

When One View Isn't Enough: Multimodal Cardiac Representation Fusion for Sudden Cardiac Arrest Risk

Jarin Prity

jarinprity438@gmail.com

Metropolitan University <https://orcid.org/0009-0007-5512-8078>

Research Article

Keywords: sudden cardiac arrest, multimodal learning, electrocardiography, cardiac magnetic resonance, echocardiography

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11261393/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.

Abstract

Risk of sudden cardiac arrest may be expressed differently by anatomy, mechanical function and electrical activity, while studies that contain all three modalities for the same patient are uncommon. We developed CardiacGuard-Net, a modular deep-learning framework that learns modality-specific representations from cardiac magnetic resonance imaging (CMR), echocardiography and twelve-lead electrocardiography (ECG), then combines them for cardiac-arrest classification. The modality-specific models achieved mean Dice 0.953 for CMR segmentation, 3.8% mean absolute error for echocardiographic ejection-fraction estimation and AUC 0.938 for 24-h arrest prediction from ECG. For multimodal evaluation, 1,204 records were assembled into a matched benchmark using age band, sex, ejection-fraction band and diagnostic category because no dataset available to the study contained all three modalities for the same patients. On this benchmark, CardiacFusion-Net achieved AUC 0.947, compared with 0.920 for late fusion. Removing the ECG, CMR or echocardiographic branch reduced AUC by 0.035, 0.026 and 0.016, respectively. A single trained model also retained discrimination when modalities were withheld, with AUCs of 0.890 for ECG alone, 0.905 for MRI+echo, 0.921 for MRI + ECG and 0.931 for echo + ECG. The multimodal results are therefore best interpreted as an internal proof-of-concept benchmark of complementary information, not as patient-level clinical validation.

1 Introduction

Sudden cardiac arrest remains difficult to predict because risk can arise from several parts of the cardiac system at once. An electrocardiogram (ECG) captures electrical activity, echocardiography describes mechanical function, and cardiac magnetic resonance (CMR) provides detailed structural information. These measurements are related, but they are not interchangeable. A heart can have preserved global function while still showing structural or electrical features associated with arrhythmic risk [3, 7, 9, 2].

This creates a practical challenge for multimodal machine learning. The most informative datasets would contain several measurements from the same patient at approximately the same stage of disease, yet such cohorts are difficult to assemble. Recent deep-learning work has shown that ECG waveforms can contain information relevant to sudden cardiac death beyond traditional measures [12], while multimodal models have begun to combine CMR with clinical and echocardiographic information for arrhythmic-risk assessment in specific patient groups [8]. Multimodal learning has increasingly been proposed as a way to combine complementary biomedical signals [16, 17]. Recent deep-learning work has also shown that ECG waveforms can contain information relevant to sudden cardiac death beyond traditional measures [12], while multimodal models have begun to combine CMR with clinical and echocardiographic information for arrhythmic-risk assessment in specific patient groups [8]. These studies strengthen the case for multimodal modelling, but they also highlight how dependent performance is on the population, endpoint and availability of genuinely linked measurements.

The present study asks a narrower question: can useful information learned separately from CMR, echocardiography and ECG be brought together into a single model, and does each view add information to the others? We designed CardiacGuard-Net as a modular framework with three modality-specific networks and a final fusion network. The model can also operate when one or more modalities are unavailable, allowing its robustness to incomplete inputs to be examined without changing the trained system.

A key limitation must be made explicit from the outset. No accessible dataset contained contemporaneous CMR, echocardiography and ECG for the same patients. We therefore constructed a matched multimodal benchmark rather than claiming a naturally paired cohort. The matching used age band, sex, ejection-fraction band and diagnostic category. This design permits an internal comparison of learned representations and modality contribution, but it cannot establish how the three signals interact within individual patients in routine care. The study is consequently framed as a biomedical-engineering proof of concept that motivates, rather than replaces, evaluation in genuinely paired multimodal cohorts.

Our objectives were to (1) build useful modality-specific cardiac representations, (2) test whether combining them improves discrimination in the matched benchmark, and (3) examine which modalities and fusion components are most important when information is added or withheld.

2 Methods

2.1 Framework and study design

CardiacGuard-Net was developed in two stages. First, CardiacMRI-Net, EchoCardio-Net and ECG-ArrhythmiaNet were trained independently on their corresponding datasets. Their learned encoders were then frozen and used as feature extractors for CardiacFusion-Net. This separation allowed the representation-learning stage to use the larger single-modality datasets while keeping the multimodal training stage restricted to the available matched benchmark.

2.2 Datasets

Four sources were used. ACDC provided 100 CMR examinations with expert annotations [1]. EchoNet-Dynamic provided 10,030 apical four-chamber echocardiographic videos with reported ejection fraction [13], and CAMUS provided 500 patients for cross-dataset assessment of cardiac structure segmentation [10]. The ECG cohort of Kwon et al. [11] comprised 47,505 ten-second twelve-lead ECGs from 25,672 patients and included a 24-h cardiac-arrest endpoint.

Because no available source contained all three modalities for the same patient, 1,204 records were assembled into a matched multimodal benchmark without replacement. Matching was performed on age band, sex, ejection-fraction band and diagnostic category. These matched sets should not be interpreted as contemporaneous patient-level multimodal observations; rather, they provide a controlled

benchmark in which the three learned representations can be compared under the same broad clinical strata.

Table 1
Datasets and their role in the study

Dataset	Size	Role in this study
ACDC	100 studies; 1,860 annotated slices	Training and evaluation of CMR segmentation
EchoNet-Dynamic	10,030 videos	Training and evaluation of echocardiographic EF estimation
CAMUS	500 patients	Cross-dataset assessment of cardiac segmentation
Kwon ECG cohort	47,505 ECGs; 25,672 patients	Training and evaluation of 24-h arrest prediction
Matched multimodal benchmark	1,204 matched record sets	Evaluation of CardiacFusion-Net and modality subsets

2.3 Modality-specific models

The three modality-specific networks were designed to produce clinically meaningful intermediate representations rather than to serve as the paper’s primary contribution.

CardiacMRI-Net uses an encoder–decoder segmentation architecture, in the general family of architectures established for biomedical image segmentation [14], with a self-adaptive attention bottleneck that combines spatial, channel and patch-level information. The resulting segmentations are used to estimate ventricular volumes and related functional measures.

EchoCardio-Net uses a spatiotemporal video network with an LV segmentation head and an ejection-fraction regression head, following the broader video-based learning approach used by EchoNet-Dynamic [13]. Ejection fraction was estimated from detected cardiac cycles and summarized across beats.

ECG-ArrhythmiaNet uses a one-dimensional residual network for twelve-lead ECGs with auxiliary physiological prediction tasks, drawing on the residual-learning paradigm [6]. Lead dropout during training permits evaluation with fewer leads without retraining.

2.4 Multimodal fusion

CardiacFusion-Net receives the frozen representations from the three modality encoders. Each representation is refined before the model considers relationships between modalities using multi-head attention [18], after which a learned modality gate combines the available representations and produces an arrest probability. During training, modality dropout randomly removes individual branches so that the same model can also be evaluated when only a subset of modalities is available.

2.5 Training and evaluation

All dataset splits were performed at the patient level where patient identifiers were available. ACDC used five-fold stratified cross-validation with the official test set held out. EchoNet-Dynamic used its official train/validation/test split, while CAMUS was used only for cross-dataset testing. The ECG cohort used a 70/15/15 train/validation/test split. The matched multimodal benchmark was divided 60/20/20 into training, calibration and test partitions. Temperature scaling was used only after model fitting and on the designated calibration partition [5]. Architectural choices and hyperparameters were fixed before final test evaluation.

Single-modality performance was assessed using Dice/IoU for segmentation, mean absolute error for ejection fraction, and AUC for ECG classification. The multimodal benchmark was evaluated using AUC-ROC, sensitivity, specificity, F1, calibration slope and component ablations. Results are reported over five random seeds. Pairwise AUC comparisons used DeLong tests [4] with Holm–Bonferroni correction.

3 Results

3.1 Single-modality performance

CardiacMRI-Net achieved a mean Dice score of 0.953 on the ACDC test set (LV 0.952, myocardium 0.955, RV 0.953). Performance remained high across the reported anatomical categories, and the attention module improved Dice relative to the attention-free baseline. Derived ejection fraction was within 2.1 percentage points of expert tracing.

EchoCardio-Net estimated echocardiographic ejection fraction with a mean absolute error of 3.8% on EchoNet-Dynamic. Performance transferred to the CAMUS data, although image quality affected segmentation performance.

ECG-ArrhythmiaNet achieved an AUC of 0.938 for 24-h arrest prediction from twelve-lead ECGs. With fewer leads, AUC decreased to 0.921 using six limb leads and 0.894 using lead I alone (Table 2). These results are presented as robustness experiments rather than evidence for wearable deployment.

Table 2
ECG arrest prediction under reduced lead availability

Metric	12 leads	6 limb leads	Lead I only
AUC-ROC	0.938	0.921	0.894
Specificity at 90% sensitivity	0.87	0.84	0.79
Positive predictive value	0.12	0.09	0.07

3.2 Multimodal fusion

On the matched multimodal benchmark, CardiacFusion-Net achieved an AUC of 0.947. This was higher than early fusion (0.901) and late fusion (0.920), and higher than the strongest single-modality result evaluated on the same benchmark (ECG, 0.890). The result is an internal benchmark comparison; because the three records in a matched set were not naturally paired within patients, the AUC should not be interpreted as prospective patient-level clinical performance.

The modality subset analysis showed a progressive increase in discrimination as views were added: MRI alone, 0.840; echo alone, 0.870; ECG alone, 0.890; MRI+echo, 0.905; MRI + ECG, 0.921; echo + ECG, 0.931; and all three modalities, 0.947 (Fig. 2). Thus, the benefit of multimodality was visible even without relying solely on the full fusion architecture.

Component ablation further showed that removing the ECG, MRI or echocardiographic branch reduced AUC by 0.035, 0.026 and 0.016, respectively. Removing cross-modal attention reduced AUC by 0.012, whereas removing modality gates reduced AUC by 0.009. These results suggest that the three views provide complementary information, while the additional value of explicit cross-modal interaction is smaller than the value of adding the modalities themselves. We do not interpret the difference as a formal decomposition of complementarity and interaction because the matched benchmark does not capture true patient-specific cross-modal dependence.

Temperature scaling produced a calibration slope of 0.94 on the multimodal benchmark. The model also retained useful discrimination when modalities were withheld, indicating that the modular design can support robustness testing without retraining separate models for every modality subset.

3.3 Interpretability and computational characteristics

Grad-CAM [15] visualizations highlighted cardiac regions and waveform segments that were anatomically or physiologically plausible within the corresponding modality. The ECG counterfactual analysis also recovered terminal QRS/ST-T morphology, including a pattern in lead aVL, consistent with the type of ECG morphology highlighted in recent deep-learning work on sudden cardiac death [12]. These findings are qualitative and are not presented as independent mechanistic validation.

End-to-end inference across all three modalities required 1.73 s per record on one GPU, while the ECG-only pathway required 0.05 s. These measurements describe the current implementation and are not intended as clinical deployment benchmarks.

4 Discussion

The main finding is straightforward: when information from anatomy, mechanical function and electrical activity was brought together in a matched multimodal benchmark, the combined representation discriminated better than any individual view and better than conventional early- or late-fusion baselines. Each modality contributed, and performance increased across the reported non-empty modality subsets.

The most important qualification is equally straightforward. The multimodal benchmark was constructed by matching records from separate sources rather than by following patients who underwent all three studies. The resulting AUC therefore measures how well these representations work together under the matching scheme used in this study. It does not establish that the same gain would occur when CMR, echocardiography and ECG are measured in the same person at the same clinical time point. This distinction is central to interpreting the paper and is why the results are presented as a proof of concept rather than a clinically validated risk model.

The ablation analysis provides a useful engineering observation. Removing any of the three modality branches lowered performance, with the largest decrease after removing ECG. Removing cross-modal attention had a smaller effect than removing an entire modality branch, suggesting that much of the observed gain comes from combining complementary representations rather than from the fusion module learning a large amount of additional interaction structure. Because the matched benchmark weakens patient-specific correlation, however, this observation should not be treated as a general statement about how much multimodal benefit is caused by interaction.

The ability to evaluate any non-empty subset with one trained model is another practical property of the framework. It is useful for studying incomplete inputs and for comparing the value of different examinations without building a separate predictor for every combination. The present results support this as a modelling capability, not as evidence that a reduced-modality pathway is ready for clinical use.

These results also place the work in a rapidly developing field. Recent ECG-only work has shown that deep learning can recover clinically meaningful information about sudden cardiac death from surface waveforms [12], and recent multimodal work in hypertrophic cardiomyopathy has demonstrated the value of combining CMR with other clinical information for arrhythmic-risk prediction [8]. Our contribution is narrower: a modular framework that combines independently learned cardiac representations and explicitly examines how much each view adds in a setting where fully paired three-modality data are not available.

4.1 Limitations

Several limitations define the scope of the conclusions. First, and most importantly, the multimodal benchmark was created by record matching rather than natural patient-level pairing. The fusion results therefore do not provide external clinical validation and should not be compared directly with patient-level prospective models as though the evaluation settings were identical. Second, the single-modality datasets differ in acquisition protocol, population and endpoint definition, and these differences carry into the matched benchmark. Third, the study is retrospective, and the ECG cohort was obtained from the original investigators rather than redistributed by the present author, which limits independent reproduction of the exact ECG experiment. Finally, interpretability analyses are qualitative and the computational timing was measured on a single implementation.

5 Conclusion

When one cardiac view is incomplete, another may carry information that remains useful. In our matched multimodal benchmark, combining CMR, echocardiography and ECG representations produced higher discrimination than individual modalities and conventional fusion baselines, and each modality made a measurable contribution. A single trained model also retained discrimination when modalities were removed. These findings support multimodal representation fusion as a useful biomedical-engineering strategy for cardiac risk modelling, while the constructed nature of the multimodal benchmark means that the results should be regarded as proof-of-concept evidence awaiting validation on genuinely paired patient data.

Declarations

CRediT authorship contribution statement

Jarin Alam Prity: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration.

Ethics statement

This study did not involve the collection of new data from human participants. The analyses used secondary, de-identified data from the publicly available ACDC, EchoNet-Dynamic and CAMUS datasets and the retrospective ECG cohort of Kwon et al. [11]. The ECG cohort was obtained from the original investigators under the conditions governing the original study. No new participant recruitment or intervention was performed.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The author declares no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

ACDC, EchoNet-Dynamic and CAMUS are available from their respective data providers. The ECG cohort used in this study was obtained from the corresponding authors of Kwon et al. [11] on reasonable request and is subject to the original data-use terms; it is not redistributed here. The matched benchmark records are derived from these source datasets and are not redistributed in a way that would bypass their respective access conditions.

Code availability

The code used for the analyses is available from the corresponding author on reasonable request, subject to any restrictions associated with the source data and the author's former research environment.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During preparation of the manuscript, an AI-assisted writing tool was used to help restructure and edit text derived from the author's draft. The author reviewed and approved the manuscript and takes responsibility for its content. No experimental results, patient data or figures were generated by the AI tool.

Acknowledgements

The author thanks Dr. Swaty Haque for sharing datasets that supported this work.

Author biography. Jarin Alam Prity is a computer science researcher working on deep learning, medical imaging, physiological signal analysis and multimodal biomedical data.

References

1. Olivier Bernard A, Lalande C, Zotti F, Cervenansky X, Yang, Pheng-Ann H (2018) Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans Med Imaging* 37(7):2514–2525
2. Alfred E, Buxton KL, Lee GE, Hafley (2007) Josephson. Limitations of ejection fraction for prediction of sudden death risk in patients with coronary artery disease. *J Am Coll Cardiol* 50(12):1150–1157
3. Sumeet S, Chugh K, Reinier C, Teodorescu A, Evanado E, Kehr MA, Samara R, Mariani (2008) Karen Gunson, and Jonathan Jui. Epidemiology of sudden cardiac death: Clinical and research implications. *Prog Cardiovasc Dis* 51(3):213–228
4. Elizabeth R, DeLong DM, DeLong, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44(3):837–845
5. Guo C, Pleiss G (2017) Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778
7. Christopher M, Kramer Jörg, Barkhausen C, Bucciarelli-Ducci SD, Flamm, Raymond J, Kim, Nagel E (2020) Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *J Cardiovasc Magn Reson* 22(1):17

8. Changxin Lai M, Yin EG, Kholmovski DM, Popescu D-Y, Lu E, Scherer E, Binka SL, Zimmerman J, Chrispin AG, Hays DM, Phelan M (2025) Roselle Abraham, and Natalia A. Trayanova. Multimodal ai to forecast arrhythmic death in hypertrophic cardiomyopathy. *Nat Cardiovasc Res* 4:891–903
9. Roberto M, Lang LP, Badano V, Mor-Avi J, Afilalo (2015) Anderson Armstrong, and Laura Ernande. Recommendations for cardiac chamber quantification by echocardiography in adults. *J Am Soc Echocardiogr* 28(1):1–39
10. Sarah Leclerc E, Smistad J, Pedrosa A, Østvik (2019) Frederic Cervenansky, and Florian Espinosa. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Trans Med Imaging* 38(9):2198–2210
11. Kwon Jmyoung, Kim K-H, Jeon K-H, Lee SY (2020) Jinsik Park, and Byung-Hee Oh. Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. *Scand J Trauma Resusc Emerg Med* 28(1):98
12. Ziad Obermeyer A, Schubert J, Ross S, Mullainathan, Lingman M (2026) An ecg biomarker for sudden cardiac death discovered with deep learning. *Nature* 655(8121):210–218
13. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, Heidenreich PA, Harrington RA, Liang DH, Ashley EA, Zou JY (2020) Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 580(7802):252–256
14. Olaf Ronneberger P, Fischer, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241
15. Ramprasaath R, Selvaraju M, Cogswell A, Das R, Vedantam D, Parikh, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626
16. Sören R, Stahlschmidt B, Ulfenborg, Synnergren J (2022) Multimodal deep learning for biomedical data fusion: A review. *Brief Bioinform* 23(2):bbab569
17. Jing R, Teoh J, Dong X, Zuo KW, Lai (2024) Khairunnisa Hasikin, and Xiang Wu. Advancing healthcare through multimodal data fusion: A comprehensive review. *PeerJ Comput Sci* 10:e2298
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN (2017) Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008

Figures

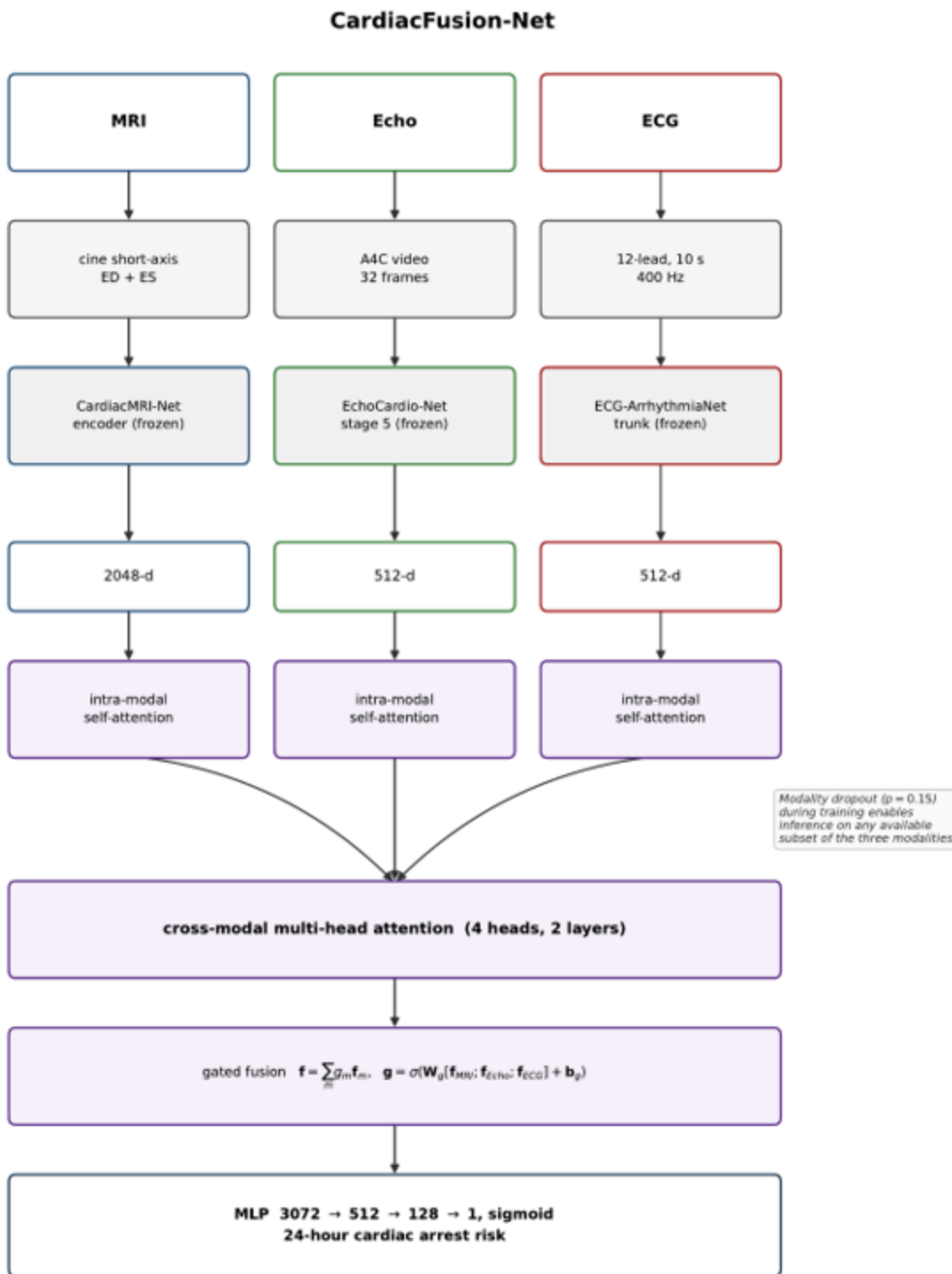


Figure 1

CardiacGuard-Net combines separate representations from CMR, echocardiography and ECG. The modality-specific encoders are trained first and then frozen during multimodal training. A cross-modal fusion stage integrates the representations, while modality dropout allows the same trained fusion model to be evaluated when one or more modalities are unavailable.

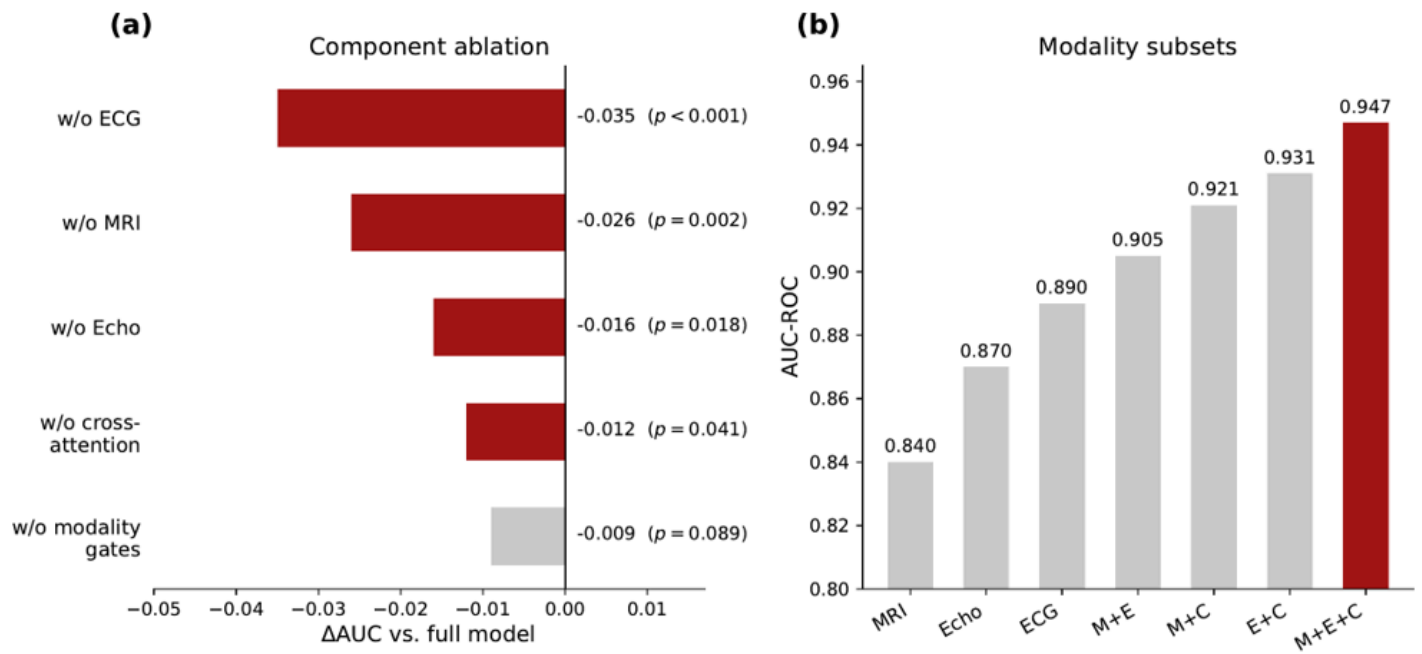


Figure 2

Component and modality-subset analysis. Removing an individual modality reduced AUC, with the largest decrease after removing ECG. The right panel shows performance for each non-empty modality subset in the matched multimodal benchmark.