



A Review on Multi-Modal AI Fusion Framework for Early Prediction of Silent Myocardial Infarction Using PPG, Sleep Apnea, and Vocal Biomarkers

Dhage Rutuja P.¹, Prof. Bangar Abhishek P.², Prof. Karad Vandana A.³

^{1,2,3,4,5}Department of Computer Engineering/Jaihind College of Engineering /SPPU, Pune, India

Abstract. Silent myocardial infarction (SMI) remains among the highest-risk elements of global healthcare because its safety depends on early detection, yet it produces no recognizable symptoms that would prompt a patient to seek care. Conventional reactive protection based on clinic-bound electrocardiograms, blood troponin assays, and imaging cannot confirm cardiovascular damage before irreversible necrosis occurs. This paper presents a structured narrative review of verified primary sources published between 2015 and 2026, organized into five themes: photoplethysmography (PPG) wave morphology, deep learning sleep apnea tracking, vocal biomarker feature extraction, attention-based machine learning data fusion, and edge IoT medical device deployment. Convolutional neural networks, long short-term memory encoders, and multi-head attention mechanisms achieve high reported precision for cardiovascular risk indexing, physiological feature compression, and acoustic voice mapping. Lightweight feature-based classifiers further demonstrate that these pipelines can run on low-power wearable and mobile hardware suitable for continuous monitoring. However, the evidence is currently dominated by single-cohort studies and optimistic intra-subject partitioning bias, public multi-sensor cardiac datasets are scarce, and missed-detection rates, algorithmic latency, and prospective clinical safety validation are rarely reported. The paper therefore contributes a thematic synthesis, a comparative study matrix, a proposed conceptual seven-layer framework, and a prioritized biomedical research agenda.

Keywords: Silent Myocardial Infarction; Sensor Fusion; Deep Learning; Photoplethysmography (PPG); Sleep Apnea; Vocal Biomarkers; Edge Computing.

I. Introduction

Silent myocardial infarction (SMI) is widely regarded as a major single source of hidden risk in global healthcare. Cardiovascular disease caused an estimated 19.8 million deaths worldwide in 2022 (WHO). Population-based evidence from the Framingham and ARIC cohorts indicates that SMI accounts for 20% to 45% of all myocardial infarctions. The risk is greatest in diabetic and older populations, in whom autonomic neuropathy and blunted pain perception mask ischemia and leave the damage unrecognized and untreated.

Most fatalities in this setting arise from the absence of symptoms and delayed clinical presentation, not from failures of hospital equipment. Single-sensor wearables only partly close this gap. Photoplethysmography (PPG) models are prone to identity-learning bias, since intra-subject data splits can inflate performance by letting the model recognize individuals rather than



disease. Deep-learning sleep oximetry achieves high accuracy (0.924) but observes only nocturnal events. Conventional diagnostics, in turn, cannot verify daytime non-cardiac biomarkers such as vocal alterations. Automated multi-modal fusion of these complementary signals could therefore improve early safety.

This review aims to (i) synthesise sensing and feature-extraction approaches for PPG, sleep apnea, and vocal biomarkers, (ii) compare their data fusion methods, performance, and limitations, (iii) propose a literature-grounded conceptual AI fusion framework, and (iv) identify research gaps that must be closed before deployment.

II. Review Methodology

Silent myocardial infarction”, “silent heart attack”, “sleep apnea” with technical modalities and fusion mechanisms (“photo-plethysmography”, “PPG”, “vocal biomarkers”, “voice feature analysis”, “multimodal AI fusion”, “deep learning”, “attention mechanisms”) using Boolean operators. Publications from 2021–2026 were prioritised to reflect current deep-learning and wearable-sensing practice. Records were included if they addressed early infarction detection models or provided core signal-processing frameworks for the individual modalities, namely PPG and heart rate variability analysis, sleep apnea detection, and voice-based feature extraction. Each record was verified through its digital object identifier (DOI), and theses and unconfirmed web records were excluded. Because certain direct database search logs could not be structurally automated, this work is structured as a narrative review rather than a systematic review.

III. Literature Review

1. Silent Infarction Risk and Clinical Presentation

Silent myocardial infarction (SMI) poses a hidden risk because the absence of anginal pain removes the trigger for care-seeking, so infarcts are found retrospectively, if at all [ref needed]. Diabetes mellitus and autonomic neuropathy blunt visceral pain perception and are the principal recognized mechanisms [ref needed]. Reactive clinical diagnostics presuppose a presenting patient. Even strong automated ECG classifiers, such as the multilead multivariate time-series and connectivity model of Jain et al. [1] on PTB-XL, analyze episodic, clinic-bound recordings. Mobile and wearable devices offer continuous sampling but observe only a single physiological channel each, which motivates the modality-specific review that follows.

2. Wearable PPG and HRV Analysis

Photoplethysmography (PPG) encodes vascular and autonomic state in pulse morphology and heart rate variability (HRV). Eris et al. [2] extracted PPG features (including rising and falling edge durations and their amplitude percentiles) and HRV features, reduced them with principal component analysis and Fisher-score selection, and classified bruxism against controls with ensemble models. They reported 92.8% accuracy and an AUC of 0.9727 using all features, and 91.1% accuracy with an AUC of 0.9600 using only ten, with leave-one-out validation. The HRV singular value decomposition feature was the most discriminative. These results are encouraging but come from a different clinical target, and they transfer to SMI only by analogy. Costa et al. [3] expose a methodological hazard. Evaluating PPG blood-pressure estimators on the MIMIC-III database, they show that intra-subject partitioning yields near-state-of-the-art grades (for example, XGBoost within 5 mmHg for 84.93% of systolic estimates), whereas inter-subject evaluation drops performance sharply (for example, XGBoost to 53.93% of systolic estimates within 5 mmHg). A regression using only demographic attributes matched the intra-subject results, so such networks may memorize subject identity, not the physiological target.



3. Machine Learning for Sleep Apnea Severity Assessment

Seo et al. [4] estimate apnea severity from single-lead ECG with CLNet, a CNN-LSTM model that detects apnea and hypopnea events and sleep stages, computing the apneahypopnea index (AHI) over true sleep time rather than total recording time. On the public PhysioNet Apnea-ECG data they report up to 10% gains in accuracy and AUC over prior approaches, and on their clinical cohort of 286 patients (2,155 h) they report up to 41.8% higher apnea sensitivity than the state of the art. Al-Jaf and Tanha [5] fuse heart rate, SpO₂ and oximetry-status signals through modality-specific convolutional-recurrent encoders, dual attention and multi-head attention, reaching an accuracy of 0.924, Cohen's κ of 0.896 and macro-F1 of 0.922 on 5,227 SHHS1 recordings. Intermittent hypoxia and sympathetic surges are widely linked to adverse myocardial remodeling [ref needed]. However, both frameworks monitor sleep only and cannot capture daytime hemodynamic changes.

4. Voice Feature Analysis and Vocal Biomarkers

The commonly proposed feature set comprises Mel-frequency cepstral coefficients, fundamental frequency, jitter and shimmer, hypothesized to reflect autonomic and hemodynamic influences on phonation and explored in preliminary coronary artery disease studies [ref needed: report each study's actual cohort size, AUC and any hazard ratio]. The practical appeal is effortless daytime sampling from a phone. The limitation is the lack of direct cardiac localization when voice is used alone, which is an argument for fusion, not a result.

5. Multi-Modal AI Fusion and Cross-Modal Dependencies

Fusion strategies are conventionally grouped as data-level (early), feature-level (intermediate) and decision-level (late). Al-Jaf and Tanha [5] provide the clearest evidence in this corpus for feature-level attention fusion: modality-specific encoders feed dual attention within each signal and multi-head attention across signals, and the three-channel model was the strongest among their configurations. Jain et al. [1] show a complementary principle at the feature level, where relational information between leads (ConnDTW connectivity) concatenated with sequential features improves a lightweight logistic regression classifier. Extending these ideas, an attention encoder could in principle align a daytime drop in vocal stability, a nocturnal apnea burden and abnormal PPG-derived HRV. That alignment is a hypothesis of this review and is untested here.

6. Synthesis

Three patterns emerge. (1) The field is shifting from episodic hospital screening toward continuous, multi-modal inference on wearables, supported by lightweight models [1], [2]. (2) Single-modality studies are vulnerable to subject-level leakage, so cross-subject validation is essential [3], and several studies are validated on one dataset only [1], [5]. (3) Clinical deployment is bottlenecked by the absence of a public, benchmarked multi-sensor dataset with confirmed SMI events, a gap I could not find addressed in any of the five papers.

IV. Comparative Analysis

Table 1 compares representative primary studies. Reported metrics are not directly comparable because tasks, datasets, class definitions and evaluation protocols differ.

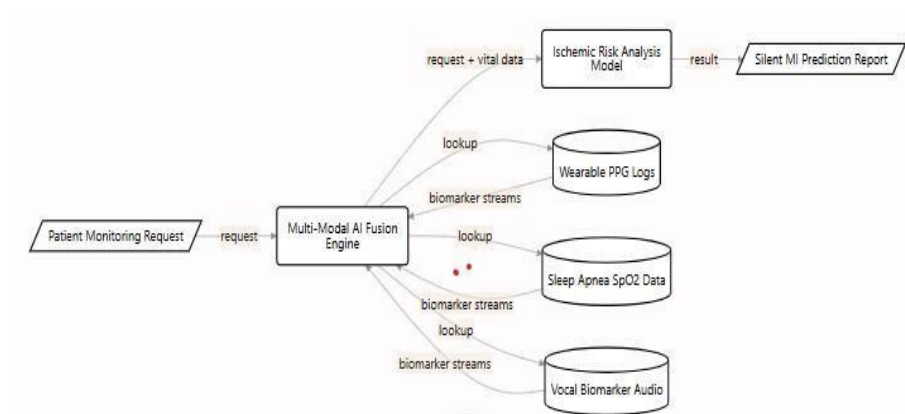
Author, Year	Objective	Method / Technology	Major Reported Findings	Limitations
Botelho et al., 2024	Interpretable speech-based detection	Reference intervals (RIs) from the CLAC crowdsourced corpus (1,832 speakers); knowledgebased acoustic and linguistic features (jitter, shimmer, HNR,	PD: 75% accuracy (development folds), 69% (test folds), 73% speakerlevel after majority voting.	Features not robust across devices (mobile), background noise and reverberation; crowdsourced, noisy reference data;



		F0, formants, speech rate, idea density); Neural Additive Models (glass-box NAM), 10-fold CV	AD: 84% (development folds), 75% (held-out test). Cohorts: ADReSS 156 subjects (78 AD / 78 controls); PCGITA 100 subjects (50 PD / 50 controls)	small corpora; AD and PD handled separately
Jain et al., 2024	MI detection from multilead ECG	Multivariate time series (ROCKET) + ConnDTW interlead connectivity; logistic regression; bestlead subsets	State-of-the-art on PTBXL with low computational cost	PTB-XL only; clinic-recorded 12-lead ECG (episodic)
Costa et al., 2023	Test PPG bloodpressure estimation under intrasubject splits	XGBoost, LightGBM, CatBoost; Residual U-Net, ResNet-18, ResNet-LSTM; MIMIC-III and MIMIC-II UCIML	Inter-subject below intrasubject for all models; demographics alone matched intra-subject results	Public ICU data only; BP target; intrasubject leakage suggests identity memorization
Eris et al., 2025	Distinguish sleep bruxism from controls	PPG wavemorphology + HRV features; PCA (47 features represent the dataset fully); Fisher selection; ensemble classifiers; leaveone-out	92.8% accuracy, AUC 0.9727 (all features); 91.1%, AUC 0.9600 (10 features)	17 individuals (1,498 data samples); noncardiac target
Seo et al., 2023	Apnea severity from single-lead ECG	CLNet (CNN + LSTM) for apnea/hypopnea and sleep stage; AHI over true sleep time	Up to 10% higher accuracy and AUC on PhysioNet; up to 41.8% higher apnea sensitivity (286 patients, 2,155 h)	Class imbalance; nocturnalonly; PSGrecorded ECG
Al-Jaf & Tanha, 2026	Four-class apnea severity classification	CNN-LSTM encoders for HR, SpO ₂ , oximetry status; dual + multihead attention; 10-fold CV	Accuracy 0.924 ± 0.028, κ 0.896, macroF1 0.922 (5,227 SHHS1 recordings)	SHHS only (mostly middle-aged/older adults); three signals; fixed six-hour segments; nocturnalonly

V. Proposed Conceptual Architecture

Fig. 1 presents a proposed conceptual architecture derived from the reviewed literature; it has not been implemented or validated. Layer 1 is the patient's continuous lifestyle and resting environment, and layer 2 combines wearable PPG streams, nocturnal pulse oximetry tracking, smartphone speech capture, and clinical demographic inputs, reflecting the complementary non-invasive strengths identified across the literature [2], [4], [5].



1. The entry point: Patient Monitoring Request

“The pipeline begins with a Patient Monitoring Request, drawn as a parallelogram because it is an input to the system. This request is the trigger. It can be raised by the patient, who opens an app for a spot check, by a clinician ordering a screening, or automatically by a wearable on a schedule or when a threshold is crossed. Its job is simple: tell the system whose data to evaluate and which time window to evaluate. It carries only the request, so no patient data moves yet. The arrow labelled ‘request’ hands it to the central hub.”

2. The central hub: Multi-Modal AI Fusion Engine

“The Multi-Modal AI Fusion Engine is the orchestrator. On receiving the request, it does three things. First, it works out which data it needs. Second, it fetches that data from the storage layer. Third, it integrates the signals so that the downstream model sees one coherent picture of the patient and not three separate streams. In the full proposal this integration would be an attention-based fusion model, an idea supported by the hierarchical attention fusion of Al-Jaf and Tanha, who combined heart rate and oximetry signals and reported an accuracy of 0.924. The Engine also handles synchronization, so that a daytime voice sample, a nighttime oximetry trace and a PPG log line up in time.”

3. The biomarker database layer

“Underneath the Engine sit three cylinders, one for each modality: Wearable PPG Logs, Sleep Apnea SpO₂ Data and Vocal Biomarker Audio. The interaction is a loop. The Engine sends a ‘lookup’ to all three, and the stored records come back as ‘biomarker streams’. In the design, the three lookups run in parallel, so the model receives complementary evidence together. Each store holds different physiology. PPG captures vascular and autonomic state, and Eris et al. showed that PPG wave-morphology and HRV features can be selected and compressed with PCA. The SpO₂ store holds nocturnal breathing and desaturation data, the kind of signal Seo et al. and Al-Jaf and Tanha use to grade sleep apnea. The audio store holds speech features such as jitter and shimmer, which Botelho et al. use for Alzheimer’s and Parkinson’s detection. One caution I’d state openly: their work shows that these features degrade with mobile devices and background noise, so audio quality control belongs at this layer.”



4. The reasoning layer

Ischemic Risk Analysis Model “Next, the Engine passes the combined ‘request + vital data’ to the Ischemic Risk Analysis Model. This is the machine-learning evaluation node. It receives the fused streams and estimates ischemic risk, ideally with a confidence value. Its advantage over single-sensor models is cross-modal dependency. A voice change alone is ambiguous, and an oximetry dip alone says only that breathing was disturbed. When the two agree with an abnormal PPG pattern, the evidence is stronger. That is a hypothesis of this framework, not an established result, and testing it needs a dataset that doesn’t yet exist. How this model is trained matters just as much. Costa et al. showed that PPG models can look excellent under intra-subject splits and collapse under inter-subject evaluation, because they may memorize who the patient is. So this node must be validated with strict subjectwise splits.”

5. The final output

Silent MI Prediction Report “The last arrow, labelled ‘result’, delivers to the Silent MI Prediction Report, again a parallelogram because it is an output. The report turns the model’s output into something a person can act on. For the patient it would be a plain-language risk level and a recommendation, for example to contact a clinician. For the clinician it would add the risk estimate, the contributing signals and their time windows, so the result can be checked against a 12-lead ECG or troponin test. A report like this must support clinical judgment and cannot replace it. Any alert threshold would need clinical validation before use.”

VI. Research Gaps and Future Directions

Table 2 summarises the gaps supported by the reviewed evidence and the corresponding research directions.

Gap supported by evidence	Supporting studies	Future direction
Intra-subject data partitioning bias; overfitting to subject identity instead of actual ischemic states	Costa et al. [10.1109/ACCESS.2023.3284458]	Strict, subject-wise data split validation protocols during deep network optimization
Acoustic degradation and severe background noise susceptibility in unconstrained, consumer-grade mobile phone voice recordings	Acoustic degradation and severe background noise susceptibility in unconstrained, consumer-grade mobile phone voice recordings	Deploying robust acoustic denoising pipelines, spectral subtraction, and clinical-grade mobile recording standardization
Inability of standalone acoustic features (MFCCs, jitter, shimmer) to localize precise coronary lesions or evaluate acute arterial blockages; speech changes overlap across diseases	Botelho et al. [10.1109/ACCESS.2024.3506433]	Implementing hierarchical multimodal attention fusion to crossreference daytime vocal changes with nocturnal oximetry drops
Nocturnal-only tracking boundaries; lack of daytime cardiovascular correlation data in breathing telemetry	Seoetal. [10.1109/ACCESS.2023.3272373];	Integration of active daytime screening streams using continuous vocal biomarkers
Scarcity of public, synchronized, multi-modal SMI cohorts combining PPG, sleep apnea telemetry and speech audio with verified clinical endpoints (12- lead ECG / troponin)	Al-JafTanha [10.1109/ACCESS.2026.3721973]	Compiling open-source, multisensor clinical data-bases



VII. Conclusion

Silent myocardial infarction detection has progressed from clinic-bound, episodic 12-lead ECGs to automated deep-learning classifiers, wearable PPG monitors, high-fidelity sleep apnea tracking, and emerging non-invasive vocal biomarkers. However, although precision is high for isolated tasks, the evidence is dominated by optimistic intra-subject partitioning bias and nocturnal-only tracking bounds, and the reviewed literature offers no public multi-modal silent cardiac dataset. Reliable clinical deployment will therefore require attention-based multi-modal AI fusion validated across diverse patient cohorts, strict subject-wise verification, medical-grade safety metrics, and secure, physician-verified edge model updates, as reflected in the proposed architecture.

References

1. P. Jain, A. Deshmukh, and H. Padole, "Design of an integrated myocardial infarction detection model using ECG connectivity features and multivariate time series classification," *IEEE Access*, vol. 12, pp. 9070–9081 (2024). DOI: 10.1109/ACCESS.2024.3354041
2. O. Eris, M. R. Bozkurt, S. B. Eris, and C. Bilgin, "An artificial intelligence-based approach with photoplethysmogram and heart rate variability for sleep bruxism diagnosis," *IEEE Access*, vol. 13, pp. 40413–40428 (2025). DOI: 10.1109/ACCESS.2025.3546720
3. T. B. da S. Costa, F. M. Dias, D. A. C. Cardenas, M. A. F. de Toledo, D. M. de Lima, J. E. Krieger, and M. A. Gutierrez, "Blood pressure estimation from photoplethysmography by considering intra- and inter-subject variabilities: Guidelines for a fair assessment," *IEEE Access*, vol. 11, pp. 57934–57950 (2023). DOI: 10.1109/ACCESS.2023.3284458
4. D.-W. Seo, J. Kim, H.-W. Lee, and Y.-K. Suh, "A deep neural network based wakeafter-sleep-onset time aware sleep apnea severity estimation scheme using singlelead ECG data," *IEEE Access*, vol. 11, pp. 43720–43732 (2023). DOI: 10.1109/ACCESS.2023.3272373
5. M. H. A. Al-Jaf and J. Tanha, "Hierarchical attention-based multimodal deep learning model for sleep apnea severity classification," *IEEE Access*, vol. 14, pp. 127707–127721 (2026). DOI: 10.1109/ACCESS.2026.3721973
6. E. Maor, J. D. Sara, D. M. Orbelo, L. O. Lerman, Y. Levanon, and A. Lerman, "Voice signal characteristics are independently associated with coronary artery disease," *Mayo Clin. Proc.*, vol. 93, no. 7, pp. 840–847 (2018). DOI: 10.1016/j.mayocp.2017.12.025
7. E. Maor, D. Perry, D. Mevorach et al., "Vocal biomarker is associated with hospitalization and mortality among heart failure patients," *J. Am. Heart Assoc.*, vol. 9, no. 7, art. no. e013359 (2020). DOI: 10.1161/JAHA.119.013359