Expert Systems With Applications 326 (2026) 132601



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa





Multi-task learning with signal denoising and self-distilled representations for wearable ECG arrhythmia detection

Zhongyu Wang [a], Maarten De Vos [b,c], Caiyun Ma [a,*], Jianghai Qian [a], Jianqing Li [a,*], Chengyu Liu [a,*]

[a] Southeast University School of Instrument Science and Engineering, the State Key Laboratory of Digital Medical Engineering, No. 2, Sipailou, Xuanwu District, Nanjing, 210096, China
[b] KU Leuven Department of Electrical Engineering, Kasteelpark Arenberg 10, Leuven, 3001, Belgium
[c] KU Leuven Department of Development and Regeneration, Kasteelpark Arenberg 10, Leuven, 3001, Belgium

ARTICLE INFO

Keywords:
Arrhythmia detection
Multi-task learning
Wearable electrocardiogram
Representation learning

ABSTRACT

The rapid development of wearable electrocardiogram (ECG) technology has enabled long-term and continuous cardiac monitoring, greatly promoting its application in arrhythmia screening. However, ECG records collected in daily life are often contaminated by motion artifacts and other noise that are difficult to suppress, which seriously hinders reliable diagnosis. To address this issue, this paper proposes a multi-task learning (MTL) framework that improves the robustness of arrhythmia detection through two auxiliary tasks: signal denoising and hypersphere representation self-distillation. A bidirectional adaptive task weight balancing strategy based on loss and gradient scaling is designed, and gradient surgery is applied to mitigate task conflicts. Experiments conducted on noisy ECG recordings across different noise types and SNR levels demonstrate that the proposed method improves both detection accuracy and robustness. Under noise interference with SNR levels ranging from -18~-6 dB, the proposed framework achieved an AUC score of 0.742 on the PTB-XL database and 0.801 on the Georgia 12-lead arrhythmia database, corresponding to improvements of 0.13 and 0.109 over direct training, respectively. These results highlight the potential of the proposed approach for reliable arrhythmia screening in wearable ECG applications.

1. Introduction

Arrhythmias are a major public health concern due to their prevalence and potential to cause severe complications such as stroke and heart failure (Ansari et al., 2023). These disorders often occur in a paroxysmal and sporadic manner, which poses significant challenges for timely diagnosis and continuous monitoring. To address this, wearable electrocardiogram (ECG) technology has emerged as a promising tool, enabling long-term, continuous, and comfortable cardiac monitoring in daily life (Nemati et al., 2011). However, since wearable ECG recordings often occur in the subject's daily life, factors such as motion can easily introduce various artifacts, significantly interfering with critical ECG waveforms. This contributes to high misdiagnosis rates and false alarms in arrhythmia detection (Tripathi et al., 2022). Therefore, using signal quality assessment (SQA) methods to screen and retain valid signals with clinical diagnostic value is crucial (Satija et al., 2017).

SQA measures the degree of noise artifacts and waveform distortion. ECG signals are generally classified into three categories: Class A (clear and largely intact waveform with minimal noise), Class B (relatively clear and partially intact waveform, with some noise tolerated), and Class C (almost all waveforms are contaminated and clinically insignificant). In clinical practice, Class A signals can be used directly for arrhythmia diagnosis, while Class B signals often require filtering before screening (Satija et al., 2017; Shahriari et al., 2018). However, when the noise frequency band overlaps significantly with the primary ECG signal band, such as electrode motion artifacts and myoelectric artifacts, filtering often proves difficult to effectively remove, limiting the performance of automated ECG diagnostic algorithms for such signals (Liu et al., 2019; Seeuws et al., 2022). Therefore, developing noise-robust arrhythmia diagnosis algorithms for wearable ECG data is crucial for reducing false alarm rates in clinical practice.

* Corresponding authors.
E-mail addresses: zhongyu@seu.edu.cn (Z. Wang), maarten.devos@kuleuven.be (M. De Vos), 101300447@seu.edu.cn (C. Ma), 220243835@seu.edu.cn (J. Qian), ljq@seu.edu.cn (J. Li), chengyu@seu.edu.cn (C. Liu).

https://doi.org/10.1016/j.eswa.2026.132601
Received 15 December 2025; Received in revised form 3 April 2026; Accepted 22 April 2026
Available online 29 April 2026
0957-4174/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Recent advances in representation learning (RL) have demonstrated strong capabilities in projecting raw biomedical signals into high-dimensional feature spaces for downstream classification tasks (Hu et al., 2022; Zhang et al., 2021). However, under severe artifact interference, learned representations are prone to drifting toward biased subspaces, thereby obscuring discriminative cardiac patterns and degrading arrhythmia detection performance. From this perspective, explicitly incorporating prior knowledge of clean ECG representations into the learning process becomes critical for improving robustness against noise-induced representation distortion. To address this challenge, we formulate wearable ECG arrhythmia detection within a multi-task learning (MTL) framework, where arrhythmia classification is treated as the primary objective, and auxiliary tasks are introduced to guide robust representation learning under noisy conditions. Specifically, we first integrate a signal denoising task based on an autoencoder architecture to reconstruct clean ECG signals from noisy inputs. This task encourages the shared backbone network to recover physiologically meaningful waveform structures and fine-grained temporal dependencies that are essential for downstream diagnosis. Beyond signal-level supervision, we further enhance representation robustness at the embedding and distributional levels by introducing a self-distilled representation transfer mechanism. Inspired by recent progress in knowledge distillation and representation alignment (Ding et al., 2023; Ma et al., 2024; Sepahvand et al., 2022), we propose a novel Hypersphere Self-Distillation (HSD) strategy, which transfers geometric priors from clean ECG representations to noisy ones under shared network parameters. By explicitly matching neighborhood relationships between paired clean and noisy representations, HSD constrains noisy embeddings to conform to a hyperspherical manifold that approximates the intrinsic geometric structure of clean ECG feature distributions. In this way, the model learns noise-invariant and clinically meaningful representations beyond pointwise reconstruction, enabling robust arrhythmia discrimination under artifact interference.

By jointly incorporating signal denoising and hypersphere self-distillation as auxiliary learning objectives, the proposed framework promotes robust representation learning across multiple scales, ranging from waveform reconstruction to global manifold regularization. This multi-level supervision effectively complements wearable ECG arrhythmia detection, particularly in scenarios where signal quality is insufficient for direct diagnosis but still retains partial clinical value. More broadly, multi-task learning has demonstrated significant advantages across computer vision, natural language processing, and biomedical signal analysis by jointly modeling multiple related tasks within a shared architecture (Khoshkangini et al., 2023; Zhang & Yang, 2017, 2022). The core principle of MTL lies in transferring knowledge across tasks via shared inductive biases, thereby improving generalization and alleviating overfitting commonly observed in single-task learning. In practice, MTL frameworks typically adopt parameter sharing, representation sharing, or regularization-based constraints to capture common structures across tasks while preserving task-specific characteristics. From an optimization perspective, a standard MTL model consists of a shared backbone and multiple task-specific heads.

However, due to inherent differences in task loss scales and gradient magnitudes, naïvely assigning equal weights (EW) to all tasks (Zhang & Yang, 2022) may bias optimization toward tasks with dominant gradients, leading to task imbalance and optimization conflicts (Caruana, 1997; Zhang & Yang, 2022). To mitigate this issue, a variety of task balancing strategies have been proposed, including loss-based weighting methods such as uncertainty modeling (Cipolla et al., 2018) and loss descent rate adaptation (Liu et al., 2019), as well as gradient-based approaches that directly regulate gradient interactions, including GradNorm (Chen et al., 2018), PCGrad (Yu et al., 2020), MGDA (Sener & Koltun, 2018), IMTL (Liu et al., 2021), and XSleepNet (Phan et al., 2022).

Nevertheless, the multi-task learning setting considered in this work fundamentally differs from conventional parallel MTL formulations. Instead of assuming equal importance across tasks, we explicitly adopt a hierarchical task formulation, in which arrhythmia detection serves as the primary objective, while signal denoising and hypersphere self-distillation are introduced solely to facilitate its optimization. Consequently, the absolute optimization quality of auxiliary tasks is not a primary concern. Under this setting, conventional task balancing strategies that aim to simultaneously optimize all tasks toward Pareto optimality may be suboptimal. Therefore, we propose a task balancing strategy that jointly considers gradient direction and loss scale, with the objective of maximizing the effective contribution of auxiliary tasks to the optimization of the primary arrhythmia detection task. The proposed framework is specifically applied to wearable ECG signals with a Class B signal quality rating, rather than fully clean Class A recordings, aligning with real-world wearable monitoring scenarios where moderately noisy yet clinically informative signals constitute the main challenge.

The contributions of this paper can be summarized as follows:

(1) This work designs a multi-task learning framework that uses signal denoising and hypersphere representation constraints to assist noisy ECG arrhythmia detection. By approximating the clean signal from fine-grained and overall distribution, the model shares the backbone to extract more discriminative and robust representations.

(2) This work adopts a bidirectional adaptive task weight balancing strategy, which simultaneously scales the values of the loss function and gradient to the same scale to prevent model task collapse. At the same time, a gradient surgery method is used to alleviate potential conflicts between the auxiliary task and the main task, thereby improving the overall optimization effect.

(3) The proposed method is tested under noise interference of different noise combinations and different signal-to-noise ratios, verifying its superior robustness in the face of noise pollution of different intensities.

Compared with existing methods for detecting noisy ECG arrhythmias, this paper innovates in three ways:

(1) Unlike traditional methods that handle denoising and classification independently, we model noisy ECG arrhythmia detection as a hierarchical multi-task learning problem.

(2) Unlike existing self-distillation methods that focus on feature imitation, we propose a self-distillation strategy with hypersphere neighborhood constraints, enabling the model to obtain a clean and robust feature space.

(3) Unlike traditional multi-task balancing methods that pursue Pareto optimality, we design an adaptive optimization strategy oriented towards the main task, where the auxiliary tasks do not pursue performance improvements in metrics.

The rest of the paper is organized as follows: Section 2 describes the related work, Section 3 details the methodology and experimental setup proposed in this study, Section 4 presents the experimental results, Section 5 discusses the results, and finally, Section 6 provides the conclusions.

2. Related work

2.1. Arrhythmia detection in deep learning

In recent years, numerous studies have investigated arrhythmia detection using deep learning on twelve-lead or reduced-lead ECGs (Gupta et al., 2022). Ylldlrlm et al. (2018) proposed a single-lead arrhythmia diagnosis strategy based on a 1-D CNN, achieving 91.33% accuracy across 17 disease categories in the MIT-BIH Arrhythmia Database. Hannun et al. (2018) designed a DNN architecture and performed arrhythmia detection on 91,232 single-lead ECGs from 53,549 patients using single-lead Holter monitoring devices, achieving an AUC score of 0.97, demonstrating that the DNN's arrhythmia detection capability reached expert-level performance. Hu et al. (2022) developed a Transformer-based neural network model, ECG DETR, that effectively captures inter-beat dependencies and can locate the position and type of heartbeats in the ECG without explicit segmentation. Zhang et al. (2020) proposed a

STA-CRNN by fusing spatiotemporal attention with a convolutional recurrent network, effectively integrating global and local features of ECG signals. Petmezas et al. (2020) used CNN-LSTM and focal loss to achieve high-precision atrial fibrillation detection on imbalanced data distribution. Yao et al. (2019) proposed an attention-based time-increment convolutional neural network (ATI-CNN) capable of processing signals of different lengths. It combines a recurrent structure and an attention mechanism to achieve the fusion of spatiotemporal information of ECG signals. Mousavi et al. (2020) designed a bidirectional recurrent neural network method HAN-ECG for predicting atrial fibrillation. It integrates three-level hierarchical attention to effectively enhance the clinical interpretability of the results. proposed an arrhythmia detection algorithm that combines fractional wavelet transform (FrWT) for preprocessing, Yule-Walker autoregressive analysis (YWARA) for feature extraction, and principal component analysis (PCA) for detection, achieving a high accuracy of 99.89% on the MIT-BIH Arrhythmia database (Gupta & Mittal, 2022). Nguyen et al. (2025) introduced an ECG foundation model employing a mixture-of-layers aggregation strategy, where features from multiple transformer layers are adaptively fused via a gating mechanism. This method enhances the representation quality by capturing information at different abstraction levels and improves generalization on both in-distribution and out-of-distribution ECG classification tasks.

However, most existing studies focus on arrhythmia detection using twelve-lead or single-lead ECGs from resting databases such as PTB-XL and MIT-BIH, where noise interference is relatively limited. In contrast, wearable ECG recordings are often contaminated by strong noise due to daily activities of the wearer. While artifacts such as baseline wander and electromyographic interference can typically be mitigated during preprocessing, certain types of noise, such as motion artifacts, are nearly impossible to eliminate completely. These artifacts introduce signal distortions that significantly compromise model performance. Therefore, improving the robustness of arrhythmia detection under severe artifact contamination remains an open and critical challenge.

2.2. ECG signal denoising

ECG signals are highly susceptible to various noises, leading to high misdiagnosis and false alarm rates. Numerous studies have developed various ECG denoising methods (Gupta et al., 2024), Gupta et al. (2024). Traditional denoising methods typically rely on filtering techniques, such as wavelet transform (Gupta, 2023; Kumar et al., 2021), empirical mode decomposition (EMD) (Kabir & Shahnaz, 2012), and adaptive filtering (Marque et al., 2005). While effective, these techniques have limitations in capturing complex, non-stationary patterns. Recently, deep learning-based methods have been introduced to address these limitations. Li et al. (2024) proposed a diffusion probabilistic modeling paradigm for ECG signal denoising. In this approach, the noise corruption process is modeled as a forward diffusion process, while signal restoration is formulated as a reverse diffusion-based generative process. By progressively transforming the noisy signal distribution toward the clean signal distribution, the method ultimately achieves ECG denoising through sampling from a denoising diffusion probabilistic model (DDPM). Singh and Pradhan (2021) used a generative adversarial network (GAN) to enable a generator to learn the patterns of ECG signals through a game-like approach, thereby generating clean ECG signals from noisy signals and achieving denoising. Antczak (2019) proposed a deep recurrent neural network (DRNN) framework for ECG signal denoising, in which stacked long short-term memory (LSTM) units were employed to model the temporal dependencies of ECG waveforms. The model was first pretrained on large-scale synthetic ECG signals generated by a dynamic ECG model and subsequently fine-tuned using real ECG recordings. This pretraining strategy enabled the network to learn generic morphological patterns of ECG signals, thereby improving denoising performance under severe noise conditions, highlighting the effectiveness of recurrent architectures in capturing long-range temporal correlations for ECG signal restoration. Samann et al. proposed the RunDAE model (Samann & Schanze, 2023), which applies a running denoising autoencoder for ECG signal denoising. In this framework, the denoising autoencoder is executed in a sliding-window manner to adapt to nonstationary noise characteristics in ECG recordings, thereby enhancing denoising robustness under time-varying noise conditions by leveraging local temporal context. These methods achieve more robust and adaptive ECG denoising by learning the intrinsic features of the ECG.

Although prior studies have demonstrated the effectiveness of various denoising strategies for ECG signal restoration, most approaches focus solely on signal reconstruction and do not explicitly optimize downstream arrhythmia classification under noisy conditions. In contrast, the proposed framework leverages denoising and representation-level self-distillation as auxiliary tasks to enhance the noise robustness of the primary arrhythmia detection task, thereby bridging the gap between signal restoration and clinical diagnostic performance Fig. 1 and Tables A.1–A.3 and B.4.

2.3. Self-distillation

Self-distillation is a special type of knowledge distillation for representation transfer algorithms. Unlike traditional knowledge distillation, which transfers intermediate representations, attention maps, and data distributions from an over-parameterized teacher model to a compact student model, self-distillation does not require a separate teacher model. Instead, it typically uses itself or the parameters of the previous iteration as the teacher to regularize the representation. Zhang et al. (2019) proposed a self-distillation method that uses the deep layers of a model as teachers and the shallow layers as students, transferring high-level semantic information to the shallow layers, thereby providing positive feedback to the entire model. Kim et al. (2021) proposed multi-generation self-distillation, using the updates of previous models to guide the backpropagation of the next model, achieving good regularization. Andonian et al. (2022) proposed a robust cross-modal representation learning framework that uses progressive self-distillation and soft image-text alignment to mitigate noisy correspondences and improve the generalization ability of visual language models. Adnan et al. (2021) introduce a clustering framework that leverages self-distillation between network branches to eliminate the need for domain-specific data augmentation, thereby improving the stability and performance of unsupervised representation learning. These studies demonstrate that self-distillation is a promising parameter-friendly regularization technique.

In this work, we constrain the neighborhood distribution of noisy signals through self-distillation to achieve robust representation learning.

2.4. Multi-task learning

Multi-task learning (MTL) has emerged as an effective paradigm that leverages shared representations across related tasks to improve generalization and reduce overfitting. Early approaches primarily focused on parameter-sharing strategies, such as hard sharing, where a common backbone network is jointly optimized for multiple tasks, and soft sharing, which allows task-specific parameters with shared regularization. To balance task contributions, several methods have been proposed, including dynamic weighting of task losses (Liu et al., 2021), gradient normalization (Chen et al., 2018), uncertainty-based task weighting (Cipolla et al., 2018; Liu et al., 2021), and formulating MTL as a multi-objective optimization problem (Sener & Koltun, 2018). Other works address task competition using game-theoretic formulations (Navon et al., 2022). More recent advances further integrate attention mechanisms and adversarial learning to capture task-specific features while preserving shared representations (Bai & Zhao, 2022).

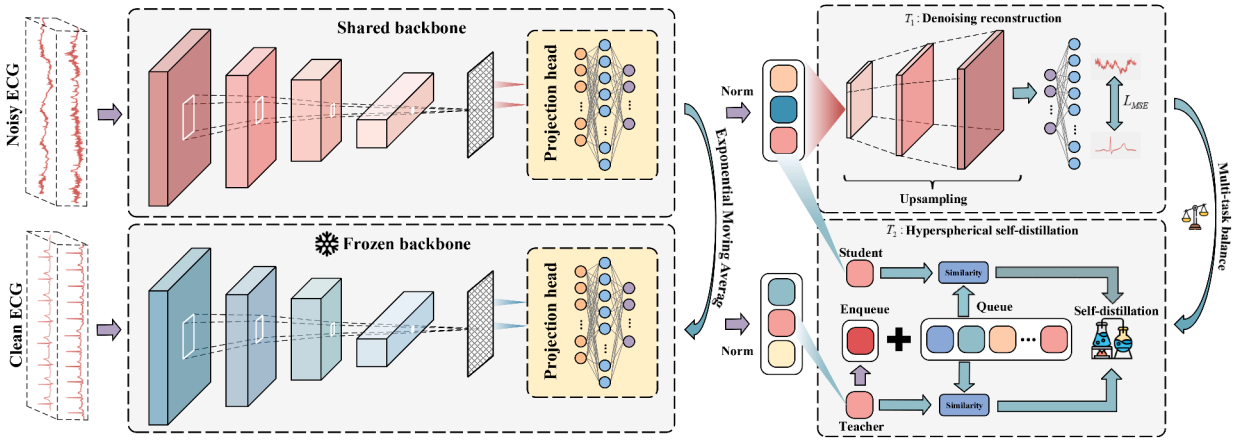


Fig. 1. Overview of the proposed multi-task learning framework based on denoising and hypersphere self-distillation assistance.

MTL has demonstrated strong potential in modeling heterogeneous yet correlated tasks across multiple domains. In medical imaging, it has been applied to simultaneously address segmentation, detection, and classification (Hering et al., 2023; Ling et al., 2024; Zhao et al., 2023), showing consistent improvements in both accuracy and efficiency. In natural language processing, MTL has been employed for joint modeling of tasks such as machine translation, syntactic parsing, and sentiment analysis, thereby significantly enhancing performance in low-resource scenarios (Chen et al., 2024). In speech and audio processing, MTL has been used to jointly optimize speech recognition, speaker identification, and emotion detection, improving robustness under noisy conditions (Sigtia et al., 2020). Furthermore, in autonomous driving and robotics, MTL has been explored for joint lane detection, object recognition, and trajectory prediction (Li et al., 2022; Song et al., 2022). These studies collectively demonstrate that MTL can effectively exploit task correlations to improve generalization, robustness, and efficiency across domains.

Despite these advances, the application of MTL in ECG analysis–particularly in arrhythmia detection–remains limited. Existing works are mostly confined to tasks such as joint heartbeat segmentation and event diagnosis (Wang et al., 2025), disease diagnosis and localization (Pan et al., 2022), or signal denoising and quality assessment (Chen et al., 2024).

In this work, we extract robust representation of noisy ECGs via a shared structure of a multi-auxiliary task constrained model, aiming to improve the diagnostic accuracy of arrhythmia detection algorithms in ECGs contaminated by artifacts.

3. Method

This section introduces the design of auxiliary tasks to improve the model's robustness against noise and the trade-off strategy for multi-tasks. In the proposed method, we denote the shared parameters of multiple tasks as θ_s and the parameters of the task head as ψ_t, where t is the number of tasks. Assuming $L_t(\theta_s, \psi_t)$ is the loss for task t, the overall loss can be expressed as $\sum_{i=1}^{t} w_i L_t(\theta_s, \psi_t)$. Generally speaking, the goal of parallel multi-task learning is to find the Pareto optimality of the model, which is defined as:

Definition 1 (Pareto optimality). A solution θ dominates a solution $\bar{\theta}$ if $L_t(\theta_s, \psi_t) \leq L_t(\bar{\theta}_s, \bar{\psi}_t)$ for all tasks t and $L(\theta_s, \psi_1, \ldots, \psi_t) \neq L(\bar{\theta}_s, \bar{\psi}_1, \ldots, \bar{\psi}_t)$.

In contrast to Pareto-based multi-objective optimization, our objective is explicitly asymmetric. We aim to optimize the primary arrhythmia detection task, while treating auxiliary tasks as regularizers that shape the representation space and optimization trajectory. An overview of the proposed framework is shown in Fig. 1. Therefore, achieving optimal performance on auxiliary tasks is neither required nor desired. Specifically, the loss function of our main task can be expressed as:

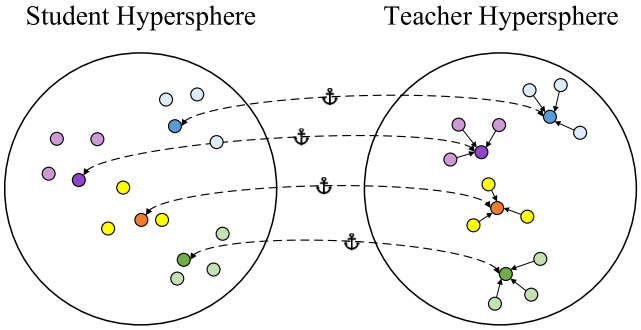


Fig. 2. Illustration of the hyperspherical self-distillation neighborhood relationship.

$$\mathcal{L}_{T_m}(\theta_s, \psi_m) = -\frac{1}{N}\sum_{N}[y \log \hat{y} + (1-y)\log(1-\hat{y})]. \tag{1}$$

Here, ψ_m is the parameter of the main task classification head, y is the ground truth of the sample, and $\hat{y}$ is the model's predicted probability of the sample. The optimization of θ_s by the auxiliary task aims at the optimal convergence of the main task.

3.1. Denoising and hypersphere self-distillation

The key to improving the noise robustness of arrhythmia diagnosis models lies in enabling them to effectively extract useful discriminative representations from disordered signals. Therefore, we use signal denoising to enable the model to capture fine-grained representations of clean components in noisy signals and utilize hypersphere constrained self-distillation to enable the model to learn inter-sample relationships in the clean signal representation, thereby improving the intra-class compactness and inter-class separability of the primary classification task.

Specifically, let signal denoising be the auxiliary task T_1 and hypersphere constrained self-distillation be the auxiliary task T_2. Assume that a batch of samples from the noise training dataset $\mathcal{D}_{noise}$ is $\{x_i, y_i\}_{i=1}^{N}$, and the corresponding clean samples from the clean signal training dataset $\mathcal{D}_{clean}$ are $\{\hat{x}_i\}_{i=1}^{N}$, N is the number of samples. Then the loss function of the auxiliary task T_1 can be expressed as:

$$\mathcal{L}_{T_1}(\theta_s, \psi_1) = \frac{1}{N}\sum_{N}||F(x_i; \theta_s, \psi_1) - \hat{x}_i||_2^2. \tag{2}$$

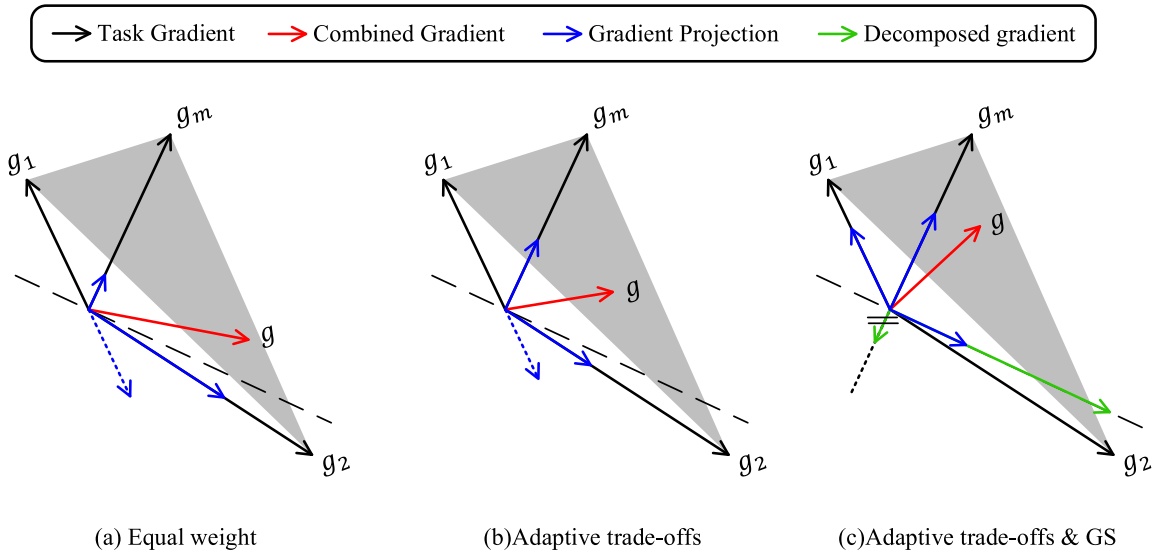


Fig. 3. Illustration of gradient space geometry in multi-task learning. Let g_m denote the gradient of the main task, and g_1, g_2 denote the gradients of auxiliary tasks T_1 and T_2, respectively. The gray background represents the linear space spanned by multi-task gradients. The black dashed line indicates the hyperplane normal to g_m. The red line shows the combined gradient, the blue line denotes its normal projection onto the single-task gradient, and the green line represents the orthogonal decomposition component. (a) Gradient space under equal weighting (EW). (b) Gradient space under adaptive gradient balancing. (c) Gradient space under the combination of adaptive balancing and gradient surgery.

Then the gradients of L_{T_1} for each component of the model are: $\theta_s \leftarrow \nabla_{\theta_s}\mathcal{L}_{T_1}$ and $\psi_1 \leftarrow \nabla_{\psi_1}\mathcal{L}_{T_1}$.

For the auxiliary task T_2, hypersphere constrained self-distillation is performed. By maximizing the mutual information and distribution consistency between the representation of the noisy signal and its corresponding clean signal on the unit hypersphere, the model refines valid information and removes invalid redundancy. Specifically, we regularize the model using the shared backbone as its own teacher. Assume that the model's representation for x_i is the student representation z_i, and the representation for $\hat{x}_i$ is the teacher representation $\hat{z}_i$. During the optimization process, the representation is first projected onto the unit hypersphere to facilitate a more compact classification. distribution within the class and a more significant inter-class interval, further, we encourage the student representation to imitate the neighborhood relationship of the teacher representation. In this process, we ensure that $\hat{z}_i$ does not affect parameter optimization by stopping the gradient propagation operation. In order to promote the convergence of the student distribution to the neighborhood relationship of the teacher distribution, the concept of MoCo (He et al., 2020) contrastive self-supervision, which has great potential in the field of representation learning, is introduced. Similar to MoCo, we design a first-in-first-out (FIFO) memory queue $Q = [q_1, q_2, \ldots, q_K]$ to store a large number of neighborhood samples of the teacher representation, where K is the number of samples in the queue. Then the neighborhood relationship P_S,P_T of the student representation and the teacher representation can be expressed as:

$$P_S[q_k] = \frac{\exp(z_i \cdot q_k/\tau^S)}{\sum_{q\sim Q}\exp(z_i \cdot q/\tau^S)}, \tag{3}$$

$$P_T[q_k] = \frac{\exp(\hat{z}_i \cdot q_k/\tau^T)}{\sum_{q\sim Q}\exp(\hat{z}_i \cdot q/\tau^T)}, \tag{4}$$

where $P_S[q_k]$ and $P_T[q_k]$ are the relationships between z_i and $\hat{z}_i$ and the neighborhood sample q_k, respectively. τ^S and τ^T represent the distillation temperatures. z_i and $\hat{z}_i$ are the unit hypersphere projections of the representations, which are calculated as $z_i = F(x_i;\theta_s)/||F(x_i;\theta_s)||_2$ and $\hat{z}_i = F(\hat{x}_i;\theta_s)/||F(\hat{x}_i;\theta_s)||_2$.

Finally, we maximize the mutual information between the neighborhood relationships of the student and teacher representations using cross-entropy loss to approximate the distribution of the noisy signal to the clean signal. The illustration of the hyperspherical self-distillation is shown in Fig. 2. The loss function can be expressed as:

$$\begin{aligned}\mathcal{L}_{T_2}(\theta_s) &= \sum_{i=1}^{N} -P_T \cdot \log P_S \\ &= \sum_{i=1}^{N}\sum_{j=1}^{K} -\frac{\exp(\hat{z}_i \cdot q_k/\tau^T)}{\sum_{q\sim Q}\exp(\hat{z}_i \cdot q/\tau^T)} \cdot \\ &\quad \log\frac{\exp(z_i \cdot q_k/\tau^S)}{\sum_{q\sim Q}\exp(z_i \cdot q/\tau^S)}.\end{aligned} \tag{5}$$

Then the gradient of the loss $\mathcal{L}_{T_2}$ with respect to each component of the model are: $\theta_s \leftarrow \nabla_{\theta_s}\mathcal{L}_{T_2}$ and $\psi_2 \leftarrow \nabla_{\psi_2}\mathcal{L}_{T_2}$. The auxiliary task T_2 directly utilizes the intermediate layer representation, so $\psi_2 = \phi$. The auxiliary task T_2 directly utilizes the representation of the intermediate layer and therefore does not have a separate task head $\psi_2 = \phi$.

3.2. Task balance and conflict regulation

In the process of parallel optimization of multiple tasks sharing the same network structure, differences in loss and gradient scales can lead to training instability. When the gradients of certain tasks dominate, model updates tend to favor a minority of tasks, weakening the positive contributions of other tasks to representation learning. When gradients conflict between tasks, this can lead to offsetting parameter updates and even task collapse, ultimately impacting the primary task with negative transfer. Therefore, designing a mechanism that dynamically balances multi-task weights and mitigates gradient conflicts is crucial for fully leveraging the value of auxiliary tasks, improving the optimization efficiency of the primary task, and enhancing overall model performance.

From a backpropagation perspective, model parameter updates depend on the calculated gradients. Therefore, this paper adopts a gradient-based task balancing approach. Specifically, we mitigate the impact of gradient scale differences by applying the logarithm to the loss, as shown below:

$$\nabla_\theta \log \mathcal{L}(\theta) = \frac{1}{\mathcal{L}(\theta)}\nabla_\theta \mathcal{L}(\theta), \tag{6}$$

this balancing feature automatically normalizes the gradients on the loss scale, making the gradients of tasks with smaller losses larger and those of tasks with larger losses smaller, thereby preventing task collapse. Furthermore, it is consistent with the maximum likelihood estimation (MLE). Assume that the total loss of T tasks on the shared parameters is:

$$\mathcal{L}_{total}(\theta) = \sum_{t=1}^{T} \mathcal{L}_t(\theta) = -\sum_{t=1}^{T} \log p(y_t|x;\theta), \tag{7}$$

where x is the model input and y_t is the observation of the input by task t. Minimizing the total loss is equivalent to minimizing the negative log-likelihood of the conditional distribution:

$$\arg\min_{\theta} \mathcal{L}_{total}(\theta) = -\arg\min_{\theta} \sum_{t=1}^{T} \log p(y_t|x;\theta) \Rightarrow \theta^*. \tag{8}$$

Furthermore, an exponential moving average (EMA) is applied to the gradient $\nabla_\theta \log \mathcal{L}(\theta)$ to reduce the impact of short-term gradient noise on training stability, and the gradient norm is taken to preserve the directionality of the gradient while eliminating scale differences.

In high-dimensional parameter estimation, independently estimating each parameter is not always optimal. Stein's paradox (, n.d.) reveals that shrinking toward the center of the whole can achieve lower overall error. Analogous to Stein's paradox (, n.d.), shrinking shared representations (information sharing) in multi-task learning can generally reduce overall error. However, when the directions of task gradients conflict significantly, forced shrinkage can cause each task to deviate from its optimal solution, resulting in increased single-task loss. Therefore, we consider auxiliary tasks that are negatively correlated with the direction of the main task gradient as conflicting and project them onto the normal hyperplane of the main task gradient. Unlike directly discarding conflicting gradients, projecting gradients helps prevent model parameters from being split.

Fig. 3 illustrates the principle of the proposed multi-task balancing strategy in the gradient space. The gray region denotes the linear space spanned by multi-task gradients, where g_m represents the main task gradient, and g_1 and g_2 denote the gradients of auxiliary tasks T_1 and T_2, respectively. The red, blue, and green lines correspond to the combined multi-task gradient, the orthogonal projection of the combined gradient onto a single-task gradient, and the decomposition of the single-task gradient along orthogonal directions, respectively. When equal weighting (EW) is applied in multi-task training, the combined gradient g in Fig. 3(a) may be dominated by the task with a larger gradient magnitude, e.g., g_2, while the contributions of other tasks are diminished, leading to negative transfer to the main task g_m. As shown in Fig. 3(b), with the adaptive balancing strategy, the gradients of different tasks contribute more equally to the combined gradient; however, conflicts with the main task gradient still occur. By applying gradient surgery, as depicted in Fig. 3(c), the conflicting task gradients are projected onto the normal hyperplane of g_m, aligning the combined gradient g closer to g_m while retaining the regularization effect of g_2, thereby mitigating inter-task conflicts. This formulation differs fundamentally from Pareto-based multi-task learning, where all tasks are treated equally. In our setting, auxiliary tasks serve as inductive biases rather than optimization targets. Therefore, enforcing Pareto optimality may lead to suboptimal solutions for the primary task. The detailed algorithm flow is shown in Algorithm 1.

Algorithm 1: The training process of the proposed method.

Input : Training Datasets $\mathcal{D}_{noise} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ and $\mathcal{D}_{clean} = \{\hat{\mathbf{x}}_i\}_{i=1}^{N}$. A parameter-sharing backbone network θ and the task heads ψ_1, ψ_2 ($\psi_2 = \phi$) and ψ_m.

Output: Convergent multi-task model.

while *not converged* **do**

- Extract mini-batches of $\mathcal{B}_{clean}$ and $\mathcal{B}_{noise}$ from the datasets $\mathcal{D}_{clean}$ and $\mathcal{D}_{noise}$.;
- Forward propagate $\mathcal{B}_{noise}$ and $\mathcal{B}_{clean}$ into the network to obtain the predictions $\hat{\mathbf{y}}_i$, representations z_i, $\hat{z}_i$ and the reconstructed signal $\hat{\mathbf{x}}$;
- Compute task-specific losses $\mathcal{L}_m$, $\mathcal{L}_{T_1}$ and $\mathcal{L}_{T_2}$ and as formulas (1), (2), (5);
- `/* Calculate the gradients of each loss. */`
- $g_t^k = \nabla_{\theta,\psi_t} \log \mathcal{L}_t^k(\theta, \psi_t), t \in [1,2]$;
- $g_m^k = \nabla_{\theta,\psi_m} \log \mathcal{L}_m^k(\theta, \psi_m)$;
- `/* Projecting conflicting gradients. */`
- **if** $\cos(g_t^k(\theta), g_m^k(\theta)) < 0$ **:**
  - $g_t^k(\theta) = g_t^k(\theta) - \frac{g_t^k(\theta) \cdot g_m^k(\theta)}{||g_m^k(\theta)||^2} \cdot g_m^k(\theta)$;
- `/* Perform EMA on the gradients. */`
- $\hat{g}_t^k = \beta \hat{g}_t^{k-1} + (1-\beta) g_t^k$;
- $\hat{g}_m^k = \beta \hat{g}_m^{k-1} + (1-\beta) g_m^k$;
- `/* Norm the gradients and update the shared parameters. */`
- $\bar{g}^k(\theta) = \alpha_k(\sum_{t=1}^{T} \frac{\hat{g}_t^k}{||\hat{g}_t^k||_2 + \epsilon} + \frac{\hat{g}_m^k}{||\hat{g}_m^k||_2 + \epsilon})$, where $\alpha_k = max||\hat{g}_t^k||_2$;
- $\theta \leftarrow \nabla_\theta \bar{g}^k(\theta)$;
- `/* Update the task head parameters. */`
- $\psi_t \leftarrow \nabla_{\psi_t} \hat{g}_t^k(\psi_t)$, $\psi_m \leftarrow \nabla_{\psi_m} \hat{g}_m^k(\psi_m)$.

end while

4. Experiments and results

4.1. Dataset and experimental settings

In this work, we use the PTB-XL and Georgia Arrhythmia (G12EC) Databases to provide disease labels. We also use the MIT-BIH Noise Stress Test Database (NSTDB) to add noise with different signal-to-noise ratios (SNRs) to clean signals to test the performance of the proposed method.

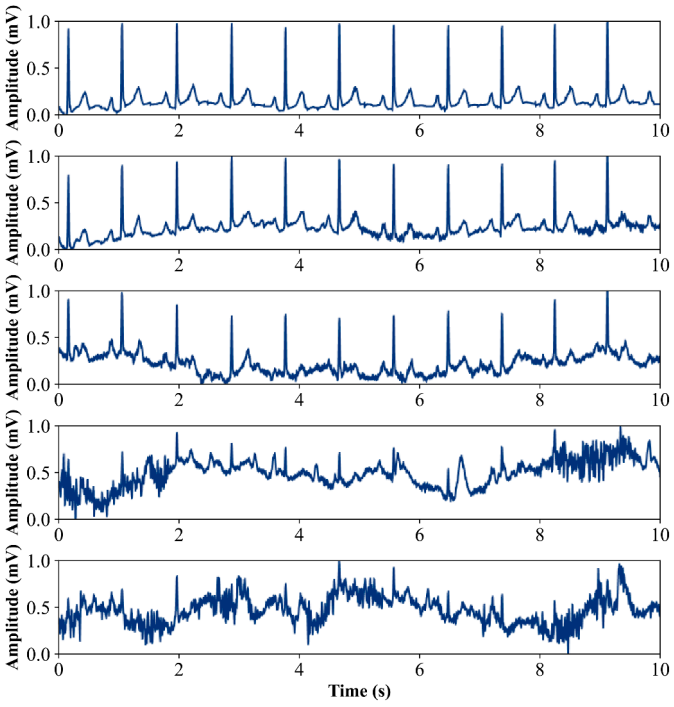


Fig. 4. Exemplary ECG signals under different signal-to-noise ratio (SNR) levels with BW, EM, and MA interference. From top to bottom: clean ECG, ECG with SNR in the range of 0~12 dB, -6~6 dB, -12~0 dB, and -18~-6 dB.

PTB-XL: The database comprises 21,799 ten-second recordings of 12-lead ECG signals, obtained from 18,869 patients diagnosed with arrhythmia in clinical practice. Each record was annotated with up to 71 descriptors related to rhythm patterns, waveform morphology, and diagnostic outcomes, with the annotations provided by at least two experienced cardiologists. For classification purposes, the dataset can be organized in different ways, including a division into five principal diagnostic categories–NORM, CD, MI, HYP, and STTC–according to established clinical standards. After discarding entries that lack a primary category label, the collection yields 21,388 ECG samples available for model training and evaluation.

G12EC: This dataset was derived from 10,344 patients diagnosed with arrhythmia in the southeastern region of the United States, with the gender distribution maintained in balance. Each sample consists of a 10-s recording from a standard 12-lead ECG. In total, the collection encompasses 27 distinct arrhythmia conditions. After discarding cases annotated with only a single label, 9407 multi-label ECG records remained and were employed for model development and evaluation.

NSTDB: The database contains 12 half-hour ECG recordings and three half-hour recordings of typical ambulatory noise, including baseline wander (bw), muscle artifact (ma), and electrode motion artifact (em). The noise signals were collected from volunteers using standard ECG equipment, and electrode motion artifact is considered the most challenging type, as it may mimic ectopic beats and is difficult to eliminate with simple filtering methods. In this study, we employed these three noise records to contaminate clean ECG signals for experimental purposes.

Mixed Dataset: To investigate the impact of different noise types, we selected electrode motion (EM) noise–the most difficult to suppress and most influential artifact–as the baseline noise and randomly sampled from the NSTDB to construct four mixed-noise types: BW + EM + MA, BW + EM, EM + MA, and EM. In each combination, the constituent noises were superimposed with equal power (1:1). To further assess the influence of noise intensity on arrhythmia detection performance, these four mixed-noise types were added to signals from both the PTB-XL and G12EC databases with signal-to-noise ratios (SNRs) sampled uniformly within the ranges of -18 to -6 dB, -12 to 0 dB, -6 to 6 dB, and 0 to 6 dB. As a result, a total of 16 noise-mixed datasets were constructed based on the PTB-XL and G12EC databases. We only use lead II in the dataset for training and testing to simulate a single-lead ECG environment. Fig. 4 shows examples of ECG signal contamination at different signal-to-noise ratios. All ECG signals in the noise mixed datasets are of quality level B.

1) Training setting: In this study, the hyperparameter was set to $\beta = 0.9$, and the Adam optimizer was employed to update the model parameters. The initial learning rate was set to 0.01 and subsequently adjusted using step decay with a decay factor of 0.1. The batch size was set to 32 in all experiments. The HSD queue length is set to 2048 samples, the representation dimension within the queue to 128, and the distillation temperatures for the student and teacher representations to 0.2 and 0.001, respectively. For the evaluation on each mixed dataset, the samples were randomly shuffled, with 80% allocated for training and the remaining 20% reserved for testing.

2) Evaluation metrics: Since the databases are multi-label annotated, the AUC score was employed in this study to evaluate model performance. It is defined as the area under the receiver operating characteristic (ROC) curve and reflects the capability of the classifier to discriminate between positive and negative samples across different decision thresholds.

The hypersphere representation was used to illustrate the alignment of class distributions. Furthermore, four metrics, namely Alignment (ALI), Intra-class compactness (INC), Inter-class separation (IES), and Davies-Bouldin Index (DBI), are used to quantify the alignment, intra-class compactness, inter-class separability, and class independence of distributions under different training strategies. The detailed meanings of the indicators are as follows:

- ALI: ALI measures the degree to which samples within each class are oriented consistently toward their class prototype. For each class, we compute the normalized class center as the mean of its sample embeddings, and then calculate the average cosine similarity between each sample and its corresponding class center. A higher ALI value indicates better intra-class alignment and more coherent directional clustering.
- INC: INC quantifies how tightly samples from the same class are clustered on the hypersphere. It is computed as the average pairwise cosine distance (or angular distance) among samples within each class. Lower INC values reflect stronger intra-class aggregation, which typically benefits classification robustness under noisy conditions.
- IES: IES evaluates the discriminability between different classes by measuring the average cosine distance between class centers. A larger IES value indicates that distinct classes are well separated on the hypersphere, reducing the risk of confusion among arrhythmia types.
- DBI: DBI is a widely used clustering validity index that jointly considers intra-class compactness and inter-class separation. For each class, we identify the most similar other class based on the ratio of the sum of their intra-class scatter to their inter-class distance. DBI is then averaged over all classes. A lower DBI value signifies a more favorable trade-off between within-class cohesion and between-class separability.

Table 1
Experimental configurations for different ablation studies.

Ablation	BCE	Denoise	HSD	Gradient surgery	Task balance
M1	✓	✗	✗	✗	✗
M2	✓	✓	✗	✗	✗
M3	✓	✗	✓	✗	✗
M4	✓	✓	✓	✗	✗
M5	✓	✓	✗	✓	✓
M6	✓	✗	✓	✓	✓
M7	✓	✓	✓	✗	✓
M8	✓	✓	✓	✓	✓

4.2. Ablation studies

To verify the effectiveness of the proposed auxiliary tasks and multi-task balancing strategy, we employed three different CNN backbones–ResNet34, MobileNet V2, and ShuffleNet V2–within the MTL framework. Each model was trained and evaluated on the mixed-noise versions of PTB-XL and G12EC databases. The ablation experiments considered the eight experimental setups shown in Table 1.

Table 2 presents the AUC scores of ResNet34 under different noise types and signal-to-noise ratio (SNR) levels. As the SNR decreases, the data distribution gradually approaches a Gaussian form, while pseudo-artifacts contaminate the signals and generate noisy labels, leading to compromised model performance. For example, under weak noise conditions (SNR 0~12 dB) with the M1 strategy, the AUC scores on PTB-XL across the four mixed-noise datasets were 0.826, 0.832, 0.822, and 0.822, while the corresponding scores on G12EC were 0.809, 0.815, 0.807, and 0.803. When the SNR was further reduced to -18~-6 dB, the AUC scores dropped to 0.625, 0.612, 0.601, and 0.610 on PTB-XL, and to 0.693, 0.735, 0.529, and 0.692 on G12EC. Similar degradation trends were observed for MobileNet V2 and ShuffleNet V2. The corresponding macro-AUPRC metric results are shown in Tables A.1–A.3.

When applying the equal-weighted denoising-assisted multi-task strategy (M2), the models exhibited performance degradation across almost all noisy datasets of PTB-XL and G12EC. This observation is consis-

Table 2
AUC scores of **ResNet34** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at varying SNR levels.

Dataset		PTB-XL							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.625	0.607	0.640	0.660	0.657	0.619	0.670	**0.701**
	−12∼0	0.741	0.728	0.731	0.758	0.728	0.748	0.771	**0.787**
	−6∼6	0.801	0.795	0.805	0.803	0.795	0.803	0.814	**0.822**
	0∼12	0.826	0.825	0.823	0.828	0.834	0.825	0.833	**0.838**
BW + EM	−18∼−6	0.612	0.673	0.708	0.701	0.718	0.692	0.713	**0.742**
	−12∼0	0.754	0.743	0.768	0.760	0.779	0.760	0.777	**0.800**
	−6∼6	0.808	0.801	0.814	0.812	0.814	0.814	0.818	**0.827**
	0∼12	0.832	0.824	0.828	0.832	0.837	0.833	0.839	**0.841**
EM + MA	−18∼−6	0.601	0.600	0.602	0.626	0.604	0.604	0.628	**0.672**
	−12∼0	0.727	0.697	0.726	0.728	0.759	0.730	0.750	**0.778**
	−6∼6	0.789	0.780	0.788	0.788	0.800	0.793	0.807	**0.813**
	0∼12	0.822	0.818	0.821	0.826	0.830	0.826	0.828	**0.837**
EM	−18∼−6	0.610	0.640	0.619	0.672	0.677	0.654	0.674	**0.717**
	−12∼0	0.749	0.743	0.751	0.755	0.761	0.731	0.761	**0.784**
	−6∼6	0.798	0.787	0.796	0.803	0.808	0.807	0.806	**0.820**
	0∼12	0.822	0.821	0.825	0.825	0.830	0.827	0.834	**0.838**
Dataset		G12EC							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.693	0.680	0.704	0.738	**0.761**	0.645	0.694	0.753
	−12∼0	0.762	0.737	0.798	0.801	0.809	0.783	0.803	**0.829**
	−6∼6	0.783	0.770	0.818	0.818	0.823	0.822	0.827	**0.852**
	0∼12	0.809	0.797	0.837	0.830	0.840	0.828	0.830	**0.865**
BW + EM	−18∼−6	0.735	0.675	0.779	0.731	0.776	0.755	0.756	**0.787**
	−12∼0	0.770	0.705	0.798	0.787	0.812	0.796	0.810	**0.833**
	−6∼6	0.793	0.771	0.825	0.824	0.826	0.827	0.826	**0.854**
	0∼12	0.815	0.786	0.824	0.835	0.837	0.821	0.828	**0.858**
EM + MA	−18∼−6	0.529	0.520	0.704	0.585	0.746	0.694	0.681	**0.765**
	−12∼0	0.747	0.735	0.792	0.786	0.802	0.779	0.778	**0.814**
	−6∼6	0.776	0.764	0.799	0.818	0.817	0.822	0.812	**0.849**
	0∼12	0.807	0.772	0.839	0.835	0.833	0.826	0.813	**0.859**
EM	−18∼−6	0.692	0.700	0.728	0.734	0.769	0.712	0.738	**0.801**
	−12∼0	0.754	0.711	0.809	0.782	0.814	0.802	0.800	**0.831**
	−6∼6	0.776	0.782	0.801	0.815	0.824	0.815	0.823	**0.849**
	0∼12	0.803	0.785	0.837	0.837	0.844	0.838	0.819	**0.858**

The results of macro-AUPRC and per-class AUPRC are shown in Appendices A.1 and B.1, respectively.

tent with Stein's paradox, which states that in high-dimensional parameter estimation, forcing shrinkage toward a shared global center leads to suboptimal solutions for individual tasks when their optimal centers differ. When using the M3 strategy, it can be observed that the model accuracy is improved to some extent compared to the M1 strategy under different signal-to-noise ratios and backbone networks. This improvement indicates that although a multi-task balancing strategy is not adopted, HSD can still effectively constrain the features of noisy signals in the representation space to the clean signal representation.

By incorporating adaptive task balancing and gradient surgery, the M5 strategy ensured that the losses and gradients of denoising and detection tasks were maintained on comparable scales. This prevented the large loss magnitude of the denoising task from dominating optimization and displacing the detection task from its optimal center. Moreover, gradient surgery enforced detection as the primary optimization target, thereby alleviating task conflicts effectively.

Further integrating hyperspherical self-distillation, the M8 strategy achieved the best performance across nearly all noisy datasets. By leveraging clean-signal representations as anchors, noisy samples were forced to align with the neighborhood structure of clean signals, effectively guiding the overall distribution toward the clean-signal manifold. Combined with the fine-grained features provided by the denoising auxiliary task, this enhanced the robustness of the model. Notably, the lower the SNR, the greater the performance gains of the proposed strategy. For instance, at SNR 0∼12 dB, the AUC improvements on PTB-XL were 0.012, 0.009, 0.015, and 0.016, while on G12EC they were 0.056, 0.043, 0.052,

Table 3
AUC scores of **ShuffleNet V2** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at varying SNR levels.

Dataset		PTB-XL							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.604	0.600	0.606	0.605	0.608	0.596	0.599	**0.626**
	−12∼0	0.671	0.667	0.674	0.678	0.688	0.633	0.643	**0.699**
	−6∼6	0.730	0.718	0.753	0.759	0.748	0.701	0.741	**0.765**
	0∼12	0.783	0.775	0.796	0.798	0.793	0.787	0.794	**0.804**
BW + EM	−18∼−6	0.611	0.600	0.634	0.614	0.611	0.630	0.610	**0.668**
	−12∼0	0.683	0.680	0.693	0.709	0.714	0.665	0.689	**0.735**
	−6∼6	0.756	0.745	0.774	0.773	0.765	0.746	0.767	**0.785**
	0∼12	0.791	0.782	0.805	0.808	0.801	0.798	0.806	**0.813**
EM + MA	−18∼−6	0.601	0.599	0.601	0.602	0.592	0.601	0.610	**0.617**
	−12∼0	0.647	0.643	0.664	0.633	0.662	0.621	0.648	**0.690**
	−6∼6	0.720	0.707	0.735	0.740	0.708	0.706	0.726	**0.747**
	0∼12	0.777	0.767	0.785	0.794	0.628	0.780	0.787	**0.798**
EM	−18∼−6	0.629	0.608	0.627	0.619	0.650	0.616	0.638	**0.666**
	−12∼0	0.678	0.668	0.673	0.691	0.691	0.677	0.675	**0.716**
	−6∼6	0.750	0.720	0.755	0.746	0.744	0.742	0.752	**0.769**
	0∼12	0.784	0.777	0.791	0.794	0.791	0.790	0.801	**0.808**
Dataset		G12EC							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.587	0.503	0.645	0.575	0.665	0.614	0.625	**0.684**
	−12∼0	0.659	0.572	0.621	0.624	0.650	0.664	0.664	**0.741**
	−6∼6	0.729	0.588	0.712	0.717	0.778	0.724	0.722	**0.785**
	0∼12	0.750	0.604	0.677	0.745	0.791	0.742	0.750	**0.811**
BW + EM	−18∼−6	0.637	0.583	0.619	0.567	0.671	0.632	0.651	**0.689**
	−12∼0	0.686	0.585	0.650	0.653	0.674	0.683	0.676	**0.745**
	−6∼6	0.731	0.625	0.702	0.729	0.693	0.725	0.740	**0.772**
	0∼12	0.755	0.659	0.764	0.662	**0.788**	0.747	0.758	0.784
EM + MA	−18∼−6	0.508	0.511	0.573	0.547	**0.662**	0.596	0.580	0.658
	−12∼0	0.662	0.566	0.664	0.613	0.697	0.647	0.667	**0.729**
	−6∼6	0.716	0.594	0.664	0.651	0.755	0.708	0.736	**0.759**
	0∼12	0.755	0.608	0.674	0.714	**0.777**	0.747	0.750	0.774
EM	−18∼−6	0.630	0.566	0.623	0.586	0.622	0.500	0.493	**0.706**
	−12∼0	0.686	0.582	0.644	0.668	**0.746**	0.671	0.688	0.740
	−6∼6	0.722	0.598	0.682	0.681	0.721	0.712	0.714	**0.770**
	0∼12	0.747	0.614	0.766	0.724	0.775	0.751	0.765	**0.797**

The results of macro-AUPRC and per-class AUPRC are shown in Appendices A.2 and B.2, respectively.

and 0.055. At lower SNR levels (-18∼-6 dB), the gains on PTB-XL reached 0.076, 0.130, 0.071, and 0.107, and on G12EC they were 0.060, 0.052, 0.236, and 0.109. These results indicate that the proposed strategy provides stronger generalization ability under severe artifact contamination.

4.3. Comparison with other multi-task balancing strategies

To demonstrate that the proposed multi-task balancing strategy is suitable for multi-task learning scenarios driven by a primary task, we compared our proposed strategy with several commonly used balancing strategies, as shown in Table 5–7. For fairness, all strategies used denoising and hypersphere self-distillation as parallel tasks, and ResNet34, ShuffleNet V2, MobileNet v2 as the backbone models. Specifically, we compared the performance with MGDA (Sener & Koltun, 2018), CAG (Liu et al., 2021), IMTL (Liu et al., 2021), and UW (Cipolla et al., 2018) under various signal-to-noise ratios. All experiments were repeated using a 5-seed dataset split.

MGDA formulates multi-task learning as a multi-objective optimization problem and searches for a Pareto-stationary descent direction by jointly minimizing all task objectives. As shown in Table 5–7, MGDA achieves a modest performance improvement over direct training on the G12EC database. However, on the PTB-XL database, its performance is markedly inferior to that of the proposed method, particularly under low SNR conditions. This can be attributed to the fact that MGDA enforces a global trade-off among all tasks and assigns equal importance

Table 4
AUC scores of **MobileNet V2** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at varying SNR levels.

Dataset		PTB-XL							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.603	0.600	0.598	0.604	0.595	0.599	0.664	**0.679**
	−12∼0	0.738	0.615	0.756	0.760	0.749	0.742	0.750	**0.777**
	−6∼6	0.798	0.786	0.804	0.807	0.787	0.801	0.806	**0.812**
	0∼12	0.823	0.814	0.825	0.820	0.821	0.827	**0.827**	0.826
BW + EM	−18∼−6	0.609	0.600	0.613	0.716	0.699	0.670	0.697	**0.723**
	−12∼0	0.756	0.753	0.755	0.776	0.763	0.758	0.771	**0.789**
	−6∼6	0.807	0.798	0.807	0.813	0.797	0.805	0.811	**0.816**
	0∼12	0.829	0.822	0.826	0.830	0.822	0.826	0.830	**0.835**
EM + MA	−18∼−6	0.603	0.597	0.604	0.646	0.597	0.616	0.647	**0.659**
	−12∼0	0.619	0.608	0.731	0.752	0.739	0.723	0.738	**0.766**
	−6∼6	0.790	0.781	0.790	0.791	0.783	0.788	0.788	**0.801**
	0∼12	0.824	0.811	0.818	0.778	0.814	0.811	0.825	**0.825**
EM	−18∼−6	0.609	0.601	0.630	0.618	0.672	0.660	0.673	**0.675**
	−12∼0	0.748	0.749	0.757	0.750	0.748	0.744	0.739	**0.764**
	−6∼6	0.796	0.788	0.801	0.801	0.786	0.796	0.793	**0.810**
	0∼12	0.828	0.812	0.825	0.827	0.819	0.823	0.828	**0.832**
Dataset		G12EC							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.536	0.532	0.678	0.696	0.517	0.696	0.695	**0.727**
	−12∼0	0.774	0.603	0.759	0.735	0.779	0.756	0.767	**0.792**
	−6∼6	0.770	0.594	0.750	0.756	0.776	0.784	0.794	**0.809**
	0∼12	0.779	0.621	0.803	0.797	0.785	0.782	0.795	**0.806**
BW + EM	−18∼−6	0.674	0.598	0.726	0.683	0.743	0.729	0.706	**0.759**
	−12∼0	0.690	0.596	0.773	0.728	0.782	0.772	0.761	**0.799**
	−6∼6	0.767	0.600	0.791	0.788	0.772	0.778	0.786	**0.810**
	0∼12	0.784	0.620	0.798	0.805	0.781	0.803	0.784	**0.823**
EM + MA	−18∼−6	0.508	0.513	0.654	0.615	0.495	0.528	0.511	**0.741**
	−12∼0	0.688	0.592	0.766	0.743	0.763	0.741	0.746	**0.791**
	−6∼6	0.752	0.596	0.787	0.750	0.767	0.780	0.784	**0.823**
	0∼12	0.782	0.613	0.780	0.772	0.791	0.542	0.792	**0.821**
EM	−18∼−6	0.671	0.589	0.726	0.631	0.736	0.645	0.723	**0.768**
	−12∼0	0.695	0.593	0.760	0.725	0.781	0.755	0.751	**0.799**
	−6∼6	0.777	0.598	0.761	0.743	0.783	0.778	0.767	**0.811**
	0∼12	0.799	0.621	0.800	0.793	0.785	0.787	0.792	**0.823**

The results of macro-AUPRC and per-class AUPRC are shown in Appendices A.3 and B.3, respectively.

to auxiliary tasks and the main classification task. In the proposed hierarchical architecture, the losses associated with denoising and hypersphere self-distillation may dominate the optimization process. Consequently, a Pareto-optimal solution biased toward auxiliary objectives may drive the shared representation away from the optimum for arrhythmia classification, leading to degraded classification performance under severe noise. CAG seeks to alleviate gradient conflicts by solving a constrained optimization problem that balances task gradients while minimizing their mutual coupling. Although CAG has demonstrated effectiveness in conventional parallel multi-task learning scenarios, the results in Table 5–7 indicate relatively poor performance on both the PTB-XL and G12EC datasets, especially at low SNR levels. This limitation is primarily due to its symmetric treatment of all tasks, where gradient conflicts are resolved without explicitly considering task priority. As a result, when auxiliary task gradients are poorly aligned in strong noise environments, they may still exert unfavorable influence on the optimization direction of the main arrhythmia detection task. IMTL introduces fairness constraints to promote balanced optimization across tasks by equalizing gradient magnitudes and preventing any single task from being suppressed. As shown in the Tables, IMTL consistently underperforms the proposed method across all SNR ranges. This outcome suggests that enforcing task fairness is suboptimal in the considered setting. In the proposed framework, denoising and hypersphere self-distillation are designed solely as auxiliary mechanisms to support classification. Assigning them equal optimization priority may impede the convergence of the main task, particularly when auxiliary

Table 5
AUC scores of **ResNet34** under various multi-task balancing strategies on PTB-XL and G12EC datasets with BW + EM + MA noise at varying SNR levels.

Method	PTB-XL				G12EC			
	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB
MGDA (Sener & Koltun, 2018)	0.599±0.005	0.729±0.018	0.757±0.006	0.782±0.008	0.754±0.026	0.806±0.014	0.842±0.007	0.838±0.008
CAG (Liu et al., 2021)	0.599±0.002	0.721±0.012	0.787±0.005	0.811±0.008	0.523±0.004	0.608±0.057	0.614±0.036	0.635±0.032
IMTL (Liu et al., 2021)	0.616±0.022	0.701±0.016	0.775±0.007	0.804±0.002	0.612±0.004	0.698±0.007	0.749±0.011	0.773±0.011
UW (Cipolla et al., 2018)	0.661±0.004	0.749±0.005	0.800±0.004	0.826±0.004	0.720±0.010	0.797±0.008	0.823±0.003	0.834±0.009
This work	**0.687±0.011***	**0.764±0.013***	**0.802±0.011**	**0.835±0.007***	**0.763±0.010***	**0.818±0.007***	**0.849±0.002***	**0.861±0.002***

* indicates statistical significance ($p < 0.05$) compared with all baseline methods.

Table 6
AUC scores of **ShuffleNet V2** under various multi-task balancing strategies on PTB-XL and G12EC datasets with BW + EM + MA noise at varying SNR levels.

Method	PTB-XL				G12EC			
	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB
MGDA (Sener & Koltun, 2018)	0.602±0.002	0.665±0.007	0.728±0.008	0.760±0.009	0.566±0.044	0.673±0.033	0.685±0.055	0.631±0.006
CAG (Liu et al., 2021)	0.594±0.001	0.603±0.004	0.613±0.006	0.627±0.008	0.502±0.016	0.542±0.022	0.543±0.035	0.570±0.036
IMTL (Liu et al., 2021)	0.603±0.002	0.650±0.012	0.742±0.005	0.783±0.009	0.557±0.006	0.631±0.018	0.675±0.025	0.742±0.008
UW (Cipolla et al., 2018)	0.608±0.002	0.672±0.011	0.736±0.005	0.787±0.005	0.621±0.006	0.704±0.024	0.763±0.010	0.778±0.013
This work	**0.628±0.003***	**0.702±0.001***	**0.762±0.002***	**0.804±0.001***	**0.640±0.004***	**0.716±0.003***	**0.774±0.002***	**0.798±0.003***

* indicates statistical significance ($p < 0.05$) compared with all baseline methods.

Table 7
AUC scores of **MobileNet V2** under various multi-task balancing strategies on PTB-XL and G12EC datasets with BW + EM + MA noise at varying SNR levels.

Method	PTB-XL				G12EC			
	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB
MGDA (Sener & Koltun, 2018)	0.600±0.003	0.692±0.009	0.722±0.005	0.742±0.022	0.547±0.027	0.773±0.014	0.807±0.015	0.825±0.010
CAG (Liu et al., 2021)	0.591±0.002	0.598±0.001	0.648±0.012	0.668±0.013	0.515±0.005	0.531±0.017	0.574±0.015	0.585±0.020
IMTL (Liu et al., 2021)	0.635±0.013	0.736±0.010	0.789±0.003	0.820±0.005	0.536±0.006	0.556±0.004	0.573±0.014	0.660±0.015
UW (Cipolla et al., 2018)	0.653±0.013	0.737±0.009	0.793±0.005	0.818±0.005	0.702±0.016	0.769±0.008	0.791±0.004	0.805±0.006
This work	**0.705±0.003***	**0.772±0.001***	**0.813±0.001***	**0.827±0.001***	**0.740±0.005***	**0.799±0.005***	**0.822±0.006***	**0.831±0.011**

* indicates statistical significance ($p < 0.05$) compared with all baseline methods.

objectives exhibit weak relevance or partial conflict with arrhythmia detection. UW dynamically adjusts task weights based on homoscedastic uncertainty, thereby implicitly normalizing losses of different scales. As reported in the Tables, UW outperforms gradient-based balancing approaches, especially at moderate SNR levels, indicating its effectiveness in mitigating loss-scale imbalance. Nevertheless, its performance remains inferior to that of the proposed method under severe noise conditions. This is mainly because uncertainty-based weighting assumes that loss magnitude reflects task importance, an assumption that does not hold in the present scenario. Under strong noise, the denoising and hypersphere self-distillation losses tend to exhibit increased variance, leading to higher assigned weights for auxiliary tasks and, consequently, reduced optimization emphasis on the main classification objective.

In contrast, the proposed method consistently achieves the best performance across all SNR levels and both datasets. Arrhythmia detection is explicitly treated as the primary task, while auxiliary tasks are not required to reach their individual optima but are constrained to provide stable and constructive support to the main task. By integrating logarithmic loss scaling, gradient normalization, and main-task-oriented gradient surgery, the proposed approach ensures that auxiliary gradients contribute positively without interfering with the optimization direction of the classification objective. This design enables more effective exploitation of auxiliary information under severe noise conditions, resulting in enhanced robustness and superior overall detection performance.

4.4. Comparison with other denoising methods

We compare our proposed method with other similar arrhythmia detection methods under noise conditions to further demonstrate its superiority. Specifically, Mallikarjunamallu and Khasim (2024) employ a two-stage processing approach. First, they use a notch filter to denoise the entire dataset, then use the denoised data for model training and testing. Khatar et al. (2024) follows the same paradigm as Mallikarjunamallu and Khasim (2024), also using a two-stage processing approach and filter denoising. However, they replace the notch filter with a fusion filter from Chebyshev, Butterworth, and Daubechies. Wu et al. (2024) utilizes a transfer learning paradigm, using a fully convolutional autoencoder to train the model for denoising noise signals. Then, they discard the decoder part and use the prior knowledge of the pre-trained encoder to fine-tune the noisy ECG/arrhythmia classification head. Park et al. (2024) also employs a two-stage processing approach, using a generative adversarial network (GAN) to align the data distribution of noisy ECGs with their corresponding clean ECGs, supplemented by a denoising reconstruction loss, to achieve the goal of using the generator as the denoising model. This study utilizes a denoising model to denoise the entire dataset before using the clean data for training and testing the arrhythmia monitoring model.

To ensure a fair and controlled comparison, all methods were evaluated using a uniform experimental protocol. Specifically, we employed the same training/validation set partitioning, the same noise injection

Table 8
Comparison with various arrhythmia detection methods under noise conditions tested on the PTB-XL and G12EC BW + EM + MA mixed noise datasets at various SNR ratios.

Method	PTB-XL				G12EC			
	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB	−18∼−6 dB	−12∼0 dB	−6∼6 dB	0∼12 dB
Mallikarjunamallu and Khasim (2024)	0.588	0.591	0.788	0.812	0.712	0.776	0.798	0.822
Khatar et al. (2024)	0.637	0.740	0.794	0.822	0.698	0.762	0.808	0.818
Wu et al. (2024)	0.690	0.764	0.802	0.826	0.731	0.767	0.782	0.818
Park et al. (2024)	0.599	0.651	0.727	0.778	0.688	0.736	0.772	0.798
This work	**0.701**	**0.787**	**0.822**	**0.838**	**0.753**	**0.829**	**0.852**	**0.865**

strategy, and a consistent preprocessing procedure for all comparison methods. The arrhythmia detection backbone model was ResNet34 in all experiments, and all other hyperparameter configurations remained consistent.

It can be seen that the proposed multi-task learning strategy achieves the best performance across all signal-to-noise ratio (SNR) ranges. Different arrhythmia detection methods show similar performance at high SNR ($0 \sim 12$ dB), while the proposed method has a more significant advantage at low SNR ($-18 \sim -6$ dB). Compared to strategies that use filters or pre-trained denoising models followed by model optimization (Khatar et al., 2024; Mallikarjunamallu & Khasim, 2024; Park et al., 2024), the proposed method improves arrhythmia detection accuracy by 0.026, 0.016, and 0.06 on the PTB-XL database at high SNR, and by 0.113, 0.064, and 0.102 at low SNR. The trend is consistent on the G12EC database. This demonstrates that the proposed multi-task learning strategy can better handle the impact of strong noise without relying on a separate, strong denoising model. While improving arrhythmia detection accuracy through representation transfer from the denoising encoder is also effective, it ignores the fact that the optimized distribution of the reconstructed denoised signal does not conform to the class separability of arrhythmia detection. The proposed method, by explicitly constraining the representation space of the noisy signal, achieves a more discriminative and interpretable representation without pursuing perfect denoising. This shows that the multi-task learning strategy for arrhythmia detection assisted by denoising and hypersphere self-distillation can effectively constrain the bias subspace and better handle the impact of artifacts at the representation level.

4.5. Representation distribution analysis

The unit circle distribution visualization provides an intuitive means to examine the structural properties of feature representations, including the uniformity and alignment of different class distributions. We employ the PTB-XL database under the BW + EM + MA mixed-noise condition with a signal-to-noise ratio (SNR) of -18∼-6, and project the features extracted by ResNet34 for different categories onto the unit circle, as shown in Fig. 5. By comparing the representations obtained using the proposed strategy with those from random initialization, it can be observed that the proposed method yields more uniformly distributed features across the hypersphere without obvious collapse. Moreover, compared with the random baseline, the proposed approach produces class-specific distributions that are more compact, forming arc-shaped clusters, while maintaining fixed and non-overlapping positions across classes. This demonstrates improved intra-class compactness and inter-class separability, which directly contributes to enhanced robustness and classification performance under severe noise conditions.

Furthermore, we quantified the alignment, intra-class compactness, inter-class separability, and the intra-/inter-class distribution balance using ALI, INC, IES, and DBI, respectively, for the random baseline and the M1-M4 strategies, as summarized in Table 9. As can be observed, the proposed method consistently achieves the best performance on the PTB-XL and G12EC mixed-noise datasets across almost all metrics, with particularly notable improvements in DBI and IES. These improvements indicate that the proposed strategy not only enhances the discriminative power of the learned representations but also yields a more robust and well-structured feature space, which directly contributes to stable arrhythmia detection under severe noise conditions.

4.6. Case study

Model interpretability plays a critical role in medical diagnosis. Effective explanation techniques not only assist clinicians in understanding the decision-making process of models but also help identify and correct potential biases and errors. Among them, Relevance-CAM (Lee et al., 2021) has been widely adopted as a model explanation method. In contrast to gradient-based approaches such as Grad-CAM, Relevance-CAM relies on Layer-wise Relevance Propagation (LRP) to propagate relevance scores back through the network, thereby quantifying the contribution of each neuron to the prediction. This approach mitigates the instability caused by gradient vanishing or explosion and provides a more causally consistent interpretation.

In this work, Relevance-CAM was employed to visualize the decision basis of four models on the same CD record, as illustrated in Fig. 6: (a) the directly trained model, (b) the multi-task model with denoising assistance, (c) the adaptive balanced multi-task model with denoising assistance, and (d) the adaptive balanced multi-task model with denoising and hyperspherical self-distillation assistance. The selected signal was corrupted with BW + EM + MA artifacts under an SNR level ranging from -18 dB∼-6 dB. For clarity, the relevance maps were overlaid on the corresponding clean ECG signal. The ECG record indicates a significant CD abnormality at the fifth heartbeat. Fig. 6(a) shows that the model's attention was almost entirely concentrated on the fourth heartbeat, while Fig. 6(b) still focused on the fourth heartbeat but distributed partial attention to other positions. In Fig. 6(c), the model assigned higher relevance to the fifth heartbeat, suggesting that the multi-task balancing strategy enabled the denoising task to guide the model toward learning fine-grained representations of clean signals; however, the relevant regions remained relatively dispersed. Fig. 6(d) demonstrates that the model allocated the highest relevance to the most prominent diseased heartbeat, while relatively ignoring weaker regions. These results indicate that the proposed auxiliary multi-task strategy, which integrates fine-grained representation learning and distribution-constrained alignment, facilitates multi-level knowledge representation.

5. Discussion

5.1. Performance analysis

5.1.1. Ablation study

The experimental results demonstrate that the proposed multi-task auxiliary learning framework significantly enhances the robustness of wearable ECG arrhythmia detection under noisy conditions. By incorporating denoising and hyperspherical self-distillation as auxiliary tasks, the model is able to capture fine-grained components of clean signals

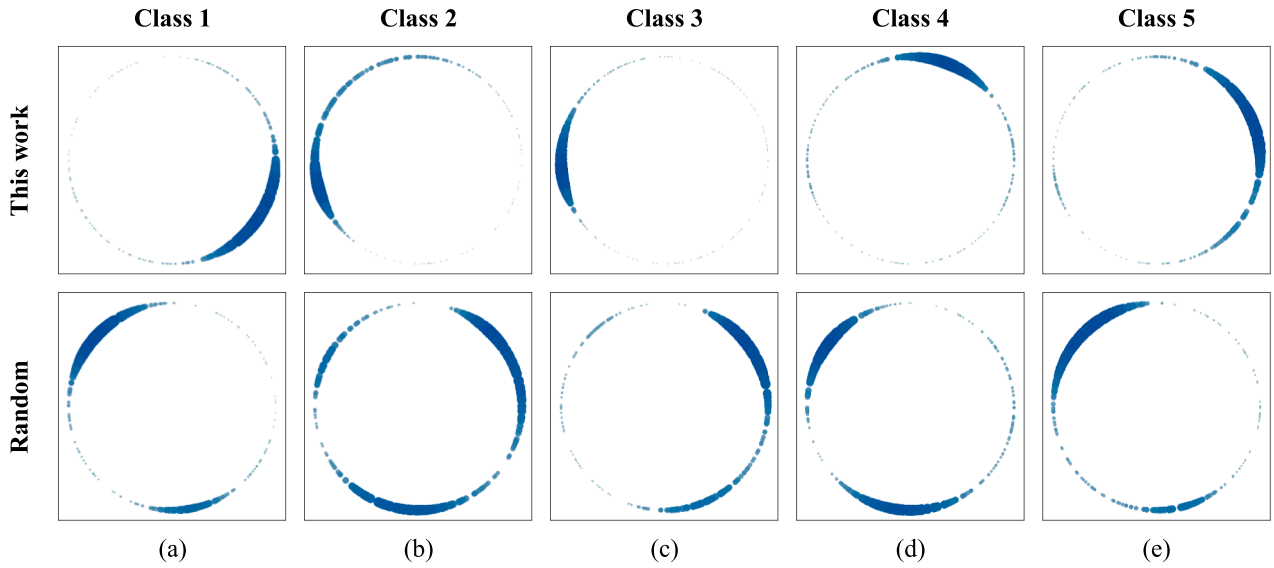


Fig. 5. Distribution of various representations on the unit hypersphere using the proposed method and random initialization on PTB-XL noise dataset($SNR: -18 \sim -6$). (a) to (e) represent categories 1 to 5 respectively.

Table 9
Representation distribution indicators of different methods on the PTB-XL and G12EC noise mixture datasets ($SNR = -18 \sim -6$).

Dataset	PTB-XL								
Method	M0	M1	M2	M3	M4	M5	M6	M7	M8
ALI↓	13.5922	6.9272	4.2721	7.2671	5.6531	1.7463	2.3181	2.3032	**1.4863**
INC↓	196.9020	39.8195	14.6042	37.8540	21.0405	1.9314	3.5064	3.2633	**1.2966**
IES↑	0.0401	0.0154	0.0384	0.1671	0.0993	0.0881	0.2325	0.2112	**0.2729**
DBI↓	20.7841	17.4753	13.2197	9.4103	**9.0378**	10.5490	11.4170	9.8315	9.5465
Dataset	G12EC								
Method	M0	M1	M2	M3	M4	M5	M6	M7	M8
ALI↓	10.7746	8.1511	6.1519	21.3153	18.1744	**2.4799**	4.0072	2.6246	3.2284
INC↓	102.7190	37.9024	25.8271	282.0689	184.2199	**3.3247**	9.8800	3.7657	5.5371
IES↑	0.0587	0.0642	0.0709	0.4357	0.4967	0.2502	0.3794	0.5716	**0.6139**
DBI↓	40.0595	10.3050	14.8001	9.3494	7.5374	7.3880	9.9710	7.5527	**6.5020**

M0: Random initialization.

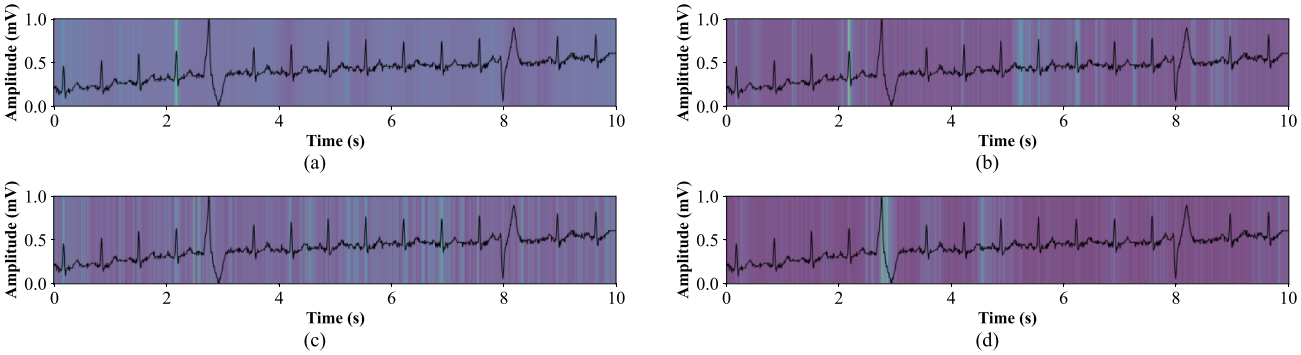


Fig. 6. Relevance-CAM visualization analysis of the ResNet-34 model on a CD recording based on four methods: (a) direct training, (b) a multi-task strategy with denoising assistance, (c) an adaptive balanced multi-task strategy with denoising assistance, and (d) an adaptive balanced multi-task strategy with denoising and hypersphere self-distillation assistance.

while simultaneously constraining the neighborhood relationships of noisy signals in the hyperspherical space to align with the clean signal distribution. This dual mechanism enables the extraction of more discriminative and noise-robust features, and shows strong adaptability across different noise types and SNR levels, as presented in Tables 2–4. Notably, the performance improvement gained from introducing hypersphere self-distillation in arrhythmia detection tasks is typically significantly higher than that obtained by relying solely on signal denoising. This result indicates that for classification tasks, the performance degradation caused by noise stems not only from the destruction of signal morphology but, more importantly, from its perturbation of the discriminative feature space structure. Hypersphere self-distillation, by

imposing constraints on the angular relationships of feature vectors, explicitly promotes the uniformity of the representation space and inter-class separability, effectively suppressing the bias introduced by noise at the representation level. This allows the model to learn robust and discriminative feature representations even under imperfect denoising or complex and variable noise conditions. Furthermore, the framework is applicable to various backbone architectures, allowing flexibility for different detection tasks and levels of difficulty. Notably, the performance gain becomes more significant at low SNR conditions, highlighting the framework's potential to maintain diagnostic reliability in complex real-world scenarios.

5.1.2. Diagnostic-level performance analysis

To further analyze diagnostic performance, we report each class of AUPRC under severe noise conditions (SNR = -18∼-6), as shown in Tables B.1–B.3 of Appendix B. Detailed explanations of disease type abbreviations can be found in Tables B.4. From the classification results, it can be observed that the performance improvement was not limited to specific categories, but rather evenly distributed across multiple arrhythmia types. Notably, on the PTB-XL dataset, the model showed more significant performance improvements in challenging and lower-performing categories (e.g., CD and HYP in PTB-XL) compared to its dominant categories. Simultaneously, the model also achieved varying degrees of performance improvement across different categories on the G12EC dataset. However, as the number of categories increases, on the G12EC dataset, the model's improvement in dominant categories is more significant than in weaker categories, raising concerns about data imbalance–that is, the model's ability to represent rare diseases or diseases with limited data is weaker.

It should also be noted that the category with the greatest performance improvement was not entirely consistent across different experimental settings. This is expected in noisy ECG classification, as different types of arrhythmias have varying sensitivities to noise patterns and data distribution.

From a diagnostic perspective, these improvements suggest that the proposed method enhances the model's ability to capture subtle and noise-sensitive ECG patterns. In particular, arrhythmia categories characterized by weak morphological deviations or irregular rhythm patterns tend to be more vulnerable to noise contamination. The observed improvements indicate that the proposed multi-task framework helps preserve clinically relevant waveform structures and discriminative temporal dependencies under severe noise conditions.

5.1.3. Comparison of task balancing strategies

Compared to equally weighted strategies that emphasize simplicity and efficiency, and most parallel MTL strategies including MGDA, CAG, IMTL, and UW, the proposed method more rationally considers the optimization priorities of different tasks. Traditional MTL strategies aim to achieve a balance or Pareto optimal solution among all objectives. Although they balance the scale of losses and gradients and suppress task conflicts, they are not suitable for the problem addressed in this paper, where the introduction of auxiliary tasks is solely to assist in optimizing the main arrhythmia classification task. In noisy ECG signals, auxiliary tasks such as signal denoising and representation space constraints often produce larger loss values and gradients, especially at low SNR levels. These gradients can dominate or distort the shared representation, leading to negative transfer. The proposed method addresses this problem by anchoring the optimization process to the main task and adjusting the auxiliary gradients based on their directional consistency with the classification objective. Therefore, auxiliary tasks are only allowed to play a role when they provide effective guidance, which explains the significant performance improvement observed in experiments under severely noisy conditions. Furthermore, Relevance-CAM-based visualization further demonstrates that, in noisy environments, the model gradually focuses on clinically relevant abnormal heartbeats rather than spurious artifacts.

5.1.4. Comparison with noise-robust methods

Compared to other methods for arrhythmia detection under noisy conditions, the proposed strategy has significant advantages, as shown in Table 8. It should be noted that the methods in Table 8 all rely on pre-trained encoders or denoising models; that is, the denoising model is trained independently first and then used as a fixed preprocessing step for classification; or the autoencoder is trained first for signal reconstruction and then transferred to the classification task. These methods implicitly assume perfect or near-perfect denoising performance and ignore the potential distribution mismatch between training and inference under unknown noise conditions. When the signal-to-noise ratio is too low, the traditional filters of Mallikarjunamallu and Khasim (2024) and Khatar et al. (2024) cannot effectively remove noise and may even discard some effective physiological information, causing such methods to almost fail. Park et al. (2024) aligns the distribution of noisy signals with that of clean signals through adversarial game theory, but due to the instability of the GAN optimization process and the higher complexity of ECG artifact noise compared to Gaussian noise, it is difficult to achieve better results. Furthermore, optimizing the denoiser or autoencoder only for signal reconstruction may not necessarily meet the goal of robust arrhythmia classification. In contrast, the method proposed in this paper differs from the above two-stage strategies. We integrate denoising and arrhythmia classification into a unified multi-task learning framework and further introduce hyperspherical self-distillation to explicitly normalize the representation space. This design enables the model to learn discriminative and noise-invariant representations even with imperfect denoising, thereby improving the model's robustness to noisy ECG signals. Our experimental results further demonstrate that treating denoising as a plug-and-play preprocessing module may lead to overly optimistic estimates of classification performance, especially when the denoiser's generalization ability to real-world noise is insufficient.

For wearable ECG applications, where artifacts such as motion noise and electromyographic interference are inevitable, the proposed framework provides a substantial improvement in robustness. By reducing false alarms and misdiagnoses, the approach directly enhances the usability of wearable devices and increases clinicians' confidence in decision-making. Therefore, the proposed framework offers a promising pathway toward more reliable arrhythmia screening in everyday scenarios.

5.2. Limitation and prospects

Despite the progress achieved in this study, several limitations remain. The experiments were conducted using arrhythmia datasets PTB-XL and G12EC in combination with synthetic noisy signals from the NSTDB database. However, real-world clinical environments may present more diverse ECG morphologies and more complex noise patterns, and variations in acquisition devices could further degrade model performance. These aspects were not comprehensively validated in this work. In addition, as the number of auxiliary tasks increases, the computational demand may hinder the deployment of the proposed framework on resource-constrained wearable devices, limiting its practical applicability. Class imbalance is also a problem that cannot be ignored. Severe class imbalance can cause the model's high accuracy to be dominated by the dominant class, while neglecting the weaker classes, thus creating an illusion of high accuracy.

Future research will therefore focus on the lightweight design of multi-auxiliary-task models for arrhythmia detection, as well as the exploration of additional auxiliary tasks (e.g., heartbeat segmentation) to further enhance the robustness of learned representations. Meanwhile, improving the model's detection accuracy for weak-class arrhythmias by introducing class-weighted loss, focal loss, weighted sampling, oversampling, undersampling, and stratified splitting is a top priority for future work. Moreover, large-scale clinical validation will be pursued to assess

the generalizability of the framework across different populations and acquisition devices.

6. Conclusion

This work proposed an adaptive multi-task learning strategy with denoising and hyperspherical self-distillation as auxiliary tasks for arrhythmia detection under strong noise interference. The framework demonstrates the ability to reduce false alarms and misdiagnoses in wearable ECG scenarios heavily affected by artifacts, making it particularly suitable for applications such as wearable ECG monitoring systems, remote cardiac health screening, early arrhythmia detection in daily-life scenarios, and mobile health (mHealth) applications, thereby offering a promising pathway to enhance the reliability of wearable ECG devices in clinical settings.

CRediT authorship contribution statement

Zhongyu Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization; **Maarten De Vos:** Resources, Writing – review & editing, Supervision, Funding acquisition; **Caiyun Ma:** Writing – review & editing, Supervision, Project administration, Funding acquisition; **Jianghai Qian:** Data curation, Writing – review & editing; **Jianqing Li:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition; **Chengyu Liu:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Funding

This work was supported in part by the National Natural Science Foundation of China (62171123, 62501148, and 62211530112), the China Postdoctoral Science Foundation (2024M760444), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX24_0382), the Postdoc toral Fellowship Program (Grade C) of the China Postdoc toral Science Foundation under Grant Number GZC20251919, the Jiangsu Funding Program for Excellent Postdoctoral Talent (2025ZB519). This research received funding from the Flemish Government (FAIR), KU Leuven (BOF) VHeart-FM and EU under grant agreement #101095384.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Macro-AUPRC results of ablation studies

A.1. Macro-AUPRC results of ResNet34

Table A.1
Macro-AUPRC of **ResNet34** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at varying SNR levels.

Dataset		PTB-XL							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.362	0.352	0.383	0.372	0.402	0.356	0.419	**0.444**
	−12∼0	0.504	0.489	0.490	0.502	0.536	0.517	0.545	**0.567**
	−6∼6	0.590	0.574	0.597	0.588	0.608	0.588	0.615	**0.626**
	0∼12	0.630	0.620	0.624	0.640	0.642	0.630	0.644	**0.652**
BW + EM	−18∼−6	0.355	0.414	0.465	0.395	0.461	0.442	0.464	**0.507**
	−12∼0	0.513	0.502	0.530	0.502	0.498	0.538	0.568	**0.589**
	−6∼6	0.593	0.584	0.605	0.564	0.615	0.610	0.623	**0.637**
	0∼12	0.638	0.625	0.632	0.616	0.652	0.644	0.657	**0.657**
EM + MA	−18∼−6	0.350	0.354	0.353	0.413	0.352	0.356	0.377	**0.418**
	−12∼0	0.483	0.433	0.478	0.506	0.459	0.492	0.510	**0.551**
	−6∼6	0.570	0.560	0.570	0.577	0.577	0.577	0.602	**0.610**
	0∼12	0.624	0.619	0.632	0.640	0.616	0.637	0.635	**0.652**
EM	−18∼−6	0.357	0.382	0.367	0.359	0.394	0.402	0.423	**0.469**
	−12∼0	0.493	0.499	0.510	0.438	0.511	0.497	0.527	**0.564**
	−6∼6	0.583	0.574	0.583	0.540	0.606	0.602	0.609	**0.626**
	0∼12	0.628	0.615	0.629	0.594	0.628	0.632	0.646	**0.653**
Dataset		G12EC							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.181	0.174	0.185	0.182	0.232	0.138	0.187	**0.246**
	−12∼0	0.226	0.208	0.282	0.283	0.282	0.235	0.277	**0.314**
	−6∼6	0.244	0.239	0.299	0.293	0.310	0.305	0.310	**0.363**
	0∼12	0.295	0.259	0.335	0.308	0.352	0.307	0.317	**0.399**
BW + EM	−18∼−6	0.212	0.169	0.238	0.184	0.203	0.213	0.229	**0.261**
	−12∼0	0.235	0.202	0.273	0.222	0.257	0.247	0.281	**0.326**
	−6∼6	0.268	0.242	0.322	0.299	0.297	0.291	0.319	**0.368**
	0∼12	0.297	0.244	0.307	0.337	0.342	0.283	0.326	**0.399**
EM + MA	−18∼−6	0.086	0.076	0.192	0.114	0.237	0.175	0.169	**0.240**
	−12∼0	0.210	0.203	0.259	0.262	0.279	0.235	0.238	**0.300**
	−6∼6	0.250	0.232	0.276	0.310	0.318	0.298	0.289	**0.359**
	0∼12	0.292	0.247	0.339	0.339	0.340	0.303	0.310	**0.386**
EM	−18∼−6	0.171	0.178	0.195	0.167	0.133	0.193	0.204	**0.278**
	−12∼0	0.224	0.193	0.282	0.222	0.189	0.261	0.263	**0.322**
	−6∼6	0.249	0.242	0.267	0.286	0.224	0.285	0.309	**0.358**
	0∼12	0.278	0.252	0.334	0.324	0.268	0.327	0.305	**0.384**

A.2. Macro-AUPRC results of ShuffleNet V2

Table A.2
Macro-AUPRC of **ShuffleNet V2** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at varying SNR levels.

Dataset		PTB-XL							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.348	0.346	0.350	0.352	0.348	0.345	0.348	0.366
	−12∼0	0.409	0.404	0.416	0.425	0.428	0.384	0.391	**0.447**
	−6∼6	0.493	0.476	0.518	0.533	0.508	0.457	0.502	**0.540**
	0∼12	0.562	0.534	0.587	0.595	0.581	0.565	0.586	**0.605**
BW + EM	−18∼−6	0.355	0.351	0.380	0.361	0.353	0.375	0.357	**0.413**
	−12∼0	0.426	0.416	0.436	0.456	0.459	0.415	0.442	**0.492**
	−6∼6	0.522	0.494	0.552	0.555	0.540	0.504	0.541	**0.567**
	0∼12	0.580	0.553	0.604	0.609	0.593	0.585	0.599	**0.615**
EM + MA	−18∼−6	0.355	0.355	0.354	0.355	0.351	0.354	0.358	**0.365**
	−12∼0	0.392	0.382	0.409	0.388	0.402	0.375	0.392	**0.438**
	−6∼6	0.474	0.458	0.492	0.514	0.459	0.462	0.486	**0.517**
	0∼12	0.560	0.528	0.571	0.584	0.390	0.562	0.573	**0.593**
EM	−18∼−6	0.374	0.357	0.376	0.369	0.391	0.365	0.390	**0.408**
	−12∼0	0.421	0.414	0.424	0.444	0.441	0.421	0.421	**0.468**
	−6∼6	0.514	0.471	0.519	0.509	0.500	0.496	0.515	**0.549**
	0∼12	0.564	0.538	0.583	0.583	0.582	0.564	0.594	**0.608**
Dataset		G12EC							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.105	0.076	0.125	0.094	**0.158**	0.113	0.120	0.155
	−12∼0	0.158	0.099	0.116	0.114	0.138	0.148	0.150	**0.197**
	−6∼6	0.195	0.106	0.176	0.174	0.229	0.194	0.198	**0.249**
	0∼12	0.223	0.111	0.131	0.207	0.254	0.226	0.234	**0.275**
BW + EM	−18∼−6	0.127	0.098	0.106	0.088	**0.164**	0.122	0.144	0.159
	−12∼0	0.168	0.105	0.126	0.133	0.178	0.167	0.166	**0.199**
	−6∼6	0.197	0.125	0.149	0.171	0.184	0.201	0.211	**0.222**
	0∼12	0.212	0.150	0.207	0.136	0.253	0.221	0.230	**0.261**
EM + MA	−18∼−6	0.077	0.075	0.099	0.086	**0.141**	0.114	0.108	0.132
	−12∼0	0.150	0.096	0.158	0.106	0.176	0.148	0.155	**0.182**
	−6∼6	0.195	0.109	0.128	0.128	**0.226**	0.186	0.205	0.218
	0∼12	0.222	0.123	0.137	0.179	0.244	0.219	0.226	**0.248**
EM	−18∼−6	0.126	0.091	0.117	0.101	0.126	0.074	0.075	**0.161**
	−12∼0	0.171	0.099	0.121	0.129	**0.210**	0.165	0.175	0.193
	−6∼6	0.198	0.108	0.157	0.155	0.199	0.191	0.196	**0.227**
	0∼12	0.216	0.113	0.209	0.188	0.237	0.222	0.226	**0.267**

A.3. Macro-AUPRC results of MobileNet V2

Table A.3
Macro-AUPRC of **MobileNet V2** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at varying SNR levels.

Dataset		PTB-XL							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.345	0.344	0.344	0.348	0.342	0.347	0.413	**0.423**
	−12∼0	0.496	0.362	0.517	0.521	0.500	0.500	0.510	**0.546**
	−6∼6	0.584	0.562	0.593	0.604	0.559	0.586	0.592	**0.606**
	0∼12	0.631	0.591	0.624	**0.636**	0.610	0.624	0.631	0.632
BW + EM	−18∼−6	0.355	0.350	0.359	0.464	0.442	0.414	0.439	**0.470**
	−12∼0	0.517	0.511	0.518	0.556	0.516	0.521	0.545	**0.568**
	−6∼6	0.604	0.570	0.599	0.608	0.567	0.595	0.606	**0.614**
	0∼12	0.624	0.618	0.634	0.645	0.612	0.636	0.638	**0.645**
EM + MA	−18∼−6	0.350	0.349	0.357	**0.414**	0.351	0.363	0.392	0.404
	−12∼0	0.370	0.365	0.484	0.509	0.481	0.475	0.493	**0.525**
	−6∼6	0.568	0.555	0.571	0.579	0.548	0.566	0.564	**0.584**
	0∼12	0.627	0.591	0.623	0.628	0.594	0.596	0.626	**0.628**
EM	−18∼−6	0.355	0.351	0.375	0.391	0.416	0.404	0.418	**0.419**
	−12∼0	0.509	0.507	0.524	0.509	0.505	0.498	0.494	**0.529**
	−6∼6	0.587	0.551	0.595	0.591	0.546	0.580	0.581	**0.613**
	0∼12	0.627	0.597	0.626	0.637	0.599	0.630	0.636	**0.644**
Dataset		G12EC							
Type	SNR(dB)	M1	M2	M3	M4	M5	M6	M7	M8
BW + EM + MA	−18∼−6	0.077	0.076	0.154	0.152	0.079	0.174	0.170	**0.209**
	−12∼0	0.223	0.101	0.221	0.191	0.241	0.221	0.226	**0.252**
	−6∼6	0.228	0.103	0.214	0.200	0.253	0.257	0.262	**0.288**
	0∼12	0.250	0.111	0.283	0.274	0.264	0.270	0.299	**0.306**
BW + EM	−18∼−6	0.160	0.096	0.187	0.153	**0.220**	0.194	0.181	0.216
	−12∼0	0.178	0.102	0.235	0.180	0.244	0.242	0.231	**0.278**
	−6∼6	0.242	0.104	0.257	0.256	0.256	0.260	0.277	**0.326**
	0∼12	0.258	0.106	0.264	0.281	0.272	0.304	0.272	**0.330**
EM + MA	−18∼−6	0.079	0.077	0.148	0.109	0.076	0.078	0.076	**0.209**
	−12∼0	0.171	0.101	0.227	0.211	0.234	0.235	0.208	**0.256**
	−6∼6	0.219	0.104	0.264	0.206	0.245	0.256	0.270	**0.337**
	0∼12	0.259	0.107	0.241	0.245	0.273	0.090	0.287	**0.340**
EM	−18∼−6	0.158	0.095	0.198	0.128	0.192	0.131	0.187	**0.235**
	−12∼0	0.177	0.101	0.219	0.169	0.240	0.219	0.215	**0.269**
	−6∼6	0.240	0.105	0.221	0.203	0.251	0.256	0.259	**0.293**
	0∼12	0.256	0.107	0.260	0.243	0.258	0.263	0.266	**0.349**

Appendix B. Per-class AUPRC results of ablation studies

B.1. Per-class AUPRC results of ResNet34

Table B.1
Per-class AUPRC of **ResNet34** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at -18∼-6 SNR level.

PTB-XL								
Type	BW + EM + MA		BW + EM		EM + MA		EM	
Setting	M1	M8	M1	M8	M1	M8	M1	M8
NORM	0.562	0.691(+0.129)	0.525	0.747(+0.222)	0.517	0.648(+0.131)	0.533	0.700(+0.167)
MI	0.355	0.387(+0.032)	0.374	0.452(+0.078)	0.365	0.386(+0.021)	0.374	0.425(+0.051)
STTC	0.436	0.474(+0.038)	0.462	0.575(+0.113)	0.465	0.499(+0.034)	0.475	0.590(+0.115)
CD	0.299	0.445(+0.146)	0.267	0.501(+0.234)	0.265	0.360(+0.095)	0.265	0.429(+0.164)
HYP	0.159	0.223(+0.074)	0.150	0.258(+0.108)	0.136	0.195(+0.059)	0.137	0.203(+0.066)
G12EC								
Type	BW + EM + MA		BW + EM		EM + MA		EM	
Setting	M1	M8	M1	M8	M1	M8	M1	M8
AVB I	0.156	0.451(+0.295)	0.137	0.401(+0.264)	0.085	0.364(+0.279)	0.147	0.412(+0.265)
AF	0.085	0.142(+0.057)	0.104	0.154(+0.050)	0.073	0.138(+0.065)	0.092	0.166(+0.074)
AFL	0.035	0.070(+0.035)	0.063	0.057(−0.006)	0.026	0.095(+0.069)	0.054	0.093(+0.039)
Brady	0.001	0.001	0.001	0.003(+0.002)	0.001	0.001	0.001	0.027(+0.026)
CRBBB	0.017	0.004(−0.013)	0.157	0.006(−0.151)	0.010	0.025(+0.015)	0.014	0.014
IRBBB	0.067	0.083(0.016)	0.066	0.122(+0.056)	0.044	0.076(+0.032)	0.068	0.197(+0.129)
LAFB	0.161	0.175(+0.014)	0.216	0.174(−0.042)	0.039	0.181(+0.142)	0.102	0.198(+0.096)
LAD	0.472	0.553(+0.081)	0.551	0.569(+0.018)	0.121	0.528(+0.407)	0.457	0.544(+0.087)
LBBB	0.177	0.236(+0.059)	0.290	0.337(+0.047)	0.038	0.371(+0.333)	0.195	0.376(+0.181)
Low QRS	0.068	0.049(−0.019)	0.077	0.065(−0.012)	0.054	0.063(+0.009)	0.058	0.060(+0.002)
NIVCD	0.041	0.060(+0.019)	0.060	0.055(−0.005)	0.024	0.074(+0.050)	0.065	0.044(−0.021)
PAC	0.068	0.078(+0.010)	0.077	0.071(−0.006)	0.081	0.075(−0.006)	0.077	0.102(+0.025)
Long QT	0.244	0.321(+0.077)	0.241	0.314(+0.073)	0.140	0.282(+0.142)	0.186	0.324(+0.138)
Q wave Abn	0.161	0.187(+0.026)	0.131	0.150(+0.019)	0.055	0.137(+0.082)	0.102	0.117(+0.015)
RAD	0.018	0.025(+0.007)	0.016	0.022(+0.006)	0.006	0.040(+0.034)	0.009	0.058(+0.049)
RBBB	0.124	0.222(+0.098)	0.158	0.456(+0.298)	0.083	0.239(+0.156)	0.150	0.445(+0.295)
SA	0.074	0.087(+0.013)	0.091	0.091	0.049	0.111(+0.062)	0.110	0.104(−0.006)
SB	0.561	0.772(+0.211)	0.608	0.707(+0.099)	0.238	0.652(+0.414)	0.545	0.913(+0.368)
SR	0.393	0.638(+0.245)	0.490	0.675(+0.185)	0.219	0.583(+0.364)	0.308	0.649(+0.341)
ST	0.693	0.876(+0.183)	0.768	0.926(+0.158)	0.156	0.896(+0.740)	0.720	0.928(+0.208)
T Abn	0.366	0.459(+0.93)	0.399	0.468(+0.069)	0.278	0.433(+0.155)	0.318	0.443(+0.125)
TWI	0.122	0.118(−0.004)	0.112	0.112	0.088	0.111(+0.023)	0.094	0.109(+0.015)
PVC	0.069	0.050(−0.019)	0.069	0.066(−0.003)	0.062	0.057(−0.005)	0.068	0.068

B.2. Per-class AUPRC results of ShuffleNet V2

Table B.2
Per-class AUPRC of **ShuffleNet V2** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at -18∼-6 SNR level.

PTB-XL								
Type	BW + EM + MA		BW + EM		EM + MA		EM	
Setting	M1	M8	M1	M8	M1	M8	M1	M8
NORM	0.539	0.568(+0.029)	0.522	0.629(+0.107)	0.527	0.547(+0.020)	0.567	0.599(+0.032)
MI	0.357	0.370(+0.013)	0.374	0.406(+0.032)	0.369	0.381(+0.012)	0.378	0.418(+0.040)
STTC	0.441	0.462(+0.021)	0.461	0.507(+0.046)	0.464	0.475(+0.011)	0.487	0.531(+0.044)
CD	0.259	0.281(+0.022)	0.267	0.337(+0.070)	0.265	0.277(+0.012)	0.288	0.319(+0.031)
HYP	0.145	0.149(+0.004)	0.150	0.187(+0.037)	0.149	0.143(−0.006)	0.151	0.174(+0.023)
G12EC								
Type	BW + EM + MA		BW + EM		EM + MA		EM	
Setting	M1	M8	M1	M8	M1	M8	M1	M8
AVB I	0.116	0.151(+0.035)	0.128	0.135(+0.007)	0.086	0.129(+0.043)	0.133	0.141(+0.008)
AF	0.090	0.077(−0.013)	0.080	0.091(+0.011)	0.068	0.077(+0.009)	0.085	0.087(+0.002)
AFL	0.035	0.034(−0.001)	0.042	0.023(−0.019)	0.024	0.041(+0.017)	0.068	0.032(−0.036)
Brady	0.001	0.005(+0.004)	0.001	0.001	0.001	0.002(+0.001)	0.001	0.025(+0.015)
CRBBB	0.007	0.005(−0.002)	0.009	0.011(+0.002)	0.005	0.017(+0.012)	0.014	0.020(+0.007)
IRBBB	0.043	0.078(+0.035)	0.049	0.077(+0.028)	0.046	0.056(+0.010)	0.057	0.056(−0.001)
LAFB	0.072	0.178(+0.106)	0.097	0.145(+0.048)	0.027	0.153(+0.126)	0.088	0.206(+0.118)
LAD	0.347	0.457(+0.110)	0.366	0.457(+0.091)	0.097	0.404(+0.307)	0.383	0.494(+0.111)
LBBB	0.062	0.188(+0.126)	0.179	0.268(+0.089)	0.042	0.157(+0.115)	0.107	0.242(+0.135)
Low QRS	0.051	0.042(−0.009)	0.049	0.059(+0.010)	0.039	0.064(+0.025)	0.040	0.080(+0.040)
NIVCD	0.033	0.044(+0.011)	0.039	0.044(+0.005)	0.021	0.038(+0.017)	0.048	0.078(+0.030)
PAC	0.076	0.071(−0.005)	0.068	0.070(+0.002)	0.067	0.066(−0.001)	0.071	0.069(−0.002)
Long QT	0.164	0.180(+0.016)	0.168	0.213(+0.045)	0.164	0.169(+0.005)	0.168	0.166(−0.002)
Q wave Abn	0.050	0.091(+0.041)	0.073	0.077(+0.004)	0.050	0.052(+0.002)	0.068	0.076(+0.008)
RAD	0.005	0.024(+0.019)	0.006	0.024(+0.018)	0.006	0.017(+0.011)	0.008	0.023(+0.015)
RBBB	0.106	0.118(+0.012)	0.126	0.150(+0.024)	0.082	0.171(+0.089)	0.145	0.109(−0.036)
SA	0.045	0.072(+0.027)	0.054	0.108(+0.054)	0.047	0.047	0.057	0.078(+0.021)
SB	0.248	0.539(+0.291)	0.348	0.389(+0.041)	0.207	0.329(+0.122)	0.297	0.446(+0.149)
SR	0.234	0.327(+0.093)	0.271	0.309(+0.038)	0.187	0.203(+0.016)	0.256	0.302(0.046)
ST	0.215	0.515(+0.300)	0.332	0.555(+0.223)	0.139	0.435(+0.296)	0.374	0.546(+0.172)
T Abn	0.265	0.233(−0.032)	0.273	0.299(+0.026)	0.231	0.250(+0.019)	0.282	0.272(−0.010)
TWI	0.099	0.092(−0.007)	0.087	0.083(−0.004)	0.084	0.103(+0.019)	0.092	0.100(+0.008)
PVC	0.058	0.046(−0.012)	0.079	0.059(−0.020)	0.044	0.056(+0.012)	0.064	0.051(−0.013)

B.3. Per-class AUPRC results of MobileNet V2

Table B.3
Per-class AUPRC of **MobileNet V2** under multiple multi-task training strategies on PTB-XL and G12EC datasets with mixed noise at -18∼-6 SNR level.

PTB-XL								
Type	BW + EM + MA		BW + EM		EM + MA		EM	
Setting	M1	M8	M1	M8	M1	M8	M1	M8
NORM	0.541	0.651(+0.110)	0.524	0.731(+0.207)	0.527	0.607(+0.080)	0.535	0.634(+0.099)
MI	0.358	0.382(+0.024)	0.376	0.412(+0.036)	0.366	0.382(+0.016)	0.370	0.410(+0.040)
STTC	0.434	0.470(+0.036)	0.459	0.526(+0.067)	0.463	0.484(+0.021)	0.470	0.537(+0.067)
CD	0.258	0.395(+0.137)	0.272	0.459(+0.187)	0.261	0.347(+0.086)	0.262	0.330(+0.068)
HYP	0.134	0.217(0.083)	0.146	0.222(+0.076)	0.133	0.198(+0.065)	0.138	0.184(+0.046)
G12EC								
Type	BW + EM + MA		BW + EM		EM + MA		EM	
Setting	M1	M8	M1	M8	M1	M8	M1	M8
AVB I	0.086	0.173(+0.087)	0.145	0.232(+0.087)	0.077	0.249(+0.172)	0.139	0.315(+0.176)
AF	0.056	0.126(+0.070)	0.090	0.149(+0.059)	0.084	0.128(+0.044)	0.078	0.126(+0.048)
AFL	0.027	0.079(+0.052)	0.078	0.088(+0.010)	0.035	0.090(+0.055)	0.046	0.148(+0.102)
Brady	0.001	0.001	0.001	0.006(+0.005)	0.001	0.002(+0.001)	0.001	0.006(+0.005)
CRBBB	0.002	0.010(+0.008)	0.028	0.008(−0.020)	0.007	0.094(+0.087)	0.014	0.009(−0.005)
IRBBB	0.038	0.058(+0.020)	0.056	0.074(+0.018)	0.041	0.069(+0.028)	0.060	0.081(+0.021)
LAFB	0.032	0.140(+0.108)	0.105	0.124(+0.019)	0.055	0.148(+0.093)	0.116	0.141(+0.025)
LAD	0.107	0.498(+0.391)	0.403	0.548(+0.145)	0.101	0.518(+0.417)	0.402	0.571(+0.169)
LBBB	0.034	0.269(+0.235)	0.240	0.214(−0.026)	0.039	0.227(+0.188)	0.179	0.296(+0.117)
Low QRS	0.042	0.059(+0.017)	0.050	0.053(+0.003)	0.038	0.054(+0.016)	0.050	0.066(+0.016)
NIVCD	0.023	0.071(+0.048)	0.042	0.069(+0.027)	0.018	0.067(+0.049)	0.049	0.051(+0.002)
PAC	0.066	0.060(−0.006)	0.066	0.069(+0.003)	0.066	0.064(−0.002)	0.067	0.088(+0.021)
Long QT	0.155	0.321(+0.166)	0.166	0.262(+0.096)	0.156	0.260(+0.104)	0.179	0.274(+0.095)
Q wave Abn	0.057	0.105(+0.048)	0.077	0.144(+0.067)	0.052	0.112(+0.060)	0.076	0.135(+0.059)
RAD	0.018	0.009(−0.009)	0.011	0.019(+0.008)	0.006	0.010(+0.004)	0.010	0.009(−0.001)
RBBB	0.085	0.140(+0.055)	0.130	0.193(+0.063)	0.073	0.164(+0.091)	0.134	0.351(+0.217)
SA	0.056	0.064(+0.008)	0.086	0.080(−0.006)	0.043	0.075(+0.032)	0.102	0.085(−0.017)
SB	0.200	0.636(+0.436)	0.492	0.535(+0.043)	0.215	0.558(+0.343)	0.469	0.581(+0.112)
SR	0.178	0.620(+0.442)	0.312	0.638(+0.326)	0.189	0.529(+0.340)	0.303	0.602(+0.299)
ST	0.152	0.798(+0.646)	0.664	0.885(+0.221)	0.143	0.820(+0.677)	0.707	0.873(+0.166)
T Abn	0.239	0.411(+0.172)	0.278	0.424(+0.146)	0.241	0.402(+0.161)	0.293	0.458(+0.165)
TWI	0.079	0.116(+0.037)	0.089	0.107(+0.018)	0.108	0.123(+0.015)	0.092	0.106(+0.014)
PVC	0.035	0.044(+0.009)	0.068	0.044(−0.024)	0.038	0.051(+0.013)	0.084	0.051(−0.033)

B.4. Explanation of the meaning of disease names

Table B.4
Disease name abbreviation details.

PTB-XL	
NORM	Normal ECG
MI	Myocardial Infarction
STTC	ST/T Change
CD	Conduction Disturbance
HYP	Hypertrophy
G12EC	
AVB I	1st degree AV block
AF	Atrial Fibrillation
AFL	Atrial Flutter
Brady	Bradycardia
CRBBB	Complete Right Bundle Branch Block
IRBBB	Incomplete Right Bundle Branch Block
LAFB	Left Anterior Fascicular Block
LAD	Left Axis Deviation
LBBB	Left Bundle Branch Block
Low QRS	Low QRS Voltages
NIVCD	Nonspecific Intraventricular Conduction Disorder
PAC	Premature Atrial Contraction
Long QT	Prolonged QT Interval
Q wave Abn	Q wave Abnormal
RAD	Right Axis Deviation
RBBB	Right Bundle Branch Block
SA	Sinus Arrhythmia
SB	Sinus Bradycardia
SR	Sinus Rhythm
ST	Sinus Tachycardia
T Abn	T wave Abnormal
TWI	T wave Inversion
PVC	Ventricular Premature Beats

References

Adnan, M., Ioannou, Y. A., Tsai, C.-Y., & Taylor, G. W. (2021). Domain-agnostic clustering with self-distillation. https://arxiv.org/abs/2111.12170.

Andonian, A., Chen, S., & Hamid, R. (2022). Robust cross-modal representation learning with progressive self-distillation. https://arxiv.org/abs/2204.04588.

Ansari, Y., Mourad, O., Qaraqe, K., & Serpedin, E. (2023). Deep learning for ECG arrhythmia detection and classification: An overview of progress for period 2017-2023. *Frontiers in Physiology*, *Volume 14 - 2023*. https://doi.org/10.3389/fphys.2023.1246746

Antczak, K. (2019). Deep recurrent neural networks for ECG signal denoising. https://arxiv.org/abs/1807.11551.

Bai, G., & Zhao, L. (2022). Saliency-regularized deep multi-task learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. https://doi.org/10.1145/3534678.3539442

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Chen, M., Li, Y., Zhang, L., Zhang, X., Gao, J., Sun, Y., Shi, W., & Wei, S. (2024). Multitask learning-based quality assessment and denoising of electrocardiogram signals. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1–13.

Chen, S., Zhang, Y., & Yang, Q. (2024). Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, *56*(12), Article 295. https://doi.org/10.1145/3663363.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., & Rabinovich, A. (2018). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. https://arxiv.org/abs/1711.02257.

Cipolla, R., Gal, Y., & Kendall, A. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. https://doi.org/10.1109/cvpr.2018.00781

Ding, X., Wang, Y., Xu, Z., Wang, Z. J., & Welch, W. J. (2023). Distilling and transferring knowledge via cGAN-generated samples for image classification and regression. *Expert Systems with Applications*, *213*(B). https://doi.org/10.1016/j.eswa.2022.119060

Gupta, V. (2023). Wavelet transform and vector machines as emerging tools for computational medicine. *J Ambient Intell Human Comput*, *14*, 4595-4605. https://doi.org/https://doi.org/10.1007/s12652-023-04582-0

Gupta, V., Mittal, M., & Mittal, V. (2024a). A simplistic and novel technique for ECG signal pre-processing. *IETE Journal of Research*, *70*(1), 815-826. https://doi.org/10.1080/03772063.2022.2135622

Gupta, V., Mittal, V., & Mittal, M. (2022). A novel frWT based arrhythmia detection in ECG signal using YWARA and PCA. *Wireless Pers Commun*, *124*, 1229-1246. https://doi.org/https://doi.org/10.1007/s11277-021-09403-1

Gupta, V., Saxena, N. K., Kanungo, A., Gupta, A., Kumar, P., & Salim (2022). A review of different ECG classification/detection techniques for improved medical applications. *International Journal of System Assurance Engineering and Management*, (p. 1037-1051). https://doi.org/10.1007/s13198-021-01548-3

Gupta, V., Sharma, A. K., Pandey, P. K., Jaiswal, R. K., & Gupta, A. (2024b). Pre-processing based ECG signal analysis using emerging tools. *IETE Journal of Research*, *70*(4), 4219-4230. https://doi.org/10.1080/03772063.2023.2202162

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2018). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, (p. 65-69). https://doi.org/10.1038/s41591-018-0268-3

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00975

Hering, A., Hansen, L., Mok, T. C. W., Chung, A. C. S., et al. (2023). Learn2Reg: Comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, *42*(3), 697–712. https://doi.org/10.1109/TMI.2022.3213983.

Hu, Q., Lin, F., Wang, B., & Li, C. (2022a). Mbrep: Motif-based representation learning in heterogeneous networks. *Expert Systems with Applications*, *190*. https://doi.org/10.1016/j.eswa.2021.116031

Hu, R., Chen, J., & Zhou, L. (2022b). A transformer-based deep neural network for arrhythmia detection using continuous ECG signals. *Computers in Biology and Medicine*, (p. 105325). https://doi.org/10.1016/j.compbiomed.2022.105325

Kabir, M. A., & Shahnaz, C. (2012). Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains. *Biomedical Signal Processing and Control*, *7*(5), 481–489. https://doi.org/https://doi.org/10.1016/j.bspc.2011.11.003

Khatar, Z., Bentaleb, D., & El Mansouri, M. (2024). Integrating advanced combined numerical filters for ECG denoising and cardiovascular disease classification using deep learning. In *Digital technologies and applications* (pp. 539–547). Cham: Springer Nature Switzerland.

Khoshkangini, R., Mashhadi, P., Tegnered, D., Lundstrom, J., & Rognvaldsson, T. (2023). Predicting vehicle behavior using multi-task ensemble learning. *Expert Systems with Applications*, *212*. https://doi.org/10.1016/j.eswa.2022.118716

Kim, K., Ji, B., Yoon, D., & Hwang, S. (2021). Self-knowledge distillation with progressive refinement of targets. https://arxiv.org/abs/2006.12000.

Kumar, A., Tomar, H., Mehla, V. K., Komaragiri, R., & Kumar, M. (2021). Stationary wavelet transform based ECG signal denoising method. *ISA Transactions*, *114*, 251–262. https://doi.org/https://doi.org/10.1016/j.isatra.2020.12.029

Lee, J. R., Kim, S., Park, I., Eo, T., & Hwang, D. (2021). Relevance-CAM: Your model already knows where to look. In *2021 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.01470

Li, H., Ditzler, G., Roveda, J., & Li, A. (2024). Descod-ECG: Deep score-based diffusion model for ECG baseline wander and noise removal. *IEEE Journal of Biomedical and Health Informatics*, *28*(9), 5081–5091. https://doi.org/10.1109/JBHI.2023.3237712

Li, K., Xiong, H., Yu, D., Liu, J., Guo, Y., & Wang, J. (2022). An end-to-end multi-task learning model for drivable road detection via edge refinement and geometric deformation. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 8641–8651. https://doi.org/10.1109/TITS.2021.3084058

Ling, Y., Wang, Y., Dai, W., Yu, J., Liang, P., & Kong, D. (2024). MTANet: Multi-task attention network for automatic medical image segmentation and classification. *IEEE Transactions on Medical Imaging*, *43*(2), 674–685. https://doi.org/10.1109/TMI.2023.3317088.

Liu, B., Liu, X., Jin, X., Stone, P., & Liu, Q. (2021). Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, *34*, 18878–18890.

Liu, C., Zhang, X., Zhao, L., Liu, F., Chen, X., Yao, Y., & Li, J. (2019a). Signal quality assessment and lightweight QRS detection for wearable ECG smartvest system. *IEEE Internet of Things Journal*, *6*(2), 1363–1374. https://doi.org/10.1109/JIOT.2018.2844090

Liu, L., Yi, L., Kuang, Z.-B., Xue, J.-H., Chen, Y., Yang, W., Liao, Q., & Zhang, W. (2021b), Towards impartial multi-task learning. In *International Conference on Learning Representations,International Conference on Learning Representations* (pp. 1–21).

Liu, S., Johns, E., & Davison, A. J. (2019b). End-to-end multi-task learning with attention. In *2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr.2019.00197

Ma, D., Zhang, K., Cao, Q., Li, J., & Gao, X. (2024). Coordinate attention guided dual-teacher adaptive knowledge distillation for image classification. *Expert Systems with Applications*, *250*. https://doi.org/10.1016/j.eswa.2024.123892

Mallikarjunamallu, K., & Khasim, S. (2024). Arrhythmia classification using noise filtering and 1d CNN. *Traitement du Signal*, *41*. https://doi.org/10.18280/ts.410416

Marque, C., Bisch, C., Dantas, R., Elayoubi, S., Brosse, V., & Pérot, C. (2005). Adaptive filtering for ECG rejection from surface EMG recordings. *Journal of Electromyography and Kinesiology*, *15*(3), 310–315. https://doi.org/https://doi.org/10.1016/j.jelekin.2004.10.001

Mousavi, S., Afghah, F., & Acharya, U. R. (2020). HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks. *Computers in Biology and Medicine*, *127*, 104057. https://doi.org/10.1016/j.compbiomed.2020.104057.

Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., & Fetaya, E. (2022). Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, *162*, 16428–16446. PMLR.

Nemati, E., Deen, M., & Mondal, T. (2011). A wireless wearable ECG sensor for long-term applications. *IEEE Communications Magazine*, *50*(1), 36-43. https://doi.org/10.1109/mcom.2012.6122530

Nguyen, P. X., Phan, H., Pham, H., Chatzichristos, C., Vandenberk, B., & De Vos, M. (2025). Exploiting a mixture-of-layers in an electrocardiography foundation model. https://arxiv.org/abs/2509.00102.

Pan, W., An, Y., Guan, Y., & Wang, J. (2022). Mca-net: A multi-task channel attention network for myocardial infarction detection and location using 12-lead ecgs. *Computers in Biology and Medicine*, *150*, 106199. https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.106199

Park, Y., Park, Y. H., Jeong, H., Kim, K., Jung, J. Y., Kim, J.-B., & Kang, D. R. (2024). Development and validation of a real-time service model for noise removal and arrhythmia classification using electrocardiogram signals. *Sensors*, *24*(16). https://doi.org/10.3390/s24165222

Petmezas, G., Haris, K., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J. A., Katsaggelos, A. K., & Maglaveras, N. (2020). Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. *Biomedical Signal Processing and Control*, (p. 102194). https://doi.org/10.1016/j.bspc.2020.102194

Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., & De Vos, M. (2022). Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5903–5915. https://doi.org/10.1109/TPAMI.2021.3070057

Samann, F., & Schanze, T. (2023). RunDAE model: Running denoising autoencoder models for denoising ECG signals. *Computers in Biology and Medicine*, *166*, 107553. https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107553

Satija, U., Ramkumar, B., & Manikandan, M. S. (2017). A review of signal processing techniques for electrocardiogram signal quality assessment. *IEEE Reviews in Biomedical Engineering*, (p. 36-52). https://doi.org/10.1109/rbme.2018.2810957

Seeuws, N., De Vos, M., & Bertrand, A. (2022). Electrocardiogram quality assessment using unsupervised deep learning. *IEEE Transactions on Biomedical Engineering*, *69*(2), 882–893. https://doi.org/10.1109/TBME.2021.3108621

Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, *31*, 525–536. Curran Associates, Inc.

Sepahvand, M., Abdali-Mohammadi, F., & Taherkordi, A. (2022). Teacher-student knowledge distillation based on decomposed deep feature representation for intelligent mobile applications. *Expert Systems with Applications*, *202*. https://doi.org/10.1016/j.eswa.2022.117474

Shahriari, Y., Fidler, R., Pelter, M. M., Bai, Y., Villaroman, A., & Hu, X. (2018). Electrocardiogram signal quality assessment based on structural image similarity metric. *IEEE Transactions on Biomedical Engineering*, *65*(4), 745–753. https://doi.org/10.1109/TBME.2017.2717876

Sigtia, S., Marchi, E., Kajarekar, S., Naik, D., & Bridle, J. (2020). Multi-task learning for speaker verification and voice trigger detection. In *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6844–6848). https://doi.org/10.1109/ICASSP40776.2020.9054760

Singh, P., & Pradhan, G. (2021). A new ECG denoising framework using generative adversarial network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 759–764. https://doi.org/10.1109/TCBB.2020.2976981

Song, T.-J., Jeong, J., & Kim, J.-H. (2022). End-to-end real-time obstacle detection network for safe self-driving via multi-task learning. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 16318–16329. https://doi.org/10.1109/TITS.2022.3149789

Stein, C. (d). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Contribution to the Theory of Statistics* , pp. 197-206). https://doi.org/10.1525/9780520313880-018

Tripathi, P. M., Kumar, A., & Komaragiri, R., et al. (2022). A review on computational methods for denoising and detecting ECG signals to detect cardiovascular diseases. *Arch Computat Methods Eng*, *29*, 1875-1914. https://doi.org/10.1007/s11831-021-09642-2.

Wang, X., Gao, H., Ma, C., Zhu, T., Yang, F., Liu, C., & Fu, H. (2025). Uncertainty-inspired multi-task learning in arbitrary scenarios of ECG monitoring. *IEEE Journal of Biomedical and Health Informatics*, *29*(6), 4046–4058. https://doi.org/10.1109/JBHI.2025.3545927

Wu, Y.-C., Jhang, B.-Y., Lee, S.-Y., & Chen, J.-Y. (2024). Fully convolution based denoise autoencoder AI accelerator for ECG arrhythmia classification. In *2024 IEEE Biomedical circuits and systems conference (bioCAS)* (pp. 1–5). https://doi.org/10.1109/BioCAS61083.2024.10798155

Yao, Q., Wang, R., Fan, X., Liu, J., & Li, Y. (2019). Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. *Information Fusion*, (p. 174-182). https://doi.org/10.1016/j.inffus.2019.06.024

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, *33*, 5824–5836.

Yldlrlm, Özal, Pławiak, P., Tan, R.-S., & Acharya, U. R. (2018). Arrhythmia detection using deep convolutional neural network with long duration ECG signals. *Computers in Biology and Medicine*, (p. 411-420). https://doi.org/10.1016/j.compbiomed.2018.09.009

Zhang, G.-Y., Zhou, Y.-R., Wang, C.-D., Huang, D., & He, X.-Y. (2021). Joint representation learning for multi-view subspace clustering. *Expert Systems with Applications*, *166*. https://doi.org/10.1016/j.eswa.2020.113913

Zhang, J., Liu, A., Gao, M., Chen, X., Zhang, X., & Chen, X. (2020). Ecg-based multiclass arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network. *Artificial Intelligence in Medicine*, (p. 101856). https://doi.org/10.1016/j.artmed.2020.101856

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. https://arxiv.org/abs/1905.08094.

Zhang, Y., & Yang, Q. (2017). An overview of multi-task learning. *National Science Review*, *5*(1), 30–43. https://doi.org/10.1093/nsr/nwx105

Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203

Zhao, Y., Wang, X., Che, T., Bao, G., & Li, S. (2023). Multi-task deep learning for medical image computing and analysis: A review. *Computers in Biology and Medicine*, *153*, 106496. https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.106496