Translation of Radar Signals into Latent Cardiac Event Space for Scalable, Annotation-free Heart Disease Diagnosis

AOXIANG YUE*, University of Science and Technology of China, China
ZHI LU*, University of Science and Technology of China, China
YANG HU, University of Science and Technology of China, China
ZHENZHEN CAO, University of Science and Technology of China, China
YAN CHEN†, University of Science and Technology of China, China

Radar-based cardiac diagnosis faces significant scalability challenges due to its heavy reliance on expert-labeled data. We propose Radar-to-Event Net (R2E-Net), a novel cross-modal framework that recasts contactless diagnosis as a semantic translation task into a latent cardiac event space. Our method leverages ECG-supervised event representations as a structural prior, projecting radar signals into this shared space via a Transformer-based encoder. This unidirectional alignment enables fully annotation-free radar diagnosis, bypassing radar label dependency while preserving physiological semantics. Moreover, we introduce a robust spatiotemporal radar encoder, which eliminates point-selection heuristics and directly learns meaningful motion patterns from beamforming data. Evaluated on a large-scale real-world dataset of 7,336 clinical recordings, R2E-Net achieves an average F1 score of 82.4% without radar labels, which can even outperform radar models trained under full supervision. Moreover, it generalizes well across age, gender, and disease subtypes, and remains robust in the presence of interfering cardiac conditions. These results demonstrate that our approach offers a scalable, accurate, and annotation-free solution for real-world cardiac monitoring.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; *Ubiquitous and mobile computing design and evaluation methods*; Empirical studies in ubiquitous and mobile computing.

Additional Key Words and Phrases: Cardiac sensing, mmWave radar, heart disease diagnosis, cross-modal learning, annotation-free

ACM Reference Format:
Aoxiang Yue, Zhi Lu, Yang Hu, Zhenzhen Cao, and Yan Chen. 2025. Translation of Radar Signals into Latent Cardiac Event Space for Scalable, Annotation-free Heart Disease Diagnosis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 237 (December 2025), 30 pages. https://doi.org/10.1145/3770693

*Both authors contributed equally to this research.
†Corresponding author.

Authors' Contact Information: Aoxiang Yue, yueaoxiang@mail.ustc.edu.cn, University of Science and Technology of China, Hefei, Anhui, China; Zhi Lu, zhilu@ustc.edu.cn, University of Science and Technology of China, Hefei, Anhui, China; Yang Hu, eeyhu@ustc.edu.cn, University of Science and Technology of China, Hefei, Anhui, China; Zhenzhen Cao, zhenzhen1003@mail.ustc.edu.cn, University of Science and Technology of China, Hefei, Anhui, China; Yan Chen, eecyan@ustc.edu.cn, University of Science and Technology of China, Hefei, Anhui, China.

ACM 2474-9567/2025/12-ART237
https://doi.org/10.1145/3770693

1 INTRODUCTION

Cardiovascular disease remains the leading cause of mortality worldwide [47, 59], yet early detection still depends primarily on in-clinic diagnostics, leaving a critical gap in population-scale screening and timely intervention. Electrocardiography (ECG), the clinical gold standard, offers precise identification of pathological events such as atrial fibrillation (AF) and sinus bradycardia (SB) [4, 33]. However, its reliance on contact-based electrodes and the need for qualified professionals restricts its use to clinical environments, limiting its accessibility in daily life settings [21, 49]. Recent research has investigated radar as a contactless modality for vital sign monitoring [1, 36, 65]. By capturing subtle chest wall motion, radar-based systems can noninvasively estimate fine-grained cardiac activity [5, 61, 70], providing surrogate measures of cardiac function and thereby offering the potential for unobtrusive cardiac monitoring outside clinical settings [63, 64]. However, existing studies have largely been conducted under controlled conditions [5, 61, 70] or focus narrowly on a limited set of specific diseases [63, 64]. Radar-based cardiac diagnostics are still under-explored in the broader real-world context, and this gap is largely due to its reliance on expert-annotated data that fundamentally limits large-scale deployment.

With limited data, supervised radar models face two interrelated challenges, each revealing a distinct pathway to spurious feature learning [61, 69]. First, as radar deployments scale to broader populations, the proliferation of abnormal categories erodes the clear distinction between normal and pathological states, increasing the complexity of the task. Spurious features, such as obvious rhythmic irregularities, sudden amplitude changes, or waveform distortions, are less reliable and fail to generalize [50]. Previous studies, which mainly involved highly regular heart patterns with only a few isolated and well-defined anomalies, may therefore struggle to reveal the pathological complexity of arrhythmia. Second, due to the rarity of specific cardiac conditions [16, 73], the problem exhibits extreme class imbalance. This scarcity makes models more susceptible to overfitting, a challenge further amplified by the nature of radar signals. Unlike ECG, which is dominated by structured, sharp transient components [44, 52], radar signals exhibit high-dimensional, non-stationary motion patterns [1]. They also contain substantial environmental and human-related information that is difficult to filter out, making models more likely to memorize incidental noise [9, 58, 62]. Consequently, scaling radar-based diagnosis may require larger and more diverse datasets, while the annotation remains a slow, resource-intensive process that typically requires specialized expertise [13, 21].

To overcome the persistent annotation bottleneck, we propose a radar-to-event network (R2E-Net) that translates radar signals into an ECG-derived latent space, learned from large-scale annotated ECG data [11, 13, 21].This shifts radar-based cardiac diagnosis from direct classification to an ECG-supervised paradigm, which only requires paired ECG signals and enables annotation-free radar data acquisition. Although hospital-collected ECG data naturally provide diagnostic labels, they are confined to clinical sites and accumulate only slowly [13, 21]. In contrast, our approach extends proactive collection to non-diagnostic settings, such as homes, elder care facilities, and community screenings. This approach can help broaden coverage to more diverse populations and environments that may be underrepresented in hospital-based collection.

Since the ECG-derived latent space can be learned as a disentangled representation of cardiac events [11, 43], supervising radar signals within this space helps suppress information unrelated to cardiac dynamics, such as environmental noise and motion artifacts. To construct this space, we train an ECG autoencoder to capture temporally localized cardiac dynamics that are both diagnostically discriminative and faithful to waveform morphology. These latent events provide a structured representation of cardiac cycles, capturing diagnostically relevant variations (e.g., arrhythmic patterns or waveform irregularities) in a way that remains consistent across patients and conditions. Radar signals, which are high-dimensional and entangled with non-cardiac motion, are mapped into this event space, encouraging the model to retain only components relevant to cardiac dynamics. Because the ECG-derived latent space inherently does not contain interference from the radar domain, we can bypass traditional point-selection heuristics, which extract the most salient physiological signals but may

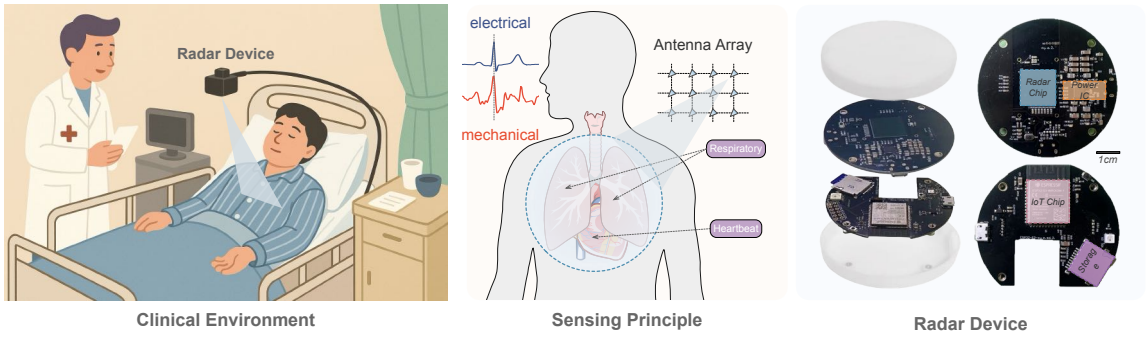


Fig. 1. Real-world data acquisition setup. Left: Clinical deployment of radar-based sensing. Middle: Radar sensing mechanism for vital signs. Right: Internal structure of the radar device.

overlook information at certain moments. Instead, we feed the complete beamforming radar data directly into the model, reducing information loss while eliminating the need for intricate preprocessing.

R2E-Net projects radar signals into ECG-derived event space through one-way semantic transfer, which is closely related to cross-modal knowledge distillation and alignment, but has fundamentally different supervision strategies. Knowledge distillation [14] typically transfers knowledge by using a pretrained model in one modality to supervise a student in another, but it relies on paired data from both modalities for effective supervision. By comparison, cross-modal alignment [8] can be annotation-free and embeds both modalities into a shared latent space. However, enforcing complete alignment can be counterproductive because the low-entropy structure of ECG signals can be distorted by the high-dimensional and redundant radar features, which often leads to spurious correspondences. R2E-Net adopts a unidirectional supervision scheme as a constraint, which preserves the ECG representation in its original diagnostic form and guides radar features to capture only clinically relevant semantics. As a result, it can enable zero-shot generalization because radar projections can be directly classified by the ECG-trained model.

Our contributions are threefold:

- **Annotation-free diagnostic paradigm.** We recast radar-based cardiac diagnosis as a latent event translation task, projecting radar signals into an ECG-derived event space and eliminating the need for radar-specific annotations.
- **End-to-end spatiotemporal radar model.** We develop a robust model that consumes beamforming data directly, avoiding point-selection heuristics and improving generalization in real-world, noise-prone environments.
- **Extensive clinical validation** We validate R2E-Net on 7,336 clinical recordings, demonstrating its annotation-free capability alongside strong diagnostic performance across diverse patients and arrhythmias.

2 RADAR-BASED MODELING FOR CARDIAC DIAGNOSIS

Our study leverages a compact linear frequency-modulated continuous-wave (FMCW) radar system for non-contact cardiac monitoring. As illustrated in Fig. 1, the device is deployed in a clinical environment, positioned near the patient's bed without requiring physical contact. The radar emits low-power signals and captures subtle chest wall displacements caused by respiration and heartbeat. Vital sign information can be computed from these to enable continuous and unobtrusive monitoring.

2.1 Sensing Principle

The transmitted signal of FMCW radar is defined as [26]:

$$s_{tx}(t) = A_{tx}\cos(2\pi f_0 t + \pi\alpha t^2), \tag{1}$$

where A_{tx} is the amplitude, f_0 is the initial frequency, and $\alpha = B/T$ is the frequency sweep slope determined by the bandwidth B and the sweep period T.

Assuming the radar signal reflects off the chest at distance $d_0 + x(t)$, with $x(t)$ representing chest displacement caused by combined cardiopulmonary motion, the received signal can be written as:

$$s_{rx}(t) = A_{rx}\cos(2\pi f_0(t-\tau) + \pi\alpha(t-\tau)^2), \tag{2}$$

where A_{rx} is the received amplitude and τ is the round-trip delay time:

$$\tau(t) = \frac{2(d_0 + x(t))}{c}. \tag{3}$$

The mixer multiplies the transmitted and reflected signals. After low-pass filtering, an intermediate-frequency signal is obtained:

$$s_{IF}(t) = A_{IF}\cos\left(2\pi f_{IF} t + \phi_{IF}(t)\right), \text{ where } f_{IF} = \alpha\tau(t),\ \phi_{IF}(t) = 2\pi f_0 \tau(t). \tag{4}$$

Since $d_0 \gg x(t)$, f_{IF} remains effectively constant, allowing the target to stay within the same range bin after FFT. The residual chest motion $x(t)$ can be extracted through phase demodulation of the signal within that range bin.

To improve spatial resolution, we employ a 2D antenna array, which facilitates the reconstruction of echo signals from multiple locations within 3D space. However, owing to sidelobe leakage, multipath propagation, and the distributed scattering properties of the human body, the recovered echoes from these spatial locations inevitably include components associated with chest motion. Notably, various pathological conditions can alter the dynamic patterns of cardiac motion, and such alterations may manifest within the spatially resolved echo signals. These spatially distributed motion signatures have the potential to serve as biomarkers, offering a noninvasive means to infer underlying cardiac abnormalities.

Previous methods prioritized locations of maximal energy or employed optimal template matching to mitigate interference, but this came at the expense of discarding spatial information. In contrast, we leverage all beamformed points, allowing the model to autonomously learn spatial attention for robust cardiac motion extraction.

2.2 Problem Formulation

We consider the task of annotation-free heart disease diagnosis using paired radar and ECG data. Let $\mathcal{D} = \{(x_i^r, x_i^e)\}_{i=1}^N$ denote the dataset, which comprises synchronously acquired radar signals $x_i^r \in \mathcal{X}_r$ and ECG signals $x_i^e \in \mathcal{X}_e$. In conventional supervised learning paradigms, the clinical label $y_i \in \mathcal{Y}$ is derived from ECG data to learn the radar-based diagnosis model.

Hospital-collected ECG data typically come with labels, as they are acquired in clinical settings where patients are routinely diagnosed during outpatient visits or hospitalization. However, for cases such as long-term Holter monitoring or overnight sleep studies, it is challenging to provide complete, segment-by-segment labels, even in hospitals. Moreover, in many scenarios, such as home monitoring, portable devices, or elder care settings, there is no ongoing access to cardiologists to provide such labels, as these environments typically lack real-time specialist oversight. This motivates us to study the annotation-free setting, where labels y_i are absent.

Since training a diagnosis model without annotations is typically infeasible, we leverage externally labeled ECG data to construct a pretrained, structured, and disentangled latent space. This latent space serves as a proxy representation into which radar features can be projected and aligned, thereby enabling annotation-free training.

Cardiac Event Space. Due to the high regularity of ECG waveforms, we define a structured cardiac event space $\mathcal{E}$ that captures recurring signal patterns. This space is inspired by canonical components such as the P wave, QRS complex, and T wave. These events correspond to distinct stages of the cardiac cycle, and different cardiac conditions can be characterized by unique temporal configurations of these events. For example, SB is reflected in prolonged intervals between QRS complexes, whereas the absence of P waves and irregular rhythms characterizes AF. In practice, the actual event representations are learned from ECG data.

To construct $\mathcal{E}$, we leverage an auxiliary ECG dataset $\mathcal{A} = \{(x_j^e, y_j)\}$, where labels are available for supervised training in the ECG domain. We first train an ECG autoencoder on $\mathcal{A}$ to capture temporally structured representations of ECG signals, and then retain only the encoder $\psi_e(\cdot)$ for downstream tasks:

$$\psi_e(x_e) = \{e_1, \ldots, e_K\},\ e_k \in \mathcal{E}. \tag{5}$$

Modern approaches allow the cardiac event space $\mathcal{E}$ to be learned either as a discrete set or as a continuous but disentangled latent space. An event-based classifier f_e is then trained by minimizing the supervised task loss:

$$\min_{f_e} \mathbb{E}_{(x^e,y)\sim\mathcal{A}} \mathscr{L}_{\text{task}}\left(f_e(\psi_e(x^e)), y\right). \tag{6}$$

Annotation-free Learning. With $\mathcal{E}$ established, the radar-based diagnosis task can be formulated as a cardiac event translation problem, which maps a radar feature sequence into an ECG-derived cardiac event space in an annotation-free manner:

$$\mathcal{T} : \{x_t\}_{t=1},\ x_t \in \mathbb{R}^d \ \rightarrow\ \{e_t\}_{t=1},\ e_t \in \mathcal{E}. \tag{7}$$

Let $\psi_r(\cdot)$ be the radar-specific encoder. The corresponding event representation from radar domain is given by:

$$v_r = \mathcal{T}(\psi_r(x^r)) = \{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K\}. \tag{8}$$

If the radar-derived event sequence v_r faithfully captures cardiac dynamics, the ECG classifier f_e can be directly applied to radar inputs:

$$f_r \triangleq f_e(\mathcal{T}(\psi_r(x^r))). \tag{9}$$

Therefore, by jointly learning the translator $\mathcal{T}$ and radar encoder ψ_r, we achieve training in an annotation-free manner. When annotations for dataset $\mathcal{D}$ are available, we can further finetune the classifier to improve performance.

Relation to Zero-Shot Learning. Our framework bears a superficial resemblance to zero-shot learning (ZSL) [22, 35, 41, 57] in its reliance on an auxiliary space to mediate model learning. In classical ZSL, labeled data from seen classes are used to learn a projection into a semantic space (e.g., attributes, linguistic embeddings), which in turn supports recognition of unseen classes. However, our formulation departs from this paradigm in two fundamental ways. First, no labeled radar data are available, rendering the standard ZSL training pipeline inapplicable. Second, the ECG domain lacks a well-defined semantic space, as many cardiac conditions share coarse and overlapping linguistic descriptors, which prevents the construction of a disentangled attribute set.

Our cardiac event space functions analogously to a semantic space, serving as a proxy representation into which radar features can be projected and aligned for diagnosis. Once radar features are aligned to this event space, our approach inherits the capacity for zero-shot generalization, enabling the recognition of novel cardiac conditions not explicitly encountered during training, provided their ECG signatures are represented within $\mathcal{E}$.

3 SYSTEM OVERVIEW

The workflow of the R2E-Net system is shown in Fig. 2 and consists of four main steps:

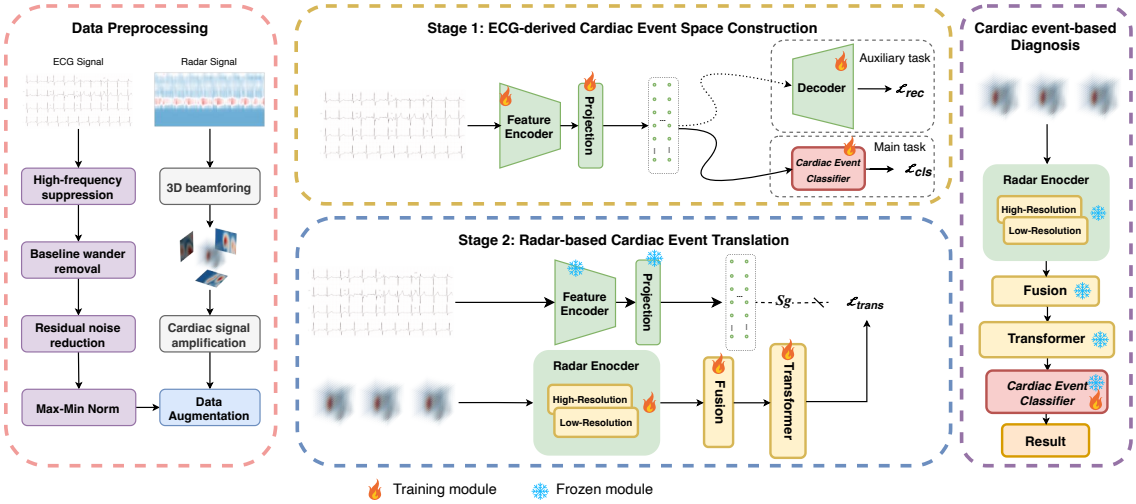


Fig. 2. R2E-Net framework. The system comprises four parts: (1) preprocessing radar and ECG signals with beamforming, denoising, and augmentation; (2) learning ECG-derived latent events via dual supervision; (3) translating radar signals into the event space using a spatiotemporal encoder and Transformer; and (4) performing annotation-free learning with a reused ECG-trained classifier.

- **Data Preprocessing.** Radar and ECG signals are first preprocessed to enhance relevant physiological information and standardize inputs. Radar frames undergo 3D beamforming to generate spatiotemporal representations, followed by a second-order difference filter to amplify cardiac motion. ECG signals are denoised using Butterworth filtering, LOESS smoothing, and non-local means, then normalized with a min–max scheme. A sliding-window augmentation strategy with class-dependent strides is applied to mitigate data imbalance.
- **ECG-derived Cardiac Event Space Construction**. A multi-layer convolutional encoder extracts temporal features from ECG signals, which are mapped to a latent cardiac event space through a projection layer. This event space is trained with classification and reconstruction objectives to preserve diagnostic semantics and maintain temporal structure.
- **Radar-based Cardiac Event Translation**. A radar-specific spatiotemporal encoder extracts motion features from radar inputs. These features are then mapped into the ECG-derived event space using a Transformer-based translator, aligning radar representations with the event space learned from ECG.
- **Cardiac event-based Diagnosis**. The disease classifier trained on ECG-derived event representations is applied to radar-translated events for diagnosis. When radar labels are unavailable, the classifier operates in the zero-shot setting. When additional annotations become available, joint fine-tuning on both radar and ECG data further enhances the diagnostic performance on radar signals.

4 SYSTEM DESIGN

4.1 Data Preprocessing

4.1.1 Radar Signals. We first apply 3D beamforming to the FMCW radar data from the 2D antenna array, steering the received signals toward specific directions to produce a voxel-wise representation of the scene (Fig. 3). At each timestamp $t \in [T]$, the beamformed output is defined as:

$$\mathrm{S}_t(r,\theta,\phi) = \boldsymbol{a}_t(r,\theta,\phi)\exp\left(j\boldsymbol{\omega}_t(r,\theta,\phi)\right) \in \mathbb{C}^{N_r \times N_\theta \times N_\phi}, \tag{10}$$

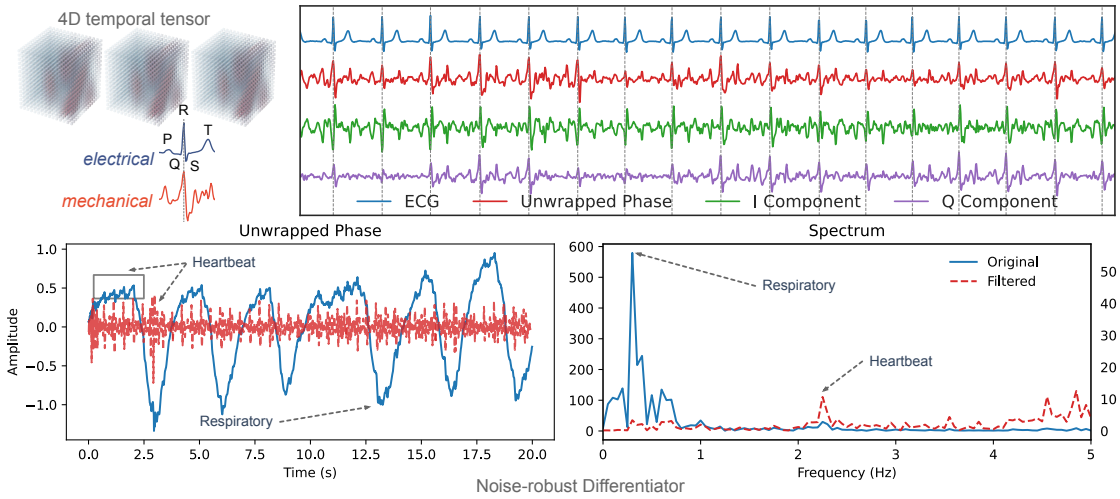


Fig. 3. Radar signal and filtering. The top left shows the 4D radar signal tensor; the top right presents synchronized ECG, unwrapped phase, I, and Q signals. The bottom left compares raw and temporally filtered phase signals. The bottom right shows the frequency spectrum before and after temporal filtering.

where r, θ, and ϕ denote the range, azimuth, and elevation grid points, respectively. The signal is represented in polar form, with amplitude $\boldsymbol{a}_t$ capturing reflection strength and phase $\boldsymbol{\omega}_t$ encoding signal delay.

This voxel-wise representation enables the extraction of subtle chest movements by isolating reflections from specific spatial locations. Chest micro-movements induce phase variations in $\boldsymbol{\omega}_t$, encoding displacement information, while fluctuations in $\boldsymbol{a}_t$ reflect changes in the radar cross-section. Respiration produces dominant periodic variations, whereas heartbeat-induced motion introduces finer, higher-frequency perturbations.

Respiration introduces slow-varying components into radar signals, which can obscure the subtler and faster cardiac micro-motions. To highlight heartbeat-related dynamics while suppressing these slower variations, we apply a second-order temporal difference operator to each beamformed radar channel, including the in-phase (I), quadrature (Q), and unwrapped phase signals. This operator emphasizes transient changes by attenuating low-frequency trends. We use the following filter [17]:

$$\ddot{s}[t] = \frac{4s[t] + (s[t-1] + s[t+1]) - 2(s[t-2] + s[t+2]) - (s[t-3] + s[t+3])}{16h^2} \tag{11}$$

where $\ddot{s}[t]$ denotes the estimated second derivative at time t, $s[t]$ is the signal at frame t, and h is the sampling interval. This filter functions as a high-pass operator, preserving rapid cardiac fluctuations while dampening slower respiratory trends.

After filtering, each radar signal yields a tensor of shape $C \times N_r \times N_\theta \times N_\phi$ at each timestep, where $C = 3$ corresponds to different channels. This representation captures spatially localized motion over time across multiple signal modalities. Fig. 3 illustrates how the second-order differentiator suppresses slow respiratory trends and amplifies heartbeat-induced fluctuations.

4.1.2 ECG Signals. The external ECG datasets used in this study are highly diverse and contain varying levels of noise, such as power line interference, electrode contact noise, motion artifacts, muscle contractions, baseline wandering, and random noise. These heterogeneous noise characteristics vary across datasets and acquisition

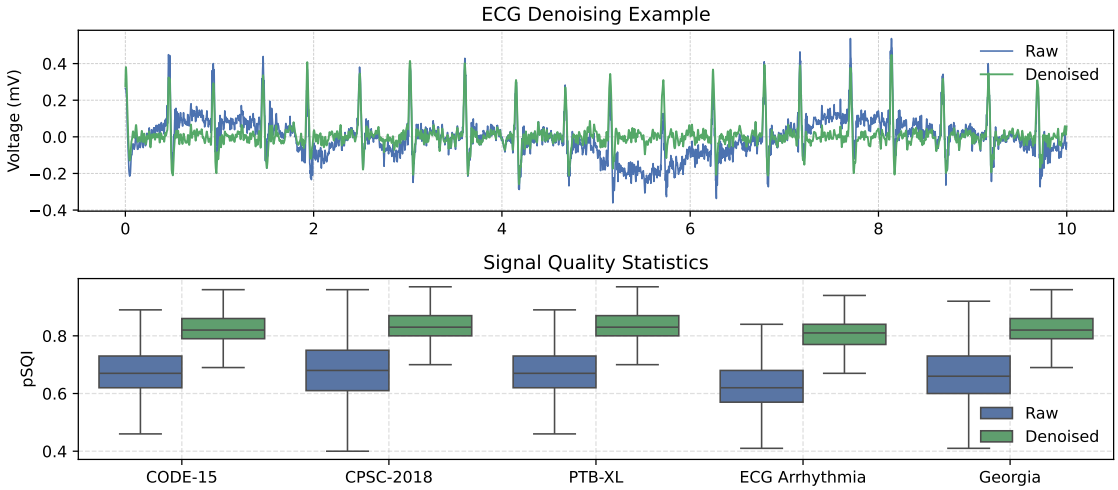


Fig. 4. Denoising performance on ECG signals. Top: example of raw and denoised ECG. Bottom: pSQI quality improvement across five datasets.

conditions, resulting in a domain shift that hinders the learning of robust and generalizable ECG representations. To address this issue, we implement a structured ECG preprocessing pipeline for noise reduction and format standardization, inspired by prior work [72], but adapted to our multi-source setting. The pre-processing consists of three sequential steps.

High-frequency suppression. Given that the frequency range of standard ECG signals lies between 0.5 Hz and 50 Hz, we apply a Butterworth low-pass filter [3] to eliminate high-frequency noise. The filter is designed with a passband at 50 Hz, a stopband at 60 Hz, a passband ripple of 1.0 dB, and a stopband attenuation of 2.5 dB.

Baseline wander removal. A LOESS smoother [7] is used to estimate and remove baseline drift. The trend is obtained via weighted least squares, with weights defined as:

$$w_i(x_0) = W\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right), \tag{12}$$

where $W(\cdot)$ denotes the tricube weight function, and Δ represents the window size. The estimated trend is then subtracted from the original ECG signal.

Residual noise reduction. To suppress residual noise while preserving critical morphological features, we apply a non-local means (NLM) filter [2, 51], adapted for univariate physiological time series. Each data point $S(i)$ is reconstructed as a weighted average of all other points based on local patch similarity:

$$S(i) = \sum_{j \in N(i)} w(i,j) D(j), \quad w(i,j) = \frac{1}{Z(i)} \exp\left(-\frac{|D_\Delta(i) - D_\Delta(j)|^2}{\lambda^2}\right), \tag{13}$$

where $D_\Delta(i)$ is a local patch centered at position i, λ controls smoothness, and $Z(i)$ is a normalization term.

Fig. 4 presents the distribution of ECG signal quality, measured by the power-based Signal Quality Index (pSQI), across multiple datasets before and after denoising. The pSQI quantifies the proportion of energy concentrated in the QRS complex relative to the entire signal. Higher pSQI values generally indicate more distinct and structured

Table 1. Amplitude statistics (mV) of Lead-I ECG signals across datasets. Maximum and minimum values are calculated by averaging the per-record extrema to reduce the impact of outliers.

Metric	CODE-15	CPSC-2018	ECG-Arrhythmia	Georgia	PTB-XL	SPH	Ours
Mean value	0.04	0.02	0.02	0.02	0.03	0.02	0.00
Standard deviation	0.20	0.11	0.10	0.12	0.13	0.09	0.08
Maximum	1.54	0.60	0.67	0.77	0.84	0.58	0.41
Minimum	-0.36	-0.25	-0.21	-0.21	-0.21	-0.20	-0.17

Algorithm 1 Sliding-window–based data augmentation strategy

Require: Paired signal (x^r, x^e) of length L; window length W; pseudo-label $y \in \{1, \ldots, C\}$; class-dependent strides $\{s_1, \ldots, s_C\}$
Ensure: Augmented sample list $\mathcal{X}_{\text{aug}}$

$\mathcal{X}_{\text{aug}} \leftarrow \emptyset$
$stride \leftarrow s_y$ ▷ stride for class y
$N \leftarrow \lfloor (L - W)/stride \rfloor$ ▷ number of windows
for $i \leftarrow 0$ **to** $N - 1$ **do**
  $offset \leftarrow \text{random}(0, stride)$
  $start \leftarrow i \times stride + offset$
  $end \leftarrow start + W$
  $x_{\text{win}} \leftarrow (x^r[start : end], x^e[start : end])$
  append x_{win} to $\mathcal{X}_{\text{aug}}$
end for
return $\mathcal{X}_{\text{aug}}$

cardiac events, thus reflecting improved signal quality. As expected, the denoising process consistently improves pSQI across all datasets, demonstrating a significant enhancement in overall signal quality.

Following denoising, we further address the statistical variability observed across datasets. As shown in Table 1, the maximum ECG amplitude ranges from 1.5355 in the CODE-15 dataset to 0.412 in the radar dataset, revealing a substantial mismatch in signal distributions. Such discrepancies suggest a potential domain shift issue, which may hinder model training and compromise generalization performance. To mitigate this, we apply min-max normalization independently to each dataset. This standardization procedure scales the amplitude distribution into a common range while preserving waveform morphology, thereby reducing cross-domain distributional divergence.

Finally, to account for differences in sampling rates among datasets, we resample each ECG signal to match the radar sampling frequency. The processed ECG signal is represented as $x^e \in \mathbb{R}^{L \times T}$, where $L = 12$ denotes the number of ECG leads.

4.1.3 Data Augmentation. The distribution of heart disease follows a typical long-tail pattern, with a highly skewed class ratio that can severely hinder the classifier's learning performance. To mitigate the class imbalance problem, we apply the following augmentation to our radar dataset.

For each radar sequence with T time steps, we employ a sliding-window technique to expand the sample size. However, this approach alone only increases the number of samples and does not fundamentally resolve the class imbalance issue. To further refine class proportions, we vary the stride length across different disease

categories. During training, we additionally adopt a random-stride strategy to inject variability and reduce the risk of overfitting.

Since balancing disease instances during data augmentation typically relies on label information, which is unavailable for unlabeled data, we employ an ECG-based classifier to generate pseudo-labels, which are then used to inform the augmentation strategy. The overall augmentation procedure is presented in Algorithm 1.

4.2 ECG-derived Cardiac Event Space Construction

To construct a semantically meaningful latent representation of cardiac activity, we first encode the ECG signal x^e into a temporally structured sequence of deep features. Given the relatively low inter-subject variability of ECG signals, we adopt a lightweight multi-layer convolutional encoder $\mathscr{E}(\cdot)$ to extract intermediate representations:

$$\tilde{v}^e = \mathscr{E}(x^e) \in \mathbb{R}^{D' \times L}, \tag{14}$$

where D' is the feature dimension and L is the temporal length of the sequence. Each vector $\tilde{v}_i$ encodes local cardiac dynamics at time step i.To obtain a sequence of latent cardiac events in semantic space, we apply a linear projection:

$$v^e = W\tilde{v}^e + b \in \mathbb{R}^{D \times L}, \tag{15}$$

where $W \in \mathbb{R}^{D \times D'}$ and $D \ll D'$. The resulting v^e represents a compact and structured sequence of latent cardiac events, serving as the semantic foundation for diagnosis.

To encourage temporal disentanglement and physiological consistency, we design a dual-objective learning strategy that jointly optimizes a classification task and a reconstruction task. Specifically, we introduce a decoder $\mathscr{D}(\cdot)$ composed of transposed convolutional blocks to reconstruct the input ECG from its latent representation:

$$\hat{x}^e = \mathscr{D}(v^e) \in \mathbb{R}^{L \times T}. \tag{16}$$

This restoration objective preserves temporal ordering and ensures the completeness of physiological information within the latent events. In parallel, to enhance sensitivity to rare and abnormal cardiac conditions, we incorporate a classification head $f_e(\cdot)$:

$$\hat{y}^e = f_e(v^e), \tag{17}$$

where $\hat{y}^e$ denotes the predicted class label for the input ECG.

The overall loss function combines both objectives:

$$\mathscr{L}_{\text{ECG}} = \lambda_1 \mathscr{L}_{\text{cls}} + \lambda_2 \mathscr{L}_{\text{rec}}, \tag{18}$$

where

$$\begin{aligned} \mathscr{L}_{\text{rec}} &= \mathbb{E}_{\mathcal{A}} \left[\|\hat{x}^e - x^e\|_2^2 \right], \\ \mathscr{L}_{\text{cls}} &= \mathbb{E}_{\mathcal{A}} \left[y^e \log \hat{y}^e + (1 - y^e) \log(1 - \hat{y}^e) \right]. \end{aligned} \tag{19}$$

This multi-task training framework enables the latent cardiac event sequence to encode temporally structured and diagnostically informative representations, serving as a unified semantic basis for cross-modal translation and downstream disease inference.

4.3 Radar-based Cardiac Event Translation

To project radar signals into the latent cardiac event space, R2E-Net introduces a radar-specific spatiotemporal encoder as shown in Fig. 5. The spatial encoder extracts localized motion patterns using stacked 3D convolutions, while the dual-scale temporal encoder captures multi-frequency dynamics. These components are followed by a Transformer-based model, which maps radar features into the ECG-guided event space.

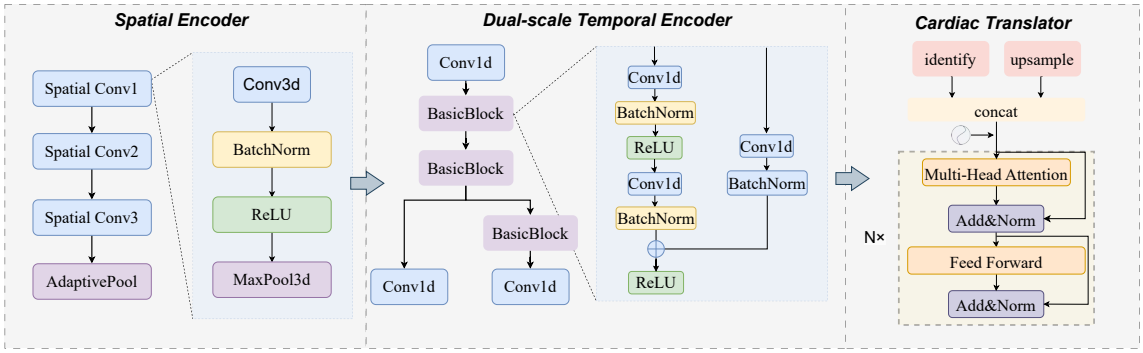


Fig. 5. Radar spatiotemporal encoder architecture. The encoder consists of three components: a spatial encoder with stacked 3D convolutions for localized motion extraction, a dual-scale temporal encoder with residual connections to capture multi-frequency dynamics, and a Transformer-based cardiac translator that projects radar features into the ECG-guided latent event space.

4.3.1 Spatial Encoding. Conventional contactless ECG methods often rely on selecting spatial locations from radar signals based on handcrafted rules or pattern matching. For instance, mmECG [5] identifies regions where temporal dynamics across range bins exhibit high internal similarity, while mmArrhythmia [70] selects fixed chest-range bins. However, such point-selection strategies are susceptible to signal quality and prone to mislocalization or partial coverage of the cardiac activity, leading to unstable measurements and degraded downstream performance.

To overcome these limitations, R2E-Net employs a fully end-to-end spatial encoder to extract cardiac-relevant features directly from raw radar frames, thereby eliminating the need for heuristic point selection. At each timestamp t, the spatial signal $x_t^r \in \mathbb{R}^{N_r \times N_\theta \times N_\phi}$, representing the radar response across range, azimuth, and elevation dimensions, is processed by a 3D convolutional encoder $\mathscr{F}_r(\cdot)$ to produce a compact spatial embedding:

$$s_t^r = \mathscr{F}_r(x_t^r) \in \mathbb{R}^D, \quad t \in [T], \tag{20}$$

where $\mathscr{F}_r$ comprises three stacked 3D convolutional layers. This process produces a sequence of spatial embeddings $\{s_t^r\}_{t=1}^T$, which preserve temporally localized spatial information and serve as input to the subsequent temporal encoder.

4.3.2 Temporal Feature Modeling. After spatial encoding, we apply a dual-scale temporal encoder to capture the dynamics of cardiac motion across time. Specifically, two convolutional branches $\mathscr{F}_l$ and $\mathscr{F}_h$ extract low- and high-level temporal features:

$$s^l = \mathscr{F}_l(s^r) \in \mathbb{R}^{D \times \frac{T}{8}}, \quad s^h = \mathscr{F}_h(s^l) \in \mathbb{R}^{D \times \frac{T}{16}}. \tag{21}$$

We then align temporal resolutions and concatenate both branches to form the final radar feature sequence:

$$v^r = v^l \oplus v^h\uparrow \in \mathbb{R}^{D \times \frac{T}{8}}, \tag{22}$$

where $\uparrow$ denotes temporal upsampling and $\oplus$ indicates channel-wise concatenation.

4.3.3 Cardiac Event Translation. To generate latent cardiac events directly from radar input, R2E-Net introduces a Transformer-based Translator $\mathcal{T}(\cdot)$, which learns to decode radar-derived representations into semantically meaningful latent cardiac event sequences. Unlike conventional alignment-based approaches that seek a shared representation space, our design treats the latent cardiac event sequence extracted from ECG as the target

Table 2. Statistics of cardiac conditions in our dataset.

Condition	Atrial Fibrillation	Sinus Bradycardia	Sinus Tachycardia	Normal	Others	Total
No. of Recordings	244	725	358	3,733	2,276	7,336

output domain. The translator maps the radar representation v^r to a translated event sequence $\hat{v}^e = \mathcal{T}(v^r)$ that approximates the structure and semantics of ECG-derived events.

To guide this translation, we apply a stop-gradient loss using the ECG latent sequence v^e as supervision:

$$\mathscr{L}_{\text{trans}} = \frac{1}{L} \sum_{t=1}^{L} \left\| \text{sg}(v_t^e) - \mathcal{T}(v^r)_t \right\|_2^2 \tag{23}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This objective ensures that the translated radar sequence inherits the temporal structure and event-level semantics defined by the ECG domain, enabling effective latent event modeling from radar without requiring explicit radar annotations.

4.4 Cardiac Event-based Diagnosis

Based on the previous translation process, radar signals are converted into latent cardiac event sequences $\hat{v}^e = (\hat{v}_1, \ldots, \hat{v}_L)$, which preserve the temporal structure and semantic granularity of ECG-derived representations. These radar-derived latent events serve as the inputs for cardiac disease classification.

To support annotation-free diagnosis, R2E-Net reuses a classifier $f(\cdot)$ originally trained on ECG latent events from auxiliary datasets:

$$\hat{y}^e = f(v^e), \quad \hat{y}^r = f(\hat{v}^e), \tag{24}$$

where v^e denotes the ECG-derived event sequence and $\hat{v}^e$ is the translated radar counterpart. During inference, the classifier is applied to $\hat{v}^e$ without any radar-side supervision.

When annotations for radar data are available, we can further jointly fine-tune the classifier on both radar-derived latent events $\hat{v}^e$ and ECG-derived latent events v^e using the standard cross-entropy loss:

$$\mathscr{L}_{\text{joint}} = \mathbb{E}_{(x^r, x^e, y)} \left[\mathcal{L}_{\text{CE}}(f(\hat{v}^e), y) + \mathcal{L}_{\text{CE}}(f(v^e), y) \right], \tag{25}$$

ensuring that the classifier not only aligns radar features with class-discriminative boundaries but also preserves diagnostic knowledge from the ECG domain, which can be viewed as a semantic regularizer in the shared latent event space.

This inference strategy transfers diagnostic knowledge from ECG to radar by operating within a shared latent cardiac-event space, thereby enabling effective classification in both annotation-free and labeled deployment scenarios.

5 EXPERIMENT SETUP

5.1 Data Collection

As illustrated in Fig. 1, we employed a self-designed compact RF device built around the TI AWR6843 radar chip. The 60 GHz radar operated at a 100 Hz sampling rate in single-chirp mode. During each session, the radar was positioned approximately 60-70 cm above the participant's chest while the participant lay on a hospital bed. A clinical 12-lead ECG was recorded concurrently, and 30 seconds of synchronized radar and ECG data were obtained for each session. All recordings were conducted in real clinical environments, including both inpatient wards and outpatient clinics, with Institutional Review Board (IRB) approval.

5.1.1 Radar Data. We constructed a dataset comprising synchronized radar signals and standard 12-lead ECG recordings from 7,336 individuals, collected in real-world clinical settings. Each sample spans 30 seconds, and the signals were recorded with a sampling rate of 100 Hz. As summarized in Table 2, rhythm annotations are provided for four primary categories: Atrial Fibrillation (AF, 244 samples), Sinus Bradycardia (SB, 725), Sinus Tachycardia (STach, 358), and Normal ECG (NECG, 3,733). In addition, 2,276 samples are labeled as "Others", representing unclassified or less common arrhythmias. The dataset is partitioned at the patient level into training, validation, and test sets using a 7:1:2 split.

5.1.2 ECG Dataset. Our method leverages external ECG datasets to construct the cardiac event space. Specifically, six publicly available 12-lead ECG datasets are utilized: CODE-15 [46], Georgia 12-Lead ECG [39], CPSC-2018 [30], PTB-XL [53], SPH [31], and ECG-Arrhythmia [71]. To ensure temporal alignment with the radar signals, all ECG recordings were resampled to 100 Hz and standardized to 1,024 time steps through either truncation or zero padding.

5.2 Baseline

Given the limited number of existing studies on radar-based cardiac diagnosis, we use two baselines for performance comparison in cardiac disease detection. The first baseline utilizes 12-lead ECG signals for arrhythmia classification, serving as a reference for the upper bound of fully automated diagnosis using contact-based measurements. The second baseline is a reproduction of mmArrhythmia [70], the first method to enable contactless arrhythmia detection in controlled laboratory environments.

5.3 Evaluation Metrics

We adopt the following metrics to evaluate the classification performance of the proposed model:

- **Accuracy**: measures the overall correctness of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{26}$$

- **Recall**: the proportion of true positives correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{27}$$

- **Precision**: the proportion of predicted positives that are true positives:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{28}$$

- F_1 **Score**: the harmonic mean of precision and recall, balancing sensitivity and precision under class imbalance:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{29}$$

- **ROC-AUC**: the Area Under the Receiver Operating Characteristic Curve:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{30}$$

  where TPR and FPR denote the true positive rate and false positive rate, respectively. It reflects the model's ability to distinguish between positive and negative classes across different decision thresholds.
- **PR-AUC**: the Area Under the Receiver Operating Characteristic Curve:

$$\text{PR-AUC} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}), \tag{31}$$

which is particularly informative for imbalanced datasets, as it focuses on the performance of the positive class.

5.4 Implementation Details

5.4.1 Hardware. Our contactless ECG monitoring platform is built on the TI AWR6843 millimeter-wave radar, integrated with the DCA1000 real-time data capture module. A virtual two-dimensional antenna array comprising 12 channels is formed using three transmitting (Tx) and four receiving (Rx) antennas. A time-division multiplexing (TDM) scheme is employed to ensure temporal orthogonality among the signals transmitted from different Tx antennas. Within each radar sensing frame, the three transmitters sequentially emit chirp signals with an interval of 45 microseconds. This configuration enables the system to capture baseband signals from all four receivers corresponding to each transmitter, thereby supporting high-resolution spatial sensing.

5.4.2 Software. Model training is performed using the Adam optimizer with a learning rate of 5×10^{-4} and a weight decay of 10^{-6}. The spatial encoder consists of three layers of 3D convolution with a kernel size of 3, each followed by Batch Normalization, ReLU activation, and pooling, producing a 64-dimensional feature representation. The dual-scale temporal encoder includes two parallel branches, each composed of four residual blocks to capture multi-scale temporal dependencies. The cardiac translator module is implemented using six Transformer blocks, each equipped with four attention heads. The model is trained for 100 epochs with a batch size of 16. All experiments are conducted using PyTorch [37] on a single NVIDIA RTX 4090 GPU.

6 PERFORMANCE EVALUATION

6.1 Overall Performance

We evaluate both the annotation-free version of R2E-Net and the version further fine-tuned with radar labels, and compare them with two baselines: an ECG-based classifier and the radar-based mmArrhythmia [70]. Table 3 summarizes the results across three arrhythmia classes: atrial fibrillation (AF), sinus bradycardia (SB), and sinus tachycardia (STach).

R2E-Net, trained solely under ECG supervision without any radar annotations (annotation-free), achieves strong performance with an average F1 score of 82.41% and a PR-AUC of 0.8987, substantially outperforming mmArrhythmia (F1: 69.81%, PR-AUC: 0.7142) across all arrhythmia types. These results highlight the model's ability to leverage latent cardiac event representations distilled from ECG data and effectively transfer them to the radar domain without direct label supervision. Notably, in the detection of atrial fibrillation (AF), a particularly challenging condition due to its temporal irregularity and inter-patient variability, R2E-Net exhibits enhanced robustness, improving the F1 score by 27.2% over the baseline method.

When further fine-tuned with radar-side annotations, R2E-Net achieves further performance gains, reaching an average F1 of 85.68% and a PR-AUC of 0.9133. The improvement over its annotation-free counterpart is particularly pronounced in AF (F1: 85.63% vs. 80.35%) and STach (85.02% vs. 81.16%), indicating that while radar-derived latent event representations already inherit discriminative structure through cross-modal alignment, supervised adaptation further sharpens disease-specific decision boundaries. Interestingly, in SB, where the pathological patterns are more stereotypical and less sensitive to sensing modality, the performance of the two variants converges, suggesting that latent event learning without explicit supervision already captures sufficiently distinctive features for reliable classification.

The ECG-based model performs direct classification on 12-lead ECG signals and serves as a reference upper bound due to its access to high-fidelity, fully supervised diagnostic features. Compared to this upper bound (F1: 90.52%), both variants of R2E-Net demonstrate competitive performance, with the fine-tuned version narrowing the average F1 gap to less than 5 percentage points. In contrast, mmArr exhibits limited diagnostic capability, particularly in AF (F1: 53.16%) and STach (F1: 71.98%), highlighting its sensitivity to signal quality and limited

Table 3. Performance comparison of R2E-Net with different baselines on arrhythmia classification. For R2E-Net, labels are only used for fine-tuning.

Class	Method	Label	Acc ↑	F1 ↑	ROC ↑	PR ↑	Rec. ↑	Prec. ↑
AF	ECG-based	✓	0.9969	0.9534	0.9996	0.9914	0.9255	0.9831
	mmArrhythmia [70]	✓	0.9748	0.5316	0.9388	0.5140	0.4884	0.5833
	R2E-Net	–	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
	R2E-Net	✓	0.9912	0.8563	0.9744	0.9069	0.7766	0.9542
SB	ECG-based	✓	0.9730	0.8710	0.9917	0.9303	0.8918	0.8511
	mmArrhythmia [70]	✓	0.9667	0.8428	0.9803	0.8863	0.8621	0.8243
	R2E-Net	–	0.9710	0.8571	0.9892	0.9101	0.8564	0.8579
	R2E-Net	✓	0.9721	0.8639	0.9822	0.9151	0.8723	0.8557
STach	ECG-based	✓	0.9897	0.8911	0.9830	0.9181	0.874	0.9087
	mmArrhythmia [70]	✓	0.9705	0.7198	0.9598	0.7425	0.7877	0.6627
	R2E-Net	–	0.9836	0.8116	0.9849	0.8981	0.7481	0.8869
	R2E-Net	✓	0.9854	0.8502	0.9953	0.9180	0.8664	0.8346
Average	ECG-based	✓	0.9865	0.9052	0.9914	0.9466	0.8971	0.9143
	mmArrhythmia [70]	✓	0.9710	0.6981	0.9596	0.7142	0.7127	0.6901
	R2E-Net	–	0.9808	0.8241	0.9868	0.8987	0.7777	0.8801
	R2E-Net	✓	0.9829	0.8568	0.9840	0.9133	0.8384	0.8815

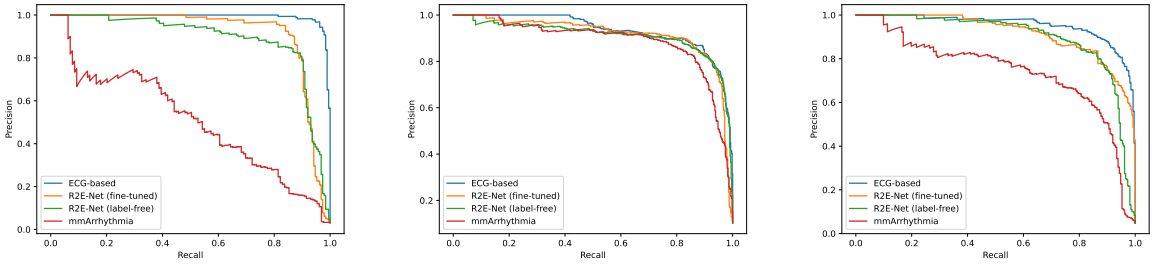
generalizability, likely due to the constrained nature of controlled laboratory settings that do not fully capture real-world complexity.

These findings collectively validate the effectiveness of latent cardiac event modeling as a modality-agnostic representation space. R2E-Net demonstrates that high-fidelity diagnosis is achievable without access to radar-side annotations, while optional fine-tuning with labeled radar data provides additional gains through domain-specific adaptation.

PR Curves: Fig. 6 curves for AF, SB, and STach across four methods. PR curves highlight the trade-off between precision and recall, providing a more informative view of a model's ability to identify positive cases, particularly in class-imbalanced scenarios.

R2E-Net trained without radar-label fine-tuning (i.e., in a label-free setting) exhibits a strong balance between precision and recall, especially in the high-recall region, indicating its ability to detect true positives while maintaining acceptable precision. With radar-side fine-tuning, the PR curve shifts upward across all arrhythmias, demonstrating simultaneous improvements in both recall and precision.

In contrast, the mmArrhythmia approach consistently shows inferior PR performance. Its precision drops sharply as recall increases, particularly in more complex conditions such as AF. The steep decline reflects its



(a) AF (b) SB (c) STach

Fig. 6. Precision-Recall (PR) curves for three arrhythmia types: atrial fibrillation (AF), sinus bradycardia (SB), and sinus tachycardia (STach).

limited capacity to maintain discriminative reliability when attempting to capture more positive instances, suggesting difficulty in modeling the temporal and spatial variability of arrhythmic patterns under noisy, real-world conditions.

The ECG-based model serves as a reference benchmark, exhibiting near-perfect PR curves that remain high across all recall levels. This underscores the superiority of fully supervised, high-fidelity ECG input in maintaining robust performance across challenging diagnostic tasks.

In summary, the PR curve analysis reinforces the effectiveness of R2E-Net, especially when fine-tuned, in balancing sensitivity and specificity in practical, label-scarce settings.

6.2 Bias in Simplified Diagnostic Settings

In real-world cardiovascular diagnosis, the inclusion of non-target arrhythmias (i.e., interfering classes) in the dataset poses a significant challenge. These interfering classes not only degrade diagnostic accuracy but may also introduce bias into performance evaluation. Most existing approaches are validated on datasets containing only a limited set of disease types. Such experimental settings may oversimplify the diagnostic task, resulting in overestimated performance that fails to reflect the complexity of real-world scenarios, where multiple arrhythmias commonly coexist. Table 4 compares model performance under two evaluation settings: one including all interfering diseases, and the other restricted to the three target conditions and healthy individuals.

Firstly, after excluding the interfering disease classes, both R2E-Net and mmArr exhibit substantial performance gains. Specifically,the average F1 score of R2E-Net improves by 8.23%, and its PR-AUC increases by 5.21%, highlighting the considerable negative impact that non-target diseases have on model performance. With these interfering conditions removed, the model is able to diagnose the target arrhythmias with significantly higher precision and recall.

For mmArrhythmia, the F1 score on AF improves markedly from 0.5316 to 0.7663, considerably enhancing its diagnostic capability in this challenging class and bringing it closer to a practical threshold for real-world use. However, it is critical to note that such improvements are observed under idealized evaluation conditions and may not accurately reflect model performance in more complex, interference-rich clinical environments.

These results clearly demonstrate that the presence of interfering diseases substantially affects the performance of cardiovascular diagnostic models. Although existing algorithms may yield promising results under simplified experimental conditions, their effectiveness often declines when evaluated in more complex, interference-rich scenarios. While R2E-Net exhibits relatively smaller performance degradation compared to mmArrhythmia, the overall drop remains non-negligible. This highlights the need for interference-aware evaluation protocols

Table 4. Performance of each method on the full and clean subsets (i.e., excluding other diseases). R2E-Net is trained using an annotation-free strategy.

Class	Method	Subset	Acc ↑	F1 ↑	ROC ↑	PR ↑	Rec. ↑	Prec. ↑
AF	R2E-Net	–	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
		✓	0.9909	0.9264	0.9893	0.9505	0.9043	0.9497
	mmArrhythmia	–	0.9748	0.5316	0.9388	0.5140	0.4884	0.5833
		✓	0.9753	0.7663	0.9691	0.7924	0.7752	0.7576
SB	R2E-Net	–	0.9710	0.8571	0.9892	0.9101	0.8564	0.8579
		✓	0.9737	0.9201	0.9926	0.9687	0.9184	0.9217
	mmArrhythmia	–	0.9667	0.8428	0.9803	0.8863	0.8621	0.8243
		✓	0.9661	0.8943	0.9831	0.9235	0.8796	0.9095
STach	R2E-Net	-	0.9836	0.8116	0.9849	0.8981	0.7481	0.8869
		✓	0.9788	0.8726	0.9871	0.9332	0.8626	0.8828
	mmArrhythmia	–	0.9705	0.7198	0.9598	0.7425	0.7877	0.6627
		✓	0.9861	0.8432	0.9915	0.8877	0.7901	0.9039
Average	R2E-Net	–	0.9808	0.8241	0.9868	0.8987	0.7777	0.8801
		✓	0.9811	0.9064	0.9897	0.9508	0.8951	0.9181
	mmArrhythmia	–	0.9710	0.6981	0.9596	0.7142	0.7127	0.6901
		✓	0.9730	0.8057	0.9750	0.8420	0.7814	0.8328

and underscores the importance of developing models with stronger robustness and adaptability for real-world deployment.

6.3 Comparison with Annotation-free Strategy

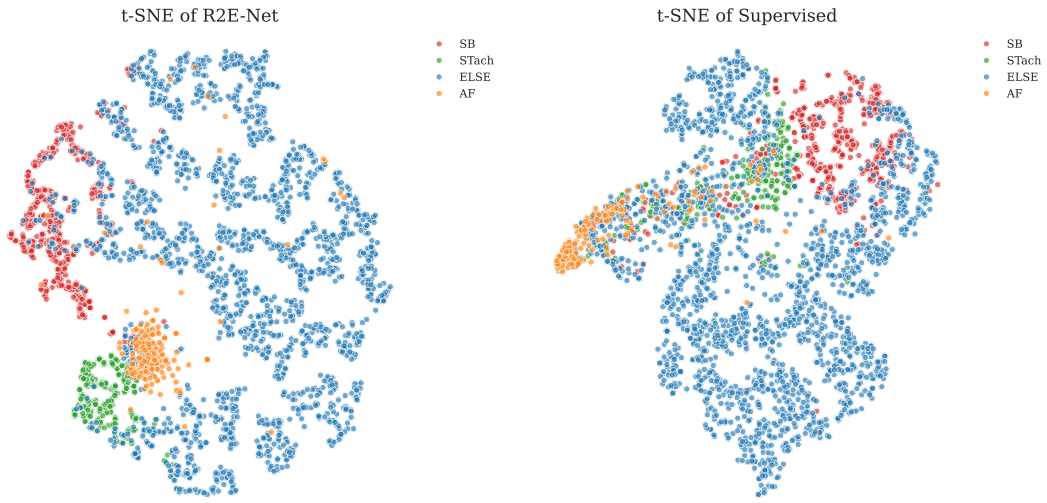
In the absence of radar annotations, a straightforward annotation-free strategy is to apply pseudo-labelling, using an existing ECG-based diagnostic model to assign labels to radar data, which can then be used for supervised training. This approach enables the construction of radar-based classifiers without manual annotation, offering a scalable alternative when labeled radar data is unavailable.

In this section, we compare this baseline strategy with R2E-Net, our proposed method that avoids direct label transfer. Instead of relying on pseudo labels, R2E-Net decouples the learning process by projecting radar signals into a latent cardiac event space derived from auxiliary ECG signals. This design bypasses the need for radar-specific labels while preserving the semantic and temporal structure of physiological events. It addresses a key limitation in radar-based diagnosis: the high cost and limited availability of labeled data, which constrain model scalability and generalizability.

Table 5 presents the experimental results comparing the pseudo-labelling method. In the classification task, the average F1 score of the pseudo-labelling method is only 0.6907, significantly lower than the 0.7848 achieved with direct supervision and the 0.8241 achieved by the R2E-Net model, which also does not rely on radar labels. Although the performance of the pseudo-labelling method is lower than that of supervised learning with true

Table 5. Performance comparison of arrhythmia classification using different labeling strategies.

Class	Strategy	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
AF	Annotation-free	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
	Pseudo-Labelling	0.9744	0.5889	0.951	0.5957	0.5372	0.6516
	Supervised	0.9788	0.7082	0.9532	0.7068	0.7553	0.6667
SB	Annotation-free	0.9710	0.8571	0.9892	0.9101	0.8564	0.8579
	Pseudo-Labelling	0.9469	0.7389	0.9367	0.7671	0.7376	0.7402
	Supervised	0.9685	0.8515	0.9822	0.9008	0.8848	0.8207
STach	Annotation-free	0.9836	0.8116	0.9849	0.8981	0.7481	0.8869
	Pseudo-Labelling	0.9795	0.7444	0.963	0.8667	0.6336	0.9022
	Supervised	0.9802	0.7947	0.9493	0.8458	0.8053	0.7844
Average	Annotation-free	0.9808	0.8241	0.9868	0.8987	0.7777	0.8801
	Pseudo-Labelling	0.9669	0.6907	0.9502	0.7432	0.6361	0.7647
	Supervised	0.9758	0.7848	0.9282	0.8178	0.8151	0.7573



(a) R2E-Net (b) Supervised

Fig. 7. t-SNE visualizations of feature spaces: (a) R2E-Net and (b) Supervised learning.

labels, this result is expected. However, R2E-Net, even without radar labels, outperforms the direct supervision strategy. The strong performance of R2E-Net can be attributed to two key aspects:

Table 6. Performance comparison of R2E-Net with AirECG on AF.

Method	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
R2E-Net	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
AirECG	0.9752	0.6250	0.9664	0.7143	0.6117	0.6389

Latent Cardiac Event Modeling: R2E-Net learns to project radar features into a latent space defined by latent cardiac event sequences derived from ECG data. This allows the radar data to inherit the semantic and temporal structure of cardiac events, even without explicit radar-specific supervision. Fig. 7 illustrates the t-SNE visualization of AF features learned by R2E-Net and a supervised learning model. As shown, although the latent space of the supervised model can partially separate AF from non-AF samples, R2E-Net achieves a more discriminative separation between AF and non-AF features. Moreover, in R2E-Net, multiple distinct feature clusters automatically emerge within the non-AF samples, which may correspond to different subtypes of heart disease—an effect not observed in the supervised model. To demonstrate this, we additionally highlighted SB and STach samples; these samples self-organized into two well-separated and compact clusters, confirming our hypothesis. This suggests that R2E-Net is more effective in capturing fine-grained ECG features, thereby enabling improved detection of various cardiac conditions.

Classifier Based on Latent Cardiac Events: The classifier in R2E-Net is built using latent cardiac event features extracted from ECG signals, enabling it to more accurately capture the physiological characteristics of heart diseases. By leveraging these features, the classifier improves diagnostic accuracy in the radar domain, even in the absence of radar-specific labels.

Based on the above analysis, it can be concluded that using a simple pseudo-labelling approach is unlikely to yield satisfactory performance in heart disease diagnosis. In contrast, R2E-Net, by decoupling representation learning from the classification process and incorporating auxiliary ECG data, achieves effective heart disease diagnosis without radar labels.

In addition to pseudo labeling, an alternative approach for annotation-free diagnosis first reconstructs ECG waveforms from radar signals and then applies conventional ECG interpretation to diagnose disease. The representative work following this pipeline is AirECG [69]. To ensure a fair comparison, we reimplemented AirECG on the same AF dataset and evaluated it alongside R2E-Net; the results are summarized in Table 6.

AirECG achieves a certain level of classification performance, yet it still falls significantly behind R2E-Net on all metrics. The reason lies in the two stage paradigm of AirECG. The radar signals, which are limited in bandwidth and contaminated by noise and motion artifacts, must first be transformed into high fidelity ECG signals, inevitably introducing reconstruction errors. These errors are then amplified during downstream classification, leading to degraded diagnostic accuracy. In contrast, R2E-Net bypasses fine grained waveform reconstruction and directly maps radar data to coarse-grained cardiac event representations, retaining only the information most critical for diagnosis. This event driven strategy effectively shields the model from uncertainties inherent in radar sensing, yielding higher robustness and clinical utility.

6.4 Model Architecture Performance

In this section, to objectively evaluate the performance of the proposed radar spatiotemporal encoder, we adopted an experimental design different from R2E-Net.

We do not use any training paradigm, but only rely on the radar spatiotemporal encoder itself for supervised learning, and compare it with the existing mmArr method. The experimental data from Table 7 shows that the

Table 7. Comparison with baseline on different radar-based arrhythmia classes.

Class	Method	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
AF	mmArrhythmia	0.9748	0.5316	0.9388	0.5140	0.4884	0.5833
	R2E-Net	0.9788	0.7082	0.9532	0.7068	0.7553	0.6667
SB	mmArrhythmiar	0.9667	0.8428	0.9803	0.8863	0.8621	0.8243
	R2E-Net	0.9685	0.8515	0.9822	0.9008	0.8848	0.8207
STach	mmArrhythmia	0.9705	0.7198	0.9598	0.7425	0.7877	0.6627
	R2E-Net	0.9802	0.7947	0.9493	0.8458	0.8053	0.7844
Average	mmArrhythmia	0.9710	0.6981	0.9596	0.7142	0.7127	0.6901
	R2E-Net	0.9758	0.7848	0.9282	0.8178	0.8151	0.7573

supervised learning approach of the radar spatiotemporal encoder has significant advantages in multiple key indicators, especially in the atrial fibrillation (AF) detection task, where its F1 score reached 70.82%, close to the clinical application performance threshold. This result fully demonstrates the effectiveness of the proposed radar spatiotemporal encoder architecture and its ability to extract useful physiological information from complex noise.

Further analysis shows that the radar spatiotemporal encoder uses an end-to-end deep learning framework to directly and automatically extract spatiotemporal features from raw radar signals, eliminating the intermediate steps of traditional methods that depend on preset signal selection algorithms. This avoids information loss and selection errors in traditional preprocessing. Compared to the mmArr method, this encoder can adaptively capture subtle physiological changes in radar signals through a data-driven feature learning mechanism, which is likely the key to its performance improvement.

6.5 Generalization of R2E-Net

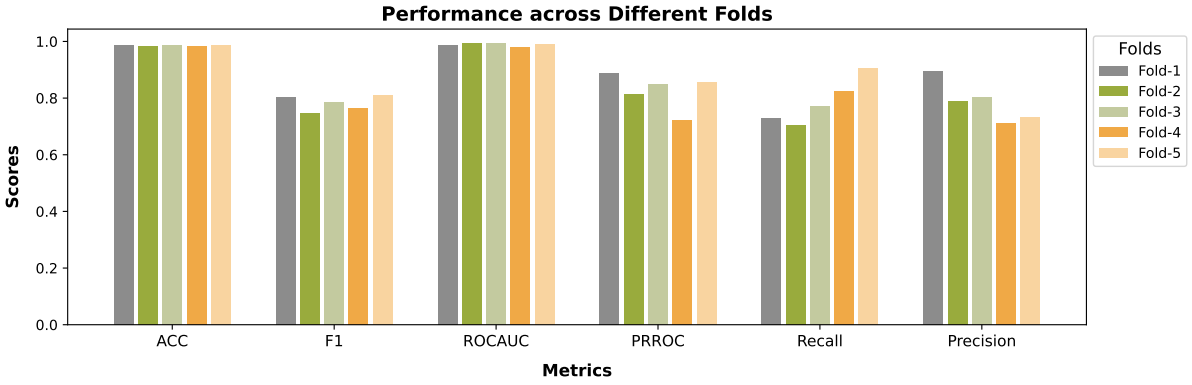


Fig. 8. Comparison of 5-Fold cross validation

In this section, we employed 5-fold cross-validation to assess the model's performance. Given that previous results showed that SB and STach, due to their overly simplistic feature representations, were insufficient to

Table 8. Performance of R2E-Net (annotation-free) on PB classification.

Class	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
PB	0.9710	0.6825	0.9256	0.6995	0.6099	0.7748

effectively evaluate the true performance of the system, we opted to perform five-fold cross-validation on AF to further investigate the generalization ability of the R2E-Net.

The results of the 5-fold cross-validation are shown in Fig. 8:

In the 5-fold cross-validation, the highest F1 score for AF is 0.8102, the lowest is 0.7449, and the average F1 score is 0.7818. This indicates that R2E-Net exhibits generalization ability in cardiac disease monitoring, maintaining relatively stable performance across different data subsets.

In terms of recall and precision, R2E-Net showed significant variability, with the two metrics exhibiting opposite trends. This phenomenon is due to the optimal threshold selection on the validation set, which caused trade-offs between recall and precision across the different folds. However, as seen in the PR curve in Fig6a, although the threshold selection may affect recall and precision, R2E-Net still managed to strike a good balance between both metrics.

The PRROC metric, which is independent of threshold selection, more accurately reflects the true performance of the model. The PRROC value for the fourth fold is 0.7224, significantly lower than the average value of 0.8252, which may be due to the data distribution in that fold. However, the PRROC values for the other four folds were all above 0.8, further validating the superior performance of R2E-Net in practical applications.

In addition to cross-validation, we further evaluated the model's demographic robustness by stratifying performance across age and gender groups. Table 9 summarizes model performance across different age and gender subgroups under a fixed 1:30 class imbalance ratio. As shown in Table 9, R2E-Net maintains consistently strong performance across most age and gender groups under a controlled class imbalance setting. While a gradual decline in F1 score is observed with increasing age, the model still achieves competitive results even in the >65 cohort, indicating that it retains effective discrimination capabilities across demographically diverse populations. The relatively stable performance across both male and female groups further supports the model's generalization capacity in realistic settings.

One notable exception is the lower F1 score observed in the <45 female group. However, this subgroup contains significantly fewer positive samples compared to other age strata, which limits the reliability of this measurement. As such, we refrain from attributing this result to a systematic bias, and instead consider it a reflection of limited statistical support.

Finally, to systematically assess R2E-Net's generalizability across a broader spectrum of arrhythmias, we selected premature beats—both premature atrial contractions (PAC) and premature ventricular contractions (PVC)—as our validation cohort. PACs and PVCs are clinically significant, serving as early harbingers of structural heart disease such as atrial fibrillation, cardiomyopathy, and heart failure, and are strongly associated with an elevated risk of malignant arrhythmias and sudden cardiac death. As reported in Table 8, R2E-Net attains an overall accuracy of 0.971 and an F1-score of 0.682 on premature-beat detection. These results demonstrate that R2E-Net transcends rhythm-specific constraints and possesses robust transferability to other cardiac arrhythmias, laying a solid foundation for its broad clinical deployment.

Taken together, the results from cross-validation, demographic-subgroup analyses, and premature-beat (PAC/PVC) experiments strongly support the conclusion that R2E-Net exhibits desirable generalization properties across a broad arrhythmic spectrum.

Table 9. Performance across different age and gender groups on AF. To ensure comparability, the positive-to-negative sample ratio was fixed at 1:30 for each group.

Age Group	Male			Female		
	F1 Score	Recall	Precision	F1 Score	Recall	Precision
<45	0.8787	1.0000	0.7836	0.7958	0.7516	0.8456
45–54	0.8571	0.8361	0.8793	0.8953	0.8652	0.9277
54–65	0.8019	0.7944	0.8095	0.7949	0.7623	0.8304
>65	0.7037	0.7724	0.6463	0.7132	0.8017	0.6424

Table 10. Performance of R2E-Net (annotation-free) on AF classification with and without ECG pre-processing. Pre-processing improves all metrics.

Data	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
Processed	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
Raw ECG	0.9836	0.7443	0.9332	0.7098	0.6968	0.7988

6.6 Ablation Study on Modeling Conditions

This section presents an ablation study to evaluate the effect of key design choices on the construction and supervision of latent cardiac event representations in R2E-Net. While these factors are not components of the model architecture, they play a critical role in shaping the performance and robustness of model

6.6.1 Impact of ECG Signal Processing. As shown in Table 1, there are significant differences in the quality, distribution characteristics, and statistical properties of data from different sources. These differences may have a profound impact on the subsequent signal processing and analysis. Therefore, to ensure consistency across various data sources, we process the ECG signals in Sec 4.1.2, reducing the potential bias introduced by source domain differences. This section will explore the specific effects of ECG signal processing and analyze its impact on performance.

As presented in Table 10, whether or not signal processing is applied directly affects the model's performance. In the absence of signal processing, due to the variability of ECG signals across different source domains, the quality and representational power of the signals are significantly reduced, leading to a notable decline in model performance. Specifically, without signal processing, the model's performance decreased by 5.92% (from 0.8035 to 0.7443). This result indicates that ECG signal processing not only effectively eliminates the differences between source domains but also significantly improves the quality and representational power of the signals, thereby optimizing the performance of downstream tasks.

6.6.2 Impact of Similarity Metrics. To evaluate how different similarity metrics affect the radar's ability to translate ECG-based latent cardiac events, we compare three variants of R2E-Net, each employing a distinct similarity metric during the supervision stage in the latent space. The role of the similarity metric is to guide the radar encoder to generate latent event representations that align with the structural characteristics of the ECG-defined reference space.

Table 11. Impact of different distance metrics in latent event translation for R2E-Net on AF classification.

Metric	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
Euclidean Distance	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
Cosine Similarity	0.9870	0.7857	0.9798	0.8528	0.7021	0.8919
KL Divergence	0.9719	0.6056	0.9437	0.5993	0.6330	0.5805

Table 12. Impact of auxiliary autoencoder for R2E-Net on AF classification.

Aux. AE	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
–	0.9822	0.6817	0.9732	0.7901	0.5638	0.8618
✓	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954

From Table 11, the baseline model using L2 distance achieves the best overall performance (F1: 0.8035, PR-AUC: 0.8878), as it jointly constrains both the magnitude and direction of feature vectors. This enables the radar features to better match the geometry of the latent space, leading to more accurate event classification. In contrast, cosine similarity only enforces directional consistency while ignoring magnitude, exhibits a slight performance drop, indicating that magnitude (often reflective of physiological intensity) is also critical for precise event-level prediction. KL Divergence matchs the distributions of radar and ECG features, shows a notable decline across all metrics. This suggests that distribution-level constraints may be overly sensitive to local deviations, thereby hindering effective alignment of radar features to the ECG latent structure.

6.6.3 Impact of Auxiliary Autoencoder. To systematically assess the contribution of the auxiliary autoencoder within R2E-Net, we conducted a targeted ablation study; the results are presented in Table 12.

The table shows that introducing the auxiliary autoencoder yields substantial improvements across all key metrics: overall accuracy rises from 0.9822 to 0.9879, F1-score increases from 0.6817 to 0.8035, and the area under the PR curve improves from 0.7901 to 0.8878. This gain can be attributed to the auxiliary autoencoder's explicit reconstruction of the ECG temporal structure, which preserves the integrity and sequential characteristics of cardiac features, delivers a clearer latent space, further facilitates alignment between radar and ECG, and ultimately enhances classification performance.

6.6.4 Impact of Segmentation Strategy. In the medical domain, data typically exhibits a pronounced long-tailed distribution pattern, which inherently leads to class imbalance problems. As evidenced by statistical analysis (see Table 2), AF as a minority class constitutes merely 3.3% of the total sample, a proportion disparity that may exert substantial impact on model performance. To address this challenge, we propose a segmentation-based data augmentation strategy (Algorithm 1). This section conducts ablation studies to systematically evaluate the influence of different segmentation stride on the performance of R2E-Net.

Experimental results indicate that smaller strides increase sample diversity and lead to a more balanced class distribution in the training data, which helps mitigate the class imbalance problem and enhances the model's generalization capabilities. Our method also demonstrates robust performance across different stride sizes, with a noticeable decline in F1, ROC, and recall only when the stride becomes large, such as 256 or 512.

Table 13. Impact of different segmentation stride sizes on AF classification.

Stride	Accuracy	F1	ROC	PR	Recall	Precision
16	0.9847	0.7792	0.9882	0.8770	0.8418	0.7614
32	0.9874	0.8034	0.9879	0.8770	0.7606	0.8512
64	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954
128	0.9861	0.7819	0.9721	0.8302	0.7340	0.8364
256	0.9836	0.7305	0.9840	0.7889	0.6489	0.8356
512	0.9836	0.7234	0.9783	0.7721	0.6330	0.8440

Table 14. Impact of the size of the ECG dataset for R2E-Net on AF classification.

No. of records	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
5,000	0.9782	0.6084	0.9712	0.7463	0.5000	0.7769
10,000	0.9827	0.7073	0.9852	0.8407	0.6170	0.8286
20,000	0.9840	0.7227	0.9893	0.8609	0.6170	0.8722
40,000	0.9854	0.7840	0.9863	0.8636	0.7819	0.7861
80,000	0.9870	0.7919	0.9841	0.8423	0.7287	0.8671
~ 160,000	0.9879	0.8035	0.9863	0.8878	0.7287	0.8954

Concurrently, it is noteworthy that an excessively small stride (e.g., 16) induces performance degradation. This phenomenon is presumably attributable to the excessive segmentation of positive instances: under this stride, the number of AF-category samples becomes approximately 1.12 times that of the remaining categories, thereby engendering over-fitting to the AF class during training and consequently compromising the overall generalization capability.

6.6.5 Impact of the size of the ECG Dataset. R2E-Net utilizes large-scale, publicly annotated ECG datasets comprising approximately 160,000 recordings to construct a latent space for cardiac events. The use of extensive ECG data yields a clearer and more compact latent structure, as well as a more efficient and accurate classifier. In this section, we examine how the size of external ECG data affects performance.

As shown in Table 14, when the number of external ECG samples increases from 5,000 to 40,000, the F1 score rises from 0.6084 to 0.7840, indicating that expanding data scale yields more compact and discriminative representations, thereby enhancing the model's diagnostic capability. However, once the sample size exceeds 40,000, all metrics plateau, demonstrating that 40,000 ECG records are sufficient to build a high-quality latent space and classifier, with additional data offering only limited marginal benefits.

6.6.6 Impact of Select-Point. Traditional radar sensing typically relies on point-selection strategies that use heuristic rules to identify the best signals for improving the signal-to-noise ratio. In real-world conditions,however,these strategies face two key drawbacks: (1) identifying the optimal points is challenging in noisy environments, and (2) discarding seemingly redundant points leads to irreversible information loss. To overcome these issues, R2E-Net replaces heuristic point selection with a spatial encoder that learns the spatial distribution of radar signals in an end-to-end manner.

We compared two alternative strategies: (i) selecting the points with the highest energy and (ii) selecting the points most correlated with the ECG signal.To ensure fairness, we selected 21 radar points, each providing I,

Table 15. Performance comparison of R2E-Net with Select-Point.

Method	Accuracy	F1 Score	ROC-AUC	PR-AUC	Recall	Precision
Spatial Encoder	0.9788	0.7082	0.9532	0.7068	0.7553	0.6667
Maximum Energy	0.9678	0.5095	0.9441	0.4282	0.4946	0.5254
Maximum Correlation	0.9606	0.4233	0.9186	0.3638	0.4255	0.4211

Q, and unwrapped phase channels, resulting in a total of 63 channels, similar to the 64 channels of our spatial encoder.

Results are summarized in Table 15.Both alternative strategies performed worse than the spatial encoder because suboptimal point selection not only led to information loss but also made it difficult to determine an appropriate number of points to balance information coverage and overfitting. In contrast, our method leverages the cardiac event space to guide the spatial encoder in filtering out irrelevant noise,resulting in a more robust representation.

7 RELATED WORK

7.1 RF-based Vital Sign Sensing

Radio frequency (RF)-based vital sign sensing has become an attractive modality for contactless health monitoring. Early studies in this field primarily focused on coarse-grained indicators such as heart rate (HR), respiration, and inter-beat intervals (IBI), demonstrating that radar signals can capture periodic cardiorespiratory motion in both single- and multi-person scenarios [1, 6, 36, 65]. Subsequent work extended this line of research to heart rate variability (HRV) and seismocardiography (SCG)-like mechanical activity [15, 55, 56, 64], enabling more fine-grained insights into cardiac function through chest-wall micro-motions.

More recent efforts have explored the reconstruction of ECG-like signals from millimeter-wave radar, aiming to bridge mechanical radar measurements with electrical cardiac activity. Some approaches investigate the correlation between cardiac mechanical motion and ECG waveforms [12, 61], while others leverage deep learning models to translate radar signals into ECG representations [5]. These include the use of pre-trained ECG autoencoders [32], multi-domain disentanglement [23], and physics-informed priors such as ordinary differential equation constraints [66, 67]. Recent denoising-based pipelines [69] further enhance morphological fidelity and robustness, and have started to address abnormal heartbeats beyond normal sinus rhythms.

To move beyond waveform reconstruction, a growing body of work investigates direct diagnosis of cardiovascular diseases using radar signals. For instance, mmArrhythmia [70] employs deep learning models to classify arrhythmia types directly from radar-derived features, bypassing ECG recovery. AirECG [69] integrates radar-to-ECG translation and abnormal beat detection into a unified diffusion-based pipeline. These approaches demonstrate the feasibility of contactless disease recognition, opening new avenues for non-invasive diagnostic systems.

However, despite promising results, most of these systems are validated under controlled lab conditions or on small-scale datasets. In real-world deployment, radar-based monitoring must contend with heterogeneous patient populations, spontaneous motion, and interfering diseases. Such factors introduce significant domain variability and can degrade diagnostic performance[10, 19, 68]. Robust, generalizable models and evaluation frameworks that reflect practical deployment scenarios remain critical research challenges.

7.2 Cross-modal Learning

Cross-modal learning aims to bridge heterogeneous modalities (such as vision and language, audio and video, or physiological signals and clinical text) by projecting them into a shared semantic space. This joint space enables knowledge transfer across modalities, facilitating multimodal reasoning, retrieval, and classification.

A seminal line of work in cross-modal representation learning explores how to align vision and language using either attention-based fusion or contrastive objectives. ViLBERT [34] introduces a dual-stream Transformer architecture, where visual and textual inputs are processed separately and interact via co-attention layers. This design enables bidirectional reasoning between image regions and language tokens while preserving modality-specific structures. In contrast, more recent approaches such as CLIP and BLIP [24, 42] adopt contrastive learning to align modality-specific encoders into a shared latent space, achieving strong zero-shot performance across diverse visual tasks. These methods scale effectively to large web corpora, shifting the focus from task-specific fine-tuning to generalized multimodal understanding.

In contrast to full alignment strategies, CMG [18, 27, 60] advocates for disentangled multimodal representation learning. By preserving modality-specific latent structures and only aligning shared components via a common codebook, CMG facilitates robust transfer even under weak or missing modality supervision.

Other works investigate cross-modal knowledge distillation (CMKD), which aims to transfer semantic structure across modalities using supervision from a more informative modality. For instance, CMKD frameworks [38, 48] project audio signals into a learned video representation space, enabling models to infer visual dynamics from audio alone. These methods demonstrate the potential of modality translation even in the absence of paired samples.

Beyond visual-text alignment, cross-modal learning has also been widely explored across other modality pairs, including speech-text[40, 54], video-audio[20, 38, 45, 48], and biosignal-language integration [25, 28, 29]. These studies demonstrate the broad applicability of cross-modal representation frameworks across diverse perceptual and physiological domains.

8 CONCLUSION

This work introduces a new paradigm for contactless cardiac diagnosis by reformulating the task as a latent event translation problem, which enables radar-based inference without requiring radar-specific annotations. By learning a structured event space from ECG and projecting radar signals into this representation, R2E-Net eliminates the need for complex transfer learning schemes, achieving robust diagnostic performance through a simple and principled cross-modal projection. Validated on 7,336 clinical recordings, the largest dataset to date in this domain, our method not only generalizes well across arrhythmias and comorbid conditions, but also consistently outperforms supervised models trained directly on radar data. Although supervision requires synchronized ECG, it does not rely on expert labeling, allowing data to be collected passively in practical settings such as community clinics. This significantly lowers the barrier for dataset construction and paves the way for scalable, annotation-free cardiac monitoring in real-world deployments.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62302471 and 62172381.

References

[1] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 2015. Smart Homes That Monitor Breathing and Heart Rate. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. 837–846. doi:10.1145/2702123.2702200

[2] A Buades, B Coll, and J Morel. 2005. A review of image denoising algorithms. *With A New* (2005).

[3] Stephen Butterworth et al. 1930. On the theory of filter amplifiers. *Wireless Engineer* 7, 6 (1930), 536–541.

[4] Hugh Calkins, Matthew R Reynolds, Peter Spector, Manu Sondhi, Yingxin Xu, Amber Martin, Catherine J Williams, and Isabella Sledge. 2009. Treatment of atrial fibrillation with antiarrhythmic drugs or radiofrequency ablation: two systematic literature reviews and meta-analyses. *Circulation: Arrhythmia and Electrophysiology* 2, 4 (2009), 349–361.

[5] Jinbo Chen, Dongheng Zhang, Zhi Wu, Fang Zhou, Qibin Sun, and Yan Chen. 2022. Contactless electrocardiogram monitoring with millimeter wave radar. *IEEE Transactions on Mobile Computing* 23, 1 (2022), 270–285.

[6] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-Robust Vital Signs Waveform Recovery via Deep Interpreted RF Sensing. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. ACM, 392–405. doi:10.1145/3447993.3483251

[7] William S Cleveland and Susan J Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association* 83, 403 (1988), 596–610.

[8] Sedigheh Eslami and Gerard de Melo. 2025. Mitigate the Gap: Improving Cross-Modal Alignment in CLIP. In *Proceedings of the The Thirteenth International Conference on Learning Representations*.

[9] Vitaly Feldman and Chiyuan Zhang. 2020. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). arXiv.

[10] Francesco Fioranelli and Julien Le Kernec. 2021. Radar sensing for human healthcare: Challenges and results. In *2021 IEEE Sensors*. IEEE, 1–4.

[11] Sam F. Friedman, Shaan Khurshid, Rachael A. Venn, Xin Wang, Nate Diamant, Paolo Di Achille, Lu-Chen Weng, Seung Hoan Choi, Christopher Reeder, James P. Pirruccello, Pulkit Singh, Emily S. Lau, Anthony Philippakis, Christopher D. Anderson, Mahnaz Maddah, Puneet Batra, Patrick T. Ellinor, Jennifer E. Ho, and Steven A. Lubitz. 2025. Unsupervised Deep Learning of Electrocardiograms Enables Scalable Human Disease Profiling. *npj Digital Medicine* 8, 1 (2025), 1–13. doi:10.1038/s41746-024-01418-9

[12] Rémi Grisot, Pierre Laurent, Claire Migliaccio, Jean-Yves Dauvignac, Mélanie Brulc, Camille Chiquet, and Jean-Paul Caruana. 2023. Monitoring of Heart Movements Using an FMCW Radar and Correlation with an ECG. *IEEE Transactions on Radar Systems* 1 (2023), 423–434. doi:10.1109/TRS.2023.3298348

[13] Xiao Gu, Zhangdaihong Liu, Jinpei Han, Jianing Qiu, Wenfei Fang, Lei Lu, Lei Clifton, Yuan-Ting Zhang, and David A. Clifton. 2025. Transforming Label-Efficient Decoding of Healthcare Wearables with Self-Supervised Learning and "Embedded" Medical Domain Expertise. *Communications Engineering* 4, 1 (2025), 135. doi:10.1038/s44172-025-00467-6

[14] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross Modal Distillation for Supervision Transfer. In *CVPR*. 2827–2836. doi:10.1109/CVPR.2016.309

[15] Unsoo Ha, Salah Assana, and Fadel Adib. 2020. Contactless Seismocardiography via Deep Learning Radars. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20)*. ACM, 1–14. doi:10.1145/3372224.3419982

[16] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nature Medicine* 25, 1 (2019), 65–69. doi:10.1038/s41591-018-0268-3

[17] Pavel Holoborodko. 2008. Smooth Noise Robust Differentiators.

[18] Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Jieming Zhu, Zhenhua Dong, and Zhou Zhao. 2024. Unlocking the potential of multimodal unified discrete representation through training-free codebook optimization and hierarchical alignment. *arXiv preprint arXiv:2403.05168* (2024).

[19] Shekh Md Mahmudul Islam. 2022. Radar-based remote physiological sensing: Progress, challenges, and opportunities. *Frontiers in Physiology* 13 (2022), 955208.

[20] Simon Jenni, Alexander Black, and John Collomosse. 2023. Audio-visual contrastive learning with temporal self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7996–8004.

[21] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. 2023. Practical Intelligent Diagnostic Algorithm for Wearable 12-Lead ECG via Self-Supervised Learning on Large-Scale Dataset. *Nature Communications* 14, 1 (2023), 3741. doi:10.1038/s41467-023-39472-8

[22] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 951–958.

[23] Bin Li, Wenlong Li, Yuchen He, Wei Zhang, and Hong Fu. 2024. RadarNet: Noncontact ECG Signal Measurement Based on FMCW Radar. *IEEE Transactions on Instrumentation and Measurement* 73 (2024), 1–9. doi:10.1109/TIM.2024.3476545

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[25] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. 2023. Frozen Language Model Helps ECG Zero-Shot Learning. In *Medical Imaging with Deep Learning (Proceedings of Machine Learning Research, Vol. 227)*, Ipek Oguz, Jack H. Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heimann, Despina Kontos, Bennett A. Landman, and Benoit M. Dawant (Eds.).

PMLR, 402–415. arXiv:2303.12311

[26] Xinrong Li, Xiaodong Wang, Qing Yang, and Song Fu. 2021. Signal Processing for TDM MIMO FMCW Millimeter-Wave Radar Sensors. *IEEE Access* 9 (2021), 167959–167971. doi:10.1109/ACCESS.2021.3137387

[27] Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2021. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438* (2021).

[28] Che Liu, Zhongwei Wan, Sibo Cheng, Mi Zhang, and Rossella Arcucci. 2024. ETP: Learning Transferable ECG Representations via ECG-text Pre-Training. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8230–8234. doi:10.1109/ICASSP48485.2024.10446742

[29] Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024. Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement. In *Forty-First International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

[30] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. 2018. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics* 8, 7 (2018), 1368–1373.

[31] Hui Liu, Dan Chen, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Yinglong Wang. 2022. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific data* 9, 1 (2022), 272.

[32] Kai-Chun Liu, Sheng-Yu Peng, Yu Tsao, Che-Yu Liu, Zhu-An Chen, Zong-Han Han, Wen-Chi Chen, Po-Quan Hsieh, You-Jin Li, Yu-Juei Hsu, et al. 2024. A Cross-Modal Autoencoder for Contactless Electrocardiography Monitoring Using Frequency-Modulated Continuous Wave Radar. *IEEE Sensors Journal* (2024).

[33] José A López-López, Jonathan AC Sterne, Howard HZ Thom, Julian PT Higgins, Aroon D Hingorani, George N Okoli, Philippa A Davies, Pritesh N Bodalia, Peter A Bryden, Nicky J Welton, et al. 2017. Oral anticoagulants for prevention of stroke in atrial fibrillation: systematic review, network meta-analysis, and cost effectiveness analysis. *bmj* 359 (2017).

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[35] Jiang Lu, Jin Li, Ziang Yan, Fenghua Mei, and Changshui Zhang. 2018. Attribute-based synthetic network (abs-net): Learning more from pseudo feature representations. *Pattern Recognition* 80 (2018), 129–142.

[36] Marco Mercuri, Ilde Rosa Lorato, Yao-Hong Liu, Fokko Wieringa, Chris Van Hoof, and Tom Torfs. 2019. Vital-Sign Monitoring and Spatial Tracking of Multiple People Using a Contactless Radar-Based Sensor. *Nature Electronics* 2, 6 (2019), 252–262. doi:10.1038/s41928-019-0258-6

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[38] Fabrizio Pedersoli, Dryden Wiebe, Amin Banitalebi, Yong Zhang, George Tzanetakis, and Kwang Moo Yi. 2022. Estimating visual information from audio through manifold learning. *arXiv preprint arXiv:2208.02337* (2022).

[39] Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D Clifford, and Matthew A Reyna. 2020. Classification of 12-Lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement* 41, 12 (2020), 124003. doi:10.1088/1361-6579/abc960

[40] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 513–520.

[41] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence* 45, 4 (2022), 4051–4070.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[43] Adityanarayanan Radhakrishnan, Sam F. Friedman, Shaan Khurshid, Kenney Ng, Puneet Batra, Steven A. Lubitz, Anthony A. Philippakis, and Caroline Uhler. 2023. Cross-Modal Autoencoder Framework Learns Holistic Representations of Cardiovascular State. *Nature Communications* 14, 1 (2023), 2436. doi:10.1038/s41467-023-38125-0

[44] Sandeep Raj and Kailash Chandra Ray. 2018. Sparse Representation of ECG Signals for Automated Recognition of Cardiac Arrhythmias. *Expert Systems with Applications* 105 (2018), 49–64. doi:10.1016/j.eswa.2018.03.038

[45] Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean-baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, et al. 2023. Zorro: the masked multimodal transformer. *arXiv preprint arXiv:2301.09595* (2023).

[46] Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. 2021. *CODE-15%: a large scale annotated dataset of 12-lead ECGs*. doi:10.5281/zenodo.4916206

[47] Gregory A Roth, Catherine O Johnson, Kalkidan Hassen Abate, Foad Abd-Allah, Muktar Ahmed, Khurshid Alam, Tahiya Alam, Nelson Alvis-Guzman, Hossein Ansari, Johan Ärnlöv, et al. 2018. The burden of cardiovascular diseases among US states, 1990-2016. *JAMA*

cardiology 3, 5 (2018), 375–389.
[48] Pritam Sarkar and Ali Etemad. 2024. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14875–14885.
[49] Mohamed Adel Serhani, Hadeel T. El Kassabi, Heba Ismail, and Alramzana Nujum Navaz. 2020. ECG Monitoring Systems: Review, Architecture, Processes, and Key Challenges. *Sensors* 20, 6 (2020), 1796. doi:10.3390/s20061796
[50] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2021. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics* 25, 5 (2021), 1519–1528. doi:10.1109/jbhi.2020.3022989
[51] Brian H Tracey and Eric L Miller. 2012. Nonlocal means denoising of ECG signals. *IEEE transactions on biomedical engineering* 59, 9 (2012), 2383–2386.
[52] M. Unser and A. Aldroubi. 1996. A Review of Wavelets in Biomedical Applications. *Proc. IEEE* 84, 4 (1996), 626–638. doi:10.1109/5.488704
[53] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data* 7, 1 (2020), 1–15.
[54] Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*. PMLR, 10937–10947.
[55] Fengyu Wang, Xiaolu Zeng, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2021. mmHRV: Contactless heart rate variability monitoring using millimeter-wave radio. *IEEE Internet of Things Journal* 8, 22 (2021), 16623–16636.
[56] Haoyu Wang, Jinbo Chen, Dongheng Zhang, Zhi Lu, Changwei Wu, Yang Hu, Qibin Sun, and Yan Chen. 2024. Contactless Radar Heart Rate Variability Monitoring Via Deep Spatio-Temporal Modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 111–115.
[57] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019), 1–37. doi:10.1145/3293318
[58] Zixin Wen and Yuanzhi Li. 2021. Toward Understanding the Feature Learning Process of Self-Supervised Contrastive Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. 11112–11122. arXiv:2105.15134 [cs] doi:10.48550/arXiv.2105.15134
[59] Elizabeth Wilkins, L Wilson, Kremlin Wickramasinghe, Prachi Bhatnagar, Jose Leal, Ramon Luengo-Fernandez, R Burns, Mike Rayner, and Nick Townsend. 2017. European cardiovascular disease statistics 2017. (2017).
[60] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. 2024. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems* 36 (2024).
[61] Chenhan Xu, Huining Li, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Xingyu Chen, Kun Wang, Ming-chun Huang, and Wenyao Xu. 2021. CardiacWave: A mmWave-based Scheme of Non-Contact and High-Definition Heart Activity Computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 135:1–135:26. doi:10.1145/3478127
[62] Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G. Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R. Junna, Aleksandar Videnovic, Terry D. Ellis, Melissa C. Lipford, Ray Dorsey, and Dina Katabi. 2022. Artificial Intelligence-Enabled Detection and Assessment of Parkinson's Disease Using Nocturnal Breathing Signals. *Nature Medicine* 28, 10 (2022), 2207–2215. doi:10.1038/s41591-022-01932-x
[63] Yuqin Yuan, Jinbo Chen, Dongheng Zhang, Ruixu Geng, Hanqin Gong, Guixin Xu, Yu Pu, Zhi Lu, Yang Hu, Dong Zhang, Likun Ma, Qibin Sun, and Yan Chen. 2025. Atrial Fibrillation Detection via Contactless Radio Monitoring and Knowledge Transfer. *Nature Communications* 16, 1 (2025), 4317. doi:10.1038/s41467-025-59482-y
[64] Bin-Bin Zhang, Dongheng Zhang, Yadong Li, Zhi Lu, Jinbo Chen, Haoyu Wang, Fang Zhou, Yu Pu, Yang Hu, Li-Kun Ma, et al. 2024. Monitoring long-term cardiac activity with contactless radio frequency signals. *Nature Communications* 15, 1 (2024), 1–11.
[65] Dongheng Zhang, Yang Hu, and Yan Chen. 2021. MTrack: Tracking Multiperson Moving Trajectories and Vital Signs with Radio Signals. *IEEE Internet of Things Journal* 8, 5 (2021), 3904–3914. doi:10.1109/JIOT.2020.3025820
[66] Yuanyuan Zhang, Runwei Guan, Lingxiao Li, Rui Yang, Yutao Yue, and Eng Gee Lim. 2024. radarODE: An ODE-embedded Deep Learning Model for Contactless ECG Reconstruction from Millimeter-Wave Radar. arXiv:2408.01672 (2024). arXiv:2408.01672 [eess] doi:10.48550/arXiv.2408.01672
[67] Yuanyuan Zhang, Rui Yang, Yutao Yue, and Eng Gee Lim. 2024. radarODE-MTL: A Multi-Task Learning Framework with Eccentric Gradient Alignment for Robust Radar-Based ECG Reconstruction. arXiv:2410.08656 (2024). arXiv:2410.08656 [eess] doi:10.48550/arXiv.2410.08656
[68] Yuanyuan Zhang, Rui Yang, Yutao Yue, Eng Gee Lim, and Zidong Wang. 2023. An overview of algorithms for contactless cardiac feature extraction from radar signals: Advances and challenges. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–20.
[69] Langcheng Zhao, Rui Lyu, Hang Lei, Qi Lin, Anfu Zhou, Huadong Ma, Jingjia Wang, Xiangbin Meng, Chunli Shao, Yida Tang, et al. 2024. AirECG: Contactless Electrocardiogram for Cardiac Disease Monitoring via mmWave Sensing and Cross-domain Diffusion Model. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–27.
[70] Langcheng Zhao, Rui Lyu, Qi Lin, Anfu Zhou, Huanhuan Zhang, Huadong Ma, Jingjia Wang, Chunli Shao, and Yida Tang. 2024. mmarrhythmia: Contactless arrhythmia detection via mmwave sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and*

Ubiquitous Technologies 8, 1 (2024), 1–25.
[71] Jianwei Zheng, Hangyuan Guo, and Huimin Chu. 2022. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet* (2022).
[72] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. 2020. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data* 7, 1 (2020), 48.
[73] Yuxi Zhou, Shenda Hong, Junyuan Shang, Meng Wu, Qingyun Wang, Hongyan Li, and Junqing Xie. 2019. K-Margin-Based Residual-Convolution-Recurrent Neural Network for Atrial Fibrillation Detection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 6057–6063. doi:10.24963/ijcai.2019/839