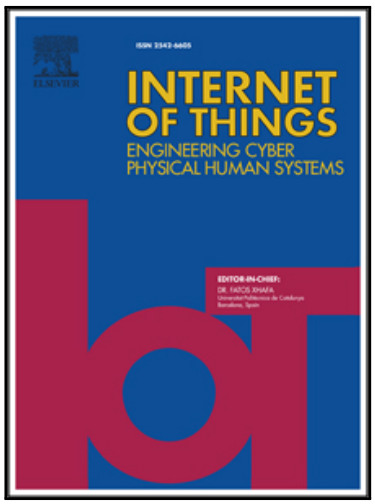
Journal Pre-proof



AVAESA: Adaptive VAE with Self-Attention and Learnable Signal Processing for Robust Radar-Based Heart Rate Estimation

Mohammad Hossein Shirazi, Sira Yongchareon, Anuradha Singh, Jing Ma

PII: S2542-6605(26)00061-2
DOI: https://doi.org/10.1016/j.iot.2026.101931
Reference: IOT 101931

To appear in: *Internet of Things*

Received date: 21 September 2025
Revised date: 30 January 2026
Accepted date: 21 March 2026

Please cite this article as: Mohammad Hossein Shirazi, Sira Yongchareon, Anuradha Singh, Jing Ma, AVAESA: Adaptive VAE with Self-Attention and Learnable Signal Processing for Robust Radar-Based Heart Rate Estimation, *Internet of Things* (2025), doi: https://doi.org/10.1016/j.iot.2026.101931

Graphical Abstract

AVAESA: Adaptive VAE with Self-Attention and Learnable Signal Processing for Robust Radar-Based Heart Rate Estimation

Mohammad Hossein Shirazi, Sira Yongchareon, Anuradha Singh, Jing Ma

Highlights

AVAESA: Adaptive VAE with Self-Attention and Learnable Signal Processing for Robust Radar-Based Heart Rate Estimation

Mohammad Hossein Shirazi, Sira Yongchareon, Anuradha Singh, Jing Ma

- Novel AVAESA architecture with dual-stream I/Q processing and self-attention for enhanced temporal modeling
- Adaptive signal processing reduces dependency on fixed filter parameters through learnable, signal-driven preprocessing
- Improved cross-domain generalization maintains performance across diverse measurement conditions with statistical validation
- Comprehensive evaluation against multiple baselines (CNN, LSTM, Bi-LSTM, TCN, VAE) establishes architectural design principles for contactless cardiac monitoring

AVAESA: Adaptive VAE with Self-Attention and Learnable Signal Processing for Robust Radar-Based Heart Rate Estimation

Mohammad Hossein Shirazi[a,*], Sira Yongchareon[a], Anuradha Singh[a], Jing Ma[a]

[a]*Department of Data Science and AI School of Engineering Computer and Mathematical Sciences Auckland University of Technology Auckland New Zealand*

Abstract

Non-contact heart rate monitoring using radar sensors offers significant advantages for healthcare and automotive applications by preserving privacy while enabling continuous physiological assessment. Current Variational Autoencoder (VAE) approaches for radar-based vital sign monitoring, while superior to traditional neural networks, suffer from fixed preprocessing assumptions and inadequate temporal modeling that limit their generalization across diverse measurement conditions. This study introduces AVAESA (Adaptive VAE with Self-Attention and Learnable Signal Processing), a novel architecture that addresses these limitations through three key innovations: dual-stream in-phase/quadrature signal processing that preserves critical phase relationships, multi-head self-attention mechanisms for enhanced temporal dependency modeling, and adaptive signal preprocessing with learnable parameters that derive frequency bands and processing weights directly from input signal characteristics. The framework was evaluated on 1,920 measurements from 10 participants across 48 measurement scenarios (4 distances × 3 angles × 4 orientations), assessing cross-scenario robustness under measurement domain shift, with Polar H10 chest strap ground truth validation. Comprehensive comparison against multiple architectures (CNN, LSTM, Bi-LSTM, TCN, VAE) with statistical significance testing demonstrates substantial performance improvements, with mean absolute error reductions

*corresponding author

Email address: hossein.shirazi@autuni.ac.nz (Mohammad Hossein Shirazi)

ranging from 17.3% under optimal conditions to 62.6% under challenging cross-domain generalization scenarios. AVAESA maintains high accuracy (correlation coefficient > 0.86, $R^2 > 0.84$) even under extreme domain shift conditions where baseline approaches exhibit degraded performance, demonstrating potential for contactless cardiac monitoring systems across diverse measurement environments through improved cross-scenario robustness.

Keywords: Non-intrusive Vital Sign Monitoring, Machine Learning, Radar, Generative Models, Attention Modules, Variational Autoencoders

1. Introduction

Heart rate monitoring represents a critical component of physiological health assessment, serving as an essential indicator for heart rate variability evaluation and early detection of cardiac abnormalities [1, 2]. The increasing demand for continuous health monitoring, particularly in applications such as driving, independent living, and patient care, necessitates innovative non-contact monitoring solutions [3, 4].

Radar sensors facilitate the acquisition of physiological signals while preserving privacy and minimizing environmental disturbances such as lighting variations, rendering them essential for ambient intelligence applications [5]. The technology operates by emitting signals that reflect off the human body, enabling detection of minute chest movements associated with cardiac activity [6]. This provides significant advantages over contact-based sensors, which may be uncomfortable during extended monitoring or for individuals with skin sensitivities [7].

The integration of radar with machine learning and deep learning has transformed signal analysis, enhancing the accuracy of vital sign acquisition [8, 9, 10]. Recent comprehensive analyses have demonstrated that machine learning approaches for radar-based vital sign monitoring achieve superior performance compared to traditional signal processing techniques across various monitoring scenarios [11, 12].

A systematic survey by Shirazi et al. [11] provides crucial insights into the effectiveness of different machine learning paradigms for radar-based vital sign monitoring. Their comprehensive analysis of 34 studies from 2020-2025 reveals that VAE architectures demonstrate effective performance in vital sign monitoring applications, achieving correlation scores of 0.9 for UWB sensors and 0.92 for FMCW sensors compared to ground truth measurements

[13]. VAE-based systems have shown strong adaptability, with implementations achieving 92.5% accuracy in respiratory rate estimation while reducing signal distortion by 15% compared to traditional approaches [14]. These results demonstrate that VAE architectures excel in noise reduction and signal quality enhancement through their ability to learn probabilistic representations of physiological patterns [15, 16]. The survey findings indicate that VAE approaches offer fundamental advantages over discriminative models through their generative modeling capabilities, enabling effective noise handling and robust signal separation [16, 17, 18, 19]. Unlike traditional neural network architectures that learn direct mappings from inputs to outputs, VAEs learn continuous latent representations that capture the underlying structure of physiological signals, leading to improved generalization and robustness to measurement artifacts [20, 19].

Despite these advances, current VAE-based approaches face three critical limitations that constrain their practical deployment. First, they rely on **fixed preprocessing parameters** (e.g., predetermined frequency bands, static filter designs) that assume specific measurement configurations, leading to poor generalization when deployment conditions differ from training scenarios [22]. Second, conventional VAE encoders employ **inadequate temporal modeling** through simple pooling or basic recurrent operations, failing to capture long-range dependencies critical for cardiac rhythm analysis [23]. Third, magnitude-only signal processing approaches discard **phase information** essential for accurate timing extraction, particularly when processing complex I/Q radar returns [24]. These architectural deficiencies result in substantial performance degradation under domain shift conditions, limiting the applicability of VAE-based approaches in uncontrolled monitoring environments.

This study addresses the challenge of **cross-scenario robustness**—the ability to maintain accurate heart rate estimation when measurement conditions (distance, angular position, subject orientation) differ substantially from training configurations. Unlike cross-subject generalization studies that evaluate performance on unseen individuals, our work evaluates whether architectural innovations enable consistent performance when *measurement geometry* changes, even when monitoring the same individuals. For example, a model trained on measurements at 40cm, 80cm, and 120cm must maintain accuracy when deployed at 160cm. This reflects practical radar deployment scenarios where systems trained in controlled laboratory conditions must operate under diverse field geometries.

Using a publicly available dataset with 1,920 measurements from 10 participants across 48 systematically varied measurement configurations (4 distances $\times$ 3 angles $\times$ 4 orientations), we demonstrate that architectural innovations in signal processing and temporal modeling substantially improve robustness to measurement domain shift. The dataset's rich scenario diversity provides appropriate experimental conditions for evaluating architectural mechanisms that enable cross-scenario generalization.

This study makes three key contributions to radar-based heart rate monitoring:

- We introduce AVAESA (Adaptive VAE with Self-Attention), a novel architecture that integrates attention-based temporal modeling and positional encoding within a variational framework. By employing dedicated self-attention layers for separate I/Q signal streams, AVAESA achieves improved accuracy in heart rate estimation compared to conventional architectures, including Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Convolutional Neural Network (CNN), Temporal Convolutional Network (TCN), and baseline VAE models.

- We implement a learnable signal processing module that eliminates reliance on fixed preprocessing parameters (predetermined frequency bands at 0.7-3.5 Hz, static multi-scale weights, fixed component fusion ratios). The adaptive mechanism derives all processing parameters directly from input signal characteristics—SNR, spectral shape, and dominant frequencies—enabling automatic adjustment to varying measurement conditions (distance, angle, orientation) without manual recalibration.

- We present comprehensive cross-scenario evaluation demonstrating AVAESA's enhanced robustness when measurement conditions differ from training configurations. Using systematic variations in distance, angular position, and subject orientation, AVAESA achieves error reductions ranging from 17.3% under matched conditions to 62.6% under extreme domain shift scenarios compared to baseline VAE approaches.

The remainder of this paper is structured as follows: Section 2 reviews related work in radar-based heart rate monitoring and VAE architectures. Section 3 presents the methodology, including dataset characteristics and

the proposed attention-enhanced VAE framework. Section 4 presents experimental results and performance comparisons. Section 5 concludes with findings and implications.

2. Related Work

Variational autoencoders provide superior generative modeling capabilities for radar-based vital sign monitoring through their ability to learn robust latent representations of physiological signals [16, 18, 25]. Unlike discriminative models that directly map noisy inputs to outputs, VAEs can potentially filter noise while preserving essential physiological characteristics through their probabilistic latent space representation[26].

Zhang et al. [13] pioneered VAE networks for extracting fine-grained heartbeat waveforms from RF sensors, achieving correlation scores of 0.9 for UWB and 0.92 for FMCW sensors. Their approach successfully addressed nonlinear signal mixing challenges, demonstrating superior performance compared to traditional EEMD [27] methods in handling signal artifacts. However, Zhang et al.'s method focused primarily on waveform reconstruction accuracy without addressing temporal consistency issues critical in continuous monitoring.

Building upon Zhang et al.'s reconstruction framework, Jang et al. [15] developed a comprehensive VAE-based signal reconstruction system targeting temporal consistency problems. Their approach transforms Doppler Cardiogram signals into synthetic ECG representations, achieving 75.5% improvement in heart rate variability consistency compared to direct measurements. While addressing Zhang et al.'s temporal limitations, this approach required synthetic ECG generation, adding computational complexity and potentially introducing artifacts.

The MoRe-Fi system by Zheng et al. [21] applied VAE architecture for respiratory monitoring, achieving remarkable robustness to extreme noise conditions including full-scale body movements while maintaining cosine similarity above 0.95. Compared to Zhang et al.'s and Jang et al.'s cardiac-focused approaches, MoRe-Fi demonstrated superior motion robustness but specialized for respiratory patterns rather than complex cardiac signals. This highlights the fundamental trade-off between signal-specific optimization and generalizability across vital signs.

These applications demonstrate that while VAE's generative modeling capabilities offer inherent advantages in learning noise-resistant representa-

tions, each implementation addresses only a subset of radar-based vital sign monitoring challenges, creating fragmented specialized solutions rather than comprehensive approaches.

2.1. Limitations of Traditional VAE and the Need for Enhanced Architectures

Despite their generative advantages, traditional VAE implementations suffer from critical limitations constraining their effectiveness in radar-based physiological monitoring. Comparative analysis reveals three fundamental shortcomings that no single method adequately addresses.

First, conventional VAE architectures rely on simple temporal aggregation methods such as mean pooling, inadequately capturing complex temporal dynamics essential for physiological signal analysis [28]. This limitation is evident across all approaches: Zhang et al.'s method struggles with temporal dependencies spanning multiple heartbeat cycles; Jang et al.'s synthetic ECG approach partially addresses this through post-processing but fails to capture temporal relationships within the VAE latent space; and Zheng et al.'s MoRe-Fi system relies on simplified temporal modeling unsuitable for complex cardiac rhythm analysis.

Second, existing VAE implementations struggle with capturing intricate underlying correlations and temporal dependencies in time series data [29]. Zhang et al.'s high correlation scores mask poor performance on complex arrhythmias, while Jang et al.'s improved temporal consistency loses subtle cardiac timing relationships present in original radar signals. Traditional VAEs face difficulties estimating underlying data distributions when complex prior distributions require assumptions differing from actual data characteristics[30].

Third, current approaches typically process radar signals in simplified ways, converting complex I/Q data to magnitude representations and losing critical phase information [31, 32]. Zhang et al. focus primarily on magnitude information, losing phase relationships critical for accurate timing; Jang et al. partially preserve phase through synthetic ECG but introduce new phase distortions; Zheng et al.'s MoRe-Fi maintains some phase information but optimizes for respiratory rather than cardiac phase patterns. Encoder architectures rely only on basic convolutional or recurrent layers, lacking advanced mechanisms to focus on relevant temporal features [33].

Recent advances propose enhanced architectures integrating additional mechanisms with VAE frameworks. The VAEAT approach demonstrates that combining VAEs with adversarial training and attention mechanisms

improves reconstruction quality and pattern recognition in complex temporal sequences [29]. The VAR-VAE framework shows that operating VAE models in latent spaces with explicit temporal dynamics significantly improves time series performance [28].

These findings motivate integrating self-attention mechanisms with VAE architectures to enhance temporal modeling capabilities while preserving generative advantages of variational inference for radar-based vital sign monitoring.

2.2. Self-Attention Mechanisms: Addressing Temporal Modeling Deficiencies

Self-attention mechanisms [34] provide a powerful solution to weak temporal modeling that constrains all previously discussed approaches. Through learnable query–key–value interactions, self-attention adaptively attends to the most informative parts of input sequences, rather than relying on fixed pooling or simplistic recurrent aggregation [35, 36]. This represents a fundamental advance over static approaches used by Zhang et al., post-hoc temporal processing by Jang et al., and specialized but limited temporal modeling in Zheng et al.'s MoRe-Fi system. Dynamic weighting mechanisms enable capture of subtle temporal dependencies across multiple time scales [37].

Research demonstrates multi-head self-attention effectiveness for complex temporal modeling, offering solutions to limitations in existing radar-based VAE approaches. Chen et al. [38] achieve state-of-the-art accuracy in physiological signal analysis through hybrid CNN–Transformer models where different attention heads specialize in distinct patterns—directly addressing I/Q processing limitations in Zhang et al.'s magnitude-focused approach. SDVS-Net employs multivariate self-attention to preserve variable dependencies across temporal dimensions [39], offering principled solutions to temporal consistency problems Jang et al. address through synthetic ECG generation. MDSAnet utilizes memory-driven self-attention mechanisms for global feature extraction while maintaining computational efficiency [40], potentially providing MoRe-Fi's motion robustness while extending to cardiac monitoring.

Positional encoding enriches this framework by providing explicit temporal order, absent in conventional VAE formulations [41]. When integrated into VAE frameworks, self-attention strengthens temporal representation while preserving probabilistic generative capacity of variational inference [42]. This combination creates architectures capable of both noise-resilient signal reconstruction and fine-grained temporal modeling, making

self-attention–enhanced VAEs compelling for radar-based vital sign monitoring that can unify specialized advantages demonstrated by Zhang et al.'s reconstruction accuracy, Jang et al.'s temporal consistency, and Zheng et al.'s motion robustness within a comprehensive framework.

3. Methodology

Table 1 summarizes the mathematical notation used throughout this section.

Table 1: Summary of mathematical notation

Symbol	Description
$\mathbf{I}, \mathbf{Q}$	In-phase and quadrature radar signals
$r(t)$	Complex radar signal, $r(t) = I(t) + jQ(t)$
n, m, t	Antenna index, chirp index, fast-time sample index
$\mathbf{z}$	Latent representation
$\boldsymbol{\mu}, \boldsymbol{\sigma}^2$	Variational posterior mean and variance
$\boldsymbol{\phi}$	Signal-derived adaptive parameters
$\mathcal{E}_I, \mathcal{E}_Q$	I-channel and Q-channel encoders
$\mathcal{D}$	Decoder network
$\mathcal{A}$	Adaptive signal processor
w_I, w_Q	Uncertainty-based fusion weights
α_k	Multi-scale convolution weights (kernel size k)
d_k	Attention key dimensionality (512)
$h, \hat{h}$	Ground truth and predicted heart rate
$\mathcal{L}_{\text{ELBO}}$	Evidence lower bound objective
$\mathcal{L}_{\text{HR}}$	Heart rate supervision loss

The AVAESA framework operates through two main components: (1) Signal Preprocessing and I/Q Domain Analysis and (2) Adaptive VAE with Self-Attention (AVAESA). The architecture is shown in Figure 1.

3.1. Signal Preprocessing and I/Q Domain Analysis

The signal preprocessing pipeline establishes the foundation for effective VAE training by preserving critical I/Q baseband information while eliminating fixed preprocessing assumptions introduced by fixed filtering assumptions. This approach maintains phase relationships essential for cardiac signal extraction through adaptive processing mechanisms.

3.1.1. Raw I/Q Baseband Preservation and Signal Quality Assessment

Traditional radar preprocessing pipelines apply immediate filtering and demodulation operations that embed fixed assumptions about cardiac frequency ranges, typically employing 0.7-3.5 Hz Butterworth filters [43, 44]. This creates measurement bias when fixed frequency bands assume a specific distance and angular configurations, phase information loss from magnitude-only processing, and non-adaptive signal conditioning through fixed filters [45].

Raw data extraction organizes interleaved I/Q samples following the complex representation:

$$r_{n,m}(t) = I_{n,m}(t) + jQ_{n,m}(t) \tag{1}$$

where n denotes the antenna index, m represents the chirp index, and t is the fast-time sample index [46]. Separate I and Q matrices are extracted for each antenna:

$$\mathbf{I}_n = [I_{n,1}(t), I_{n,2}(t), \ldots, I_{n,M}(t)]^T \tag{2}$$

$$\mathbf{Q}_n = [Q_{n,1}(t), Q_{n,2}(t), \ldots, Q_{n,M}(t)]^T \tag{3}$$

where M represents the total number of chirps in the measurement sequence.

Cross-correlation analysis verifies proper quadrature relationship between I and Q components:

$$\rho_{IQ} = \frac{\sum_t (I(t) - \bar{I})(Q(t) - \bar{Q})}{\sqrt{\sum_t (I(t) - \bar{I})^2 \sum_t (Q(t) - \bar{Q})^2}} \tag{4}$$

where $\bar{I}$ and $\bar{Q}$ represent the mean values of I and Q components. Hilbert transform verification provides theoretical validation:

$$Q_{\text{hilbert}}(t) = \mathcal{H}\{I(t)\} = I(t) * \frac{1}{\pi t} \tag{5}$$

where $*$ denotes convolution and $\mathcal{H}$ represents the Hilbert transform operator. The correlation validation:

$$\rho_{\text{hilbert}} = \text{corr}(Q(t), Q_{\text{hilbert}}(t)) > 0.7 \tag{6}$$

3.1.2. Range-Doppler Processing, Target Detection, and Background Subtraction

Range-Doppler processing isolates target signals through two-dimensional FFT operations:

$$S_n(f_r, f_d) = \mathrm{FFT}_d\{\mathrm{FFT}_r\{r_{n,m}(t)\}\} \tag{7}$$

where f_r and f_d represent range and Doppler frequencies. Target bin selection identifies the range corresponding to maximum cardiac reflection energy:

$$\mathrm{bin}_{\mathrm{target}} = \arg \max_{b \in [b_{\min}, b_{\max}]} \sum_m |S_n(b, m)|^2 \tag{8}$$

SNR estimation for antenna selection computes cardiac-specific quality metrics:

$$\mathrm{SNR}_n = \frac{\sigma^2_{\mathrm{signal}}}{\mu_{\mathrm{noise}} + \epsilon} \tag{9}$$

where $\sigma^2_{\mathrm{signal}}$ represents signal variance, μ_{noise} is the mean noise level, and $\epsilon = 10^{-8}$ prevents division by zero.

The loopback filter removes static clutter influence:

$$c_n(t) = \beta c_{n-1}(t) + (1 - \beta) r_n(t) \tag{10}$$

where $c_n(t)$ represents the estimated static background, $r_n(t)$ denotes the received signal frame, and $\beta = 0.9$ is the forgetting factor. The background-subtracted signal:

$$r_n^-(t) = r_n(t) - c_n(t) \tag{11}$$

CFAR threshold estimation:

$$\tau_{\mathrm{noise}}(t) = \alpha \cdot \frac{1}{N_{\mathrm{ref}}} \sum_{k \in \mathcal{N}(t)} |r_k^-(t)|^2 \tag{12}$$

where α is the threshold multiplier, N_{ref} is the number of reference cells, and $\mathcal{N}(t)$ represents the neighborhood around the test cell.

3.1.3. Adaptive Signal Processing and Data Organization

The adaptive processor implements learnable transformation functions:

$$\mathbf{h}_{\text{processed}} = \mathcal{F}_{\text{adaptive}}(\mathbf{I}, \mathbf{Q}; \boldsymbol{\theta}) \tag{13}$$

where $\boldsymbol{\theta}$ represents learnable parameters including frequency band boundaries, multi-scale processing weights, and component fusion coefficients. The adaptive processor employs specialized neural network architectures with three main components: AdaptiveFrequencyLearner, multi-scale convolutional processors, and an adaptive component fusion network for I/Q integration, as shown in Figure 2.

The AdaptiveFrequencyLearner utilizes multi-layer perceptron with GELU activations and Softplus output to ensure positive frequency bounds. Multiple signal representations capture different cardiac mechanics aspects:

$$\begin{aligned} r_{\text{mag}}(t) &= \sqrt{I^2(t) + Q^2(t)} \\ r_{\text{phase}}(t) &= \arctan 2(Q(t), I(t)) \\ r_{\text{real}}(t) &= I(t) \\ r_{\text{imag}}(t) &= Q(t) \end{aligned} \tag{14}$$

Each representation undergoes specialized processing through dedicated convolutional networks with multiple kernel sizes (3, 7, 15, 31 samples) to capture different temporal scales of cardiac activity. The scale selection network employs linear layers with GELU activation to learn optimal weights for each temporal scale. The component processors implement separate convolutional pathways for magnitude, phase, real, and imaginary signal components, each employing batch normalization and GELU activation for stable gradient flow.

Attention-based fusion combines the processed representations using learned weights:

$$\mathbf{h}_{\text{final}} = \sum_{i=1}^{4} \alpha_i \cdot \mathcal{C}_i(r_i(t)) \tag{15}$$

where α_i are learned attention weights optimized during training to minimize heart rate estimation error, and $\mathcal{C}_i$ represents component-specific neural networks that apply personalized processing to each signal representation.

The HDF5 architecture organizes data where each participant and measurement combination contains two primary data groups. The first group

stores raw I and Q baseband signals as separate datasets, while the second group contains processed heart signals for validation purposes. The float32 precision maintains sufficient dynamic range for neural network training while minimizing storage requirements. The HDF5 format provides compressed storage, random access patterns for mini-batch training, parallel loading capabilities for multi-GPU training configurations, and chunk-based I/O for handling large temporal sequences. Algorithm 1 provides the complete implementation of the signal preprocessing pipeline, integrating I/Q validation, range-Doppler processing, target detection, and HDF5 data organization for VAE training.

Algorithm 1 Radar Signal Processing for VAE Training Data

Require: Binary radar file $\mathbf{B}$, participant ID p, measurement ID m
Ensure: Raw I/Q signals $\mathbf{I}_{\text{raw}}, \mathbf{Q}_{\text{raw}} \in \mathbb{R}^L$

$\mathbf{I}_{\text{rx}}, \mathbf{Q}_{\text{rx}} \leftarrow \text{ParseBinary}(\mathbf{B})$ // Extract multi-antenna I/Q
ValidateIQ($\mathbf{I}_{\text{rx}}, \mathbf{Q}_{\text{rx}}$) // Hilbert transform validation
for $a = 0$ to $N_{\text{rx}} - 1$ **do**
    $\mathbf{F}_a \leftarrow \text{RangeFFT}(\mathbf{I}_{\text{rx}}[a] + j\mathbf{Q}_{\text{rx}}[a])$ // Range processing
    $b_a \leftarrow \text{FindTargetBin}(\mathbf{F}_a)$ // Target detection
    $\boldsymbol{\phi}_a \leftarrow \text{PhaseExtract}(\mathbf{F}_a[:, b_a])$ // Phase signal
    $\text{SNR}_a \leftarrow \text{QualityMetric}(\boldsymbol{\phi}_a)$ // Signal quality
end for
$a^* \leftarrow \arg\max_a \text{SNR}_a$ // Best antenna selection
$\mathbf{I}_{\text{raw}}, \mathbf{Q}_{\text{raw}} \leftarrow \mathbf{I}_{\text{rx}}[a^*], \mathbf{Q}_{\text{rx}}[a^*]$ // Raw signal extraction
SaveHDF5($\mathbf{I}_{\text{raw}}, \mathbf{Q}_{\text{raw}}, p, m$) // VAE-compatible storage
return $\mathbf{I}_{\text{raw}}, \mathbf{Q}_{\text{raw}}$

3.2. Adaptive VAE with Self-Attention (AVAESA)

Traditional VAE implementations for radar-based vital sign monitoring suffer from fundamental rigidity: they apply fixed preprocessing parameters (predetermined frequency bands such as 0.7–3.5 Hz), static channel fusion weights, and uniform temporal aggregation regardless of measurement conditions. This inflexibility creates brittleness under domain shift—parameters optimized for one measurement geometry (e.g., 1-meter frontal positioning) perform poorly under different conditions (e.g., 3-meter angular positioning) due to range-dependent signal attenuation, angle-dependent scattering patterns, and orientation-dependent multipath interference. Furthermore,

deterministic latent representations create discontinuous feature spaces that fail to generalize to unseen physiological patterns, and temporal dependencies are inadequately modeled through simple recurrent architectures [40, 36].

AVAESA addresses these limitations through a multi-level adaptive framework where processing parameters and fusion weights are derived dynamically from input signal characteristics rather than fixed *a priori*. This adaptation operates across three interconnected levels: (1) signal-level adaptation through learnable preprocessing parameters, (2) representation-level adaptation via uncertainty-weighted I/Q fusion, and (3) temporal-level adaptation using multi-head self-attention. By conditioning processing decisions on observed signal properties—spectral content, channel reliability, and temporal patterns—rather than fixed measurement assumptions, AVAESA creates a measurement-invariant processing pipeline that maintains robust performance across diverse scenarios.

3.2.1. Signal-Adaptive Processing and Measurement-Invariant Feature Extraction

The signal-adaptive processor implements the first level of adaptation by transforming raw I/Q signals through learnable parameters derived from current signal characteristics rather than scenario-specific configurations. Unlike conventional approaches that apply predetermined bandpass filters, AVAESA analyzes each input signal's spectral content, signal-to-noise ratio, and frequency characteristics to compute optimal processing parameters. For high-SNR measurements from short distances, the processor selects narrower frequency bands centered on strong cardiac harmonics; for low-SNR measurements from longer distances, it adapts by broadening frequency bands and emphasizing lower-frequency components less affected by range attenuation. Critically, these decisions occur independently for each input sample based on its measured characteristics, reducing sensitivity to deployment conditions while maintaining physiological adaptability:

$$\mathbf{s}_{\text{processed}} = \mathcal{A}(\mathbf{I}, \mathbf{Q}; \boldsymbol{\phi}(\mathbf{I}, \mathbf{Q})) \tag{16}$$

where $\boldsymbol{\phi}(\mathbf{I}, \mathbf{Q})$ represents signal-derived adaptive parameters and $\mathcal{A}$ implements multi-scale processing. The adaptive parameter extraction operates through frequency analysis:

$$\boldsymbol{\phi}(\mathbf{I}, \mathbf{Q}) = \mathcal{F}_{\text{adaptive}}\left(\mathcal{F}_{\text{freq}}(\mathbf{I} \oplus \mathbf{Q})\right) \tag{17}$$

where $\oplus$ denotes channel concatenation, $\mathcal{F}_{\text{freq}}$ performs frequency analysis through convolutional layers, and $\mathcal{F}_{\text{adaptive}}$ maps signal characteristics to processing parameters.

Multi-scale processing captures different temporal scales of cardiac activity:

$$\mathcal{A}(\mathbf{I}, \mathbf{Q}; \boldsymbol{\phi}) = \sum_{k \in \{7,15,31,63\}} \alpha_k \cdot \text{Conv}_k(\mathbf{s}_{\text{component}}) \tag{18}$$

where α_k are learned weights derived from $\boldsymbol{\phi}$ and Conv_k represents convolution with kernel size k. Signal components are extracted as:

$$\begin{aligned} \mathbf{s}_{\text{mag}} &= |\mathbf{I} + j\mathbf{Q}| \\ \mathbf{s}_{\text{phase}} &= \angle(\mathbf{I} + j\mathbf{Q}) \\ \mathbf{s}_{\text{real}} &= \mathbf{I} \\ \mathbf{s}_{\text{imag}} &= \mathbf{Q} \end{aligned} \tag{19}$$

Attention-based component fusion combines processed representations using learned weights:

$$\mathbf{s}_{\text{final}} = \sum_{i=1}^{4} \beta_i \cdot \mathcal{C}_i(\mathbf{s}_i) \tag{20}$$

where β_i are attention weights and $\mathcal{C}_i$ represents component-specific processing networks.

3.2.2. IQVAE Methodology and Dual-Stream Architecture

The IQVAE framework maintains separate processing pathways for in-phase and quadrature components throughout the entire network architecture, as illustrated in Figure 3. The variational framework operates on the principle that cardiac signals extracted from radar reflections exist within a continuous latent space where similar physiological states map to nearby representations.

The encoder networks map input I/Q sequences to distributional parameters in the latent space:

$$q_\phi(\mathbf{z}|\mathbf{I}, \mathbf{Q}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{I}, \mathbf{Q}), \text{diag}(\boldsymbol{\sigma}^2(\mathbf{I}, \mathbf{Q}))) \tag{21}$$

where ϕ represents learnable encoder parameters, $\mathbf{z}$ denotes the latent representation, and $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the learned mean and variance functions.

The variational objective optimizes the Evidence Lower BOund (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{I},\mathbf{Q})}[\log p_\theta(\mathbf{I},\mathbf{Q}|\mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{I},\mathbf{Q})\|p(\mathbf{z})) \tag{22}$$

where the first term ensures accurate signal reconstruction while the KL divergence term regularizes the latent space to follow a standard Gaussian prior.

Each encoder stream implements hierarchical feature extraction architecture combining convolutional layers with enhanced residual blocks:

$$\mathbf{f}_{\text{I}} = \mathcal{E}_{\text{I}}(\mathbf{I};\phi_{\text{I}}) \tag{23}$$

$$\mathbf{f}_{\text{Q}} = \mathcal{E}_{\text{Q}}(\mathbf{Q};\phi_{\text{Q}}) \tag{24}$$

where $\mathcal{E}_{\text{I}}$ and $\mathcal{E}_{\text{Q}}$ represent the I and Q encoder networks with learnable parameters ϕ_{I} and ϕ_{Q}. The parallel architecture enables each stream to learn specialized representations: the I channel captures direct reflection patterns from cardiac motion, while the Q channel encodes phase-shifted information containing timing relationships critical for rhythm analysis.

3.2.3. Self-Attention Mechanisms and I/Q Feature Fusion

The self-attention mechanism provides the third level of adaptation through dynamic temporal feature aggregation. Unlike fixed temporal pooling operations (e.g., mean or max pooling), attention mechanisms learn to identify which temporal segments contain robust cardiac information and which are dominated by noise or respiratory interference. When processing measurements with strong respiratory artifacts in certain time steps, attention weights automatically decrease for those corrupted segments while emphasizing cleaner portions of the signal. This temporal selectivity adapts dynamically based on observed signal patterns in each measurement window.

The attention mechanism operates on temporal sequences of encoded features, enabling the network to model relationships between distant time points [39]:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{25}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent query, key, and value matrices derived from the temporal feature sequence with dimensionality $d_k = 512$. The scaling factor $\sqrt{d_k}$ prevents attention weights from becoming too sharp, maintaining gradient stability during training.

Positional encoding provides temporal context through sinusoidal functions:

$$\mathrm{PE}(\mathrm{pos}, 2i) = \sin\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right) \tag{26}$$

$$\mathrm{PE}(\mathrm{pos}, 2i+1) = \cos\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right) \tag{27}$$

where pos denotes the temporal position, i represents the dimension index, and $d_{\mathrm{model}} = 512$ is the feature dimension. Multi-head attention enables the network to attend to different aspects of the temporal sequence simultaneously, with AVAESA employing 8 attention heads [39]:

$$\mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_8)\mathbf{W}^O \tag{28}$$

where each attention head focuses on different temporal patterns, such as short-term cardiac cycles, respiratory modulation, or heart rate variability trends.

After parallel encoding with self-attention, the I and Q feature sequences undergo temporal aggregation through learned attention weights:

$$\mathbf{f}_{\mathrm{I,agg}} = \sum_{t=1}^{T} \alpha_{\mathrm{I,t}} \mathbf{f}_{\mathrm{I,t}} \tag{29}$$

$$\mathbf{f}_{\mathrm{Q,agg}} = \sum_{t=1}^{T} \alpha_{\mathrm{Q,t}} \mathbf{f}_{\mathrm{Q,t}} \tag{30}$$

where $\alpha_{\mathrm{I,t}}$ and $\alpha_{\mathrm{Q,t}}$ are computed through softmax normalization of mean attended features across the feature dimension, and $\mathbf{f}_{\mathrm{I,t}}$ and $\mathbf{f}_{\mathrm{Q,t}}$ represent the attention-enhanced features at temporal position t.

The aggregated features are mapped to latent distributional parameters:

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{I}} &= \mathbf{W}_{\mu,\mathrm{I}} \mathbf{f}_{\mathrm{I,agg}} + \mathbf{b}_{\mu,\mathrm{I}} \\ \log \boldsymbol{\sigma}_{\mathrm{I}}^2 &= \mathbf{W}_{\sigma,\mathrm{I}} \mathbf{f}_{\mathrm{I,agg}} + \mathbf{b}_{\sigma,\mathrm{I}} \end{aligned} \tag{31}$$

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{Q}} &= \mathbf{W}_{\mu,\mathrm{Q}} \mathbf{f}_{\mathrm{Q,agg}} + \mathbf{b}_{\mu,\mathrm{Q}} \\ \log \boldsymbol{\sigma}_{\mathrm{Q}}^2 &= \mathbf{W}_{\sigma,\mathrm{Q}} \mathbf{f}_{\mathrm{Q,agg}} + \mathbf{b}_{\sigma,\mathrm{Q}} \end{aligned} \tag{32}$$

The second level of adaptation emerges in the fusion of I and Q channel representations. The fusion mechanism derives from precision-weighted

averaging in Bayesian inference: when combining estimates from multiple sources with different uncertainties, optimal combination weights each source inversely proportional to its variance (i.e., proportional to its precision $\tau = 1/\sigma^2$). Rather than combining channels with fixed weights, AVAESA learns to predict uncertainty (variance) for each channel's latent representation separately. The exponential formulation $w = \exp(-\sigma^2)$ provides a numerically stable approximation that preserves this inverse relationship: channels exhibiting lower variance—indicating more consistent, reliable features—receive exponentially higher fusion weights, while high-variance channels are automatically downweighted. This mechanism proves essential when one channel becomes corrupted by measurement artifacts or interference: the architecture automatically relies more heavily on the cleaner channel without requiring explicit artifact detection rules. The fusion weights emerge from the data itself, adapting to the signal quality of each measurement:

$$w_{\mathrm{I}} = \exp(-\mathrm{mean}(\boldsymbol{\sigma}_{\mathrm{I}}^2)), \quad w_{\mathrm{Q}} = \exp(-\mathrm{mean}(\boldsymbol{\sigma}_{\mathrm{Q}}^2)) \tag{33}$$

$$\mathbf{z} = \frac{w_{\mathrm{I}}\mathbf{z}_{\mathrm{I}} + w_{\mathrm{Q}}\mathbf{z}_{\mathrm{Q}}}{w_{\mathrm{I}} + w_{\mathrm{Q}}} \tag{34}$$

where $\mathbf{z}_{\mathrm{I}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{I}}, \boldsymbol{\sigma}_{\mathrm{I}}^2)$ and $\mathbf{z}_{\mathrm{Q}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{Q}}, \boldsymbol{\sigma}_{\mathrm{Q}}^2)$ are the reparameterized latent variables sampled using the reparameterization trick for gradient-based optimization.

The alignment loss ensures consistency between I and Q latent distributions using the Wasserstein distance:

$$\mathcal{L}_{\mathrm{align}} = W_2(q_\phi(\mathbf{z}_I|\mathbf{I}), q_\phi(\mathbf{z}_Q|\mathbf{Q})) = \|\boldsymbol{\mu}_I - \boldsymbol{\mu}_Q\|_2 + \|\boldsymbol{\sigma}_I - \boldsymbol{\sigma}_Q\|_2 \tag{35}$$

where W_2 denotes the 2-Wasserstein distance between Gaussian distributions, computed as the sum of mean and standard deviation differences between the I and Q latent posteriors.

The integration of these three adaptive mechanisms—signal-level parameter derivation, representation-level uncertainty weighting, and temporal-level attention—creates a measurement-invariant processing pipeline. By conditioning decisions on observed signal properties rather than fixed assumptions, AVAESA learns *how to adapt* its processing based on signal characteristics, explaining its superior cross-scenario generalization demonstrated in Section 4.

3.2.4. Training Objective

The AVAESA training objective combines the variational ELBO with task-specific supervised losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \gamma(\mathcal{L}_{\text{KL}}^{I} + \mathcal{L}_{\text{KL}}^{Q}) + \eta\mathcal{L}_{\text{align}} + \lambda_{\text{HR}}\mathcal{L}_{\text{HR}} + \lambda_{\text{spectral}}\mathcal{L}_{\text{spectral}} \quad (36)$$

where $\mathcal{L}_{\text{recon}}$ is the reconstruction loss combining MSE and smoothness constraints, $\mathcal{L}_{\text{KL}}^{I}$ and $\mathcal{L}_{\text{KL}}^{Q}$ are KL divergence terms for I and Q latent distributions, $\mathcal{L}_{\text{align}}$ is the Wasserstein distance alignment between I/Q distributions (Equation 35), and $\mathcal{L}_{\text{HR}} = \text{MSE}(\hat{h}, h_{\text{gt}})$ provides direct supervision on heart rate predictions. The spectral loss $\mathcal{L}_{\text{spectral}}$ enforces frequency-domain consistency between reconstructed and target signals within the cardiac band. Hyperparameters are set to $\gamma = 1.5$, $\eta = 10^{-6}$, $\lambda_{\text{HR}} = 20$, and $\lambda_{\text{spectral}} = 10$ based on validation performance.

Heart rate $\hat{h}$ is extracted from reconstructed signals via spectral peak detection within participant-specific frequency bounds learned during training (Section 3.2), rather than fixed frequency ranges. This hybrid objective ensures the latent space captures both general cardiac signal structure through the variational framework and task-relevant features through supervised HR loss, improving estimation accuracy while preserving noise-filtering capabilities inherent to VAE architectures.

Algorithm 2 provides the complete forward pass procedure for AVAESA, integrating signal-adaptive processing, dual-stream encoding with self-attention, uncertainty-weighted latent fusion, and heart rate estimation.

3.3. Architectural Parameter Selection

This subsection validates three critical architectural parameters under challenging cross-domain conditions to demonstrate systematic optimization rather than arbitrary parameter selection. Critically, all validation experiments employ an unseen scenario protocol: models are trained exclusively on angular position (0°, 30°, 45°) and subject orientation (front, back, left, right) data, while distance measurements (40, 80, 120, 160 cm) are completely held out for testing. This means the distance domain remains entirely unseen during training, representing a rigorous evaluation of cross-domain generalization capability rather than interpolation within known measurement conditions.

Latent Space Dimensionality: Table 2 demonstrates that 128 dimensions achieve optimal performance (MAE = 2.66 BPM) on the unseen distance test set. Lower dimensions (16-32) lack representational capacity for

Algorithm 2 AVAESA

Require: I/Q sequences $\mathbf{I}, \mathbf{Q} \in \mathbb{R}^{B \times T \times L}$
Ensure: Heart rate $\hat{h} \in \mathbb{R}^B$, reconstructed signals $\hat{\mathbf{s}} \in \mathbb{R}^{B \times T \times L}$

Initialize adaptive processors $\mathcal{A}$, encoders $\mathcal{E}_I, \mathcal{E}_Q$, decoder $\mathcal{D}$
Initialize self-attention modules $\mathcal{SA}_I, \mathcal{SA}_Q$
{Signal-Adaptive Processing - eliminates fixed parameter assumptions}
for $t = 1$ to T **do**
  $\boldsymbol{\phi}_t \leftarrow \text{SignalAnalysis}(\mathbf{I}_{:,t,:}, \mathbf{Q}_{:,t,:})$ {Extract SNR, spectral shape, dominant freqs}
  $[f_{low}, f_{high}] \leftarrow \text{LearnedFreqBounds}(\boldsymbol{\phi}_t)$
  $[\alpha_7, \alpha_{15}, \alpha_{31}, \alpha_{63}] \leftarrow \text{LearnedScaleWeights}(\boldsymbol{\phi}_t)$ {Multi-scale temporal kernels}
  $\mathbf{s}_t^{(I)}, \mathbf{s}_t^{(Q)} \leftarrow \mathcal{A}(\mathbf{I}_{:,t,:}, \mathbf{Q}_{:,t,:}; \boldsymbol{\phi}_t)$ {Process 4 components: mag/phase/real/imag}
end for
{Parallel I/Q Channel Encoding with Self-Attention}
$\mathbf{f}_I \leftarrow \mathcal{E}_I(\{\mathbf{s}_t^{(I)}\}_{t=1}^T)$ {I-channel CNN encoding}
$\mathbf{f}_Q \leftarrow \mathcal{E}_Q(\{\mathbf{s}_t^{(Q)}\}_{t=1}^T)$ {Q-channel CNN encoding}
$\mathbf{h}_I \leftarrow \mathcal{SA}_I(\text{PosEnc}(\mathbf{f}_I))$ {8-head self-attention on I features}
$\mathbf{h}_Q \leftarrow \mathcal{SA}_Q(\text{PosEnc}(\mathbf{f}_Q))$ {8-head self-attention on Q features}
{Variational Latent Space}
$\boldsymbol{\mu}_I, \boldsymbol{\sigma}_I^2 \leftarrow \text{Linear}(\mathbf{h}_I)$ {I-channel latent parameters}
$\boldsymbol{\mu}_Q, \boldsymbol{\sigma}_Q^2 \leftarrow \text{Linear}(\mathbf{h}_Q)$ {Q-channel latent parameters}
$\mathbf{z}_I \sim \mathcal{N}(\boldsymbol{\mu}_I, \boldsymbol{\sigma}_I^2)$, $\mathbf{z}_Q \sim \mathcal{N}(\boldsymbol{\mu}_Q, \boldsymbol{\sigma}_Q^2)$ {Reparameterization trick}
{Uncertainty-Weighted Latent Fusion}
$w_I \leftarrow \exp(-\text{mean}(\boldsymbol{\sigma}_I^2))$, $w_Q \leftarrow \exp(-\text{mean}(\boldsymbol{\sigma}_Q^2))$ {Lower variance $\rightarrow$ higher confidence}
$\mathbf{z} \leftarrow \frac{w_I \mathbf{z}_I + w_Q \mathbf{z}_Q}{w_I + w_Q}$ {Adaptive I/Q combination based on uncertainty}
{Temporal Decoding and Heart Rate Estimation}
$\hat{\mathbf{s}} \leftarrow \mathcal{D}(\mathbf{z})$ {ConvTranspose decoding to heart signals}
$\hat{h} \leftarrow \text{SpectralHREstimation}(\hat{\mathbf{s}})$ {FFT peak detection in cardiac band}
return $\hat{h}, \hat{\mathbf{s}}$

Table 2: Architectural parameter validation under cross-domain generalization (trained on angular/orientation, tested on unseen distance scenarios)

Parameter Configuration	MAE (BPM)	R^2	Pearson r
Latent Space Dimensionality			
16 dimensions	2.91 ± 0.44	0.813	0.904
32 dimensions	3.18 ± 0.52	0.734	0.872
64 dimensions	2.68 ± 0.33	0.840	0.917
128 dimensions	$\mathbf{2.66 \pm 0.31}$	**0.841**	**0.919**
256 dimensions	3.55 ± 0.61	0.717	0.853
Attention Head Configuration			
4 heads	3.28 ± 0.49	0.764	0.875
8 heads	$\mathbf{2.66 \pm 0.31}$	**0.841**	**0.919**
16 heads	3.10 ± 0.47	0.759	0.879
I/Q Processing Alternatives			
Dual-Stream I/Q	$\mathbf{2.66 \pm 0.31}$	**0.841**	**0.919**
Magnitude-Only	3.07 ± 0.45	0.801	0.901
Concatenated I/Q	3.27 ± 0.53	0.740	0.868

complex cardiac patterns, while higher dimensions (256) exhibit overfitting to training scenarios (33.5% performance degradation), limiting cross-domain generalization to unseen measurement conditions.

Attention Head Configuration: The 8-head configuration outperforms alternatives by 18.9% over 4 heads and 16.5% over 16 heads when generalizing to unseen distance measurements. Four heads provide insufficient diversity to capture multiple temporal patterns, rhythm variations, and I/Q relationships that transfer across measurement domains. Sixteen heads introduce excessive complexity without performance benefits, indicating overfitting to training domain characteristics that do not generalize to the held-out distance scenarios.

I/Q Processing Alternatives: Dual-stream I/Q processing with dedicated encoders achieves optimal performance (MAE = 2.66 BPM, r = 0.919) on unseen distance data, outperforming alternative design choices. Magnitude-only processing shows 15.4% degradation (MAE = 3.07 BPM) by discarding phase information entirely, which contains essential timing relationships for accurate cardiac rhythm extraction. Concatenated I/Q processing degrades

performance by 22.9% (MAE = 3.27 BPM) because early channel fusion prevents the network from learning specialized representations for each component. The contribution of dual-stream architecture to AVAESA's full system is further quantified through ablation analysis in Section 4.2.6, which demonstrates that removing this component causes 33.1% performance degradation.

This validation establishes AVAESA's optimal configuration: 128-dimensional latent space, 8 attention heads, and dual-stream I/Q processing. Importantly, these parameters were selected based on performance on completely unseen distance scenarios, confirming that the architectural choices provide optimal balance between representational capacity and cross-domain generalization robustness rather than merely fitting to training data characteristics.

3.4. Experimental Design and Evaluation

The evaluation employs the Sadeghi et al. [47] dataset, a publicly available mm-Wave FMCW radar vital sign monitoring dataset comprising measurements from 10 participants with Polar H10 chest strap ground truth validation. As detailed in Table 3, the dataset includes systematic measurement scenarios across four distances (40-160 cm), three angular positions (0°-45°), and four subject orientations, with 4 repetitions per scenario totaling 1,920 one-minute measurements.

Table 3: Dataset measurement scenarios and experimental configuration

Parameter	**Values**	**Count**
Participants	P1-P10	10
Distance	40, 80, 120, 160 cm	4
Angular Position	0°, 30°, 45°	3
Orientation	Front, Back, Left, Right	4
Repetitions	-	4
Duration	1 minute	-
Total Measurements	-	1,920

Three cross-domain generalization scenarios assess AVAESA's robustness under measurement domain shift. In all experiments, the same 10 participants appear in both training and testing sets under different measurement geometries: (1) distance generalization using a 3:1 train-test split per distance (e.g., train on 40/80/120cm, test on 160cm), (2) angular generalization training on distance/orientation measurements and testing on angular positions, and (3) multi-domain generalization training exclusively on orientation data

and testing on combined angular/distance scenarios. AVAESA is compared against traditional VAE baselines using conventional encoder-decoder architectures without adaptive signal processing, self-attention mechanisms, or IQVAE methodology. Performance evaluation employs mean absolute error (MAE) to quantify estimation accuracy and the Pearson correlation coefficient (r) to assess the linear relationship between predicted and ground truth heart rates. Experiments were conducted on an NVIDIA GeForce RTX 4080 GPU (16.7 GB VRAM) with CUDA acceleration.

4. Results

This section presents a comprehensive experimental evaluation of the proposed AVAESA framework through systematic comparison with contemporary neural network architectures and traditional VAE implementations. The evaluation methodology establishes both the superiority of VAE-based approaches over discriminative architectures and the enhanced cross-domain generalization capabilities of AVAESA compared to conventional VAE implementations.

All experimental protocols utilize identical HDF5 data repositories with standardized preprocessing pipelines, ensuring that observed performance differentials reflect genuine architectural innovations rather than data manipulation artifacts. The evaluation encompasses three distinct measurement scenarios representing realistic deployment configurations: radial distance variations (40, 80, 120, 160 cm), angular orientations (0°, 30°, 45°), and subject orientations (front, back, left, right). Ground truth validation employs synchronized Polar H10 chest strap monitors with performance quantification through mean absolute error (MAE), Pearson correlation coefficients, and Bland-Altman analysis.

4.1. Baseline Architecture Comparison

To establish a comprehensive baseline for evaluating AVAESA's contributions, we first compare five prominent neural network architectures for radar-based heart rate estimation: LSTM, Bi-LSTM, CNN, TCN, and VAE. The selection of these architectures is motivated by the systematic survey of Shirazi et al. [11], which identified these as frequently employed deep learning approaches in radar-based vital sign monitoring.

Contemporary Transformer-based architectures (Informer, Autoformer, Temporal Fusion Transformer) were considered but excluded from direct

comparison for two reasons. First, no established implementations exist for radar-based cardiac monitoring in the current literature, and naive application of general-purpose Transformers to radar I/Q signals would require substantial architectural modifications (input tokenization, positional encoding schemes for physiological periodicity) that constitute independent research contributions beyond fair baseline comparison. Second, our primary research question concerns whether *adaptive* signal processing and uncertainty-weighted fusion improve cross-domain robustness within VAE frameworks—a question best answered by controlled comparison against non-adaptive VAE variants rather than entirely different architectural families. As the field matures and radar-specific Transformer implementations emerge, comparative evaluation would further contextualize AVAESA's contributions.

While CNN architectures are predominantly utilized for classification tasks in the surveyed literature, we evaluate their performance in regression-based heart rate estimation to provide comprehensive architectural comparison. TCN is included as a modern temporal modeling alternative that employs dilated causal convolutions to capture long-range dependencies without the sequential processing constraints of recurrent architectures. The inclusion of TCN, which achieves strong baseline performance (MAE = 2.13 BPM), provides a non-recurrent temporal modeling baseline that contextualizes AVAESA's contributions relative to modern convolutional approaches.

Table 4 presents comprehensive performance metrics across all five baseline architectures. The evaluation demonstrates a clear performance hierarchy, with VAE achieving superior performance across all metrics, followed by TCN, Bi-LSTM, LSTM, and CNN architectures. Notably, TCN substantially outperforms both LSTM variants, demonstrating that dilated convolutions provide more effective temporal modeling than recurrent processing for radar-based cardiac signals.

Table 4: Baseline architecture comparison under comprehensive training conditions (all scenarios). Significance tests compare against VAE; $^{***}p < 0.001$, $^{**}p < 0.01$.

Architecture	**MAE (BPM)**	**R^2**	**Pearson r**	**Mean Bias (BPM)**	**LoA Range (BPM)**
CNN	$10.67 \pm 1.89^{***}$	-0.922	0.391	-5.67	50.56
LSTM	$5.44 \pm 0.92^{***}$	0.327	0.609	-1.19	32.34
Bi-LSTM	$5.49 \pm 0.88^{***}$	0.285	0.626	-2.24	32.53
TCN	$2.13 \pm 0.41^{**}$	0.883	0.942	-0.50	13.34
VAE	$\mathbf{1.68 \pm 0.29}$	**0.911**	**0.955**	**-0.32**	**11.73**

4.1.1. CNN Architecture Performance

The CNN baseline demonstrates the poorest performance among evaluated architectures, achieving MAE of 10.67 BPM with negative R^2 (-0.922), indicating performance worse than a naive mean predictor. Figure 4 (1) illustrates the weak correlation between predicted and ground truth heart rates (Pearson r = 0.391), while Figure 4 (2) reveals substantial systematic bias (-5.67 BPM) and wide limits of agreement (±25.28 BPM range). Per-participant analysis in Figure 4 (3) shows highly variable performance across individuals, with MAE ranging from 5.00 to 15.83 BPM.

The poor CNN performance reflects fundamental architectural limitations for physiological signal regression. While convolutional layers effectively extract spatial features for classification tasks, the fixed receptive fields and pooling operations discard temporal ordering information critical for cardiac rhythm analysis. The architecture's inability to model long-range temporal dependencies results in failure to capture periodic cardiac patterns, explaining the negative R^2 value indicating systematic prediction failures.

4.1.2. LSTM and Bi-LSTM Architecture Performance

LSTM and Bi-LSTM architectures demonstrate substantially improved performance compared to CNN, achieving comparable results with MAE of 5.44 and 5.49 BPM respectively. Both architectures exhibit moderate correlation with ground truth (r = 0.609 for LSTM, r = 0.626 for Bi-LSTM) and positive R^2 values (0.327 and 0.285 respectively), indicating meaningful predictive capacity.

Figure 5 (1) and Figure 6 (1) demonstrate moderate linear relationships between predictions and ground truth, while Bland-Altman analyses (Figure 5 (2) and Figure 6 (2)) reveal reduced systematic bias compared to CNN but still substantial limits of agreement (±16.17 and ±16.27 BPM respectively). Per-participant analysis (Figure 5 (3) and Figure 6 (3)) shows improved consistency compared to CNN, though individual error ranges remain substantial.

The LSTM architectures' capacity for sequential modeling enables capture of temporal dependencies in cardiac signals, explaining their superior performance compared to CNN. The bidirectional processing in Bi-LSTM provides marginal improvements in correlation (r = 0.626 vs 0.609) by incorporating future context, though this advantage is limited for real-time heart rate estimation. However, both LSTM variants exhibit fundamental limitations: (1) sequential processing creates information bottlenecks for long-

range dependencies, (2) deterministic hidden states lack uncertainty quantification, and (3) fixed recurrent operations cannot adapt to varying signal characteristics across measurement scenarios.

4.1.3. TCN Architecture Performance

The TCN baseline achieves strong performance with MAE of 2.13 BPM, $R^2 = 0.883$, and Pearson correlation r = 0.942, positioning it as the second-best architecture after VAE. Figure 7 (1) illustrates the strong linear relationship between predicted and ground truth heart rates, with the best-fit line (y = 0.93x + 5.34) closely approximating the ideal agreement line. The Bland-Altman analysis in Figure 7 (2) reveals minimal systematic bias (-0.50 BPM) and relatively narrow limits of agreement (±6.67 BPM range), indicating consistent estimation across the physiological heart rate range. Per-participant analysis in Figure 7 (3) demonstrates moderate variability across individuals, with MAE ranging from 0.50 BPM (P10) to 4.50 BPM (P2).

TCN's substantial improvement over LSTM architectures (60.8% MAE reduction compared to LSTM, 61.2% compared to Bi-LSTM) demonstrates the effectiveness of dilated causal convolutions for cardiac signal temporal modeling. The exponentially increasing dilation factors enable TCN to capture long-range temporal dependencies spanning multiple cardiac cycles without the information bottleneck inherent in recurrent hidden state propagation. Unlike LSTM's sequential processing that compresses all historical information into fixed-size hidden states, TCN maintains direct pathways to distant temporal positions through its hierarchical dilated structure, enabling more effective modeling of cardiac rhythm patterns.

The architecture's parallel computation across temporal positions also provides more stable gradient flow during training compared to the backpropagation-through-time required for LSTM variants, contributing to improved convergence and final performance. The narrow limits of agreement (13.34 BPM) compared to LSTM (32.34 BPM) and Bi-LSTM (32.53 BPM) indicate that TCN's temporal modeling produces more consistent estimates across diverse measurement conditions.

However, TCN's deterministic convolutional feature extraction lacks the probabilistic noise handling capabilities inherent in VAE's variational inference framework. The 21.1% performance gap between TCN (2.13 BPM) and VAE (1.68 BPM) reflects this fundamental architectural difference: while TCN effectively models temporal patterns through fixed convolutional oper-

ations, VAE's probabilistic latent space representation provides implicit regularization and noise filtering through the variational objective. Additionally, TCN's fixed dilation patterns apply uniform temporal attention regardless of signal quality, whereas VAE's learned latent representations can adaptively weight informative signal segments. These limitations motivate AVAESA's integration of self-attention mechanisms with the VAE framework, combining adaptive temporal modeling with probabilistic inference.

4.1.4. VAE Architecture Superiority

VAE demonstrates improved performance compared to all discriminative architectures, achieving MAE of 1.68 BPM with strong correlation ($r = 0.955$, $R^2 = 0.911$). The 69.1% error reduction compared to the best LSTM baseline (5.44 → 1.68 BPM) and narrow limits of agreement (±5.87 BPM) establish VAE as the optimal baseline architecture for radar-based heart rate estimation.

The VAE's superior performance stems from three fundamental advantages over discriminative models. First, the probabilistic latent space representation enables robust noise handling through variational inference, where the encoder learns to map noisy observations to continuous distributions rather than deterministic points. This probabilistic formulation naturally regularizes the model against overfitting and provides implicit uncertainty quantification. Second, the generative modeling framework learns the underlying structure of cardiac signals through reconstruction objectives, enabling effective separation of physiological patterns from measurement artifacts. Unlike discriminative models that learn direct input-output mappings potentially contaminated by spurious correlations, VAE must learn meaningful latent representations that capture cardiac signal structure to enable accurate reconstruction. Third, the continuous latent space provides smooth interpolation between training examples, enabling better generalization to measurement conditions with characteristics intermediate between training scenarios.

Based on this comprehensive baseline evaluation, we establish VAE as the optimal foundation architecture for radar-based heart rate estimation. The subsequent sections evaluate AVAESA—an enhanced VAE architecture incorporating adaptive signal processing and self-attention mechanisms—demonstrating further performance improvements and, critically, superior cross-domain generalization capabilities. While VAE achieves excellent performance under comprehensive training conditions, the following analysis reveals substantial

performance degradation when measurement conditions differ from training scenarios, motivating the architectural enhancements implemented in AVAESA.

4.2. AVAESA Framework Performance Evaluation

Having established VAE as the optimal baseline architecture, this subsection presents a comprehensive comparison of the proposed AVAESA framework against conventional VAE implementations under both standard and challenging cross-domain generalization conditions. The evaluation protocol employs two distinct training strategies: comprehensive multi-scenario training utilizing all measurement configurations, and systematic leave-one-scenario-out training to assess cross-domain generalization robustness.

The leave-one-scenario-out experimental protocol comprises three distinct configurations designed to evaluate generalization across different measurement modalities: (1) training on angular and orientation data while testing exclusively on distance scenarios, (2) training on distance and orientation data while testing exclusively on angular scenarios, and (3) training exclusively on orientation data while testing on combined angular and distance scenarios. This systematic evaluation framework demonstrates AVAESA's robustness when deployment conditions diverge significantly from training scenarios, representing a critical requirement for practical radar-based vital sign monitoring systems.

Table 5 presents comprehensive performance comparisons between VAE and AVAESA under cross-domain generalization conditions, where training and testing domains exhibit systematic differences in measurement configurations.

AVAESA achieves performance improvements across all evaluation conditions. Under comprehensive training, AVAESA demonstrates 17.3% MAE improvement (1.68 $\rightarrow$ 1.39 BPM) with systematic enhancements across all metrics: R^2 increases from 0.911 to 0.943, Pearson correlation improves from 0.955 to 0.971, and Bland-Altman limits of agreement narrow by 19.8% (11.73 $\rightarrow$ 9.41 BPM range).

4.2.1. Comprehensive Training Performance Analysis

Under standard training conditions utilizing all measurement scenarios, both VAE and AVAESA achieve high accuracy, with AVAESA demonstrating superior performance across all evaluation metrics. This comprehensive

Table 5: Cross-domain performance comparison between VAE and AVAESA under different domain shift conditions. Significance tests compare AVAESA against VAE within each scenario; $^{***}p < 0.001$, $^{**}p < 0.01$.

Training/Testing Configuration	**Architecture**	**MAE (BPM)**	**R^2**	**Pearson r**	**Mean Bias (BPM)**	**LoA Range (BPM)**
All Scenarios	VAE	1.68 ± 0.29	0.911	0.955	-0.32	11.73
All Scenarios	AVAESA	$\mathbf{1.39 \pm 0.22}^{**}$	**0.943**	**0.971**	**-0.24**	**9.41**
Distance Test	VAE	7.11 ± 1.24	0.231	0.554	0.48	37.74
Distance Test	AVAESA	$\mathbf{2.66 \pm 0.31}^{**}$	**0.841**	**0.919**	**-0.48**	**17.10**
Angular Test	VAE	2.90 ± 0.58	0.687	0.836	-1.00	28.04
Angular Test	AVAESA	$\mathbf{1.65 \pm 0.26}^{**}$	**0.912**	**0.956**	**-0.30**	**15.00**
Angular+Distance Test	VAE	8.18 ± 1.47	0.024	0.366	0.61	43.59
Angular+Distance Test	AVAESA	$\mathbf{3.11 \pm 0.42}^{**}$	**0.738**	**0.865**	**0.91**	**22.41**

training represents the optimal scenario, where the model has access to data from all possible measurement configurations during training.

AVAESA exhibits enhanced correlation characteristics, with a perfect linear relationship (r = 0.971, R^2 = 0.943), compared to VAE's excellent baseline performance (r = 0.955, R^2 = 0.911), as illustrated in Figure 8. The 3.2% improvement in coefficient of determination indicates additional cardiac signal variance capture through self-attention temporal modeling and adaptive signal processing mechanisms. Per-participant analysis demonstrates the consistent advantages of AVAESA, with 9 out of 10 participants achieving sub-2 BPM error compared to 8 out of 10 for the standard VAE. Additionally, the maximum individual error is reduced by 25% (4.00 → 3.00 BPM), as shown in Figure 9. Figure 10 further illustrates the distribution of estimation errors across participants, highlighting the improved reliability, robustness, and consistency of AVAESA.

AVAESA's performance advantage under comprehensive training conditions stems from its architectural components. The self-attention mechanism enables the model to capture long-range temporal dependencies in cardiac signals that traditional VAE encoders miss. The adaptive signal processing layer dynamically adjusts to varying signal characteristics across different measurement scenarios, while the multi-scale feature extraction captures both fine-grained cardiac patterns and broader physiological trends. These architectural advantages result in more robust and accurate cardiac signal

reconstruction, as evidenced by the improved correlation and reduced per-participant variability, as well as the consistently narrower estimation errors across scenarios [48].

4.2.2. Single-Scenario Generalization Analysis

Cross-domain evaluation under single-scenario constraints reveals AVAESA advantages when training and testing conditions exhibit domain shift. This evaluation includes two critical scenarios: angular generalization (training on distance+orientation, testing on angular) and distance generalization (training on angular+orientation, testing on distance). These represent realistic deployment scenarios where the monitoring system encounters measurement conditions not present during training.

Angular generalization testing demonstrates AVAESA's robust orientation invariance with minimal performance degradation (MAE: 1.39 → 1.65 BPM, 18.7% increase) compared to VAE's considerable degradation (MAE: 1.68 → 2.90 BPM, 72.6% increase), as illustrated in Figure 11. AVAESA maintains high correlation ($r = 0.956$, $R^2 = 0.912$), indicating effective capture of angular-invariant cardiac features. Figure 13 further illustrates the reduced estimation errors and improved consistency across orientations, with a narrower error range ($\pm 14.02 \rightarrow \pm 7.50$ BPM) and decreased systematic bias (-1.00 → -0.30 BPM).

Distance generalization reveals notable performance differences between architectures. AVAESA maintains acceptable performance (MAE: 2.66 BPM, R^2: 0.841) while VAE exhibits severe performance degradation (MAE: 7.11 BPM, R^2: 0.231). The 62.6% MAE improvement demonstrates AVAESA's enhanced signal extraction under degraded radar conditions, preserving 84% cardiac signal variance compared to VAE's critical 23% retention. Per-participant analysis shows consistent AVAESA advantages, with individual errors ranging from 1.58 to 4.04 BPM compared to VAE's 3.92–13.13 BPM range, representing a 69.2% reduction in maximum error, as demonstrated in Figure 12. Figure 13 further illustrates the narrower error distribution ($\pm 18.87 \rightarrow \pm 8.55$ BPM), highlighting the improved reliability and distinction between high- and low-accuracy estimations.

AVAESA's single-scenario generalization performance results from its adaptive architectural design for cross-domain robustness. The self-attention mechanism learns to focus on cardiac signal features that remain consistent across different measurement modalities, effectively filtering out scenario-specific artifacts. The adaptive signal processing layer automatically adjusts

its parameters based on input signal characteristics, enabling robust performance even when encountering unseen measurement conditions. The multi-scale feature extraction ensures that both angular and distance-invariant cardiac patterns are captured, providing redundant pathways for accurate signal reconstruction when one measurement modality is unavailable during training.

4.2.3. Multi-Scenario Generalization Analysis

The most challenging experimental condition—combined angular and distance generalization with training exclusively on orientation data—demonstrates AVAESA's robustness under extreme domain shift conditions. In this scenario, the model is trained only on subject orientation measurements (front, back, left, right) and tested on completely different measurement modalities involving angular positions (0°, 30°, 45°) and distances (40, 80, 120, 160 cm). This represents the most realistic and demanding deployment challenge where the monitoring system must operate under multiple unseen measurement conditions simultaneously.

Under extreme domain shift conditions, AVAESA maintains considerable cardiac signal modeling capability ($R^2 = 0.738$, $r = 0.865$) while VAE experiences critical performance degradation ($R^2 = 0.024$, $r = 0.366$), as illustrated in Figure 14. The preservation of 73.8% variance explanation, compared to VAE's severely compromised performance (2.4%), demonstrates the architectural advantages of self-attention temporal modeling and adaptive signal processing. The 62.0% MAE improvement (8.18 $\rightarrow$ 3.11 BPM) highlights the substantial enhancement in estimation accuracy and reliability of the system. AVAESA achieves individual participant accuracy within the 1.48–7.15 BPM range, maintaining high reliability even under the most challenging conditions, while VAE exhibits uniformly poorer performance with maximum errors approaching 15 BPM, as shown in Figure 15. Figure 16 further illustrates the narrower error distribution ($\pm 21.80 \rightarrow \pm 11.21$ BPM), highlighting AVAESA's improved robustness and consistency under extreme domain shifts.

The multi-scenario generalization performance of AVAESA under extreme domain shift conditions demonstrates the synergistic effect of its architectural innovations. When faced with multiple unseen measurement modalities simultaneously, the self-attention mechanism proves crucial in identifying the core cardiac signal patterns that transcend specific measurement configurations. The adaptive signal processing layer dynamically reconfigures itself to

handle the compound complexity of multiple domain shifts, while the variational framework provides robust uncertainty quantification that prevents overfitting to limited training scenarios. This architectural resilience enables AVAESA to maintain meaningful performance even when deployment conditions differ dramatically from training scenarios, a critical capability for real-world radar-based monitoring systems.

4.2.4. Cross-Domain Performance Summary

The systematic cross-domain evaluation establishes AVAESA's enhanced generalization capabilities across all measurement modalities and complexity levels. This analysis employs the leave-one-scenario-out experimental protocol described in Section 3.3, where training datasets exclude specific measurement conditions (distance, angular, or orientation scenarios) that are reserved exclusively for testing. Performance improvements range from 17.3% under optimal conditions to 62.6% under challenging single-scenario generalization, with the most significant advantages emerging when training and testing conditions exhibit substantial domain shift.

Table 6 presents quantitative evidence of AVAESA's superior cross-domain performance across four experimental configurations. The comprehensive evaluation reveals a clear performance hierarchy where AVAESA's advantages become more pronounced as domain shift complexity increases.

Table 6: Comprehensive cross-domain performance analysis

Scenario	**VAE MAE (BPM)**	**AVAESA MAE (BPM)**	**MAE Improv.**	**VAE R^2**	**AVAESA R^2**	**LoA Improv.**
Comprehensive	1.68 ± 0.29	1.39 ± 0.22	17.3%	0.911	0.943	19.8%
Angular Test	2.90 ± 0.58	1.65 ± 0.26	43.1%	0.687	0.912	46.5%
Distance Test	7.11 ± 1.24	2.66 ± 0.31	62.6%	0.231	0.841	54.7%
Multi-Scenario	8.18 ± 1.47	3.11 ± 0.42	62.0%	0.024	0.738	48.6%

The superior generalization performance of AVAESA stems from three specific architectural mechanisms that enable robust cross-domain feature extraction. First, the multi-head self-attention mechanism captures long-range temporal dependencies that remain invariant across measurement conditions by learning attention patterns that focus on cardiac signal characteristics independent of radar positioning or subject orientation. Unlike conventional VAE encoders that rely on local convolutional features sensitive

to measurement-specific artifacts, self-attention dynamically weights temporal positions based on physiological relevance rather than spatial proximity, enabling the model to identify consistent cardiac patterns even when signal amplitude or phase characteristics change due to domain shift.

Second, the adaptive signal processing layer addresses the critical limitation of fixed preprocessing parameters that embed measurement-specific biases into the feature extraction pipeline. Through learnable frequency band selection and multi-scale temporal analysis, this component automatically adjusts to signal characteristics present in unseen measurement conditions. The adaptive parameter extraction in Equation 17 derives processing parameters $\boldsymbol{\phi}(\mathbf{I}, \mathbf{Q})$ directly from current signal properties rather than pre-defined fixed parameter assumptions , enabling the system to maintain optimal signal conditioning even when deployment scenarios differ substantially from training conditions.

Third, the uncertainty-weighted latent fusion mechanism (Equations 33 -34) provides robust integration of I and Q channel information by dynamically adjusting the relative contributions based on signal quality estimates. This enables the model to maintain accurate heart rate estimation even when one channel experiences degradation due to unfavorable measurement geometry or environmental interference, a common occurrence under cross-domain deployment conditions.

Under comprehensive training with abundant data diversity, both architectures achieve high accuracy, though AVAESA maintains consistent advantages across all metrics in Table 6. However, as evaluation progresses from single-scenario to multi-scenario generalization challenges, AVAESA's architectural innovations become increasingly critical for maintaining system functionality. The most striking finding is AVAESA's ability to prevent critical performance degradation under extreme domain shift conditions, where VAE performance drops substantially ($\mathbf{R^2} = 0.024$) while AVAESA maintains high and meaningful accuracy ($\mathbf{R^2} = 0.738$).

These architectural advantages enable AVAESA to maintain meaningful performance (MAE $\leq$ 3.11 BPM) even under the most challenging multi-scenario generalization conditions, representing a fundamental advancement in practical radar-based vital sign monitoring systems that enables reliable deployment in diverse real-world environments where measurement conditions cannot be precisely controlled or predicted.

Failure Case Analysis: Examination of per-participant performance reveals systematic patterns in estimation failures. Under comprehensive

training (Figure 9), participant P2 exhibited the highest error for both architectures (VAE: 4.00 BPM, AVAESA: 3.00 BPM), suggesting individual-specific signal characteristics that challenge model generalization. Under cross-domain conditions (Figure 12), distance domain shift produced the most severe degradation, with VAE errors exceeding 13 BPM for multiple participants while AVAESA maintained sub-5 BPM performance for 8 of 10 individuals. The multi-scenario evaluation (Figure 15) revealed that VAE performance collapsed uniformly across all participants (MAE range: 5.50–14.67 BPM), whereas AVAESA exhibited greater inter-participant variance (MAE range: 1.48–7.15 BPM), indicating that AVAESA's adaptive mechanisms provide differential benefit depending on individual signal quality. These failure patterns suggest that remaining estimation errors stem primarily from (1) low SNR conditions at extended distances and (2) individual physiological variations in cardiac signal morphology rather than systematic architectural limitations.

4.2.5. Attention Weight Visualization

To address interpretability of AVAESA's self-attention mechanisms, Figure 17 visualizes the learned attention weights produced by the adaptive signal processor. The attention mechanism operates on four signal representations derived from the raw complex I/Q radar data: magnitude ($|I + jQ|$), phase ($\angle(I + jQ)$), real component (I), and imaginary component (Q). Each representation captures a distinct aspect of cardiac-induced chest wall motion. Specifically, magnitude encodes displacement amplitude, phase provides fine-grained temporal information, while the raw I/Q components preserve the complete complex signal structure and phase-sensitive characteristics.

Figures 17(a) and (b) show the normalized time-domain I- and Q-channel radar signals, which exhibit a quadrature phase relationship and jointly encode subtle cardiac motion. The corresponding frequency-domain representation in Figure 17(c) confirms that dominant signal energy lies within the physiological cardiac band (0.5–4.0 Hz), with prominent components near 60 and 90 BPM, validating the cardiac origin of the observed motion patterns.

The learned attention weights shown in Figure 17(d) reveal that AVAESA assigns non-uniform and structured importance across signal representations. In contrast to fixed or heuristic preprocessing pipelines, the model adaptively emphasizes signal components based on their relevance for cardiac feature extraction. The numerical attention values, overlaid on the attention map,

indicate that the imaginary (Q) and magnitude representations receive the highest aggregate importance, while phase and real (I) components contribute complementary but lower-weighted information. This distribution reflects the strong contribution of amplitude-driven motion information and phase-sensitive quadrature components to robust heart-rate estimation.

By learning these attention weights directly from data, AVAESA eliminates the need for hand-crafted decisions regarding signal representation selection and instead discovers optimal representation combinations in a fully data-driven manner. The structured attention patterns further demonstrate meaningful interactions between complex signal components rather than arbitrary feature weighting. Overall, the visualization confirms that AVAESA's self-attention mechanism learns physiologically meaningful representations, providing interpretable evidence for its robust performance under cross-domain and cross-scenario measurement conditions.

4.2.6. Component Ablation Analysis

To quantify each architectural component's contribution to cross-domain generalization, we conducted systematic ablation experiments under the Distance Test scenario—the most challenging single-domain condition where models train exclusively on angular position and subject orientation data while testing on unseen distance measurements. While Section 3.3 validated individual design choices in isolation to justify architectural decisions, this ablation systematically removes each component from the complete AVAESA framework to quantify its contribution within the integrated system. Each ablation removes exactly one component while preserving all others. Statistical significance was assessed using paired t-tests with Bonferroni correction across all test sequences ($n = 480$). Results are summarized in Table 7.

Table 7: Component ablation study under distance domain shift. All comparisons against Full AVAESA; $^{***}p < 0.001$, $^{**}p < 0.01$.

Variant	**MAE (BPM)**	$\mathbf{R^2}$	**Pearson** r	**ΔMAE**
Full AVAESA	2.66 ± 0.31	0.841	0.919	—
No Uncertainty Weighting	$3.18 \pm 0.38^{**}$	0.798	0.896	+19.5%
No Dual-Stream I/Q	$3.54 \pm 0.42^{***}$	0.807	0.904	+33.1%
No Adaptive Preprocessing	$3.85 \pm 0.51^{***}$	0.712	0.857	+44.7%
No Self-Attention	$4.62 \pm 0.67^{***}$	0.648	0.821	+73.7%
Baseline VAE	$7.11 \pm 1.24^{***}$	0.231	0.554	+167.3%

Removing self-attention and substituting mean pooling produces the largest degradation (+73.7%), confirming that dynamic temporal weighting is essential for identifying cardiac-relevant segments while suppressing measurement-specific artifacts. The corresponding R^2 reduction (0.841 $\rightarrow$ 0.648) demonstrates that fixed pooling cannot capture the long-range dependencies required for cross-domain robustness. Replacing learnable preprocessing with fixed Butterworth filters (0.7–3.5 Hz) causes the second-largest degradation (+44.7%), validating that signal-adaptive parameter selection—based on SNR, spectral shape, and dominant frequency—enables generalization across measurement geometries where fixed parameters fail. Substituting a shared encoder for separate I/Q pathways degrades performance by 33.1%; the preserved correlation ($r = 0.904$) but reduced R^2 (0.807) indicates that single-stream processing retains gross signal structure but loses timing information encoded in I/Q phase relationships. Finally, replacing learned uncertainty weighting with fixed equal-weight fusion causes the smallest degradation (+19.5%), suggesting adaptive weighting becomes critical primarily when channel quality diverges under adverse conditions.

To assess component interactions, we conducted a joint ablation removing both self-attention and adaptive preprocessing. The resulting MAE (6.43 BPM) exceeds the predicted additive degradation (5.81 BPM) by 10.7%, indicating positive synergy: self-attention benefits from cleaner input representations that adaptive preprocessing provides.

The baseline VAE employs identical latent dimensionality ($d = 64$) and training configuration but uses a single convolutional encoder on concatenated I/Q channels, fixed preprocessing, and mean pooling. Its substantially higher error (MAE = 7.11 BPM) confirms that conventional variational architectures lack the adaptive mechanisms necessary for cross-domain generalization. Collectively, these results establish a clear architectural hierarchy: temporal modeling via self-attention forms the foundation for robustness, adaptive preprocessing enables measurement-invariant features, dual-stream processing preserves phase information, and uncertainty weighting handles asymmetric channel degradation.

5. Discussion

This study presents the first comprehensive evaluation of enhanced VAE architectures for radar-based heart rate estimation. Our findings establish AVAESA's clear superiority over conventional VAE approaches while demon-

strating substantial improvements in cross-domain generalization capabilities.

5.1. Architectural Innovations and Performance Analysis

AVAESA's fundamental innovation lies in three integrated architectural mechanisms that address critical limitations of conventional VAE approaches. **First**, the adaptive signal processor (Section 3.2.1) eliminates fixed preprocessing assumptions by deriving frequency bands, multi-scale weights, and component fusion parameters directly from input signal characteristics—SNR, spectral shape, and dominant frequencies—rather than applying predetermined 0.7-3.5 Hz filters. **Second**, the dual-stream I/Q encoder with uncertainty-weighted fusion (Section 3.2.3) preserves critical phase information through separate I and Q processing pathways, dynamically adjusting channel contributions based on learned variance estimates. **Third**, the self-attention mechanism (Section 3.2.2) captures long-range temporal dependencies across multiple scales—beat-to-beat variations, respiratory modulation, and heart rate variability trends—that conventional pooling operations discard. This architectural distinction becomes increasingly critical under domain shift conditions, where traditional approaches degrade due to dependence on training-scenario features, while AVAESA's adaptive mechanisms maintain robust performance by learning measurement-invariant cardiac representations.

5.2. Cross-Domain Generalization and Robustness Analysis

The deployment of radar-based physiological monitoring systems in uncontrolled environments necessitates robust cross-domain generalization capabilities to mitigate the deleterious effects of domain shift on predictive accuracy [49]. Our comprehensive evaluation demonstrates that AVAESA exhibits improved domain adaptation characteristics compared to conventional VAE architectures across multiple generalization scenarios.

Domain shift in radar-based cardiac monitoring manifests through spatiotemporal variations in signal propagation characteristics, including range-dependent attenuation effects, angular-dependent scattering patterns, and multipath interference variations. These phenomena fundamentally alter the statistical properties of received signals, creating significant challenges for models trained under constrained laboratory conditions. AVAESA's architectural innovations—particularly the uncertainty-weighted I/Q fusion (Equations 33-34) and self-attention temporal modeling (Equation 25)—demonstrate

effectiveness in learning domain-invariant representations that transcend these measurement-specific perturbations.

Quantitative analysis shows AVAESA's performance retention under domain shift conditions. Under distance generalization, AVAESA preserves 84.1% of the explained variance ($R^2 = 0.841$) with a Pearson correlation of $r = 0.919$, representing a 264% improvement in variance explanation compared to the baseline VAE ($R^2 = 0.231$, $r = 0.554$). The mean absolute error reduction from 7.11 BPM to 2.66 BPM constitutes a 62.6% performance improvement, demonstrating significant enhancement in predictive accuracy.

Similarly, under angular domain shift conditions, AVAESA maintains robust performance with $R^2 = 0.912$ and $r = 0.956$, while VAE performance degrades substantially ($R^2 = 0.687$, $r = 0.836$). The MAE improvement from 2.90 BPM to 1.65 BPM represents a 43.1% error reduction, confirming the architecture's capacity to generalize across diverse measurement geometries. These findings support AVAESA's potential for deployment in naturalistic monitoring scenarios where geometric constraints cannot be controlled.

5.3. Design Validation and Architectural Optimization

Our systematic design validation studies provide crucial insights into optimal architectural configurations for radar-based physiological monitoring. Three key findings emerge from this analysis.

First, the 128-dimensional latent space represents the optimal balance between representational capacity and generalization robustness. Both under-capacity and over-capacity configurations show significant performance degradation, confirming the importance of careful dimensionality selection.

Second, the 8-head attention configuration validation demonstrates that multi-head attention requires precise calibration for physiological signal processing. This configuration enables simultaneous modeling of multiple temporal patterns, including beat-to-beat variations, respiratory modulation, and heart rate variability trends.

Third, separate I/Q processing strategy validation confirms the critical importance of preserving complex signal structure for cardiac feature extraction. The 15.4% performance advantage over magnitude-only processing demonstrates that phase information contains essential timing relationships for accurate heart rate estimation. This finding has broader implications for radar signal processing in physiological monitoring applications.

5.4. Scope and Future Directions

This study establishes AVAESA's architectural advantages for cross-scenario robustness within a defined experimental scope. The following discussion characterizes this scope and identifies promising directions for extending these architectural innovations toward broader deployment contexts.

Dataset Scale and Cross-Subject Generalization. Our evaluation employs 10 participants across 48 measurement scenarios, focusing on architectural robustness to measurement geometry changes rather than inter-subject generalization. This design choice enables systematic isolation of architectural effects on cross-scenario performance, independent of inter-individual physiological variability.

To directly address leave-one-subject-out (LOSO) evaluation, we conducted preliminary experiments training on 7 participants and testing on 3 held-out individuals (P8, P9, P10). Both architectures exhibited substantially degraded performance: VAE achieved MAE = 16.87 BPM with Pearson $r = -0.214$, while AVAESA achieved MAE = 15.93 BPM with Pearson $r = 0.478$. The negative correlation for VAE indicates complete failure to capture cross-subject cardiac patterns, whereas AVAESA's positive correlation suggests preserved directional tracking despite poor absolute accuracy. Per-participant analysis revealed high variance (AVAESA: P8 = 6.35 BPM, P9 = 26.00 BPM, P10 = 15.87 BPM), confirming that individual physiological differences—chest wall thickness, cardiac morphology, respiration patterns—dominate when training data lacks sufficient subject diversity. These results establish that cross-subject generalization represents a fundamentally distinct challenge from cross-scenario robustness, requiring substantially larger cohorts or explicit domain adaptation techniques.

Extending evaluation to larger participant cohorts (30+ individuals) with demographic stratification would establish population-level generalizability. Few-shot learning and meta-learning approaches represent promising directions for enabling rapid personalization to new individuals with minimal calibration data, potentially achieving acceptable performance with limited per-user measurements. Such techniques would complement AVAESA's cross-scenario robustness by addressing cross-subject adaptation, forming a comprehensive solution for deployment across diverse populations and measurement conditions.

Static Conditions and Motion Robustness. Current evaluation protocols involve controlled, static postures enabling systematic architectural

validation without confounding motion artifacts. Real-world deployment scenarios including ambulatory monitoring, in-vehicle sensing, and home healthcare require robustness to body motion, respiratory dynamics, and postural changes. Future research should conduct systematic motion artifact testing under standardized protocols (walking at varied speeds, sitting with natural movement, driving on varied road conditions). AVAESA's adaptive signal processing framework provides theoretical motion resilience through real-time, signal-driven parameter adjustment, but empirical validation remains necessary. Integration with complementary technologies—motion compensation algorithms derived from accelerometer data, IMU sensor fusion for movement-aware signal conditioning, or multi-radar configurations enabling motion-invariant beamforming—could further enhance robustness. Motion-augmented training datasets incorporating labeled activity states would enable development of motion-adaptive preprocessing strategies.

Heart Rate Estimation and Dataset Selection. This study focuses on heart rate estimation as the primary evaluation target. Heart rate monitoring presents distinct measurement challenges compared to respiratory rate, including smaller chest displacement amplitudes and more subtle signal characteristics, making it an appropriate test case for architectural validation. The Sadeghi et al. dataset [47] provides systematic variation across 48 measurement scenarios (4 distances $\times$ 3 angles $\times$ 4 orientations), making it well-suited for evaluating cross-scenario robustness. The dataset includes heart rate ground truth from synchronized Polar H10 chest strap monitors. AVAESA's architectural principles—adaptive signal preprocessing, uncertainty-weighted multi-stream fusion, and attention-based temporal modeling—operate on periodic physiological signals and generalize naturally to other vital signs including respiratory rate, blood pressure, and heart rate variability. Future work should extend evaluation to multi-parameter monitoring, including simultaneous respiratory rate, blood pressure, and heart rate variability estimation, to validate AVAESA's architectural versatility across multiple physiological signals. Such extensions would demonstrate the generalizability of adaptive signal preprocessing and attention-based temporal modeling beyond cardiac signals, while potentially enabling improved accuracy through joint optimization of correlated vital sign parameters.

Baseline Comparisons and Emerging Architectures. Our baseline selection (VAE, LSTM, Bi-LSTM, CNN, TCN) follows the systematic survey by Shirazi et al. [11], which identified LSTM, CNN, and VAE among the most prevalent architectures in radar-based vital sign monitoring lit-

erature. TCN is included as a modern temporal modeling alternative that employs dilated causal convolutions to capture long-range dependencies without sequential processing constraints. Contemporary Transformer architectures (Informer, Autoformer, Temporal Fusion Transformer) currently lack established implementations for radar-based cardiac monitoring, preventing direct comparison. As the field matures and Transformer-based radar implementations emerge, comparative evaluation would strengthen understanding of AVAESA's relative contributions. Promising research directions include hybrid architectures combining AVAESA's adaptive, measurement-invariant signal preprocessing with Transformer backbones. Such combinations could leverage Transformer's global self-attention mechanisms for very long-range dependencies (minutes to hours) while preserving AVAESA's robust handling of measurement domain shift. Architectural search methods (Neural Architecture Search, Efficient Architecture Search) could systematically explore the design space of adaptive preprocessing coupled with modern temporal modeling approaches.

Computational Efficiency and Edge Deployment. AVAESA contains 41.8M parameters compared to 4.0M for baseline VAE, reflecting additional complexity from dual-stream I/Q encoders with dynamic convolutions, multi-head self-attention mechanisms, signal context extraction, and adaptive preprocessing modules. Per-batch processing time increases approximately 4× (532 ms vs. 127 ms on RTX 4080 GPU), reflecting the computational overhead of attention operations and dynamic filter routing. Nevertheless, inference latency remains well within real-time requirements for 60-second monitoring windows updated at typical clinical intervals. Deployment on resource-constrained platforms—battery-powered wearables, automotive-grade embedded systems, or low-power IoT edge devices—may require model compression. Structured pruning could reduce attention heads (from 8 to 4–6 with minimal accuracy loss based on Table 2), while post-training quantization and knowledge distillation offer additional pathways for reducing computational demands while preserving cross-scenario robustness.

Controlled Laboratory to Naturalistic Deployment. The dataset comprises measurements collected under standardized laboratory protocols: 1-minute signal windows, single-subject monitoring, and controlled electromagnetic conditions. This controlled experimental design enables systematic architectural evaluation by isolating measurement geometry effects from environmental confounds, providing an appropriate testbed for validating cross-scenario robustness mechanisms. Extending evaluation to naturalistic

deployment scenarios would address practical operational challenges beyond laboratory conditions. Multi-person monitoring scenarios (vehicle cabins with multiple occupants, shared living spaces) would require person-specific signal separation capabilities, potentially achievable through MIMO radar beamforming or learned spatial filtering techniques. Variable measurement durations spanning rapid vital sign checks (5-10 seconds) to extended continuous monitoring (5+ minutes) would necessitate adaptive windowing strategies beyond fixed 1-minute windows. Environmental robustness testing under electromagnetic interference (Wi-Fi/Bluetooth devices, automotive electronics), multipath propagation from metallic objects, and thermal variations would establish operational boundaries for real-world deployment. Such evaluations would guide system design decisions for diverse applications ranging from smart home elderly care to clinical intensive care monitoring.

This experimental scope provides controlled conditions necessary for systematic validation of architectural innovations while establishing clear pathways for extending AVAESA's cross-scenario robustness toward diverse real-world applications. The identified research directions collectively address the translation from architectural demonstration to practical deployment across healthcare, automotive, and ambient intelligence domains.

5.5. Reproducibility and Open Science

To ensure reproducibility, we provide complete methodological specifications enabling independent replication. All architectural details, hyperparameters, and training configurations are documented in Sections 3.2 and 3.3. The publicly available dataset [47] utilized in this evaluation ensures that researchers can validate and extend these findings using identical ground truth measurements and experimental conditions.

Implementation code and trained model weights will be released upon acceptance at a public repository. The release will include: (1) complete PyTorch implementation of the AVAESA architecture, (2) trained model checkpoints for all experimental configurations reported in this study, (3) preprocessing scripts compatible with the Sadeghi et al. dataset, and (4) evaluation scripts to reproduce all reported metrics and figures.

6. Conclusion

This study presents AVAESA as an enhanced VAE architecture for radar-based heart rate estimation through systematic architectural evaluation. AVAESA

introduces significant architectural innovations including multi-head self-attention for temporal modeling, signal-driven adaptive preprocessing that derives frequency bands and processing weights from input characteristics, and dual-stream I/Q encoding with uncertainty-weighted fusion that achieve performance improvements ranging from 17.3% under optimal conditions to 62.6% under challenging domain shift scenarios. Most significantly, AVAESA maintains meaningful performance (MAE $\leq$ 3.11 BPM) under extreme generalization conditions where traditional approaches fail.

The comprehensive validation studies confirm optimal configurations of a 128-dimensional latent space, 8-head attention mechanisms, and separate I/Q processing strategies, providing practical guidance for robust contactless monitoring systems. AVAESA's cross-domain generalization addresses the deployment challenge of maintaining accuracy when operational conditions differ from training scenarios, enabling practical deployment in healthcare facilities, smart vehicles, and home monitoring systems. These capabilities are achieved through superior noise handling, adaptive signal processing, and attention-based temporal modeling that effectively generalizes across diverse measurement conditions.

Declaration of Interest

The authors, Mohammad Hossein Shirazi, Sira Yongchareon, Anuradha Singh, and Jing Ma, hereby declare that they have no known financial or personal relationships that could inappropriately influence or bias the content of this manuscript. The research was conducted independently, and no external entity had any role in study design, data collection, analysis, interpretation, or manuscript preparation.

References

[1] R. u. S. Ahmad, W. U. Khan, M. S. Khan, P. Cheung, Emerging rapid detection methods for the monitoring of cardiovascular diseases: Current trends and future perspectives, Materials Today Bio 32 (2025) 101663. doi:10.1016/j.mtbio.2025.101663.

[2] N. Oliveira, F. Ribeiro, A. Alves, M. Teixeira, F. Miranda, J. Oliveira, Heart rate variability in myocardial infarction patients: Effects of exercise training, Revista Portuguesa de Cardiologia 32 (9) (2013) 687–700. doi:10.1016/j.repc.2013.02.010.

[3] L. N. Nguyen, P. Susarla, A. Mukherjee, M. L. Cañellas, C. Álvarez Casado, X. Wu, O. Silvén, D. B. Jayagopi, M. B. López, Non-contact multimodal indoor human monitoring systems: A survey, Information Fusion 110 (2024) 102457. doi:https://doi.org/10.1016/j.inffus.2024.102457.

[4] M. Alizadeh, G. Shaker, J. D. Almeida, P. Morita, S. Safavi-Naeini, Remote monitoring of human vital signs using mm-wave fmcw radar, IEEE Access 7 (2019) 54958–54968. doi:10.1109/ACCESS.2019.2912956.

[5] P. Mathurkar, A. Gaikwad, Advancements in non-contact radar-based techniques for vital sign detection – a review, in: Proc. 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2023, pp. 1–6. doi:10.1109/ICCUBEA58933.2023.10392271.

[6] Y. Wang, et al., Remote monitoring of human vital signs based on 77-ghz mm-wave fmcw radar, Sensors 20 (2020). doi:10.3390/s20102999.

[7] A. Gharamohammadi, A. Khajepour, G. Shaker, In-vehicle monitoring by radar: A review, IEEE Sensors Journal 23 (21) (2023) 25650–25672. doi:10.1109/jsen.2023.3316449.

[8] F. Khan, S. Azou, R. Youssef, P. Morel, E. Radoi, Ir-uwb radar-based robust heart rate detection using a deep learning technique intended for vehicular applications, Electronics 11 (16) (2022). doi:10.3390/electronics11162505.

[9] H. Chang, C. Hsu, W. Chung, Fast acquisition and accurate vital sign estimation with deep learning-aided weighted scheme using fmcw radar, in: IEEE 95th Vehicular Technology Conference (VTC2022-Spring), 2022, pp. 1–6.

[10] S. Junaid, A. Imam, A. Shuaibu, S. Basri, G. Kumar, Y. Surakat, A. Balogun, M. Abdulkarim, Artificial intelligence, sensors and vital health signs: A review, Applied Sciences 12 (22) (2022). doi:10.3390/app122211475.

[11] M. H. Shirazi, S. Yongchareon, A. Singh, J. Ma, A survey on machine learning approaches for vital sign monitoring using radar, MeasurementIn press (2025).

[12] S. Mehrdad, F. Shamout, Y. Wang, et al., Deep learning for deterioration prediction of covid-19 patients based on time-series of three vital signs, Scientific Reports 13 (1) (2023) 9968.

[13] S. Zhang, T. Zheng, Z. Chen, J. Luo, Can we obtain fine-grained heartbeat waveform via contact-free rf-sensing?, in: IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, 2022, pp. 1759–1768.

[14] G. Mauro, M. D. C. Diez, J. Ott, L. Servadei, M. Cuellar, Few-shot user-adaptable radar-based breath signal sensing, Sensors 23 (2) (2023). doi:10.3390/s23020804.

[15] Y. Jang, J. Sim, J. Yang, N. Kwon, Improving heart rate variability information consistency in doppler cardiogram using signal reconstruction system with deep learning for contact-free heartbeat monitoring, Biomedical Signal Processing and Control 76 (2022). doi:10.1016/j.bspc.2022.103691.

[16] Y. Xia, C. Chen, M. Shu, R. Liu, A denoising method of ecg signal based on variational autoencoder and masked convolution, Journal of Electrocardiology 80 (2023) 81–90. doi:https://doi.org/10.1016/j.jelectrocard.2023.05.004.

[17] Y. Wang, T. Zhao, X. Wang, Fine-grained heartbeat waveform monitoring with rfid: A latent diffusion model, HumanSys '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 86–91. doi:10.1145/3722570.3726891.
URL https://doi.org/10.1145/3722570.3726891

[18] A. Jamal, R. K. Ramasamy, J. Abdullah, Generative ai respiratory and cardiac sound separation using variational autoencoders (vaes), Computer Sciences & Mathematics Forum 10 (1) (2025) 9–0. doi:10.3390/cmsf2025010009.

[19] Z. Nowroozilarki, B. J. Mortazavi, R. Jafari, Variational autoencoders for biomedical signal morphology clustering and noise detection, IEEE Journal of Biomedical and Health Informatics 28 (1) (2024) 169–180. doi:10.1109/JBHI.2023.3320585.

[20] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[21] T. Zheng, Z. Chen, S. Zhang, C. Cai, J. Luo, More-fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar, in: 19th ACM Conference on Embedded Networked Sensor Systems, 2021, pp. 111–124.

[22] Z. Wang, C. Wang, Y. Li, Variational autoencoder based on knowledge sharing and correlation weighting for process-quality concurrent fault detection, Engineering Applications of Artificial Intelligence 133 (2024) 108051. doi:https://doi.org/10.1016/j.engappai.2024.108051.

[23] H. Haule, I. Piper, P. Jones, C. Qin, T.-Y. M. Lo, J. Escudero, Vae-if: Deep feature extraction with averaging for fully unsupervised artifact detection in routinely acquired icu time-series, Computers in Biology and Medicine 186 (2025) 109610. doi:https://doi.org/10.1016/j.compbiomed.2024.109610.

[24] Y. Takida, W.-H. Liao, C.-H. Lai, T. Uesaka, S. Takahashi, Y. Mitsufuji, Preventing oversmoothing in vae via generalized variance parameterization, Neurocomputing 509 (2022) 137–156. doi:https://doi.org/10.1016/j.neucom.2022.08.067.

[25] D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, arXiv preprint arXiv:1401.4082 (2014).

[26] R. Wei, C. Garcia, A. El-Sayed, V. Peterson, A. Mahmood, Variations in variational autoencoders - a comparative evaluation, IEEE Access 8 (2020) 153651–153670.

[27] M. E. Torres, M. A. Colominas, G. Schlotthauer, P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4144–4147.

[28] R. M. Leushuis, Probabilistic forecasting with var-vae: Advancing time series forecasting under uncertainty, Information Sciences 713 (2025) 122184. doi:https://doi.org/10.1016/j.ins.2025.122184.

[29] S. He, M. Du, X. Jiang, W. Zhang, C. Wang, Vaeat: Variational autoeencoder with adversarial training for multivariate time

series anomaly detection, Information Sciences 676 (2024) 120852. doi:https://doi.org/10.1016/j.ins.2024.120852.

[30] L. Sánchez, N. Costa, I. Couso, O. Strauss, Integrating imprecise data in generative models using interval-valued variational autoencoders, Information Fusion 114 (2025) 102659. doi:https://doi.org/10.1016/j.inffus.2024.102659.

[31] S. Islam, O. Boric-Lubecke, V. Lubecke, A. Moadi, A. Fathy, Contactless radar-based sensors: Recent advances in vital-signs monitoring of multiple subjects, IEEE Microwave Magazine 23 (7) (2022) 47–60. doi:10.1109/mmm.2022.3140849.

[32] Y.-X. Lu, Y. Ai, Z.-H. Ling, Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement, Neural Networks 189 (2025) 107562. doi:https://doi.org/10.1016/j.neunet.2025.107562.

[33] M. H. Mohd Noor, A. O. Ige, A survey on state-of-the-art deep learning applications and challenges, Engineering Applications of Artificial Intelligence 159 (2025) 111225. doi:https://doi.org/10.1016/j.engappai.2025.111225.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[35] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open 3 (2022) 111–132. doi:https://doi.org/10.1016/j.aiopen.2022.10.001.

[36] J. Li, X. Wang, Z. Tu, M. R. Lyu, On the diversity of multi-head attention, Neurocomputing 454 (2021) 14–24. doi:https://doi.org/10.1016/j.neucom.2021.04.038.

[37] H. Zhou, J. Li, S. Zhang, S. Zhang, M. Yan, H. Xiong, Expanding the prediction capacity in long sequence time-series forecasting, Artificial Intelligence 318 (2023) 103886. doi:10.1016/j.artint.2023.103886.

[38] H. Chen, X. Zhang, Z. Guo, N. Ying, M. Yang, C. Guo, Actnet: Attention based cnn and transformer network for respiratory rate estimation, Biomedical Signal Processing and Control 96 (2024) 106497. doi:10.1016/j.bspc.2024.106497.

[39] X. Tian, S. Huang, J. Xiao, H. Wang, Y. Liu, Sdvs-net: A spatial dilated convolution and variable self-attention network for multivariate long-term time series forecasting, Neurocomputing 619 (2025) 129148. doi:https://doi.org/10.1016/j.neucom.2024.129148.

[40] J. Liang, Y. Gao, Lightweight memory-driven self-attention for hyperspectral image classification with cnn-transformer cross-feature fusion, Neurocomputing 651 (2025) 130998. doi:https://doi.org/10.1016/j.neucom.2025.130998.

[41] B. Li, W. Zhang, M. Lu, Integrated codec decomposed transformer for long-term series forecasting, Neural Networks 188 (2025) 107484. doi:10.1016/j.neunet.2025.107484.

[42] S. B. Veesam, A. R. Satish, S. Tupakula, Y. Chinnam, K. Prakash, S. Bansal, M. R. I. Faruque, Design of an integrated model with temporal graph attention and transformer-augmented rnns for enhanced anomaly detection, Scientific Reports 15 (1) (2025) 2692. doi:10.1038/s41598-025-85822-5.

[43] K. S. Erer, Adaptive usage of the butterworth digital filter, Journal of Biomechanics 40 (13) (2007) 2934–2943. doi:10.1016/j.jbiomech.2007.02.019.

[44] M. He, Y. Nian, Y. Gong, Novel signal processing method for vital sign monitoring using fmcw radar, Biomedical Signal Processing and Control 33 (2017) 335–345. doi:10.1016/j.bspc.2016.12.008.

[45] M. Mercuri, T. Torfs, M. Rykunov, S. Laureti, M. Ricci, F. Crupi, Analysis of signal processing methods to reject the dc offset contribution of static reflectors in fmcw radar-based vital signs monitoring, Sensors 22 (24) (2022). doi:10.3390/s22249697.

[46] F. Weishaupt, I. Walterscheid, O. Biallawons, J. Klare, Vital sign localization and measurement using an fmcw mimo radar, in: 19th International Radar Symposium (IRS), 2018.

[47] E. Sadeghi, A. Chiumento, P. Havinga, mm-wave fmcw radar vital sign monitoring dataset: Diverse physiological scenarios, 4TU.ResearchData (2024).

[48] B. W. Nelson, C. A. Low, N. Jacobson, P. Areán, J. Torous, N. B. Allen, Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research, npj Digital Medicine 3 (1) (2020) 90. doi:10.1038/s41746-020-0297-4.

[49] C. Chen, Y. Yang, M. Liu, Z. Rong, S. Shu, Regularized joint self-training: A cross-domain generalization method for image classification, Engineering Applications of Artificial Intelligence 134 (2024) 108707. doi:10.1016/j.engappai.2024.108707.

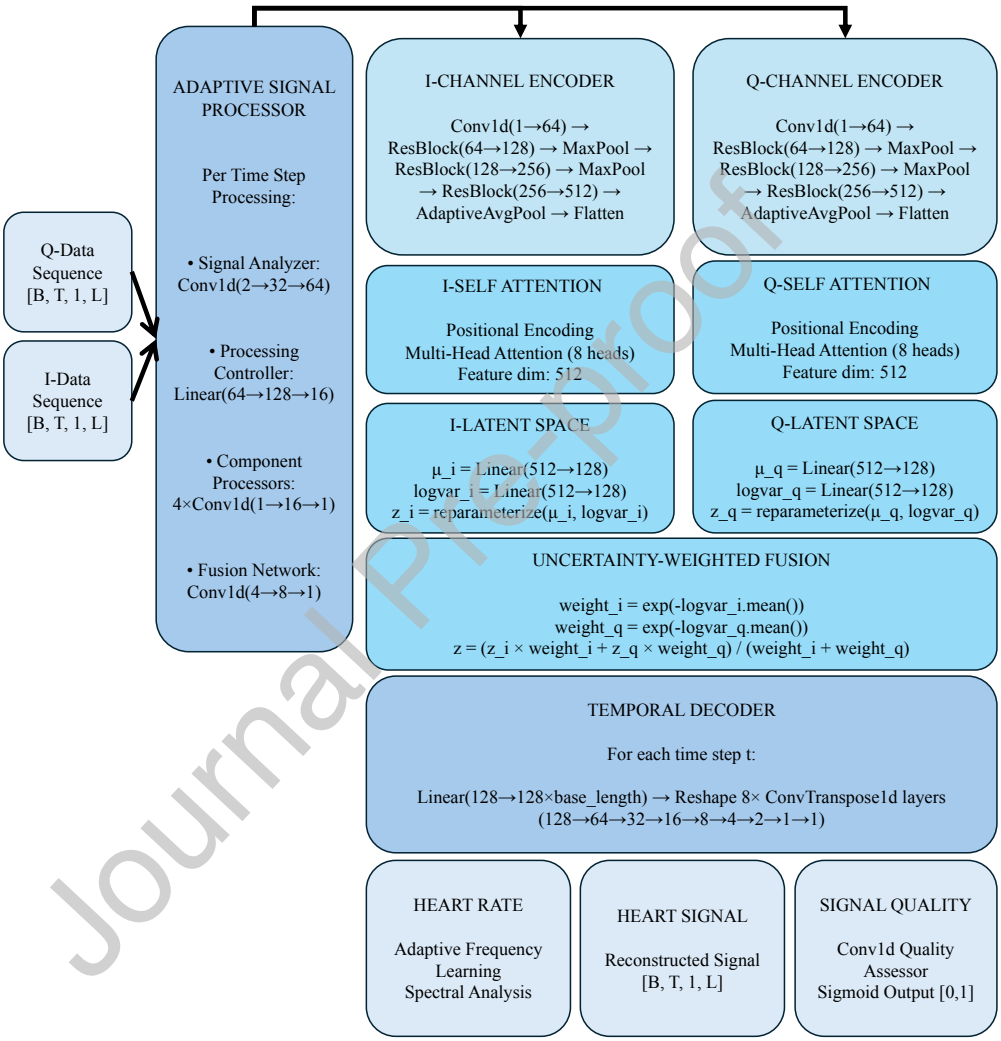


Figure 1: AVAESA architecture

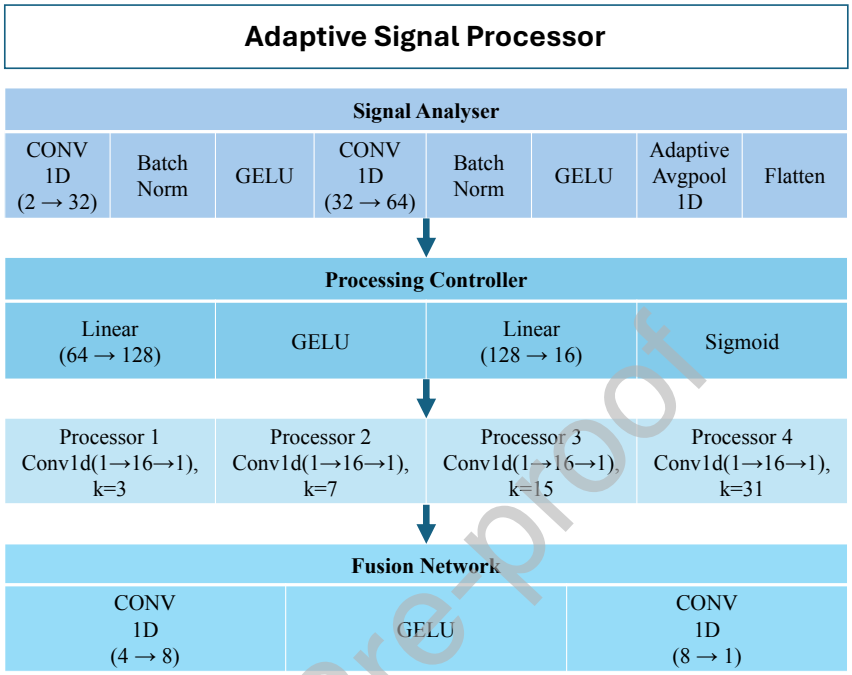


Figure 2: Adaptive Signal Processor architecture

Encoder	Decoder
Input (1 X 2560)	Linear (128 → 128x10) Reshape
Conv1d (1,64) + BatchNorm + GELU	ConvTranspose1d (128,64) + BatchNorm + GELU
Conv1d (64,128) + BatchNorm + GELU	ConvTranspose1d (64,32) + BatchNorm + GELU
MaxPool1d (Stride = 2)	ConvTranspose1d (32,16) + BatchNorm + GELU
Conv1d (128,256) + BatchNorm + GELU	ConvTranspose1d (16,8) + BatchNorm + GELU
MaxPool1d (Stride = 2)	ConvTranspose1d (8,4) + BatchNorm + GELU
Conv1d (256,512) + BatchNorm + GELU	ConvTranspose1d (4,2) + BatchNorm + GELU
Adaptive AvgPool1d	ConvTranspose1d (2,1) + BatchNorm + GELU
Flatten + Linear 512 → μσ (128 dim)	ConvTranspose1d (1,1)
	Heart Signal (1 x 2560)

Figure 3: AVAESA encoder and decoder architecture

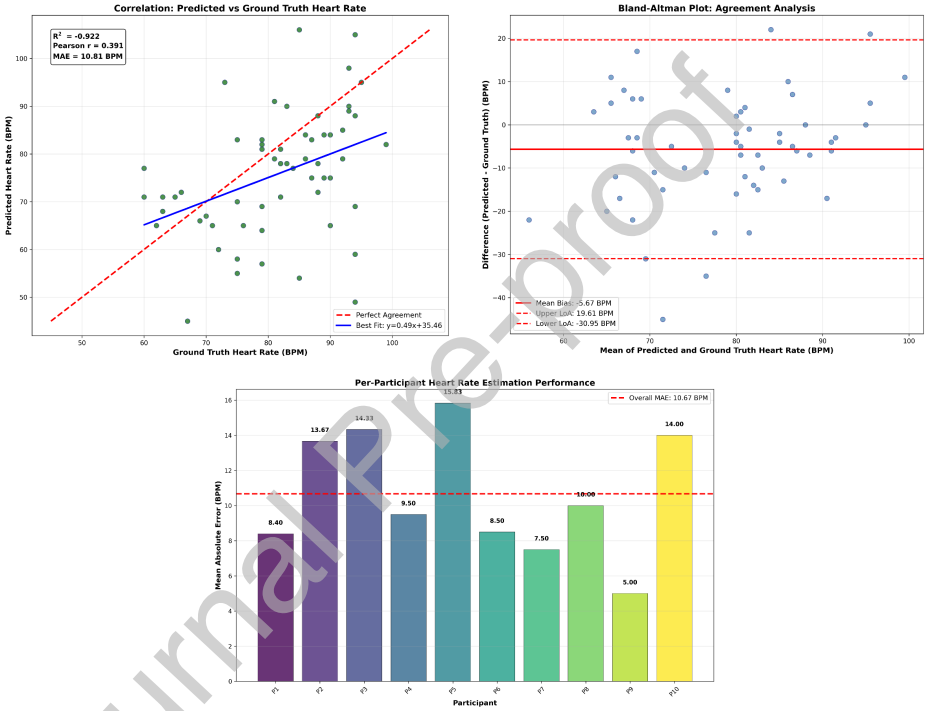


Figure 4: CNN architecture performance: (1) Correlation analysis showing R^2 = -0.922, Pearson r = 0.391, MAE = 10.81 BPM; (2) Bland-Altman analysis showing Mean Bias = -5.67 BPM, LoA range = 50.56 BPM; (3) Per-participant MAE showing high variability (5.00-15.83 BPM)

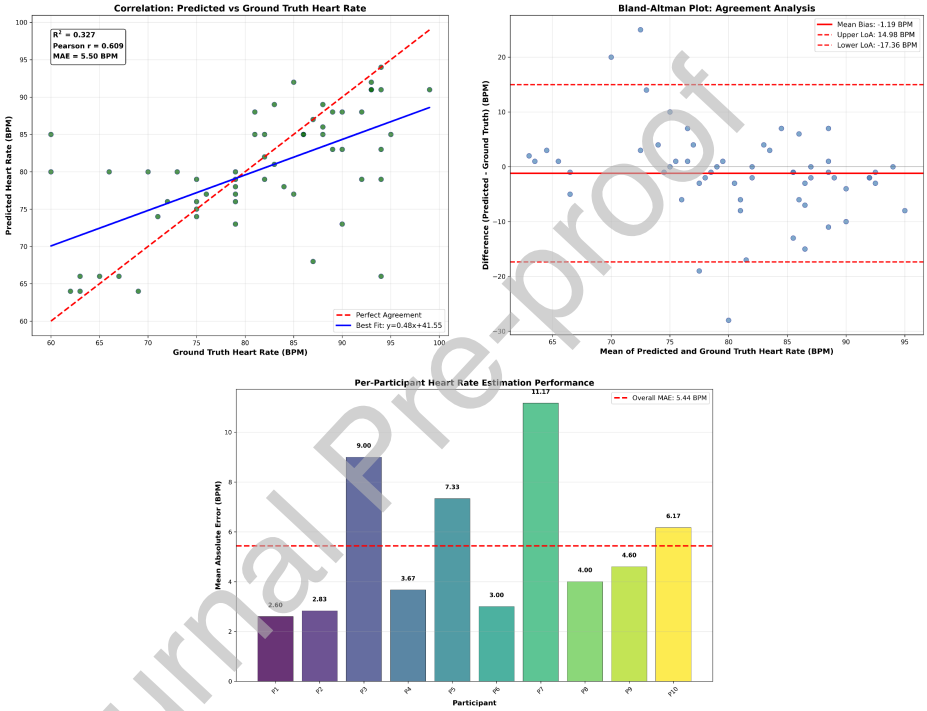


Figure 5: LSTM architecture performance: (1) Correlation analysis showing $R^2 = 0.327$, Pearson r = 0.609, MAE = 5.50 BPM; (2) Bland-Altman analysis showing Mean Bias = -1.19 BPM, LoA range = 32.34 BPM; (3) Per-participant MAE showing moderate variability (2.60-11.17 BPM)

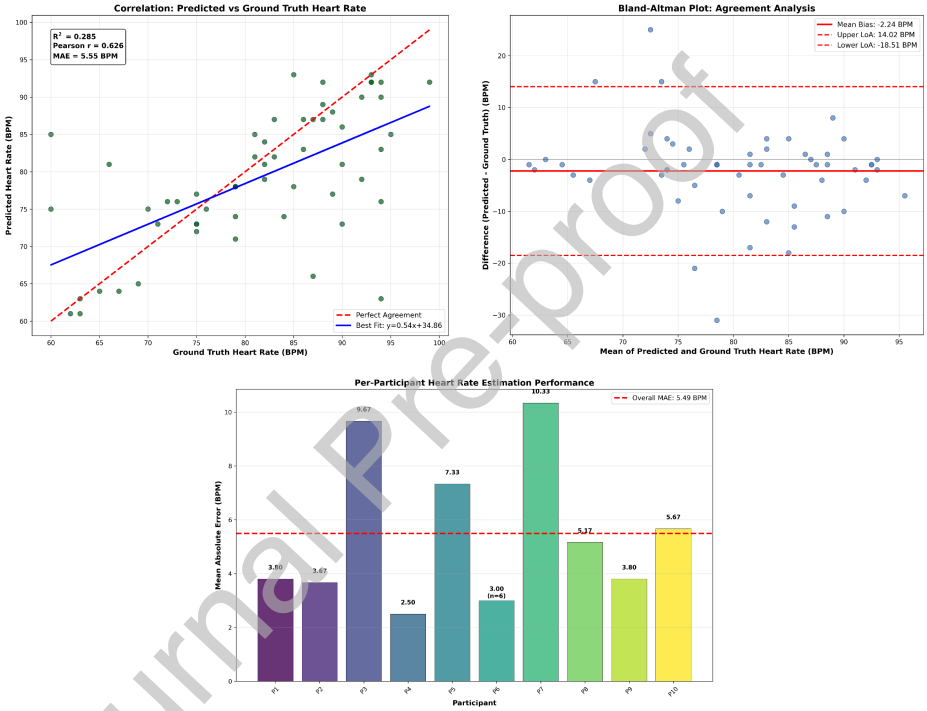


Figure 6: Bi-LSTM architecture performance: (1) Correlation analysis showing R^2 = 0.285, Pearson r = 0.626, MAE = 5.55 BPM; (2) Bland-Altman analysis showing Mean Bias = -2.24 BPM, LoA range = 32.53 BPM; (3) Per-participant MAE showing moderate variability (2.50-10.33 BPM)

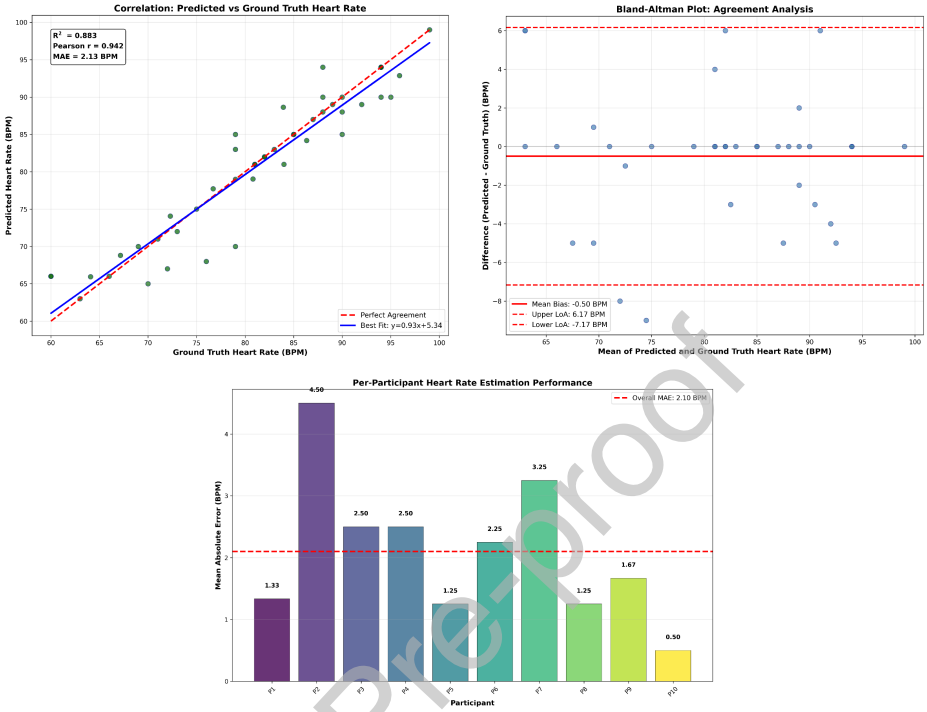


Figure 7: TCN architecture performance: (1) Correlation analysis showing $R^2 = 0.883$, Pearson r = 0.942, MAE = 2.13 BPM; (2) Bland-Altman analysis showing Mean Bias = -0.50 BPM, LoA range = 13.34 BPM; (3) Per-participant MAE showing variability (0.50-4.50 BPM)

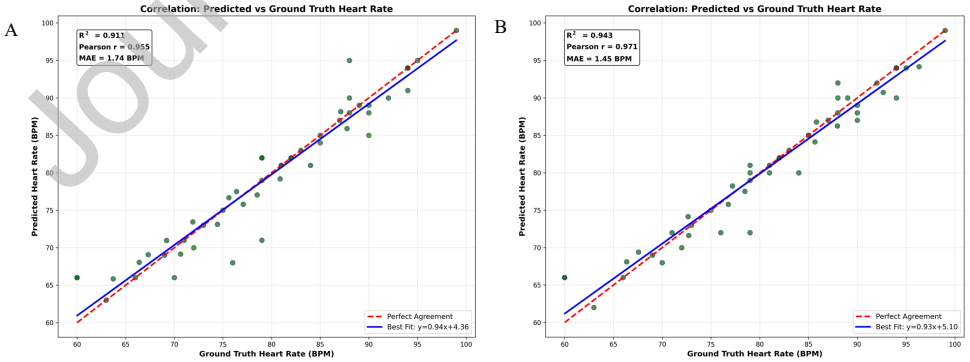


Figure 8: Comprehensive training correlation analysis comparing A) VAE and B) AVAESA architectures

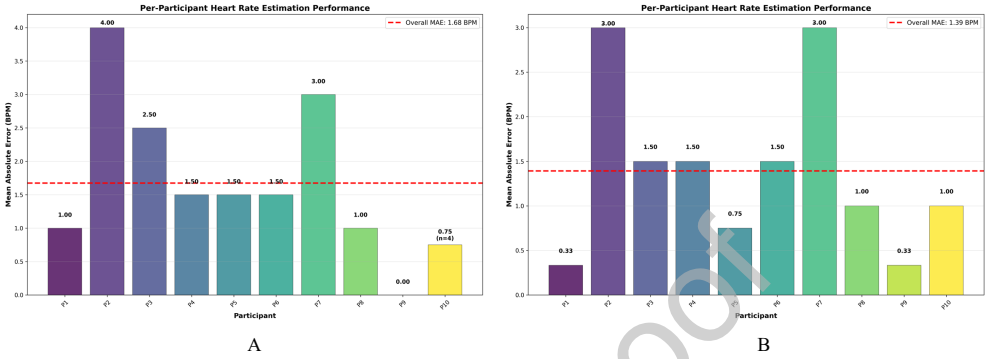


Figure 9: Individual participant performance comparison under comprehensive training conditions A)VAE, B)AVAESA. Each bar represents the mean MAE across 4 measurement repetitions per scenario.

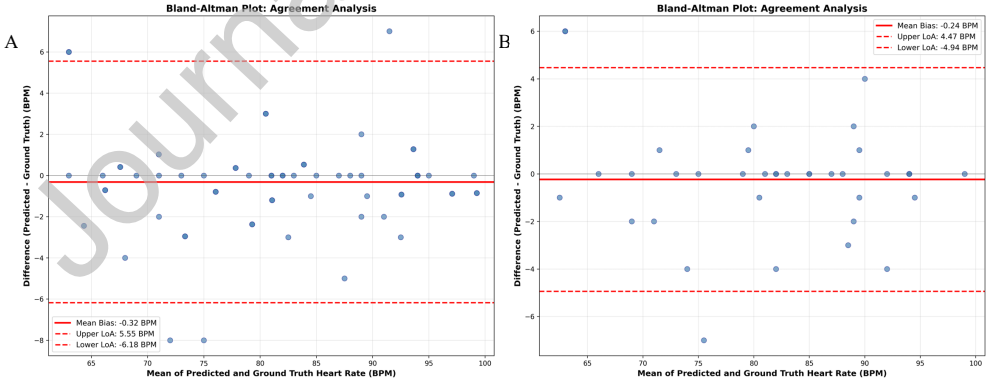


Figure 10: Bland-Altman analysis A)VAE, B)AVAESA

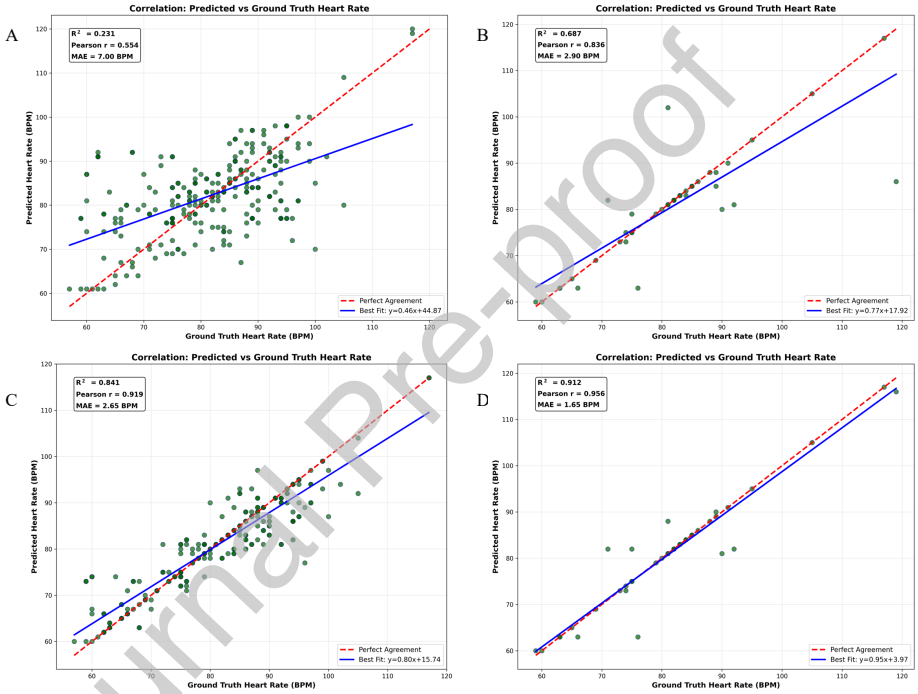


Figure 11: Cross-domain correlation analysis for single-scenario generalization testing A) Distance test VAE, B)Angular test VAE, C)Distance test AVAESA, D)Angular test AVAESA

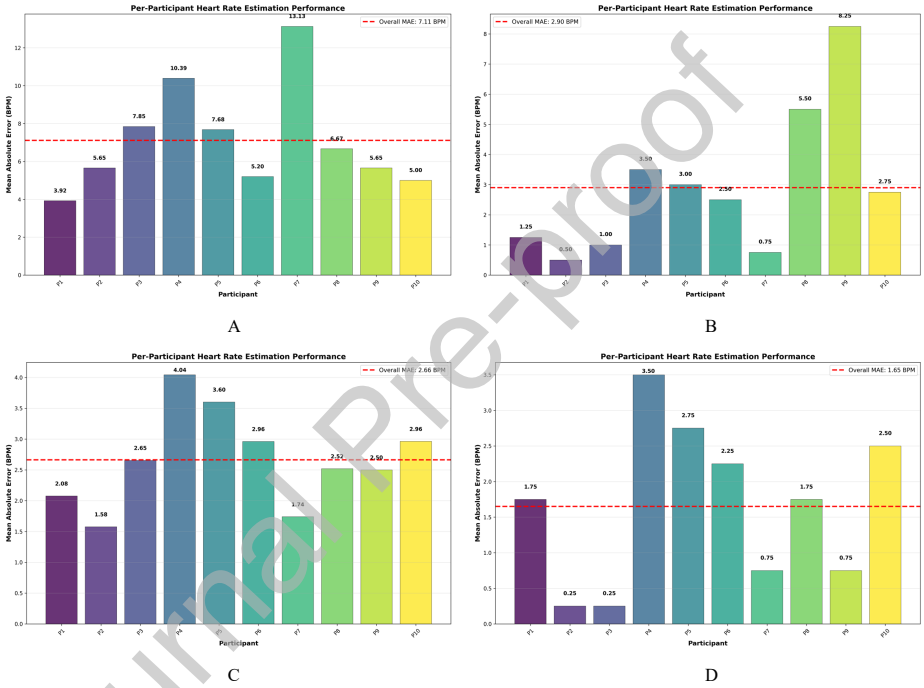


Figure 12: Individual participant accuracy across single-scenario generalization conditions A) Distance test VAE, B)Angular test VAE, C)Distance test AVAESA, D)Angular test AVAESA. Each bar represents the mean MAE across 4 measurement repetitions per scenario.

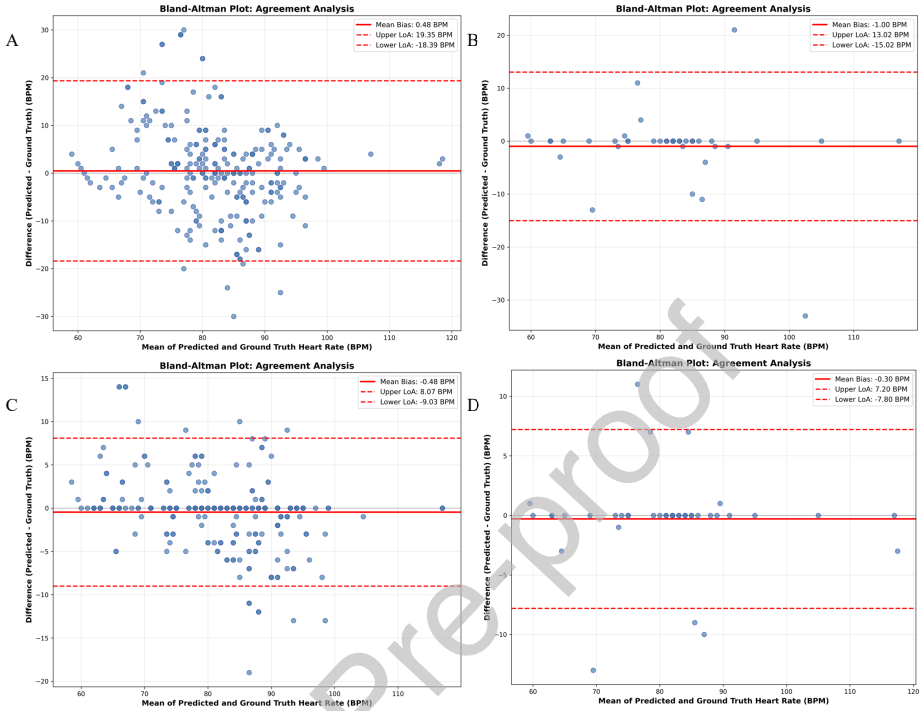


Figure 13: Single-scenario generalization challenges A) Distance test VAE, B)Angular test VAE, C)Distance test AVAESA, D)Angular test AVAESA

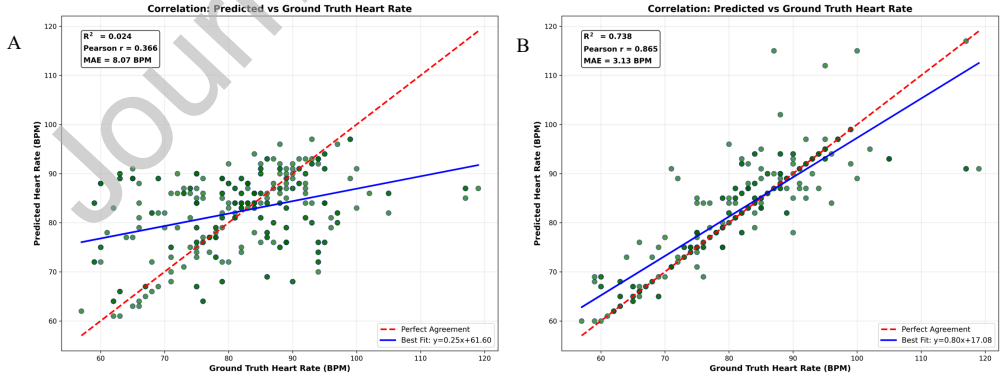


Figure 14: Multi-scenario generalization correlation comparison under extreme domain shift A)VAE, B)AVAESA

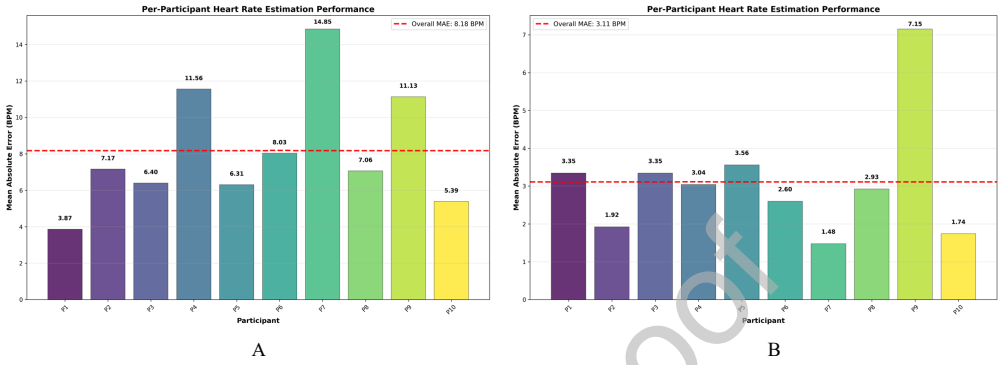


Figure 15: Individual participant accuracy under multi-scenario generalization challenges A)VAE, B)AVAESA. Each bar represents the mean MAE across 4 measurement repetitions per scenario.

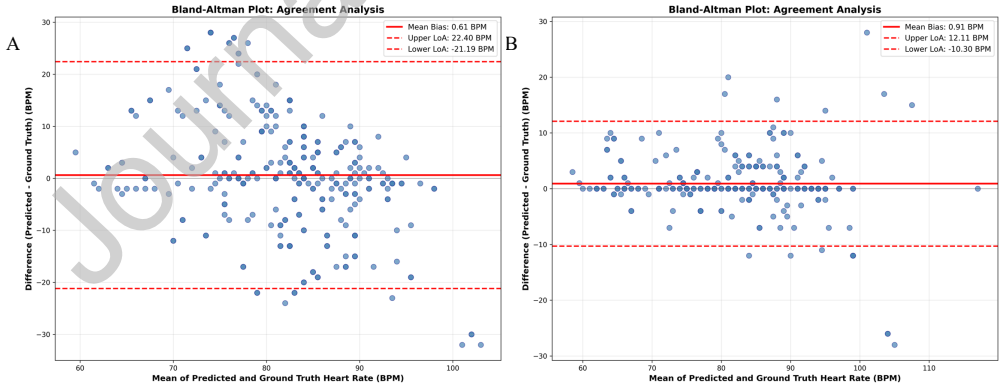


Figure 16: Multi-scenario domain shift A)VAE, B)AVAESA

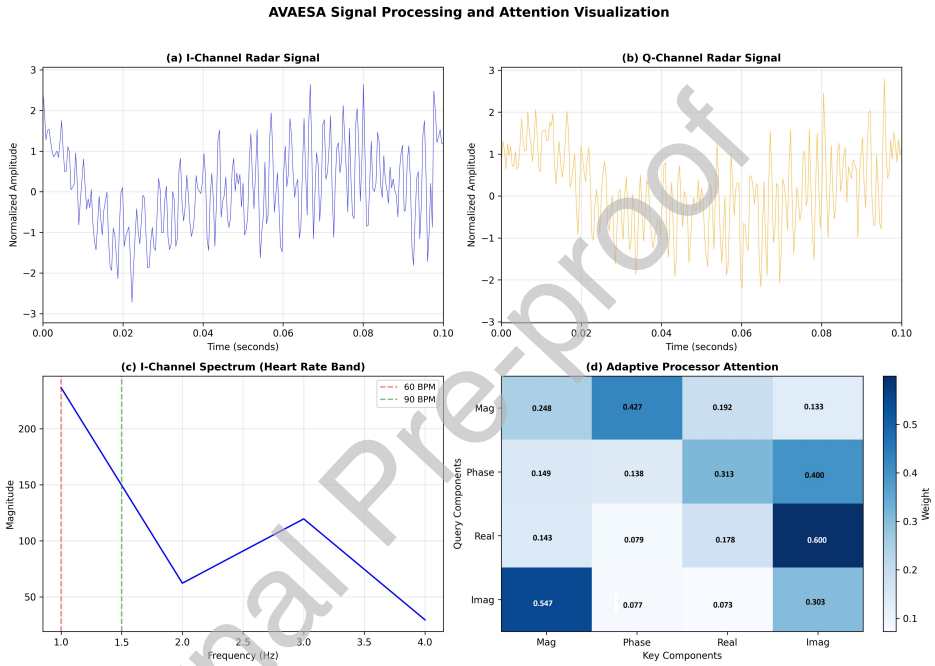


Figure 17: AVAESA signal processing and attention visualization. (a) Normalized I-channel radar signal showing chest displacement patterns induced by cardiac activity. (b) Normalized Q-channel radar signal with a 90° phase offset relative to the I-channel. (c) Frequency spectrum of the I-channel within the cardiac band (0.5–4.0 Hz), with reference lines at 60 BPM (1.0 Hz) and 90 BPM (1.5 Hz). (d) Learned attention weights from the adaptive signal processor, with numerical values indicating the relative contributions of magnitude, phase, real, and imaginary signal representations during cardiac feature extraction.