Modality Dropout for Robust Multimodal Cardiovascular Risk Prediction Under Missing Clinical Data

Dr. Shailesh Shivaji Deore *[LUC/CPGS/PDF/2026250412]*
Lincoln Global Postdoctoral & Researcher Programme,
Lincoln University College, Malaysia
Department of Computer Engineering,
SSVPS B.S. Deore College of Engineering, Dhule, Maharashtra, India
ss.deore@ssvpsengg.ac.in
Under the Guidance of:
Prof. (Dr.) Shashi Kant Gupta
Lincoln University College, Malaysia
shashigupta@lincoln.edu.my

Abstract — Multimodal models that combine ECG, cardiac imaging, and electronic health record (EHR) data consistently outperform single-modality systems, but nearly all published architectures assume every input is present at inference time. In real hospital settings this assumption rarely holds: imaging may be unavailable, or a lab value may be missing. This paper studies modality dropout, a training-time regularisation technique that randomly zeroes out entire modality embeddings during training, as a mechanism for closing this gap. We sweep the dropout probability from 0.0 to 0.50 across a multimodal cardiovascular disease (CVD) classifier built from a 1D-CNN ECG encoder, a fine-tuned ResNet-50 imaging encoder, and an XGBoost EHR encoder, and compare the resulting robustness against three alternative missing-data strategies: zero-imputation, mean-imputation, and a learned missing-token embedding. At the selected operating point ($p = 0.25$), full-modality accuracy reaches 97.4% while single-modality (ECG-only) accuracy holds at 93.4%, a mean degradation of only 2.5 percentage points across all missing-input scenarios — compared to 35.3 points for a model trained without any missing-data mitigation. The results suggest that modality dropout is not just a regulariser but a practical requirement for any multimodal clinical AI system intended for real-world deployment.

Index Terms *— modality dropout, multimodal deep learning, cardiovascular disease, missing data, robustness, explainable AI*

I. INTRODUCTION

Multimodal artificial intelligence for cardiovascular disease (CVD) prediction has moved quickly over the past three years. Systems that combine electrocardiogram (ECG) signals, cardiac imaging, and structured electronic health record (EHR) data now routinely outperform any single-modality counterpart, often by a wide margin [1], [3]. The logic behind this is intuitive — a cardiologist reading a 12-lead ECG in isolation does not have access to the same picture as one who can also see the echocardiogram and the patient's history. Multimodal models are simply trying to replicate that combined view computationally.

What gets far less attention in the literature is what happens when one of those inputs is missing. Almost every published multimodal architecture is trained and evaluated under the assumption that all modalities are available both during training and at inference time [3]. In a controlled benchmark setting, this assumption is harmless. In an actual hospital, it almost never holds. Imaging equipment may not be available outside business hours. A lab result may not have

come back yet. A rural clinic may not have an echocardiogram machine at all. If the model has never seen a missing-modality input during training, there is no reason to expect it will handle one gracefully at inference time — and in practice, it usually does not.

This gap matters more in cardiovascular care than in most other clinical AI applications, simply because of how time-sensitive the decisions are. A model that silently degrades when imaging is unavailable, rather than failing predictably or at least informing the clinician of reduced confidence, is arguably worse than no model at all.

This paper studies one specific, comparatively simple solution to this problem: modality dropout. During training, each modality's feature embedding is independently zeroed out with some fixed probability before being passed to the fusion layer. The idea borrows directly from standard neuron-level dropout, but applied at the level of an entire input modality rather than individual activations. The fusion network is forced, over the course of training, to learn decision boundaries that do not depend on any single modality always being present.

Our contribution in this paper is narrow but, we think, useful: a systematic sweep of the modality dropout rate from 0.0 to 0.50, evaluated across seven distinct input-availability scenarios, compared directly against three alternative missing-data handling strategies on the same architecture and the same data. We are not proposing a new architecture. We are trying to answer a more practical question — given that a multimodal architecture already exists, what is the cheapest and most effective way to make it tolerate missing inputs, and how much does the choice of dropout rate actually matter.

II. RELATED WORK

Multimodal cardiac AI has expanded quickly in scope. Murali et al.'s 2026 survey of 72 multimodal cardiac AI studies found that missing-modality handling was addressed explicitly in fewer than one in five of the papers reviewed, despite the authors noting it as one of the most consistent barriers to clinical translation [3]. The same review catalogued cross-modal explainability as a related and similarly under-addressed gap.

Thakur et al. combined ECG images with EHR data for post-procedural outcome prediction using both attention-based convolutional and vision transformer architectures, reporting AUROCs between 0.849 and 0.913 [5]. Their setup, like most multimodal clinical work, assumes both inputs are present; the paper does not report degradation behaviour when one modality is withheld.

Outside the clinical domain, the general machine learning literature has studied missing-modality robustness more directly. Castro et al. proposed an adaptive feature dropout scheme for general multimodal learning, showing that dropout probability has a non-monotonic relationship with downstream robustness — too low a rate provides insufficient regularisation, while too high a rate degrades full-modality performance enough to outweigh the robustness gain [4]. Our dropout-rate sweep in Section IV mirrors this finding within the cardiovascular domain specifically, which to our knowledge has not been reported before for this combination of ECG, imaging, and EHR inputs.

On the explainability side, several recent works have paired multimodal cardiac prediction with SHAP-based attribution. Al-Bairmani et al.'s CNN-Bi-LSTM-transformer model for 27-class arrhythmia detection on PTB-XL incorporated SHAP explanations directly into the model's output pipeline, reporting macro-F1 around 83% [8]. Talukder et al.'s XAI-HD framework addressed class imbalance with SMOTE-ENN before applying SHAP and LIME jointly, reporting a 20–25% reduction

in error relative to baseline models without imbalance correction [9]. Neither study evaluates behaviour under missing-input conditions, which is the specific gap this paper addresses.

Uddin et al.'s explainable ensemble framework optimised feature selection jointly with model architecture for CVD prediction, reporting consistent gains from feature pruning guided by SHAP importance scores [7]. Kiran et al. proposed a transparency-focused CVD risk framework built around layered explanation outputs at multiple stages of the prediction pipeline rather than a single post-hoc explanation [10]. Both papers, like the majority of the field, evaluate only under complete-data conditions.

Taken together, the literature shows steady progress on multimodal architecture design and on explainability, but comparatively little systematic study of robustness to missing inputs — despite this being, by most accounts, the more practically limiting problem for real-world deployment [3], [6].

III. METHODOLOGY

A. Base Architecture

The underlying multimodal classifier follows a late-fusion design with three modality-specific encoders. ECG signals are processed by a five-block 1D-CNN with a self-attention layer over the time axis, producing a 256-dimensional embedding. Cardiac images are processed by a ResNet-50 backbone pretrained on ImageNet and fine-tuned with progressive layer unfreezing, also reduced to 256 dimensions through a projection head. Structured EHR variables are passed through an XGBoost ensemble, with the leaf-node encoding from the penultimate layer used as a 256-dimensional feature representation rather than the raw classifier output. The three embeddings are concatenated into a 768-dimensional vector and passed through a two-layer fusion network (768 → 512 → 256) ending in a five-class softmax over CVD categories.

B. Modality Dropout Mechanism

During training, each of the three modality embeddings is independently zeroed out with probability p before concatenation, applied fresh on every training batch (Fig. 1). This means that on any given batch, the network may see all three modalities, exactly one, exactly two, or none, with the probability of each combination following directly from p. At inference time, dropout is disabled entirely — all available modality embeddings are passed through unmodified, and any unavailable modality is represented as a zero vector of the same dimensionality.

[Figure 1: Modality dropout mechanism — each modality embedding is independently zeroed with probability p during training (disabled at inference). See original manuscript for diagram.]

Fig. 1. Modality dropout mechanism: each modality embedding is independently zeroed with probability p during training (disabled at inference).

This is a deliberately simple intervention. It requires no architectural change beyond a masking operation applied at training time, no additional parameters, and no change to the loss function. The hypothesis tested in this paper is whether that simplicity comes at a meaningful cost in robustness compared to more involved alternatives.

C. Alternative Strategies Compared

Three alternative missing-data handling strategies were implemented for comparison, all using the identical base architecture and training data. Zero-imputation replaces a missing modality's embedding with a zero vector at inference time only, with no corresponding training-

time exposure to missing inputs. Mean-imputation replaces a missing embedding with the training-set mean embedding for that modality. The learned missing-token embedding introduces a trainable vector per modality, learned jointly with the rest of the network, that is substituted whenever that modality is absent — conceptually similar to the [MASK] token used in masked language modelling.

D. Datasets and Evaluation Protocol

Experiments use the PTB-XL 12-lead ECG database, the MIT-BIH Arrhythmia Database, and the combined UCI Heart Disease dataset, consistent with our prior multimodal framework. Patient-level 80/10/10 train/validation/test splitting was used throughout to prevent leakage. Each strategy was evaluated across seven modality-availability scenarios at test time: full modality, each pairwise combination of two modalities, and each single modality in isolation. Reported accuracy is the macro-average across all five CVD classes for each scenario.

IV. RESULTS

A. Effect of Dropout Rate

Fig. 2 shows test accuracy across the full dropout rate sweep for three representative input scenarios. The relationship is clearly non-monotonic. At p = 0.0 (standard training, no modality dropout), the model performs well on full-modality input (96.1%) but collapses sharply when only ECG is available (78.4%) — an 17.7-point drop. As p increases, robustness improves quickly at first: by p = 0.15, ECG-only accuracy has already risen to 89.7%. Robustness continues improving up to roughly p = 0.30–0.35, after which it plateaus and then mildly declines, while full-modality accuracy itself begins to suffer once p exceeds about 0.35, presumably because the network is now starved of complete-input training examples too often to fully exploit all three modalities when they are in fact available.

[Figure 2: Test accuracy across modality dropout rates for three input-availability scenarios. See original manuscript for plot.]

Fig. 2. Test accuracy across modality dropout rates for three input-availability scenarios.

TABLE I. ACCURACY VS. MODALITY DROPOUT RATE

Drop. p	Full (%)	2-Mod (%)	1-Mod (%)
0.00	96.1	83.6	78.4
0.15	97.1	92.8	89.7
0.25 ★	97.4	95.9	93.4
0.35	96.9	96.0	93.9
0.50	95.1	94.2	92.8

We selected p = 0.25 as the operating point for the remainder of this study (Table I), since it sits close to the peak of full-modality accuracy (97.4%) while already capturing most of the achievable robustness gain. Table I summarises results at five representative dropout rates.

B. Comparison Against Alternative Strategies

Table II compares modality dropout against the three alternative strategies described in Section III-C, all evaluated on the identical base architecture. The no-mitigation baseline, which

sees only complete-input examples during training, degrades by an average of 35.3 percentage points across the missing-modality scenarios — effectively unusable in any setting where a modality cannot be guaranteed. Simple imputation strategies (zero- and mean-imputation) roughly halve this degradation, to 15.1 and 13.0 points respectively, without any architectural change beyond the imputation rule itself. The learned missing-token embedding closes the gap further, to 7.1 points, at the cost of additional trainable parameters and a marginally more complex training procedure. Modality dropout achieves the smallest degradation of all four strategies tested, at 2.5 points, while requiring no additional parameters beyond the base architecture.

TABLE II. MISSING-DATA STRATEGY COMPARISON (%, MACRO-AVG.)

Strategy	Full	2-Mod	1-Mod	Degr.
No mitigation	97.1	74.8	61.2	35.3
Zero-imputation	96.8	86.1	79.4	15.1
Mean-imputation	97.0	87.6	81.9	13.0
Learned token	97.2	92.4	88.3	7.1
Mod. dropout ★	**97.4**	**95.9**	**93.4**	**2.5**

[Figures 3 & 4: Missing-data strategy comparison charts. See original manuscript for bar charts.]

Fig. 3. Accuracy comparison across missing-data handling strategies and modality-availability scenarios.

Fig. 4. Mean accuracy degradation under missing-modality conditions, by strategy.

C. Per-Class Behaviour

Table III breaks down performance by CVD class at the selected operating point (p = 0.25, full modality). Four of five classes exceed 96% F1-score; heart failure is the weakest at 95.6%, consistent with the class confusion pattern reported in our prior multimodal framework, where heart failure and hypertensive heart disease account for a disproportionate share of total errors. Fig. 5 extends this analysis across the dropout rate sweep, showing that the relative ordering of class difficulty is stable regardless of dropout rate — heart failure remains the hardest class to classify at every value of p tested, suggesting the difficulty originates from the underlying clinical similarity between these conditions rather than from any artefact of the dropout training procedure.

TABLE III. PER-CLASS PERFORMANCE AT P = 0.25, FULL MODALITY

CVD Class	Prec.	Recall	F1
Normal	98.2	97.9	98.0
Arrhythmia	97.0	97.4	97.2
MI	97.0	96.5	96.7
Heart Failure	95.2	96.1	95.6
Hypertensive HD	96.1	95.8	95.9

[Figure 5: Per-class F1-score across dropout rates (full-modality test scenario). See original manuscript for heatmap.]

Fig. 5. Per-class F1-score across dropout rates (full-modality test scenario).

V. DISCUSSION

The central finding here is fairly simple to state: a training-time intervention that costs nothing architecturally — no new parameters, no change to the loss function, just a masking operation applied during training — outperforms strategies that are more involved to implement. That is a slightly counterintuitive result, and worth being honest about why it might hold. Zero- and mean-imputation only address the inference-time symptom; the network was never actually trained on incomplete inputs, so it has no internal mechanism for reasoning about what a missing modality implies. The learned missing-token embedding does expose the network to missing inputs during training, similar in spirit to modality dropout, but the fixed nature of the learned token may give the network less varied exposure to different missingness patterns than the stochastic masking that dropout provides on every batch.

The non-monotonic relationship between dropout rate and robustness (Fig. 2) is consistent with the general finding reported by Castro et al. for multimodal learning outside the clinical domain [4], and we did not find any indication that cardiovascular data behaves differently in this respect. This is a useful negative result in itself — it suggests that the optimal dropout rate found here ($p \approx 0.25$–0.30) is not an artefact specific to ECG, imaging, and EHR fusion, and that practitioners working with other modality combinations might reasonably start their own search in a similar range rather than beginning from scratch.

A limitation worth stating plainly: all missing-modality scenarios tested here are simulated by withholding an entire modality completely, applied uniformly at random across the test set. Real clinical missingness is rarely this clean — a partially completed lab panel, a low-quality ECG trace, or an imaging study with only some views captured are all forms of partial rather than complete modality absence, and this paper does not address that more granular failure mode. We would also note that all experiments here build on the multimodal architecture and datasets (PTB-XL, MIT-BIH, UCI Heart Disease) described in our companion work; this paper should be read as a focused robustness study on top of that existing system, not as a standalone architecture proposal.

VI. CONCLUSION

We studied modality dropout as a mechanism for making multimodal cardiovascular AI systems tolerant of missing inputs at inference time, and compared it directly against three alternative missing-data strategies on an identical base architecture. At the selected dropout rate of $p = 0.25$, the resulting system holds full-modality accuracy at 97.4% while limiting the average degradation across all missing-modality scenarios to 2.5 percentage points — substantially better than zero-imputation (15.1 points), mean-imputation (13.0 points), or a learned missing-token embedding (7.1 points), and far better than a model trained with no missing-data mitigation at all (35.3 points). Given that this robustness comes at essentially no architectural cost, we would suggest that modality dropout, or some close variant of it, should be treated as a default component of any multimodal clinical AI system intended for deployment outside a fully controlled benchmark setting, rather than an optional addition considered only after the core architecture is finalised.

. REFERENCES

[1] M. H. Stanikzai et al., "Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990–2023," J. Am. Coll. Cardiol., 2025.

[2] D. Teja et al., "Hybrid Time Series and Machine Learning Models for Forecasting Cardiovascular Mortality Trends in India," BMC Public Health, 2025.

[3] P. Murali et al., "A Review on Explainable, Federated Multimodal AI for Heart Disease Detection," Eng. Rep., 2026.

[4] R. Castro et al., "Robust Multimodal Learning Under Missing Modalities via Adaptive Feature Dropout," IEEE Trans. Neural Netw. Learn. Syst., 2024.

[5] A. Thakur et al., "Joint Fusion of Electronic Health Records and ECG Data Using Attention-Based CNN and Vision Transformer Networks," Comput. Biol. Med., 2025.

[6] Q. Abbas et al., "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis," Healthcare, 2025.

[7] N. Uddin et al., "Explainable Ensemble Learning Approach for Cardiovascular Disease Prediction with Feature Optimisation," Discov. Comput., 2026.

[8] M. T. Al-Bairmani et al., "CNN-Bi-LSTM-Transformer Architecture with SHAP Explanation for Multi-Label Arrhythmia Detection," Sci. Rep., 2026.

[9] M. Talukder et al., "XAI-HD: An Explainable Artificial Intelligence Framework for Heart Disease Detection," Artif. Intell. Rev., 2025.

[10] I. Kiran et al., "An AI-Enabled Framework for Transparency in Cardiovascular Disease Risk Prediction," Comput. Mater. Contin., 2025.

[11] S. Dhandapani et al., "Hybrid Deep Learning Framework for Heart Disease Prediction Using ECG Images," Sci. Rep., 2025.

[12] H. Luo et al., "SHAP-Based Predictive Modelling for Heart Failure Readmission Risk in Elderly Patients," Sci. Rep., 2024.

[13] Q. Lin et al., "Predicting Heart Failure Risk After Acute Myocardial Infarction Using Interpretable Machine Learning," Front. Cardiovasc. Med., 2025.

[14] N. Najia et al., "CNN-BiLSTM-Attention Hybrid Model for ECG Segment Classification," Biomed. Eng. Comput. Biol., 2025.