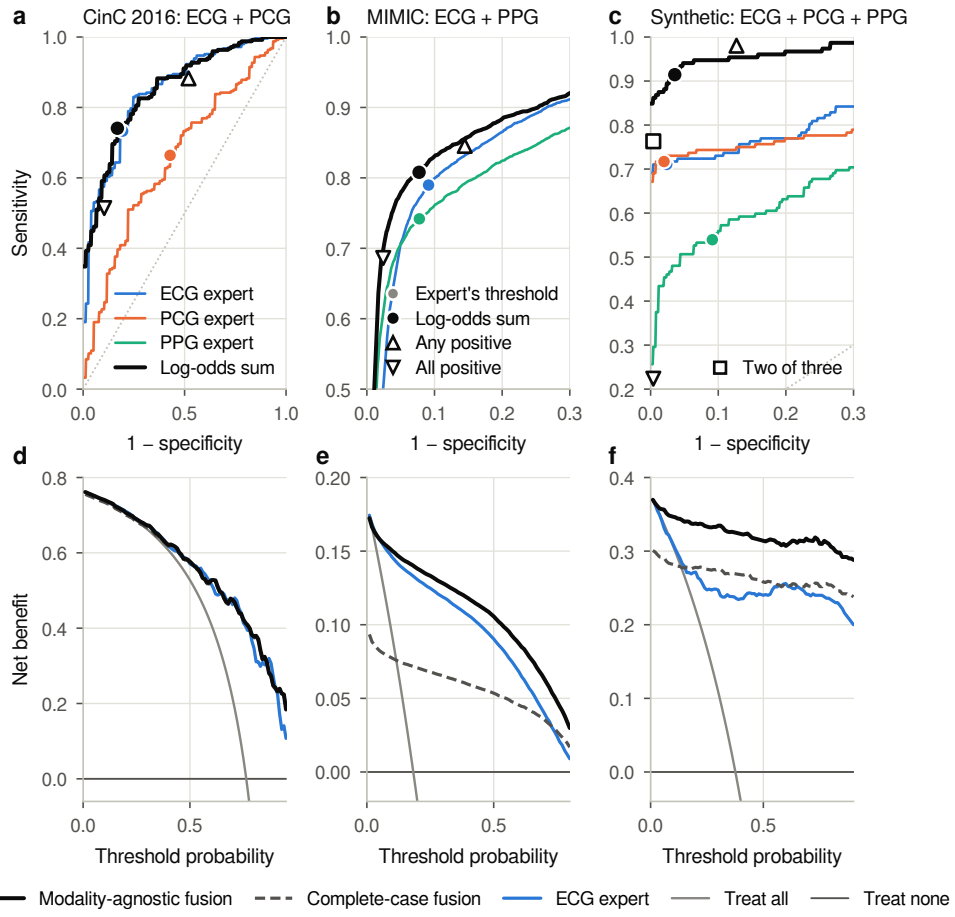
instead of counting votes, gained +0.019 (0.014–0.024) on MIMIC, close to the log-odds sum (+0.021), and matched ECG on CinC 2016 (−0.001, −0.026 to 0.024).

As decision rules at the experts' own thresholds, hard votes trade sensitivity against specificity (Table 4). On MIMIC, declaring AF when either expert was positive raised sensitivity from 0.790 (ECG) to 0.845 at a specificity of 0.855, whereas requiring both experts to agree raised specificity to 0.976 and the positive predictive value from 0.553 to 0.802, flagging 10.7% of segments instead of 17.9%, at a sensitivity of 0.686. The log-odds sum improved sensitivity (+1.8 percentage points) and specificity (+1.4 points) together. On CinC 2016, the either-positive rule reached a sensitivity of 0.883 at a specificity of 0.481, the both-positive rule a specificity of 0.896 at a sensitivity of 0.514, and the log-odds sum changed neither significantly (+0.8 and +2.6 points). On the synthetic benchmark, a two-of-three vote reached a specificity of 0.996 at a sensitivity of 0.763.

Decision rules at the experts' operating points, on the records with both signals (CinC 2016: 324 recordings; MIMIC: 33,538 segments). Each expert votes positive when its calibrated probability exceeds its Youden threshold fitted on the training folds. The hard vote chooses its vote threshold by Youden on the training folds as well; it selected the OR rule in 1 of 5 CinC folds and in 5 of 5 MIMIC folds. Sens: sensitivity; Spec: specificity; PPV: positive predictive value; Flagged: share of segments called positive.

	CinC 2016 (ECG + PCG)			MIMIC (ECG + PPG)			
2-4 Decision rule	Sens	Spec	PPV	Sens	Spec	PPV	Flagged
ECG expert	0.733	0.805	0.923	0.790	0.909	0.553	17.9%
PCG / PPG expert	0.664	0.571	0.832	0.742	0.922	0.577	16.1%
Either expert positive (OR)	0.883	0.481	0.845	0.845	0.855	0.455	23.2%
Both experts positive (AND)	0.514	0.896	0.941	0.686	0.976	0.802	10.7%
Hard vote, threshold by Youden	0.607	0.831	0.920	0.845	0.855	0.455	23.2%
Soft vote (mean probability)	0.704	0.857	0.941	0.803	0.923	0.599	16.8%
Log-odds sum	0.741	0.831	0.934	0.808	0.923	0.599	16.8%

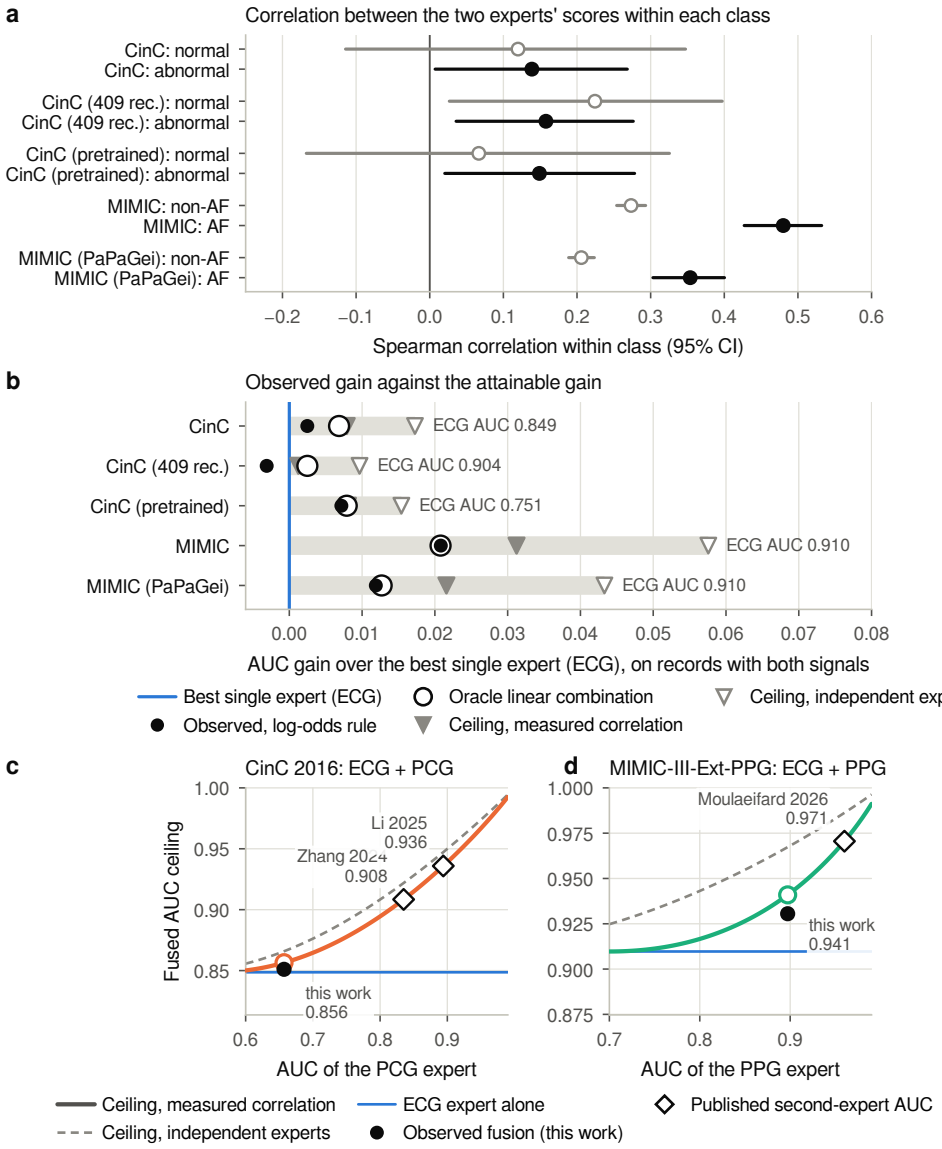
On the ROC curves, the log-odds sum operated above and to the left of the ECG expert on MIMIC and the synthetic data, whereas the OR and AND rules moved along the curves (Fig. 5a–c). In decision curve analysis (Fig. 5d–f), the modality-agnostic model had a higher net benefit than ECG alone on MIMIC at every threshold probability from 0.07 to 0.85; at 0.20, the gain was equivalent to 7.5 more AF segments detected per 1,000 without more false positives (95% CI 5.8–9.0). Complete-case fusion, which leaves segments without usable PPG unassessed, fell below ECG alone at every threshold from 0.01 to 0.68, by 60 detections per 1,000 segments at 0.20 (55–65). On CinC 2016 the decision curves of fusion and ECG nearly coincided, and on the synthetic data the modality-agnostic model led from 0.05 to 0.99, by 66 net true positives per 1,000 records at 0.20 (40–93).



Operating points and clinical utility. ***a–c*** *ROC curves of the experts and the log-odds sum on the records with every signal, with the decision rules at the experts' thresholds (Table 4 for the clinical studies): each expert, any expert positive, all experts positive, two of three positive, and the log-odds sum at its own threshold; the axes of b and c are truncated.* ***d–f*** *Decision curves on every record with an ECG: net benefit of the modality-agnostic model, the ECG expert and complete-case fusion, whose unscored records are not flagged, with treating every record or no record as positive for reference.*

4.4 Redundancy bounds the fusion gain

The experts' errors were positively correlated within classes in both clinical studies (Table [tab:redundancy], Fig. 6a). On MIMIC, a non-AF segment that looked more like AF to the ECG expert also tended to look more like AF to the PPG expert ($\rho = 0.27$, 0.25–0.29), and the dependence was stronger among AF segments ($\rho = 0.48$, 0.43–0.53); both were positive in every outer fold. On CinC 2016 the correlations were smaller (0.12 and 0.14), but the PCG expert was weak.



Redundancy between experts and the attainable gain. ***a*** *Fold-centred within-class Spearman correlation between the two experts' scores (95% subject-cluster bootstrap CI; filled markers: abnormal or AF class, open markers: normal or non-AF class).* ***b*** *AUC gain over the best single expert on records with both signals: observed (log-odds rule, out-of-fold), in-sample oracle linear combination of the same calibrated scores, and equal-variance binormal ceilings with the measured class-averaged correlation and under conditional independence. The grey bar spans the gain that independent experts would allow.* ***c,d*** *Binormal fusion ceiling as a function of the second expert's AUC, at the measured class-averaged correlation (solid) and for independent experts (dashed), with our second expert (open circle; filled dot: observed out-of-fold fusion) and single-signal AUCs reported elsewhere (diamonds) (Zhang et al. 2024; Li et al. 2025; Moulaeifard, Aston, et al. 2026).*

These correlations explain the size of the gains (Fig. 6b). Had the MIMIC experts been conditionally independent, the binormal ceiling would have been 0.967, a gain of 0.058 over ECG; with the measured correlations it fell to 0.941. The in-sample oracle linear combination of the two calibrated scores reached 0.931, and the out-of-fold log-odds rule matched it (0.931), realising 36% of the gain that independence would allow. On CinC 2016 the independence ceiling (0.866) left little room above ECG (0.849) because the PCG expert was weak, and the oracle (0.855) and the best rule (0.855) were close to the correlation-aware ceiling (0.856). In every real-data analysis the best calibrated rule came within 0.003 AUC of the oracle. The combiners were therefore close to optimal for linear fusion of these experts; what limited the gain was that the experts made correlated errors.

How much a stronger second expert could add depends on its AUC and on how its errors correlate with those of the ECG expert (Fig. 6c,d). At our measured correlation, a PCG expert with the AUC reported by Zhang et al. (Zhang et al. 2024) (0.835) would raise the ECG + PCG ceiling from 0.856 to 0.908, and one at the level reported by Li et al. with beat-level splits (Li et al. 2025) (0.894) to 0.936. On MIMIC, a PPG expert at the AUC reported for a deep network trained on the full corpus (Moulaeifard, Aston, et al. 2026) (0.960) would raise the ceiling from 0.941 to 0.971. For ECG–PCG fusion the lever is therefore the PCG expert, not the combiner.

The Shapley interaction on the d'^2 scale agreed: it was negative on MIMIC (−2.41, −2.66 to −2.19) and on CinC 2016 (−0.28, −0.63 to −0.05), indicating redundancy, and positive for ECG and PCG on the synthetic data (1.99, 0.06–4.38), where manifestation was independent by design. On the synthetic data the within-class correlations were positive among normal records but close to zero or negative among abnormal ones, and the fused AUCs exceeded the binormal ceilings, as expected for scores that are mixtures of manifesting and non-manifesting cases rather than binormal.

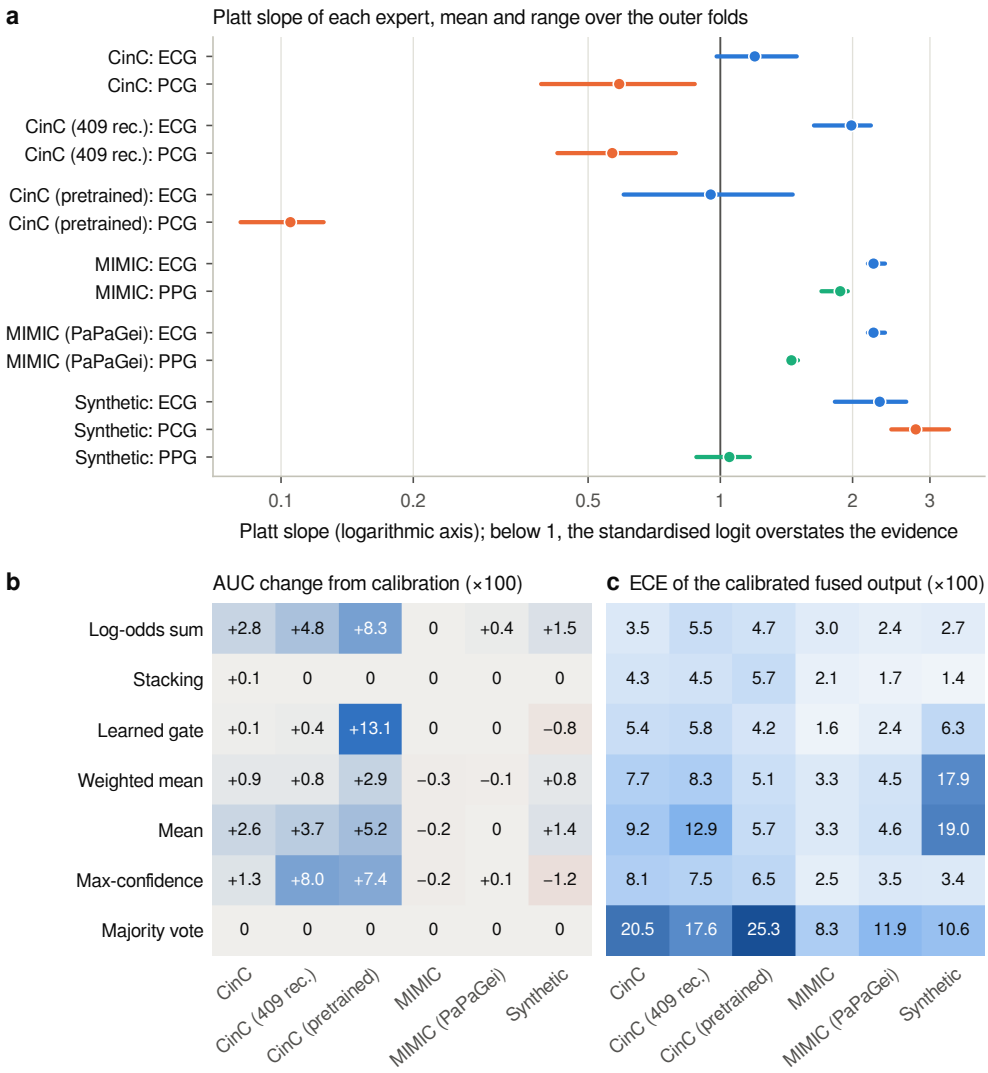
	Within-class ρ (95% CI)		Best	Binormal ceiling		Oracle	Observed
2-3 Study	normal / non-AF	abnormal / AF	single	independent	correlated	linear	log-odds (best)
CinC 2016	0.12 (−0.11 to 0.35)	0.14 (0.01–0.27)	0.849	0.866	0.856	0.855	0.851 (0.855)
CinC 2016, 409 rec.	0.22 (0.03–0.40)	0.16 (0.04–0.28)	0.904	0.913	0.905	0.906	0.900 (0.903)
CinC 2016, pretrained	0.07 (−0.17 to 0.33)	0.15 (0.02–0.28)	0.751	0.767	0.759	0.759	0.759 (0.759)
MIMIC	0.27	0.48	0.910	0.967	0.941	0.931	0.931

	Within-class ρ (95% CI)		Best	Binormal ceiling		Oracle	Observed
	(0.25–0.29)	(0.43–0.53)					(0.931)
MIMIC, PaPaGei PPG	0.21 (0.19–0.22)	0.35 (0.30–0.40)	0.910	0.953	0.931	0.922	0.922 (0.922)

4.5 Calibration acts as implicit expert weighting

The Platt slopes showed that the experts' standardised logits were on different scales relative to their discrimination (Table [tab:calibration], Fig. 7a). On CinC 2016 the PCG slope (0.59) was half the ECG slope (1.20): before calibration the PCG expert's logits were too large for the information they carried, so summing raw logits overweighted the weak expert. Calibration restored the balance, and removing it reduced the log-odds AUC by 0.028 on the primary cohort and by 0.048 on all 409 recordings. With the frozen pretrained PCG encoder the slope was 0.11 and the loss 0.083. On MIMIC the slopes of the ECG (2.23) and PPG (1.87) experts were similar and calibration did not change the AUC of any rule by more than 0.003 on the segments with both signals; it was nevertheless essential for the log-odds sum once segments without PPG were scored (Section 4.2). Stacking, which learns one weight per expert, was unaffected in every study; the learned gate, whose weights depend on the logit magnitude, lost 0.131 without calibration when expert scales differed nearly tenfold.

Calibration also governed the reliability of the fused probabilities (Fig. 7c). Without it, the ECE of the log-odds rule was 0.581 on MIMIC and 0.449 on CinC 2016, against 0.030 and 0.035 with it. Averaging calibrated probabilities (mean and weighted mean) produced under-confident outputs when several experts agreed, with an ECE of 0.190 for the mean of the three synthetic experts, whereas the log-odds sum (0.027) and stacking (0.014) remained calibrated. Majority voting produced the least calibrated outputs in both clinical studies.



Calibration. ***a*** *Platt slope of each expert (fold mean and range; logarithmic axis); a slope below 1 means the expert's standardised logit overstates its evidence.* ***b*** *Change in the AUC of the fused output (all signals) when per-expert calibration is applied, for each rule and study.* ***c*** *Expected calibration error of the calibrated fused output. Values in* ***b*** *and* ***c*** *are multiplied by 100.*

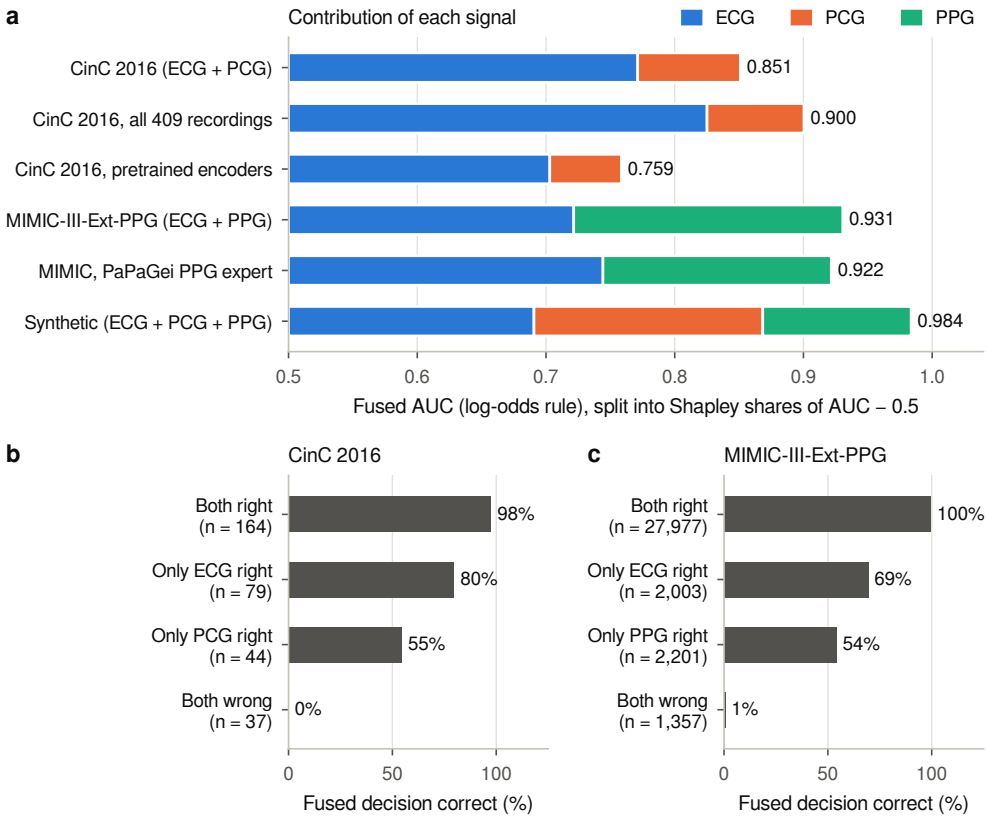
	Platt slope		Log-odds AUC		Log-odds ECE		Gate / stacking AUC	
2-3 Study	ECG	other	calib.	uncalib.	calib.	uncalib.	calib.	uncalib.
CinC 2016	1.20 (0.98–1.49)	0.59 (0.39–0.87)	0.851	0.823	0.035	0.449	0.854 / 0.854	0.853 / 0.853
CinC 2016, 409 rec.	1.99 (1.63–2.20)	0.57 (0.43–0.79)	0.900	0.853	0.055	0.347	0.903 / 0.902	0.899 / 0.901

	Platt slope		Log-odds AUC		Log-odds ECE		Gate / stacking AUC	
CinC 2016, pretrained	0.95 (0.60–1.46)	0.11 (0.08–0.13)	0.759	0.676	0.047	0.614	0.756 / 0.756	0.625 / 0.756
MIMIC	2.23 (2.17–2.37)	1.87 (1.70–1.95)	0.931	0.931	0.030	0.581	0.930 / 0.930	0.931 / 0.930
MIMIC, PaPaGei PPG	2.23 (2.17–2.37)	1.45 (1.43–1.50)	0.922	0.918	0.024	0.585	0.922 / 0.922	0.922 / 0.922

4.6 What fusion does: decision-level explanations

Exact Shapley values split the fused discrimination nearly evenly between ECG and PPG on MIMIC (ECG 0.222, 0.214–0.230; PPG 0.209, 0.202–0.216; 51% and 49% of $\mathrm{AUC} - 0.5$), but assigned 77% to ECG on CinC 2016 (ECG 0.271, PCG 0.080; Fig. 8a). On the synthetic data, ECG (0.191) and PCG (0.178) contributed similarly and PPG less (0.115), in line with their single-signal AUCs. A Shapley share is a fair division of the fused AUC, not the gain a signal adds: on CinC 2016 the PCG share was 0.080 although adding PCG to ECG changed the AUC by only +0.002.

The disagreement analysis shows how the rule used the experts (Fig. 8b,c). On MIMIC, when exactly one expert was right (4,204 segments), the fused decision was right in 61.5% of cases: 69.4% when only ECG was right and 54.2% when only PPG was right. Fusion introduced 102 errors where both experts were right and corrected 9 of 1,357 segments both got wrong, so its gain came from re-ranking records on which the experts disagreed rather than from correcting shared errors. On CinC 2016 the fused decision favoured the ECG expert (79.7% correct when only ECG was right against 54.5% when only PCG was right) and corrected none of the 37 records that both experts got wrong.

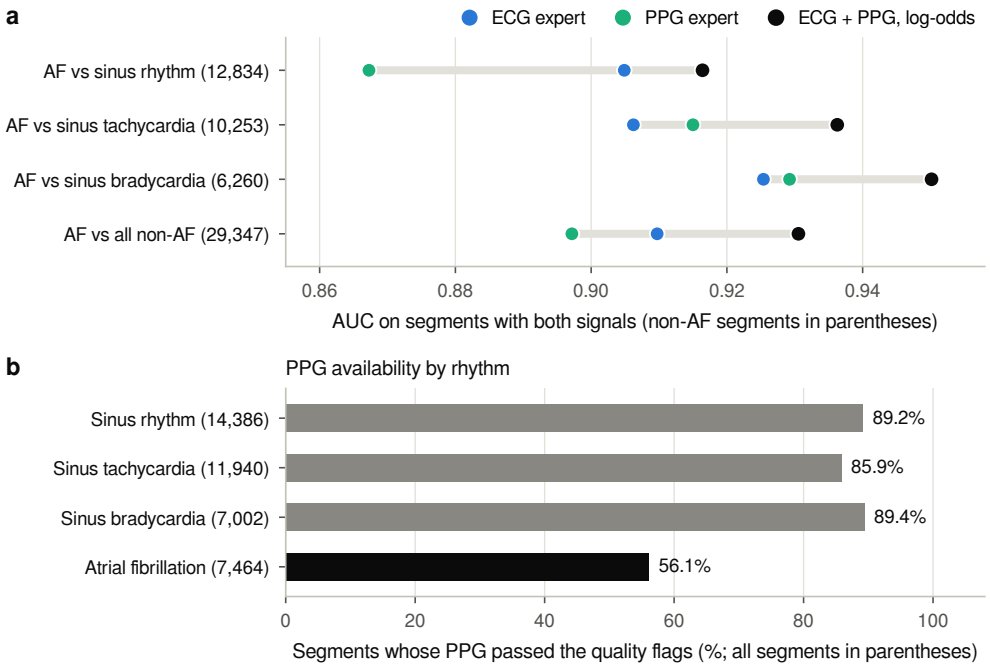


Decision-level explanations (log-odds rule). ***a*** *Fused AUC split into exact Shapley shares of* $\mathrm{AUC} - 0.5$ *for each study; the bar ends at the fused AUC.* ***b,c*** *Accuracy of the fused decision for records grouped by which experts classified them correctly at their own thresholds (CinC 2016 and MIMIC-III-Ext-PPG, all-signal records).*

4.7 Where PPG adds to ECG

Stratifying the negatives by rhythm located the benefit of PPG (Fig. 9a). On the segments with both signals, the PPG expert separated AF from sinus tachycardia (0.915) and sinus bradycardia (0.929) at least as well as the ECG expert (0.906 and 0.925), but separated AF from normal sinus rhythm less well (0.867 against 0.905). Fusion gained most where the two experts were comparable: 0.936 against sinus tachycardia and 0.950 against sinus bradycardia, against 0.916 against sinus rhythm.

Section 4.2 analyses the segments on which PPG was missing (Fig. 9b).



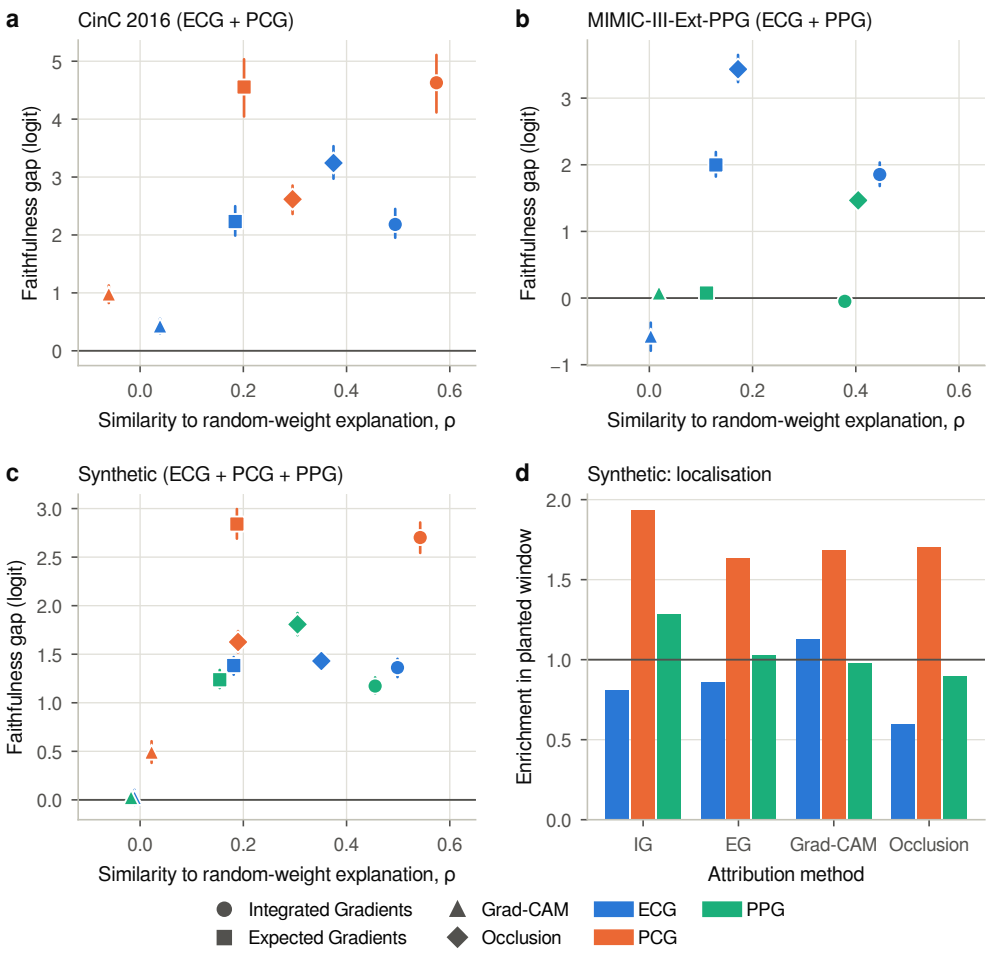
*Atrial fibrillation in MIMIC-III-Ext-PPG. **a** AUC of AF (4,191 segments) against each negative rhythm for the single experts and the fused model (log-odds rule), on segments with both signals. **b** Share of segments whose PPG passed the quality flags, by rhythm.*

4.8 Signal-level explanations

The four attribution methods differed in how faithful and how model-dependent they were (Table [tab:xai], Fig. 10). For the ECG and PCG experts, Integrated Gradients, Expected Gradients and occlusion were faithful by a wide margin; for example, Expected Gradients gave gaps of 2.00 logit (1.82–2.19) for the MIMIC ECG expert and 4.55 (4.05–5.03) for the CinC PCG expert. Grad-CAM was the least faithful method for every ECG and PCG expert and pointed at the wrong samples for the MIMIC ECG expert (gap −0.57, −0.79 to −0.37). The randomisation check separated the gradient methods: Integrated Gradients of trained and random-weight networks were moderately correlated (ρ 0.38–0.57), indicating that they partly reflect the input signal itself, whereas Expected Gradients depended much more on the learned parameters (ρ 0.11–0.20). Expected Gradients was the only method that was both faithful and model-dependent for the ECG and PCG experts, and we use it for the profiles below. The PPG expert was the exception: only occlusion was faithful (gap 1.47, 1.36–1.58); Expected Gradients (0.07) and Grad-CAM (0.08) were close to zero and Integrated Gradients was slightly negative (−0.05). Removing individual samples of a pulse wave apparently does not remove the evidence that the PPG expert uses for AF, which occlusion of 0.2-s windows does.

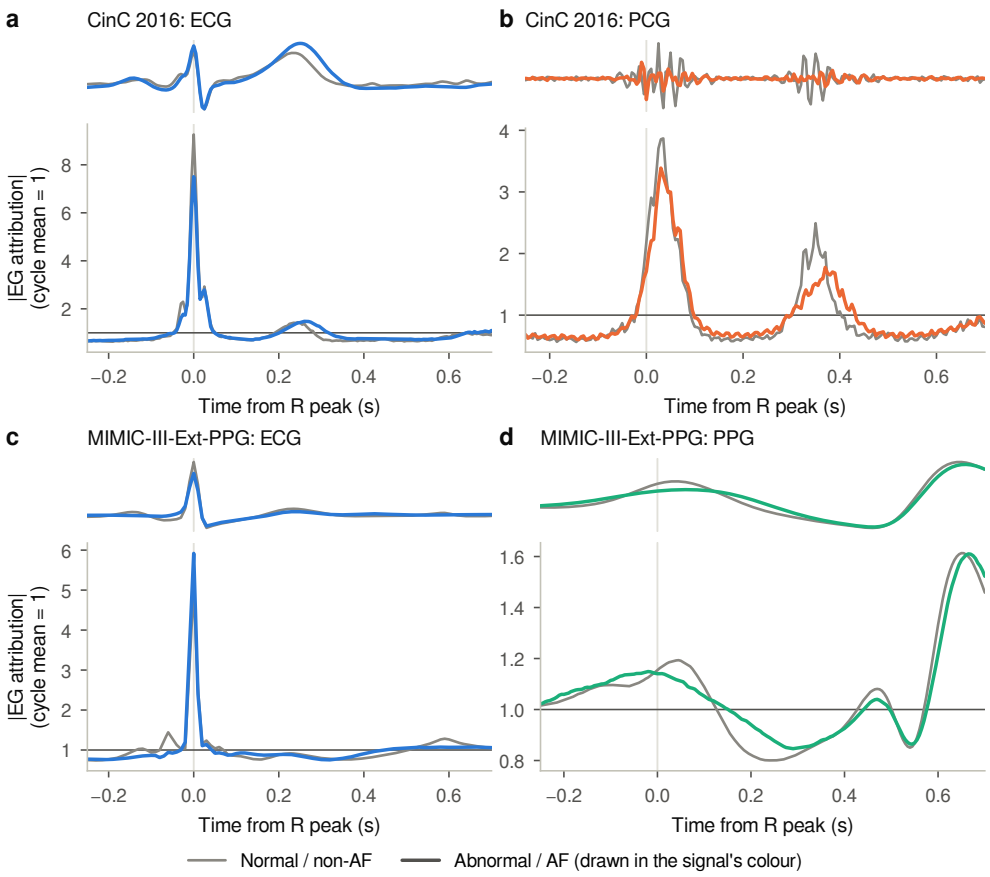
		IG		EG		Grad-CAM		Occlusion	
3-4 Study	Expert	Gap (95% CI)	ρ	Gap (95%	ρ	Gap (95% CI)	ρ	Gap (95%	ρ

		IG		EG		Grad-CAM		Occlusion	
				CI)				CI)	
CinC 2016	ECG	2.18 (1.95–2.45)	0.49	2.23 (1.99–2.50)	0.18	0.43 (0.30–0.55)	0.04	3.24 (2.97–3.53)	0.37
	PCG	4.63 (4.12–5.11)	0.57	4.55 (4.05–5.03)	0.20	0.98 (0.82–1.13)	−0.06	2.62 (2.36–2.85)	0.30
MIMIC	ECG	1.85 (1.68–2.03)	0.45	2.00 (1.82–2.19)	0.13	−0.57 (−0.79 to −0.37)	0.00	3.43 (3.24–3.65)	0.17
	PPG	−0.05 (−0.08 to −0.02)	0.38	0.07 (0.05–0.10)	0.11	0.08 (0.00–0.15)	0.02	1.47 (1.36–1.58)	0.40
Synthetic	ECG	1.36 (1.26–1.45)	0.50	1.38 (1.29–1.47)	0.18	0.03 (−0.05 to 0.10)	−0.01	1.43 (1.35–1.52)	0.35
	PCG	2.70 (2.54–2.86)	0.54	2.84 (2.69–3.00)	0.19	0.49 (0.38–0.60)	0.02	1.63 (1.52–1.74)	0.19
	PPG	1.17 (1.09–1.26)	0.46	1.24 (1.15–1.34)	0.15	0.02 (−0.03 to 0.08)	−0.02	1.81 (1.70–1.92)	0.30



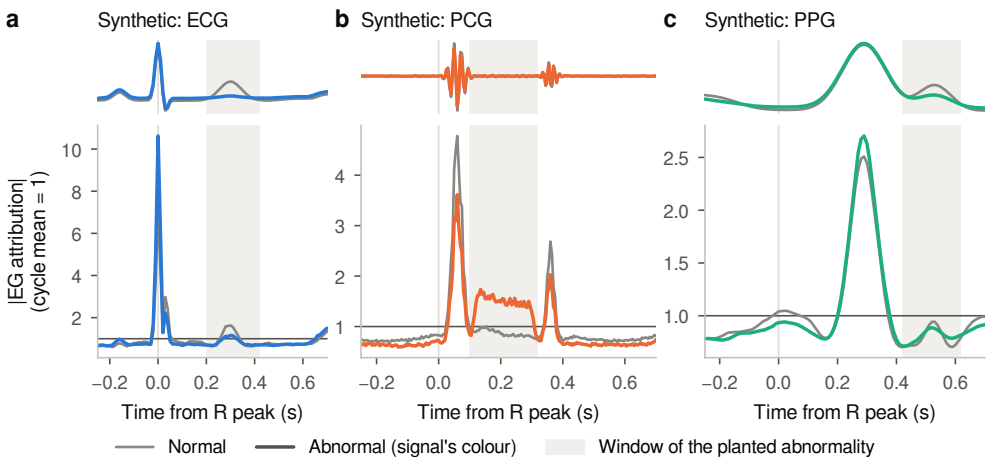
Testing the signal-level explanations. ***a–c*** *Faithfulness gap (deletion test, 95% CI) against similarity to the explanation of a random-weight network; good explanations lie at the upper left.* ***d*** *Synthetic data: enrichment of attribution in the window holding the planted abnormality (abnormal over normal records; 1 means no enrichment).*

Beat-aligned profiles related the attributions to the cardiac cycle (Fig. 11). The CinC 2016 ECG expert attended to the QRS complex, and the PCG expert to the first and second heart sounds, with its attribution peaking 30–35 ms after the R peak, where S1 falls; attribution in systole and diastole, where murmurs occur, was below the cycle average. The MIMIC ECG expert also concentrated on the QRS complexes, whose timing carries the beat-to-beat (RR-interval) irregularity of AF; in non-AF segments it additionally attributed evidence to the 100 ms before the QRS, which in sinus rhythm contains the end of the P wave and the PR segment (attribution peak 60 ms before the R peak), a region that is featureless in AF. The PPG expert's attribution was spread over the pulse, highest around the systolic peak of the following pulse and similar in both classes, consistent with evidence carried by pulse timing rather than by a localised waveform feature.



Beat-aligned Expected Gradients attribution for the clinical experts, averaged over beats aligned on the ECG R peak (CinC 2016: 2,276 normal and 7,207 abnormal ECG beats, 2,276 and 7,151 PCG beats; MIMIC: 1,292,683 non-AF and 321,213 AF ECG beats, 1,132,308 and 168,445 PPG beats). Upper strips: mean z-scored waveform; lower panels: |attribution| relative to the cycle mean (horizontal line at 1 = uniform attention).

On the synthetic data, where the planted windows are known, only the PCG explanations localised the abnormality, with enrichment between 1.63 and 1.94 for all four methods (Fig. 10d, Fig. 12). The ECG attributions were not enriched in the T-wave window (Expected Gradients 0.86, Integrated Gradients 0.81; only the unfaithful Grad-CAM exceeded 1, at 1.13). The profiles show why (Fig. 12a): the planted abnormality removes or inverts the T wave, and the expert attributed evidence to the T wave of normal records, where it is present, rather than to its absence in abnormal ones. An enrichment score that expects evidence in abnormal records penalises explanations of abnormalities that consist of something missing. The PPG attributions concentrated on the systolic peak rather than on the damped dicrotic wave, with modest enrichment for Integrated Gradients (1.29) and none for Expected Gradients (1.03), although the expert reached an AUC of 0.766.



Beat-aligned Expected Gradients attribution on the synthetic benchmark, with the window that holds the planted abnormality shaded. Abnormal records include positives in which the abnormality did not manifest in that signal (30% for ECG and PCG, 50% for PPG).

5 Discussion

5.1 When a second signal helps

The three studies span the cases a late-fusion system meets in practice. When the second expert was nearly as discriminative as the first and their errors were only partly correlated (ECG and PPG for AF), fusion improved discrimination by about 0.02 AUC with a tight confidence interval. When the second expert was weak (PCG for abnormal heart sounds), no combiner improved on the ECG expert, although every reasonable combiner protected against the weak expert and none was significantly worse. When the signals carried independent evidence by construction (synthetic data), fusion added 0.10 AUC, and the three signals together gave the highest AUC of any combination, although a third signal added little to a pair that already discriminated very well (ECG + PCG, AUC 0.976; Table [tab:combinations]). The redundancy analysis turns this description into an explanation. In all real-data analyses the best calibrated rule matched the in-sample oracle linear combination of the same scores, so better combiners would not have helped; the gap to the independence ceiling was due to the experts failing on the same records. For ECG and PPG this is physiologically plausible: both signals detect AF chiefly through beat-timing irregularity, so a segment with ectopic beats or noise that misleads one expert is likely to mislead the other, and correlation was strongest among AF segments, where the evidence is shared. Measuring the within-class correlation of candidate experts, which requires only their out-of-fold scores, is a cheap way to predict whether acquiring an additional signal is worth it. Where one expert is weak, the projection shows that a better expert, not a better combiner, is the lever: a PCG expert at a published level would raise the ECG–PCG ceiling by 0.059 to 0.087 AUC over our ECG expert, against 0.007 with ours.

The rhythm-stratified analysis adds a clinically relevant nuance. PPG separated AF from sinus tachycardia and sinus bradycardia as well as the ECG did, but was weaker than ECG against normal sinus rhythm, and fusion helped most where the experts were comparable. Tachycardia and bradycardia were included among the negatives deliberately, to stop heart rate alone from separating the classes; the pooled AUC of a PPG detector therefore depends on the mix of negative rhythms, and rhythm-stratified results should accompany pooled ones.

5.2 One model for whatever signals are usable

A model that needs every signal can only be evaluated, and used, on the records that have them. Our MIMIC results show why this matters: quality control removed PPG mostly during AF, so a complete-case ECG + PPG model would have been silent on 43.9% of the AF segments, the records on which a detector is most needed, and its reported AUC would describe a population with less than a third of the AF prevalence of the excluded segments. The modality-agnostic model needs no imputation and no separate models: the same calibrated rule scores a record from whichever experts are available, so it covers every record, falls back to the best available signal and improves on ECG alone wherever PPG is present. Because each subset is evaluated separately (Fig. 2), its behaviour under any pattern of missing signals can be reported before deployment, and each decision can be explained exactly by the experts that contributed to it. Intermediate-fusion networks can be made robust to missing inputs by modality dropout (Neverova et al. 2016), by architectures for missing modalities (Ma et al. 2021, 2022), by using a single-signal branch alone with the missing input zeroed (Zhang et al. 2024) or by a shared encoder that accepts either signal (Guo et al. 2025), but these mechanisms must be built into training and validated for each missingness pattern. Calibration is a precondition for this design rather than a refinement: without it, the same rule matched calibrated fusion on complete records yet fell below ECG alone once records scored from different signals were pooled, a failure that complete-case evaluation cannot reveal. Late fusion obtains the property by construction, at the price of not modelling interactions between signals below the decision level; for our experts, the redundancy analysis suggests that little was lost, because the calibrated rules matched the oracle linear combination of the experts' scores.

5.3 Simple, calibrated combiners suffice

The log-odds sum, stacking and the learned gate were statistically indistinguishable in every study, while two popular rules were harmful as usually implemented. Majority voting reduced each expert to a binary opinion and lost to ECG alone in both clinical studies; with two experts it yields only three scores, and most of its deficit disappeared when ties were broken by the experts' confidence, while soft voting, which averages calibrated probabilities, nearly matched the best rule. Hard votes remain useful as decision rules: requiring both ECG and PPG to flag AF cut the share of segments flagged from 17.9% to 10.7% and raised the positive predictive value from 0.553 to 0.802, which may suit alarm reduction in monitoring, whereas declaring AF when either signal flags it raised sensitivity, which may suit screening. Max-confidence picked one expert per record and was the

weakest rule on synthetic data. The log-odds sum has no parameters beyond the calibration maps and the prior, is the Bayes-optimal combiner under conditional independence and remained well calibrated, so we recommend it as the default late-fusion rule, with stacking when experts are strongly correlated. Calibration was not optional for fixed rules: the Platt slopes revealed tenfold differences in logit scale between experts, and summing uncalibrated logits gave the weak or over-confident expert too much weight. This mechanism, calibration as implicit expert weighting, explains why calibration mattered on CinC 2016, where scales differed, and not on the MIMIC segments with both signals, where they did not, and why stacking, which learns its own weights, was immune. Calibration has a second role in a modality-agnostic model: it puts records scored from different subsets of signals on one probability scale, and on MIMIC this role, not the weighting, was decisive (Section 4.2). Calibrated outputs matter beyond fusion, because clinical decisions use the predicted probability itself (Van Calster et al. 2019).

5.4 Explanations must be tested

Explanations at two levels answered different questions. Exact modality Shapley values and the disagreement analysis described how the decision used each signal, and were cheap and assumption-free because the late-fusion model is defined on every subset of signals. Shapley shares must, however, be read as a division of the fused discrimination rather than as added value: on CinC 2016 the PCG expert received 23% of the fused AUC above chance while adding nothing to ECG. At the signal level, the methods differed more than is usually acknowledged. Grad-CAM, a common choice in physiological-signal studies, ranked last on faithfulness for all ECG and PCG experts and failed outright for the MIMIC ECG expert, consistent with its coarse temporal resolution after pooling. Integrated Gradients was faithful but partly input-driven, and Expected Gradients was the best compromise for the ECG and PCG experts. For the PPG expert only occlusion passed the deletion test, which suggests that its evidence is distributed over the pulse train rather than located in individual samples. The synthetic benchmark exposed a further pitfall: a correct detector of a missing T wave produced attributions that a ground-truth localisation score judged wrong, because the evidence lay in the T wave of normal records. Faithfulness tests, randomisation checks and ground truth answer different questions, and we recommend reporting all three (Adebayo et al. 2018; Rong et al. 2022; Ghassemi et al. 2021).

5.5 Clinical implications

For screening devices that combine signals, such as ECG-enabled stethoscopes (Bachtiger et al. 2022) and smartwatches (Perez et al. 2019), a modality-agnostic model offers graceful degradation: the same model returns a calibrated decision from whichever signals pass quality control. Decision curves show what is at stake: on MIMIC, leaving the segments without usable PPG unassessed cost far more net benefit than adding PPG gained. Because a diagnosis of AF requires ECG documentation (Van Gelder et al. 2024), PPG is best viewed as a complement to the ECG rather than a replacement; our results show where it adds most, against sinus tachycardia and bradycardia. Quality control is not

neutral: PPG segments were rejected 3.7 times more often during AF, presumably because irregular pulse intervals and amplitudes are hard to distinguish from artefacts, so whether a signal is available carries information about the label (Groenwold 2020; Little and Rubin 2019). A model trained or evaluated without accounting for this can learn missingness as a shortcut, which differs between sites and devices; the log-odds sum, our primary rule, cannot learn this shortcut because it has no availability-specific terms, whereas stacking can, through its mask offsets, which fitting on subset-augmented rows only attenuates. Reporting the coverage of the positive class makes the problem visible, and a deployed system should monitor rejection rates by rhythm.

5.6 Comparison with previous work

Table [tab:prior] compares ECG–PCG fusion on training-a. Under one protocol (388 recordings, five-fold cross-validation by recording), Zhang et al. re-implemented four published fusion methods and two baselines (Zhang et al. 2024): the fused models improved on their own ECG versions by 0.004 to 0.023 AUC, the decision-level Dempster–Shafer fusion of J. Li et al. (Li et al. 2022) by 0.005, and no difference was tested. Our gain of +0.002 (95% CI −0.022 to 0.027) lies within this range, and its confidence interval shows that gains of this size cannot be distinguished from zero with a few hundred recordings. Reported absolute values range from an AUC of 0.81 to above 0.99, depending on the experts and on the splitting unit; splits of beat segments or windows place parts of the same recording in training and test sets, which inflates performance (Saeb et al. 2017; Kapoor and Narayanan 2023). Our experts were deliberately small networks trained from scratch inside nested cross-validation: our ECG expert lies within the range of the re-implemented ECG models (0.81–0.95) and our PCG expert just below that of the re-implemented PCG models (0.67–0.84), so our absolute fused AUC is lower than the best published values while the gain from fusion is comparable. On their private valve-disease cohort, in which PCG was the stronger signal, the late-fusion baseline of Zhang et al. added 0.018 to the PCG model (Zhang et al. 2024), and Calzoni et al. reported about 0.03 over ECG with an audio-pretrained PCG branch (Calzoni et al. 2025), consistent with the projection that a stronger PCG expert is the route to an ECG–PCG gain.

Table [tab:priorppg] lists ECG–PPG studies of AF. Most used subsets of MIMIC-III, reported accuracy rather than AUC and did not state whether training and test data were separated by patient (Aldughayfiq et al. 2023; Malone et al. 2025; John 2025; John et al. 2025); where single-signal results were given, fusion improved on ECG by 1.0 to 1.4 percentage points of accuracy (Malone et al. 2025; John 2025), the same direction as our patient-grouped gain of +0.021 AUC. SiamAF (Guo et al. 2025) improved each single-signal model by training one encoder on both signals, an alternative route to signal-agnostic inference that classifies either signal but does not combine them. A PPG-only network trained on the full MIMIC-III-Ext-PPG corpus reached an AUC of 0.96 on its predefined folds (Moulaeifard, Aston, et al. 2026), against 0.897 for our small PPG expert trained on a sample within nested cross-validation, which indicates the room for stronger experts within the same framework. None of these studies reported performance on the records that lacked a signal.

@L3.8cmL4.5cmL2.8cmcccc@ Study & Fusion & Validation (splitting unit) & ECG & PCG & ECG + PCG & Δ vs ECG

P. Li et al. 2021 (P. Li et al. 2021) & deep features, genetic selection, SVM & 5-fold CV (recordings) & 0.920 & 0.742 & 0.943 & +0.023
Hettiarachchi et al. 2021 (Hettiarachchi et al. 2021) & two-branch CNN, feature level & 5-fold CV (recordings) & 0.860 & 0.700 & 0.881 & +0.021
J. Li et al. 2022 (Li et al. 2022) & decision level (Dempster–Shafer) & 5-fold CV (recordings) & 0.807 & 0.672 & 0.812 & +0.005
H. Li et al. 2021 (H. Li et al. 2021) & three-branch CNN, early & 5-fold CV (recordings) & 0.927 & 0.831 & 0.937 & +0.010
Early-fusion baseline (Zhang et al. 2024) & input channels & 5-fold CV (recordings) & 0.951 & 0.835 & 0.955 & +0.004
Late-fusion baseline (Zhang et al. 2024) & feature concatenation & 5-fold CV (recordings) & 0.951 & 0.835 & 0.956 & +0.005
CPDNet (Zhang et al. 2024) & progressive dense and decision level & 5-fold CV (recordings) & 0.951 & 0.835 & 0.973 & +0.022

W. P. Li et al. 2025 (Li et al. 2025) & cross-modal attention & 5-fold CV (1-s beats) & 0.992 & 0.894 & 0.997 & +0.005
Reehana and Siddique Ibrahim 2025 (Reehana and Siddique Ibrahim 2025) & feature level & CV (8-s windows) & 92.1%^{a} & 88.1%^{a} & 98.4%^{a} & +6.3 pp^{a}
Calzoni et al. 2025 (Calzoni et al. 2025) & late, audio-pretrained PCG branch & see article & ≈0.93^{b} & ≈0.85^{b} & 0.964 & ≈+0.03
Bargarai et al. 2026 (Bargarai et al. 2026) & decision level, quality-aware & 10-fold CV (n.r.) & n.r. & n.r. & 95.6%^{a} & n.r.

Primary cohort (324 recordings with both signals) & late, log-odds sum & nested 5×4 CV (recordings) & 0.849 & 0.657 & 0.851 & +0.002 (−0.022 to 0.027)
All 409 recordings (403 with both signals) & late, log-odds sum & nested 5×4 CV (recordings) & 0.903 & 0.654 & 0.900 & −0.003 (−0.015 to 0.009)

@L2.7cmL3.2cmL2.3cmL2.7cmcccc@ Study & Data & Validation & Fusion & Metric & ECG & PPG & ECG + PPG
Aldughayfiq et al. 2023 (Aldughayfiq et al. 2023) & MIMIC PERform & hold-out (unit n.r.) & input level, CNN–BiLSTM & accuracy (%) & n.r. & n.r. & 95
Malone et al. 2025 (Malone et al. 2025) & MIMIC-III matched subset & train–test split (unit n.r.) & signal level, quality-weighted & accuracy (%) & 94.8 & 69.5 & 96.2
John 2025 (John 2025) & MIMIC-III subset & n.r. & 1-D CNN sensor fusion & accuracy (%) & 98.85 & 98.52 & 99.85
John et al. 2025 (John et al. 2025) & MIMIC-III subset & n.r. & multilevel CNN with quality indices & accuracy (%) & n.r. & n.r. & 99.33
Guo et al. 2025 (Guo et al. 2025) & training 28,539 patients (UCSF); external test 126

patients (UCLA) & external test & shared encoder for either signal[a] & AUC & 0.890 / 0.927[a] & 0.918 / 0.924[a] & –
Moulaeifard et al. 2026 (Moulaeifard, Aston, et al. 2026) & MIMIC-III-Ext-PPG & predefined folds & PPG only & AUC & – & 0.96 & –
This work, both signals & MIMIC-III-Ext-PPG sample (33,538 segments, 5,523 patients) & nested 5×4 CV (patients) & late, log-odds sum & AUC & 0.910 & 0.897 & 0.931
This work, every segment[b] & 40,792 segments, 5,642 patients & nested 5×4 CV (patients) & modality-agnostic, log-odds sum & AUC & 0.923 & – & 0.938

5.7 Limitations

Several limitations qualify these findings. First, CinC 2016 does not release subject identifiers, so recordings of the same patient (409 recordings from 121 patients) can fall in training and test folds; absolute AUCs on this dataset may be optimistic, although paired comparisons share the same splits. The label is an abnormal heart-sound annotation, a proxy for cardiac disease, and the cohort is small, so fold-to-fold variation was large and differences of a few hundredths between runs should not be over-read. Second, MIMIC-III-Ext-PPG is an intensive-care cohort; the AF prevalence is that of our sample, rhythm labels are annotated at the event time and shared by the segments of an event, and age, unit, monitor and medication were not adjusted for, so no claim of clinical performance is made. Third, the PCG–PPG and three-signal results come from synthetic data, because no public dataset combines the three signals with a disease label; they test the method, not the physiology. Fourth, every analysis used a single cross-validation seed, and the experts were small by design; stronger experts, such as PPG networks trained on the full corpus (Moulaeifard, Aston, et al. 2026), would change absolute values and possibly the correlation structure. Fifth, the binormal ceilings rest on distributional assumptions; we therefore also reported the empirical oracle, which is optimistic, and the projections additionally assume that a stronger expert would share our experts' error correlation. Sixth, the comparisons with published results are indirect: cohorts, metrics and protocols differ, and several studies did not report how training and test data were separated. Finally, the pretrained encoders did not transfer well to training-a (Appendix 7), and the PaPaGei embeddings were pretrained on data that overlap MIMIC-III, so the pretrained-encoder analyses are sensitivity analyses only.

5.8 Future work

Four extensions follow. A stronger PCG expert is the most direct route to an ECG–PCG fusion gain, and the projection gives the gain to expect. Datasets that record all three signals with clinical labels are needed to test three-signal fusion on real data. Intermediate fusion should be compared with the calibrated late-fusion baseline under identical splits and paired tests, with the redundancy analysis indicating how much any architecture can gain. Finally, experts trained to be diverse, for example by penalising correlated errors, may raise the ceiling that redundancy imposes.

6 Conclusion

Calibrated late fusion of ECG, PCG and PPG yields one modality-agnostic model: trained once, it scores every record from whichever signals pass quality control, falls back gracefully when a signal is missing and can be explained exactly at the decision level. This matters because missingness is informative: a model requiring both ECG and PPG could not have assessed 43.9% of the AF segments, whereas the modality-agnostic model scored all of them and improved on ECG alone. The benefit of fusion is governed by the redundancy of the experts rather than by the choice of combiner: adding PPG to ECG improved AF detection by about 0.02 AUC, adding a weak PCG expert to ECG did not help, in line with the small gains reported on the same data, and simple calibrated rules matched an oracle combination of the same scores. When all three signals carried independent evidence, as in our synthetic benchmark, the full combination discriminated best, with diminishing returns once two signals were already strong. Calibration should be treated as part of the model, because it is what puts records scored from different signals on one scale, hard voting replaced by soft voting or kept for screening and confirmation operating points, and explanations reported together with tests of their faithfulness. Measuring within-class expert correlation offers a direct way to decide whether a new signal, a stronger expert or a more complex fusion architecture is likely to pay off.

The aggregate result tables (CSV) from which every number, table and figure of this article was generated, and the analysis scripts, are provided as supplementary material.

We thank the creators and maintainers of PhysioNet, the PhysioNet/CinC 2016 Challenge, MIMIC-III and MIMIC-III-Ext-PPG for making their data available.

Declarations

- **Funding** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

- **Competing interests** The authors declare no competing interests.

- **Ethics approval and consent to participate** Not applicable. This study is a secondary analysis of publicly available, de-identified data from PhysioNet: the CinC 2016 data are openly available, and MIMIC-III-Ext-PPG was accessed under the PhysioNet credentialed health data use agreement. The original data collections were approved by the institutional review boards of the contributing institutions, with consent procedures described in the dataset publications (Liu et al. 2016; Johnson et al. 2016; Moulaeifard, Kutscher, et al. 2026). The study was conducted in accordance with the Declaration of Helsinki.

- **Consent for publication** Not applicable.

- **Data availability** The CinC 2016 data are available from PhysioNet (Goldberger et al. 2000; Pollard et al. 2026). MIMIC-III-Ext-PPG is available from PhysioNet to credentialed users who sign its data use agreement (Moulaeifard, Charlton, et al. 2026). The synthetic data are generated by the code. Only aggregate results derived from MIMIC are reported.
- **Code availability** The pipeline (latefusion, version 0.8.1) and the analysis scripts will be made publicly available upon publication.
- **Author contribution** MD Tasnim Alam: conceptualisation, methodology, software, formal analysis, visualisation, writing – original draft. Zakir Hossain: methodology, writing – review and editing. Mohammad Ali Moni: supervision, conceptualisation, writing – review and editing.
- **Use of AI tools** See Section [sec:genai].

7 Sensitivity analyses

Table 5 reports three sensitivity analyses. On all 409 training-a recordings, analysed with an earlier release of the same pipeline (version 0.5.3, identical from-scratch training path), the experts and fused models scored higher than on the primary cohort, partly because the 80 validation-folder duplicates were easier (ECG AUC 0.940 on them against 0.880 on the other recordings within that run) and partly because the experts trained on more data; fusion again did not improve on ECG. With the fine-tuned ECG encoder and the frozen PCG encoder on the primary cohort, both experts were weaker than their from-scratch counterparts and fusion again matched ECG. On MIMIC, the logistic head on PaPaGei-S embeddings gave a weaker PPG expert than the from-scratch network, and the fusion gain over ECG shrank accordingly, consistent with the redundancy analysis (Table [tab:redundancy]).

Sensitivity analyses: pooled out-of-fold AUC (95% CI) of the single experts and of the log-odds and best rules, and the paired difference of the log-odds rule from the ECG expert (95% CI; Holm-adjusted p).

	CinC, all 409	CinC, pretrained	MIMIC, PaPaGei
ECG expert	0.902 (0.871–0.930)	0.751 (0.689–0.808)	0.923 (0.915–0.930)
PCG / PPG expert	0.656 (0.597–0.713)	0.607 (0.527–0.686)	0.843 (0.832–0.855)
Fused, log-odds	0.900 (0.868–0.929)	0.759 (0.702–0.811)	0.922 (0.911–0.932)
Fused, best rule	0.903 (learned gate)	0.759 (log-odds sum)	0.922 (learned gate)
ΔAUC, log-odds vs	−0.003 (−0.015 to	+0.007 (−0.030 to	+0.012 (0.008–

	CinC, all 409	CinC, pretrained	MIMIC, PaPaGei
ECG	0.009)	0.044)	0.016)
Holm-adjusted p	1.00	1.00	0.014

8 Reproducibility

Experts were trained with seed 42 and deterministic algorithms where available; the predictions of every outer and inner model, the fitted calibration and fusion parameters and all evaluation tables were saved by the pipeline. The scripts provided with this article recompute the within-class correlations, the fusion ceilings, the oracle linear combination, the disagreement tables, the rhythm-stratified analysis, the deployment, voting and combination analyses, the robustness analyses and decision curves, all figures and all tables from those outputs, and fill every number in this manuscript from the resulting tables.

Acosta, Julián N., Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. 2022. "Multimodal Biomedical AI." *Nature Medicine* 28: 1773–84. https://doi.org/10.1038/s41591-022-01981-2.

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. "Sanity Checks for Saliency Maps." *Advances in Neural Information Processing Systems* 31: 9505–15.

Aldughayfiq, Bader, Farzeen Ashfaq, N. Z. Jhanjhi, and Mamoona Humayun. 2023. "A Deep Learning Approach for Atrial Fibrillation Classification Using Multi-Feature Time Series Data from ECG and PPG." *Diagnostics* 13 (14): 2442. https://doi.org/10.3390/diagnostics13142442.

Allen, John. 2007. "Photoplethysmography and Its Application in Clinical Physiological Measurement." *Physiological Measurement* 28 (3): R1–39. https://doi.org/10.1088/0967-3334/28/3/R01.

Bachtiger, Patrik, Camille F. Petri, Francesca E. Scott, et al. 2022. "Point-of-Care Screening for Heart Failure with Reduced Ejection Fraction Using Artificial Intelligence During ECG-Enabled Stethoscope Examination in London, UK: A Prospective, Observational, Multicentre Study." *The Lancet Digital Health* 4 (2): e117–25. https://doi.org/10.1016/S2589-7500(21)00256-9.

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2): 423–43. https://doi.org/10.1109/TPAMI.2018.2798607.

Bargarai, Faiq A. Mohammed, Sagvan Ali Saleh, and Abdulkadir Sengur. 2026. "An Explainable Quality-Aware ECG–PCG Fusion for Cardiovascular Disease Detection Using

Robust Feature Modeling." *Diagnostics* 16 (14): 2296. https://doi.org/10.3390/diagnostics16142296.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Calzoni, A., M. Savardi, M. Silvestri, S. Benini, and A. Signoroni. 2025. "Bimodal ECG and PCG Cardiovascular Disease Detection: Exploring the Potential and Modality Contribution." *Journal of Medical Systems* 49 (1): 113. https://doi.org/10.1007/s10916-025-02245-5.

Chakir, F., A. Jilbab, C. Nacir, and A. Hammouch. 2020. "Recognition of Cardiac Abnormalities from Synchronized ECG and PCG Signals." *Physical and Engineering Sciences in Medicine* 43 (2): 673–77. https://doi.org/10.1007/s13246-020-00875-2.

Charlton, P. H., J. Allen, R. Bailon, S. Baker, J. A. Behar, et al. 2023. "The 2023 Wearable Photoplethysmography Roadmap." *Physiological Measurement* 44 (11): 111001. https://doi.org/10.1088/1361-6579/acead2.

Clifford, Gari D., Chengyu Liu, Benjamin Moody, et al. 2016. "Classification of Normal/Abnormal Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016." *Computing in Cardiology* 43: 609–12.

Collins, Gary S., Karel G. M. Moons, Paula Dhiman, et al. 2024. "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods." *BMJ* 385: e078378. https://doi.org/10.1136/bmj-2023-078378.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics* 44 (3): 837–45. https://doi.org/10.2307/2531595.

Ding, C., Z. Guo, C. Rudin, et al. 2024. "Learning from Alarms: A Robust Learning Approach for Accurate Photoplethysmography-Based Atrial Fibrillation Detection Using Eight Million Samples Labeled with Imprecise Arrhythmia Alarms." *IEEE Journal of Biomedical and Health Informatics* 28 (5): 2650–61. https://doi.org/10.1109/JBHI.2024.3360952.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Elgendi, Mohamed. 2016. "Optimal Signal Quality Index for Photoplethysmogram Signals." *Bioengineering* 3 (4): 21. https://doi.org/10.3390/bioengineering3040021.

Erion, G., Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. 2021. "Improving Performance of Deep Learning Models with Axiomatic Attribution Priors and Expected Gradients." *Nature Machine Intelligence* 3: 620–31. https://doi.org/10.1038/s42256-021-00343-w.

Field, Christopher A., and Alan H. Welsh. 2007. “Bootstrapping Clustered Data.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 (3): 369–90. https://doi.org/10.1111/j.1467-9868.2007.00593.x.

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. 2021. “The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care.” *The Lancet Digital Health* 3 (11): e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, et al. 2000. “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.” *Circulation* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Grabisch, Michel, and Marc Roubens. 1999. “An Axiomatic Approach to the Concept of Interaction Among Players in Cooperative Games.” *International Journal of Game Theory* 28 (4): 547–65. https://doi.org/10.1007/s001820050125.

Groenwold, Rolf H. H. 2020. “Informative Missingness in Electronic Health Record Systems: The Curse of Knowing.” *Diagnostic and Prognostic Research* 4: 8. https://doi.org/10.1186/s41512-020-00077-0.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of machine learning research, vol. 70: 1321–30.

Guo, Zhicheng, Cheng Ding, Duc Do, et al. 2025. “Improving Atrial Fibrillation Detection Using a Shared Latent Space for ECG and PPG Signals.” *Harvard Data Science Review* 7 (1).

Hannun, Awni Y., Pranav Rajpurkar, Masoumeh Haghpanahi, et al. 2019. “Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network.” *Nature Medicine* 25 (1): 65–69. https://doi.org/10.1038/s41591-018-0268-3.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–78. https://doi.org/10.1109/CVPR.2016.90.

Hettiarachchi, R., U. Haputhanthri, K. Herath, et al. 2021. “A Novel Transfer Learning-Based Approach for Screening Pre-Existing Heart Diseases Using Synchronized ECG Signals and Heart Sounds.” *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5. https://doi.org/10.1109/ISCAS51556.2021.9401093.

Hinton, Geoffrey E. 2002. “Training Products of Experts by Minimizing Contrastive Divergence.” *Neural Computation* 14 (8): 1771–800. https://doi.org/10.1162/089976602760128018.

Holm, Sture. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics* 6 (2): 65–70.

Hooker, Sara, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. “A Benchmark for Interpretability Methods in Deep Neural Networks.” *Advances in Neural Information Processing Systems* 32.

Huang, Shih-Cheng, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. 2020. “Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines.” *Npj Digital Medicine* 3: 136. https://doi.org/10.1038/s41746-020-00341-z.

Ismail Fawaz, Hassan, Benjamin Lucas, Germain Forestier, et al. 2020. “InceptionTime: Finding AlexNet for Time Series Classification.” *Data Mining and Knowledge Discovery* 34: 1936–62. https://doi.org/10.1007/s10618-020-00710-y.

Jacobs, Robert A., Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. “Adaptive Mixtures of Local Experts.” *Neural Computation* 3 (1): 79–87. https://doi.org/10.1162/neco.1991.3.1.79.

John, Arlene. 2025. “Sensor Fusion Using 1D-CNNs in Atrial Fibrillation Detection and Decision Support.” *2025 IEEE Medical Measurements and Applications (MeMeA)*. https://doi.org/10.1109/MeMeA65319.2025.11068069.

John, Arlene, Keshab K. Parhi, Barry Cardiff, and Deepu John. 2025. “MLFusion: Multilevel Data Fusion Using CNNs for Atrial Fibrillation Detection.” *Computers in Biology and Medicine* 189: 109901. https://doi.org/10.1016/j.compbiomed.2025.109901.

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, et al. 2016. “MIMIC-III, a Freely Accessible Critical Care Database.” *Scientific Data* 3: 160035. https://doi.org/10.1038/sdata.2016.35.

Kapoor, Sayash, and Arvind Narayanan. 2023. “Leakage and the Reproducibility Crisis in Machine-Learning-Based Science.” *Patterns* 4 (9): 100804. https://doi.org/10.1016/j.patter.2023.100804.

Karimi, Sajjad, Amit J. Shah, Gari D. Clifford, and Reza Sameni. 2025. *Cross-Learning Between ECG and PCG: Exploring Common and Exclusive Characteristics of Bimodal Electromechanical Cardiac Waveforms*. arXiv preprint arXiv:2506.10212.

Kazemnejad, Arsalan, Peiman Gordany, and Reza Sameni. 2021. *EPHNOGRAM: A Simultaneous Electrocardiogram and Phonocardiogram Database (Version 1.0.0)*. PhysioNet. https://doi.org/10.13026/tjtq-5911.

Kingma, Diederik P., and Jimmy Ba. 2015. “Adam: A Method for Stochastic Optimization.” *3rd International Conference on Learning Representations (ICLR)*.

Kittler, Josef, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. 1998. “On Combining Classifiers.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (3): 226–39. https://doi.org/10.1109/34.667881.

Kuncheva, Ludmila I. 2014. *Combining Pattern Classifiers: Methods and Algorithms*. 2nd ed. Wiley.

Li, H., X. Wang, C. Liu, et al. 2019. "Dual-Input Neural Network Integrating Feature Extraction and Deep Learning for Coronary Artery Disease Detection Using Electrocardiogram and Phonocardiogram." *IEEE Access* 7: 146457–69. https://doi.org/10.1109/ACCESS.2019.2943197.

Li, H., X. Wang, C. Liu, P. Li, and Y. Jiao. 2021. "Integrating Multi-Domain Deep Features of Electrocardiogram and Phonocardiogram for Coronary Artery Disease Detection." *Computers in Biology and Medicine* 138: 104914. https://doi.org/10.1016/j.compbiomed.2021.104914.

Li, J., L. Ke, Q. Du, X. Chen, and X. Ding. 2022. "Multi-Modal Cardiac Function Signals Classification Algorithm Based on Improved D-S Evidence Theory." *Biomedical Signal Processing and Control* 71: 103078. https://doi.org/10.1016/j.bspc.2021.103078.

Li, P., Y. Hu, and Z.-P. Liu. 2021. "Prediction of Cardiovascular Diseases by Integrating Multi-Modal Features with Machine Learning Methods." *Biomedical Signal Processing and Control* 66: 102474. https://doi.org/10.1016/j.bspc.2021.102474.

Li, Wei Peng, Joon Huang Chuah, Guo Jeng Tan, Chengyu Liu, and Hua-Nong Ting. 2025. "A Progressive Attention-Based Cross-Modal Fusion Network for Cardiovascular Disease Detection Using Synchronized Electrocardiogram and Phonocardiogram Signals." *PeerJ Computer Science* 11: e3038. https://doi.org/10.7717/peerj-cs.3038.

Li, Yanghao, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. 2018. "Adaptive Batch Normalization for Practical Domain Adaptation." *Pattern Recognition* 80: 109–17. https://doi.org/10.1016/j.patcog.2018.03.005.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. 3rd ed. Wiley.

Liu, C., D. Springer, Q. Li, et al. 2016. "An Open Access Database for the Evaluation of Heart Sound Algorithms." *Physiological Measurement* 37 (12): 2181–213. https://doi.org/10.1088/0967-3334/37/12/2181.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30: 4765–74.

Ma, Mengmeng, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. "Are Multimodal Transformers Robust to Missing Modality?" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18177–86.

Ma, Mengmeng, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. "SMIL: Multimodal Learning with Severely Missing Modality." *Proceedings of the AAAI Conference on Artificial Intelligence* 35: 2302–10. https://doi.org/10.1609/aaai.v35i3.16330.

Malone, Shane, Barry Cardiff, Deepu John, and Arlene John. 2025. “Signal-Quality-Aware Multisensor Fusion for Atrial Fibrillation Detection.” *Healthcare Technology Letters* 12 (1): e12121. https://doi.org/10.1049/htl2.12121.

Moody, Benjamin, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. 2020. *MIMIC-III Waveform Database Matched Subset (Version 1.0)*. PhysioNet. https://doi.org/10.13026/c2294b.

Moulaeifard, Mohammad, Philip J. Aston, Peter H. Charlton, and Nils Strodthoff. 2026. *Deriving Health Metrics from the Photoplethysmogram: Benchmarks and Insights from MIMIC-III-Ext-PPG*. arXiv preprint arXiv:2603.21832.

Moulaeifard, Mohammad, Peter H. Charlton, and Nils Strodthoff. 2026. *MIMIC-III-Ext-PPG: A PPG Benchmark Dataset for Cardiorespiratory Analysis (Version 1.1.0)*. PhysioNet. https://doi.org/10.13026/r6k1-xt76.

Moulaeifard, Mohammad, M. Kutscher, Philip J. Aston, et al. 2026. “MIMIC-III-Ext-PPG, a PPG-Based Benchmark Dataset for Cardiovascular and Respiratory Signal Analysis.” *Scientific Data* 13: 668. https://doi.org/10.1038/s41597-026-07335-8.

Naeini, Mahdi Pakdaman, Gregory F. Cooper, and Milos Hauskrecht. 2015. “Obtaining Well Calibrated Probabilities Using Bayesian Binning.” *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2901–7.

Neverova, Natalia, Christian Wolf, Graham Taylor, and Florian Nebout. 2016. “ModDrop: Adaptive Multi-Modal Gesture Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (8): 1692–706. https://doi.org/10.1109/TPAMI.2015.2461544.

Oliveira, J., F. Renna, P. D. Costa, et al. 2022. “The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification.” *IEEE Journal of Biomedical and Health Informatics* 26 (6): 2524–35. https://doi.org/10.1109/JBHI.2021.3137048.

Orphanidou, C., T. Bonnici, P. Charlton, D. Clifton, D. Vallance, and L. Tarassenko. 2015. “Signal-Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring.” *IEEE Journal of Biomedical and Health Informatics* 19 (3): 832–38. https://doi.org/10.1109/JBHI.2014.2338351.

Parcalabescu, Letitia, and Anette Frank. 2023. “MM-SHAP: A Performance-Agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks.” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2023.acl-long.223.

Paszke, Adam, Sam Gross, Francisco Massa, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” *Advances in Neural Information Processing Systems* 32: 8024–35.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.

Pepe, Margaret Sullivan. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pereira, Tania, Nate Tran, Kais Gadhoumi, et al. 2020. “Photoplethysmography Based Atrial Fibrillation Detection: A Review.” *Npj Digital Medicine* 3: 3. https://doi.org/10.1038/s41746-019-0207-9.

Perez, Marco V., Kenneth W. Mahaffey, Haley Hedlin, et al. 2019. “Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation.” *New England Journal of Medicine* 381 (20): 1909–17. https://doi.org/10.1056/NEJMoa1901183.

Pillai, Arvind, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. 2025. “PaPaGei: Open Foundation Models for Optical Physiological Signals.” *The Thirteenth International Conference on Learning Representations (ICLR)*.

Platt, John C. 1999. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.” In *Advances in Large Margin Classifiers*, edited by Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans. MIT Press.

Pollard, T., B. E. Moody, L. Lehman, et al. 2026. “PhysioNet as a Global Platform for Biomedical Research.” *Nature Health* 1 (8): 792–95. https://doi.org/10.1038/s44360-026-00096-z.

Reehana, S. K., and S. P. Siddique Ibrahim. 2025. “Multi-Class: Spectral-Spatial Temporal Pyramid Network and Multi-Class Classifier-Based Cardiovascular Disease Classification.” *Frontiers in Physiology* 16: 1650134. https://doi.org/10.3389/fphys.2025.1650134.

Reyna, M. A., N. Sadr, E. A. Perez Alday, et al. 2021. “Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021.” *Computing in Cardiology* 48. https://doi.org/10.23919/CinC53138.2021.9662687.

Rong, Yao, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. 2022. “A Consistent and Efficient Evaluation Strategy for Attribution Methods.” *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of machine learning research, vol. 162: 18770–95.

Roth, Gregory A., George A. Mensah, Catherine O. Johnson, et al. 2020. “Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update from the GBD 2019 Study.” *Journal of the American College of Cardiology* 76 (25): 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.

Saeb, Sohrab, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. 2017. “The Need to Approximate the Use-Case in Clinical Machine Learning.” *GigaScience* 6 (5): gix019. https://doi.org/10.1093/gigascience/gix019.

Samek, Wojciech, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. “Evaluating the Visualization of What a Deep Neural Network Has Learned.” *IEEE Transactions on Neural Networks and Learning Systems* 28 (11): 2660–73. https://doi.org/10.1109/TNNLS.2016.2599820.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–26. https://doi.org/10.1109/ICCV.2017.74.

Shapley, Lloyd S. 1953. “A Value for n-Person Games.” In *Contributions to the Theory of Games II*, edited by Harold W. Kuhn and Albert W. Tucker, vol. 28. Annals of Mathematics Studies. Princeton University Press.

Su, J. Q., and J. S. Liu. 1993. “Linear Combinations of Multiple Diagnostic Markers.” *Journal of the American Statistical Association* 88 (424): 1350–55.

Sun, Chengfa, Changchun Liu, Xinpei Wang, Yuanyuan Liu, and Shilong Zhao. 2024. “Coronary Artery Disease Detection Based on a Novel Multi-Modal Deep-Coding Method Using ECG and PCG Signals.” *Sensors* 24 (21): 6939. https://doi.org/10.3390/s24216939.

Sun, Chengfa, Xiaolei Liu, Changchun Liu, et al. 2024. “Enhanced CAD Detection Using Novel Multi-Modal Learning: Integration of ECG, PCG, and Coupling Signals.” *Bioengineering* 11 (11): 1093. https://doi.org/10.3390/bioengineering11111093.

Sun, Xu, and Weichao Xu. 2014. “Fast Implementation of DeLong’s Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves.” *IEEE Signal Processing Letters* 21 (11): 1389–93. https://doi.org/10.1109/LSP.2014.2337313.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. “Axiomatic Attribution for Deep Networks.” *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of machine learning research, vol. 70: 3319–28.

Szegedy, Christian, Wei Liu, Yangqing Jia, et al. 2015. “Going Deeper with Convolutions.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Tison, Geoffrey H., José M. Sanchez, Brandon Ballinger, et al. 2018. “Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch.” *JAMA Cardiology* 3 (5): 409–16. https://doi.org/10.1001/jamacardio.2018.0136.

Torres-Soto, Jessica, and Euan A. Ashley. 2020. “Multi-Task Deep Learning for Cardiac Rhythm Detection in Wearable Devices.” *Npj Digital Medicine* 3: 116. https://doi.org/10.1038/s41746-020-00320-4.

Van Calster, Ben, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. 2019. “Calibration: The Achilles Heel of Predictive Analytics.” *BMC Medicine* 17: 230. https://doi.org/10.1186/s12916-019-1466-7.

Van Gelder, Isabelle C. et al. 2024. “2024 ESC Guidelines for the Management of Atrial Fibrillation Developed in Collaboration with the European Association for Cardio-Thoracic Surgery (EACTS).” *European Heart Journal* 45 (36): 3314–414. https://doi.org/10.1093/eurheartj/ehae176.

Varma, Sudhir, and Richard Simon. 2006. “Bias in Error Estimation When Using Cross-Validation for Model Selection.” *BMC Bioinformatics* 7: 91. https://doi.org/10.1186/1471-2105-7-91.

Vickers, Andrew J., and Elena B. Elkin. 2006. “Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.” *Medical Decision Making* 26 (6): 565–74. https://doi.org/10.1177/0272989X06295361.

Wang, Weiyao, Du Tran, and Matt Feiszli. 2020. “What Makes Training Multi-Modal Classification Networks Hard?” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12695–705. https://doi.org/10.1109/CVPR42600.2020.01271.

Wolpert, David H. 1992. “Stacked Generalization.” *Neural Networks* 5 (2): 241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.

Youden, William J. 1950. “Index for Rating Diagnostic Tests.” *Cancer* 3 (1): 32–35.

Zeiler, Matthew D., and Rob Fergus. 2014. “Visualizing and Understanding Convolutional Networks.” *Computer Vision – ECCV 2014* (Cham), Lecture notes in computer science, vol. 8689: 818–33. https://doi.org/10.1007/978-3-319-10590-1_53.

Zhang, Haobo, Peng Zhang, Fan Lin, et al. 2024. “Co-Learning-Assisted Progressive Dense Fusion Network for Cardiovascular Disease Detection Using ECG and PCG Signals.” *Expert Systems with Applications* 238: 122144. https://doi.org/10.1016/j.eswa.2023.122144.