Information Fusion 122 (2025) 103178

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Multi-teacher self-distillation based on adaptive weighting and activation pattern for enhancing lightweight arrhythmia recognition

Zhongyu Wang [a], Caiyun Ma [a], Shuo Zhang [a,b], Minghui Zhao [a], Yu Liu [c], Lina Zhao [a], Tingting Zhu [c], 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, Nanjing, 210096, China

[b] Southeast University, School of Biological Science and Medical Engineering, the State Key Laboratory of Digital Medical Engineering, No. 2, Sipailou, Xuanwu District, Nanjing, 210096, China

[c] University of Oxford, Department of Engineering Science, Wellington Square, Oxford OX1 2JD, Oxford, UK

ARTICLE INFO

Keywords:
Knowledge distillation
Self-distillation
Electrocardiogram
Arrhythmia detection

ABSTRACT

Wearable electrocardiogram (ECG) devices, with their outstanding advantages of comfort and portability, play a vital role in daily arrhythmia monitoring outside the hospital. However, the embedded CPU used in most devices greatly limits the deployment of high-performance models. Therefore, it is increasingly important to develop lightweight neural networks and reduce computing requirements to achieve edge deployment of wearable devices. Knowledge distillation (KD) offers a promising solution for compressing and deploying lightweight neural networks by transferring knowledge from complex teacher models to enhance the performance of compact student models. However, conventional KDs give less thought to selecting strong and accessible teachers for students, which can lead to suboptimal outcomes. To mitigate such limitations, in this study, we propose a multi-teacher self-distillation (MTSD) framework to improve the performance of lightweight arrhythmia detection models in wearable ECG monitoring. Specifically, we first leverage representations from teacher models via similarity of activation patterns in the intermediate layer, to capture inter-category and inter-channel relationships, which then incorporates an MTSD framework to ensure the correctness and acceptability of teacher supervision. Furthermore, the self-distillation framework facilitates knowledge sharing across different layers within the model, thereby enhancing overall performance. Extensive experiments conducted on three medical signal datasets demonstrate the superiority of the proposed method over existing state-of-the-art distillation methods, achieving the AUC/accuracy by 0.922, 0.908 and 87.05%. Notably, the model processed a 12-lead 10-s ECG signal in only 1 ms on an NVIDIA Jetson Orin NX.

1. Introduction

With the rise of wearable ECG devices for arrhythmia monitoring, their portability and real-time monitoring capabilities are gaining widespread attention. Currently, long-term ECG monitoring relies on electrocardiogram signals captured on the human body surface by wearable ECG devices, including 12-lead wearable ECG undershirts and wearable ECG patches. However, their limited computational resources challenge the deployment of complex neural networks. Thus, compressing and optimizing lightweight neural networks for efficient high-performance applications on these devices is becoming increasingly crucial. Knowledge distillation (KD) [1] stands out as a promising solution for model compression technique that enhances the performance of lightweight models by transferring the insights, i.e., internal representations, of an over-parameterized teacher model to a compact student model [2].

Traditional KD relies on pre-trained large models to guide the training of student models, known as offline distillation [3–5]. To achieve more advanced distillation results, previous studies have proposed a novel distillation model where two or more networks dynamically learn from each other in a peer-to-peer manner during the training process, termed online distillation [6–8]. This enables multiple networks to emulate each other's predicted probabilities, facilitating a two-way transfer of knowledge. Unlike offline distillation, online distillation does not depend on a pre-trained teacher model but often produces superior results. A specialized case within online distillation is referred to as self-distillation (SD) [9,10], eliminating the need for an

* Corresponding authors.
E-mail addresses: zhongyu@seu.edu.cn (Z. Wang), 101300447@seu.edu.cn (C. Ma), zs_techo@seu.edu.cn (S. Zhang), zmh56@seu.edu.cn (M. Zhao), yu.liu@eng.ox.ac.uk (Y. Liu), zhaolina0808@126.com (L. Zhao), tingting.zhu@eng.ox.ac.uk (T. Zhu), ljq@seu.edu.cn (J. Li), chengyu@seu.edu.cn (C. Liu).

https://doi.org/10.1016/j.inffus.2025.103178
Received 20 November 2024; Received in revised form 11 March 2025; Accepted 1 April 2025
Available online 10 April 2025

additional second network. Instead, the model's output serves as the target, guiding the model to self-learn and enhance overall generalization performance. In terms of training objectives, KDs typically use both real and soft labels of the teacher model to jointly supervise the optimization of the student model, and some studies have used intermediate layer features to allow students to learn pixel-level feature representations for more complete knowledge [11]. However, wearable ECG signals are inherently non-smooth and contain substantial heterogeneous noise, making it challenging for networks to effectively capture critical information. Consequently, the loss of information during model compression may be more pronounced when traditional online distillation methods are directly applied to ECG signal-based arrhythmia detection algorithms.

1.1. Motivation and contribution

In online distillation, there are three key issues to consider for transferred knowledge: ***(i) whether the teacher's guidance is correct, (ii) whether the knowledge is easy for the students to learn, and (iii) whether the knowledge can represent the high-level semantic information of the teacher's model and ensure the stability of the KD process***. Specifically, since in online distillation, the teacher model and the student model are optimized simultaneously in a mutually reinforcing manner, for some instances, the teacher model may not be able to predict correctly, and consequently, the student model guided by that soft label will produce sub-optimal results. Subsequently, the gap in model architecture is another barrier that prevents well-represented student model learning. Seyed et al. [12] illustrated that a larger teacher model does not result in a better student model and introduced assistants to bridge the gap between teacher and student model architectures. Essentially, the difference in model capacity represents a gap in representational power, and for overly large teacher models, it is difficult to make a compact student model effectively capture the high-level representations in the teacher model. Finally, transferring knowledge by aligning the middle layer representation of the teacher–student model may not converge in online distillation. Inseop et al. [7] illustrated that the online distillation process is not suitable due to the flexible changes in the feature maps that occur in each training round, and that directly aligning the feature maps would have a negative effect on the optimization of the model. Additionally, the middle layer feature maps of the model have redundant information and are not highly condensed knowledge.

At present, some methods have been proposed to try to guide students to get the right teacher guidance [13–15] or make the transferred knowledge more acceptable [16,17]. However, these methods often focus on only a single dimension. Although these methods have achieved certain results in their respective goals, they fail to take into account both the correctness and acceptability of knowledge at the same time, which limits the distillation effect. Therefore, how to achieve the coordinated optimization of correctness and acceptability in knowledge distillation is still a key issue that needs to be solved urgently.

To address the above challenges, we propose a novel solution to enable the student model to learn the teacher model's class activation pattern instead of aligning the feature map, using adaptive weighting to assign greater learning bias to correctly guided teachers and easier-to-learn knowledge, and combining it with a self-distillation framework to further attenuate the knowledge loss due to differences in model architectures.

The contributions of this article can be summarized as follows:

(1) We propose a multi-branch self-distillation framework that leverages deep layers as teacher supervision for shallow layers, thereby mitigating architectural discrepancies and simplifying the training process.

(2) We investigate a strategy to assign the correct teacher and easy-to-learn knowledge to the student model through coefficient-weighted distillation loss, adaptively adjusting the importance of the two weighting coefficients at different training stages by optimizing the learning process.

(3) Instead of overly redundant feature maps to supervise the optimization of the student model, the dependency between category and channel activation is refined by capturing the differences in activation patterns of different categories in the middle layer of the teacher model.

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. Deep learning based arrhythmia detection

To date, numerous studies have demonstrated the reliability of deep learning (DL) methods for ECG signal analysis and ECG arrhythmia detection. A Transformer-based deep neural network, ECG DETR, was proposed by Hu et al. [18] to perform beat-by-beat location and category detection of continuous single-lead ECGs by capturing the dependencies between heartbeats, achieving good generalization ability. Yao et al. [19] implemented the processing of ECG signals of different lengths using an attention-based time-incremental convolutional neural network (ATI-CNN). Petmezas et al. [20] proposed a hybrid CNN-LSTM neural network for capturing the temporal dynamics of ECG signals and processing unbalanced training data through focal loss, which is important for helping clinicians screen for atrial fibrillation. Soltanieh et al. [21] evaluated the effectiveness of multiple self-supervised (SSL) methods including SimCLR, BYOL, and SwAV for ECG representation learning on several popular arrhythmia datasets. To address the difficulty of detecting arrhythmic diseases based on basic ECG features, a novel deep neural network, DeepArr, was proposed by Midani et al. [22]. The network uses a sequential fusion approach to combine feed-forward and recurrent networks as a means of capturing relevant representations of arrhythmic features of ECG signals, and its effectiveness has been demonstrated on the MIT-BIH arrhythmia dataset. In addition, several studies [23–26] have used DL to diagnose arrhythmias. However, most studies have paid less attention to the computational requirements of the algorithms and the usability of wearable ECG for clinical deployment.

2.2. Knowledge distillation

KD is an effective method for achieving model compression by transferring representations from high-capacity teacher models to compact student models.

2.2.1. Knowledge from intermediate layer

Recently, many studies have utilized representations from intermediate layers of models to transfer knowledge. KD employing intermediate layers can generally be categorized into two types: feature-based and relation-based.

Deep neural networks excel in learning hierarchical feature representations with increasing levels of abstraction [27]. In feature-based KD, the output feature maps from intermediate layers are used to train student models under the guidance of knowledge from powerful teacher models. Zagoruyko et al. [28] proposed using squared CNN intermediate layer feature maps to generate attention maps, improving student model performance by mimicking the attention patterns of the teacher model. Passban et al. [29] introduced a multi-layer information fusion mechanism, enhancing traditional KD methods by considering information from all layers of the teacher model using an attention mechanism. To mitigate issues arising from the direct matching of layers with different semantics, Chen et al. [30] suggested

using an attention mechanism to dynamically assign appropriate target layers from the teacher model to each student layer. Zhang et al. [10] proposed a self-distillation framework with a multi-branch structure, where shallow and deep branches serve as student and teacher models, respectively. Wang et al. [31] proposed a method to improve knowledge distillation by regularizing the feature direction and norm. They found that traditional knowledge distillation methods mainly focus on global feature matching while ignoring the local structural information of features. By constraining the direction and norm of features at the same time, the performance of the student model was significantly improved.

Relation-based KD explores relationships between different layers of feature maps. Liu et al. [5] introduced generative adversarial networks to optimize student models by distinguishing feature maps from both student and teacher models using an additional discriminator, ensuring consistent feature distributions. Park et al. [32] aimed to capture interrelationships between samples and proposed distance and angular distillation losses to penalize structural differences in these relationships. Peng et al. [33] proposed Correlation Consistency for Knowledge Distillation (CCKD), utilizing a generalized kernel approach based on Taylor series expansion to capture inter-instance correlations and convey instance-level and inter-instance correlation information to student models.

However, direct matching of feature maps in online distillation may not be suitable, as capturing only sample correlations risks losing local and specific category information.

2.2.2. Multi-teacher distillation

In the context of multi-teacher distillation approaches, a critical aspect is how to effectively integrate information from multiple teachers. Previous studies [34] treated multiple teachers equally, overlooking the varying importance of different teachers. Seyed et al. [12] proposed the use of assistants to bridge the gap between high-capacity teacher models and student models, albeit at the cost of increased computational complexity. Li et al. [35] introduced an adaptive weighting factor approach to fusing feature maps from multiple teachers, prioritizing stronger teachers with higher weights, but did not sufficiently address the student's acceptance of knowledge from overly dominant teachers. Shi et al. [13] proposed a multi-instructor weighted distillation loss approach, encouraging student models to allocate more learning weights to teachers who are more dissimilar, yet dissimilar teachers may not necessarily represent correct guidance.

In essence, integrating multi-teacher KD entails addressing two fundamental challenges: ensuring correct guidance from teachers and ensuring that the knowledge provided by teachers is accessible and beneficial for students.

2.3. Knowledge distillation in diverse applications

The idea of knowledge distillation is also widely used in diverse fields. In the processing of dirty labels, Cheng et al. proposed a label denoising method based on dataset distillation [36]. They used dataset distillation technology to generate compact datasets and effectively remove noisy labels. Zhang et al. [37] proposed an online knowledge distillation framework based on parameter mixture (OKDPH) that explicitly estimates the curvature of the loss landscape by constructing a hybrid weight model (HWM), thereby promoting flatter minima and improving the generalization performance of the model. To address the problem of imperfect annotations in medical image segmentation, Fang et al. [38] proposed a reliable mutual distillation framework to combat annotation noise through the collaboration of two segmentation models. The framework uses the complementary knowledge between the models to clean up the training data, and significantly improves the segmentation performance under noisy annotations by enforcing consistency constraints and reliability-aware sample selection strategies. In recent years, deep learning has been widely used in industrial information fields such as fault diagnosis [39–42]. Feng et al. [39] have achieved accurate monitoring and prediction of the health status of gears by building a high-fidelity digital model and realizing real-time interaction with the physical system. This method also has great potential in the field of medical monitoring. By building a digital model of the patient's heart, the electrophysiological activity and mechanical movement of the heart are simulated in real time, so as to more accurately monitor the occurrence and development of arrhythmias. However, digital twin technology places high demands on the real-time performance and computational efficiency of the model, which is highly consistent with the goals of the knowledge distillation model compression algorithm [43,44]. Li et al. [40] proposed a cross-modal zero-shot diagnosis framework based on non-contact sensing data (such as infrared thermal imaging and acoustic data), which achieved efficient fault diagnosis through global and local feature fusion. Cross-modal data fusion analysis provides a comprehensive data perspective for the model, but also increases the computational requirements. Therefore, it is extremely important to perform cross-modal distillation [45–47] to obtain a compact and sufficient representation.

3. Method

This section first provides an overview of the proposed MTSD, followed by a description of the multi-teacher supervised adaptive weighting strategy and the techniques for capturing dependencies between middle layer categories of the teacher model and channel activation patterns.

3.1. Overview

We integrate a multi-teacher SD framework, illustrated in Fig. 1. SD is a variant of online distillation that eliminates the additional model and relies on internal feedback to improve generalization and compress the model. In this framework, a deep network acts as the teacher, and a shallow network as the student. By instructing the shallow network to mimic the deep network, high-level semantic information from deeper layers transfers to shallower layers, refining internal representations and bolstering the overall backbone network. Additionally, as both networks share the same architecture, characterization deepens layer-by-layer, enhancing the student model's representation while improving the teacher model's performance. Furthermore, the trained shallow network can operate independently of the deep layers, directly handling reasoning tasks.

In the SD framework depicted in Fig. 1, there are four branch structures denoted as B_i, where $i = 1, 2, 3, 4$. Each branch B_i consists of a sequence of bottleneck layers followed by fully connected layers and a softmax output layer. Let q_i, for $i = 1, 2, 3, 4$, denote the output of the softmax layer corresponding to the first through fourth classifiers, with q_4 representing the deepest layer.

This framework operates as a multi-teacher distillation system, where the shallow network serves as the student and the deeper networks act as its teachers. Specifically, when B_1 serves as the student, $B_{2,3,4}$ are its teachers. Similarly, when B_2 is the student, $B_{3,4}$ serve as its teachers, and when B_3 is the student, B_4 is its sole teacher. The optimization of each student model is constrained by three loss functions: the cross-entropy loss $\mathcal{L}_{CE}$ for the ground truth, the KL divergence loss $\mathcal{L}_{KL}$ between the teacher model's predictions and the student model's outputs, and the loss $\mathcal{L}_{ASM}$ associated with the proposed activation similarity map (ASM), which captures the intermediate-level categories of the teacher model through channel activation patterns. Among them, $\mathcal{L}_{KL}$ and $\mathcal{L}_{ASM}$ constitute the distillation loss of multi-teacher supervision, and the losses of different teachers are distributed by weighting with adaptive coefficients. The specific algorithm will be described in the following.

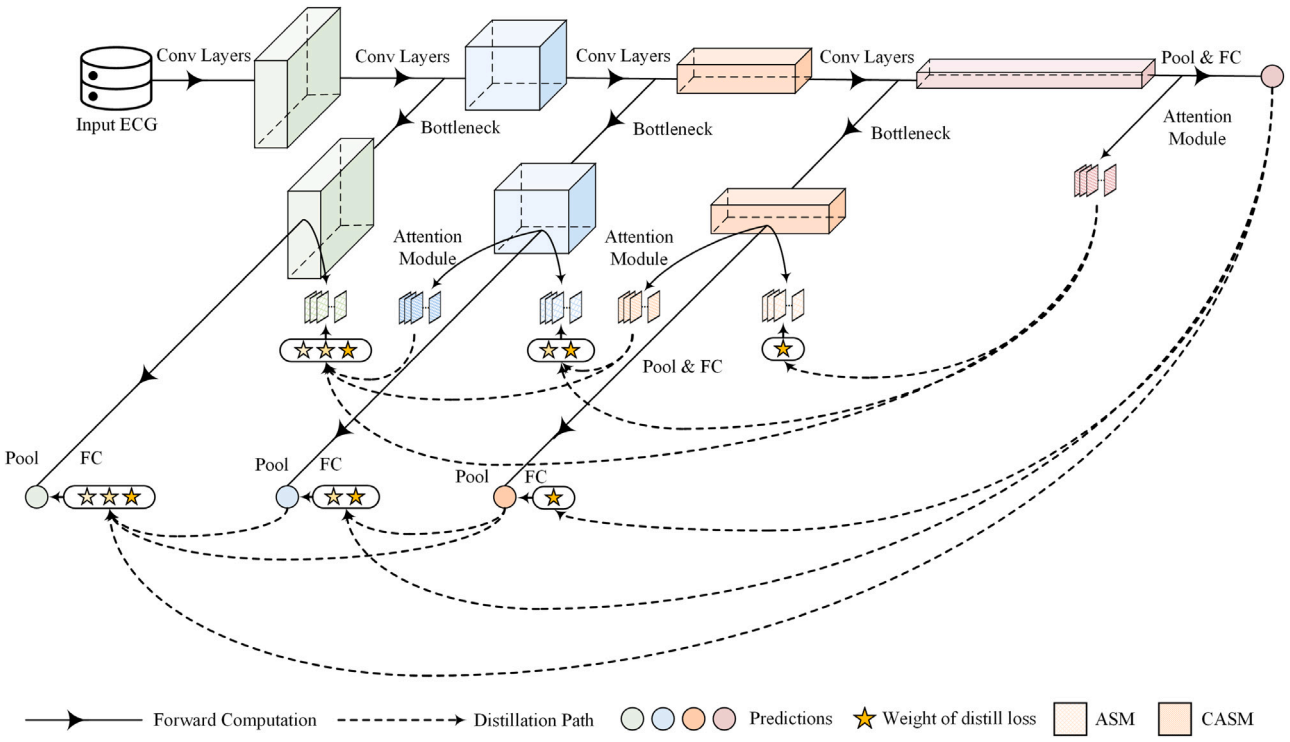

Fig. 1. An overview of the proposed multi-teacher self-distillation framework.

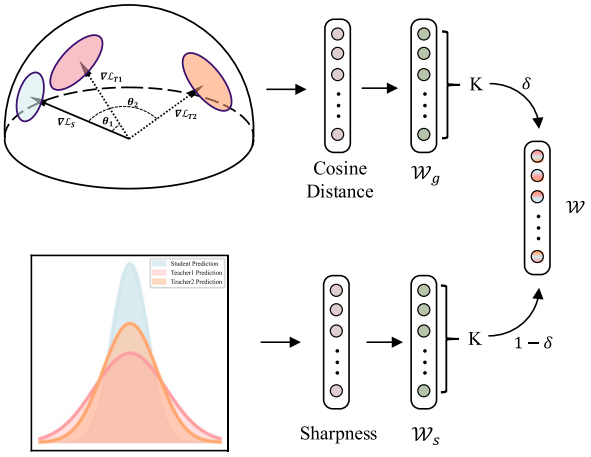

Fig. 2. Adaptive weighting factor fusion for distillation loss based on gradient direction and prediction uncertainty.

3.2. Adaptive multi-teacher guidance

This subsection introduces the proposed method of adaptive weighting for multi-teacher distillation loss coefficients, as illustrated in Fig. 2.

In multi-teacher online distillation, multiple models act as supervisors for optimizing student models. Not all supervisory signals from these teachers contribute equally to optimization; some may even mislead. Hence, it becomes crucial to assign greater loss weights to teachers providing strong positive guidance, smaller weights to weaker ones, and eliminate guidance that hinders student model optimization. We assess the benefit of distillation loss based on gradient directionality, analyzing knowledge distillation through the lens of multi-task learning.

In the KD process, the cross-entropy loss of the student model serves as the primary task, while the remaining distillation losses function as auxiliary tasks. If the gradient direction of an auxiliary task loss aligns with that of the main task loss, it indicates a beneficial impact on the main task. Conversely, orthogonal or reverse gradients suggest a detrimental effect. Fig. 3 illustrates the gradient directions and optimization objectives of auxiliary task losses supervised by different teacher models, where the main task is the cross-entropy loss of the student model. In Fig. 3(a), auxiliary task loss 2 should be discarded if its gradient opposes that of the main task loss. In Fig. 3(b), both auxiliary task losses exhibit similar gradient directions to the main task loss and share the same optimization objective, thus warranting weight assignment.

Therefore, we employ cosine distance to measure the similarity of gradient directions between different losses. The formula for cosine distance is given by:

$$D = \min(1 - \cos(\nabla \mathcal{L}_{CE}, \nabla \mathcal{L}_{KD}), 1). \tag{1}$$

Subsequently, a higher weight is assigned to the auxiliary task loss that exerts a greater impact on the optimization of the main task (i.e., the loss with a smaller cosine distance). The weight is calculated as:

$$\mathcal{W}_g^i = \frac{1}{K-1}\left(1 - \frac{\exp(D_i)}{\sum_{j=1}^{K} \exp(D_j)}\right), \tag{2}$$

where $\mathcal{W}_i^g$ denotes the weight of the distillation loss from the ith teacher model at a given time, where $i = 1, 2, \ldots, K$, and K is the total number of teacher models. D_i represents the cosine distance between the gradient of the distillation loss from the ith teacher model and the main task loss of the student model.

Simultaneously, we acknowledge that student models should prioritize learning from accessible rather than overly challenging sources. As discussed in [12], excessively large teacher models may not suit compact student models because of their greater parameter count and resulting representational power, contrasting with the capabilities of compact models. Hence, it becomes necessary to assess the disparity between teacher and student models and allocate students to teachers from whom learning is more feasible.

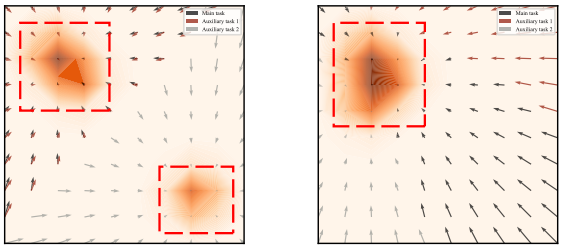

Fig. 3. Schematic representation of the direction of the loss gradient and optimization objective for the auxiliary and main tasks in the backpropagation process.

When training models for classification tasks, larger models typically produce sharper and more accurate predictive distributions compared to smaller models [48–50]. In contrast, smaller models often yield flatter and less accurate distributions. To quantify model capacity, we utilize the uncertainty of predictive distributions. Specifically, the entropy of predictions serves as a measure of uncertainty:

$$U = -\frac{1}{N}\sum_{N}\sum_{i=1}^{C} p_i \log(p_i), \tag{3}$$

where S denotes the uncertainty of the model's prediction, N is the number of samples in a batch, C is the number of categories, and p_i represents the probability of the model's prediction for the ith category. A higher value of S indicates a flatter prediction distribution, suggesting weaker model characterization. The disparity between different models can thus be formulated as:

$$\mathcal{G} = \left|U_T - U_S\right|. \tag{4}$$

Based on the gap $\mathcal{G}$ between the teacher and student models, higher weights were assigned to the distillation loss of teacher models with smaller gaps, as shown in the formula below:

$$\mathcal{W}_s^i = \frac{1}{K-1}\left(1 - \frac{\exp(\mathcal{G}_i)}{\sum_{j=1}^{K}\exp(\mathcal{G}_j)}\right), \tag{5}$$

where $\mathcal{W}_i^s$ denotes the weight of the distillation loss for the ith teacher model, with $i = 1, 2, \ldots, K$, where K is the number of teacher models. A smaller gap between the student and teacher models indicates easier knowledge transfer, resulting in a higher assigned weight.

To fuse the two sets of weight coefficients and adaptively adjust their importance across different training phases, we introduce a trainable parameter δ that is updated after each iteration using an optimizer. δ is initialized randomly within the range (0, 1). The fusion weighting coefficient is then given by:

$$\mathcal{W}^i = \sigma_s(\delta)\frac{\mathcal{W}_g^i}{2} + (1 - \sigma_s(\delta))\frac{\mathcal{W}_s^i}{2}, \tag{6}$$

where σ_s is the sigmoid function.

3.3. Channel-enhanced activation similarity map

Given the significant variation in activation across channels for different categories of samples, we aimed to enhance the category-informed representation of the teacher model by capturing these activation patterns. A straightforward approach to achieve this is by computing the ASM of the feature maps between samples, denoted as follows:

$$ASM = \frac{F_T \cdot F_T^{\top}}{\|F_T \cdot F_T^{\top}\|_2}, \tag{7}$$

where F_T denotes the intermediate layer feature map of the teacher model, with dimensions $b \times c \times l$, where b is the batch size, c is the number of channels, and l is the feature-length. ASM represents the sample channel activation pattern similarity of the teacher model, with dimensions $b \times b$.

Compared to feature maps, ASM emphasizes the activation patterns of samples across different categories. Samples within the same category typically exhibit similar activation patterns and higher correlation values, whereas samples from different categories show lower correlation values.

However, we acknowledge that the ASM obtained in this manner may lack global information from the feature map. Therefore, we introduce the Efficient Channel Attention (ECA) module [51], which aims to capture inter-channel dependencies based on global information and enhance category-specific information across different channels of the feature map. This results in the feature map representation of the teacher model as:

$$F_T^E = F_T \cdot \sigma(w * (g(F_T))), \tag{8}$$

here, $g(\cdot)$ denotes global average pooling, w represents the convolution kernel, which is defined as a set of trainable parameters, $*$ denotes the convolution operation, and σ signifies the sigmoid function.

Subsequently, the sample channel-enhanced activation similarity map (CASM), computed based on the modulated feature map, can be expressed as:

$$CASM = \frac{F_T^E \cdot F_T^{E\top}}{\|F_T^E \cdot F_T^{E\top}\|_2}, \tag{9}$$

where F_T^E represents the intermediate layer feature map of the teacher model after being enhanced by the ECA module.

We aim for the student model to acquire improved category information from the teacher model for enhanced optimization results. Therefore, we align the ASM of the student model directly with the $CASM$ of the teacher model as shown in Fig. 4. The loss is formulated as:

$$\mathcal{L}_{ASM} = \frac{1}{N}\sum_{i=1}^{N}\|ASM_S - CASM_T\|_2^2, \tag{10}$$

where ASM_S represents the ASM of the middle layer of the student model, $CASM_T$ denotes the CASM of the middle layer of the teacher model, and N indicates the number of pixel points in the ASM.

It is noteworthy that ASM has fewer pixels and exhibits less flexibility during the optimization process, thereby enhancing the stability of the distillation. Additionally, significant differences exist in the feature maps of models with varying architectures and configurations. To measure feature similarity across models and layers, we utilize centered kernel alignment (CKA) [52], as depicted in Fig. 5(a) and (b). Direct alignment can make it challenging for the student model to learn, whereas ASM mitigates these differences, as shown in Fig. 5(c) and (d), thereby enhancing the acceptance of knowledge.

3.4. Multi-teacher self-distillation

Vanilla KD entails extracting representational knowledge from a pre-trained, over-parameterized model (referred to as the teacher model) and transferring it to a compact model (referred to as the student model) through end-to-end training. The student model is supervised by both actual labels and predictions from the teacher model. Specifically, the loss function of the student model is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE}(\sigma(y_s), \mathbf{y}) + T^2\mathcal{L}_{KL}(\sigma(y_t/T), \sigma(y_s/T)). \tag{11}$$

here, $\mathcal{L}_{CE}$ represents the cross-entropy loss computed against ground-truth labels, $\mathcal{L}_{KL}$ denotes the KL divergence loss based on the teacher's output, and T signifies the distillation temperature. y_s and y_t denote logits from the student and teacher models, respectively, with $\sigma(\cdot)$ representing the softmax function. In this work, we further introduce the activation similarity map loss $\mathcal{L}_{ASM}$ and the multi-teacher distillation loss adaptive weighting strategy.

The overall loss function can be decomposed into four parts:

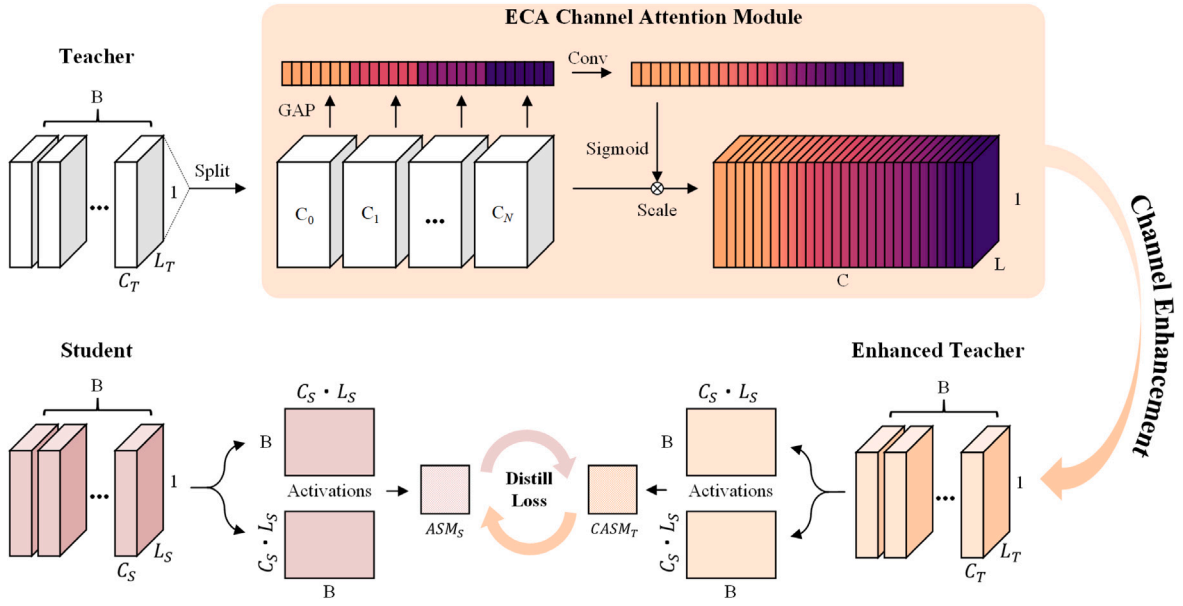

Fig. 4. Distillation of activation similarity map enhanced by ECA module.

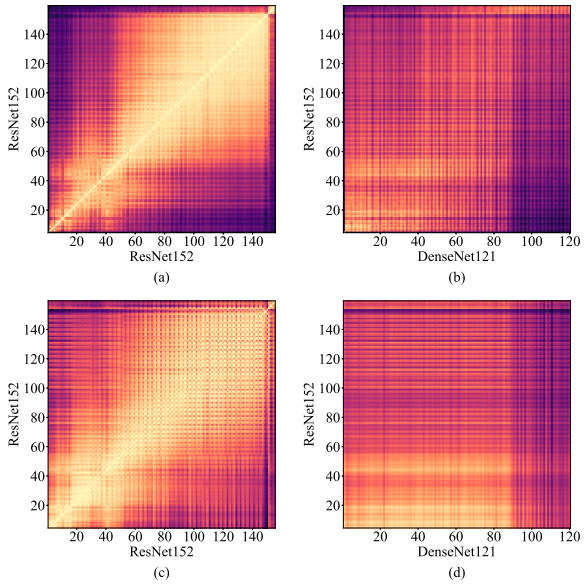

Fig. 5. Similarity heatmaps of intermediate features and activation similarity maps from different layers measured by CKA. (a), (b) measure the similarity of features in the middle layer of the model, (c), (d) measure the similarity of the model activation similarity map. The horizontal and vertical axes are the number of layers of the model.

Algorithm 1: The training process

Input : Training Dataset $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$. A backbone network with four branches B_i, $i = (1, 2, 3, 4)$. Three ECA channel attention modules and a weighting adjustment parameter δ.

Output: Four convergent branching models.

while B_1 *to* B_4 *are not converged* **do**
- Sample a mini-batch $\mathcal{B}$ from Dataset D;
- Forward propagation $\mathcal{B}$ into the network to obtain the predictions q_i, SM_i and ASM_i, $i = 1, 2, 3, 4$;
- Calculate the cross-entropy loss $\mathcal{L}_{CE}$, the KL divergence loss $\mathcal{L}_{KL}$ and the ASM loss $\mathcal{L}_{ASM}$ as Equation (10, 11);
- Calculation of the cosine distance $\mathcal{D}$ of the gradient for different distillation losses with cross-entropy losses as Equation (1);
- Calculate the uncertainty U in the predictions of the teacher model and the student model and the gap between the teacher and student models $\mathcal{G}$ as Equation (3, 4);
- Calculation of weighting factors for different loss elements $\mathcal{W}_{KL}$ and $\mathcal{W}_{ASM}$ as Equation (2, 5, 6);
- Calculate the overall loss L_{total} as Equation (12, 13, 14, 15, 16) and update the parameter;

end while

- Cross-entropy loss of the prediction of the deepest branch B_4 with the true label $\mathbf{y}$:

$$\mathcal{L}_1 = \mathcal{L}_{CE}(q_4, \mathbf{y}). \tag{12}$$

- The sum of the cross-entropy loss of the first three branches' predictions and true label $\mathbf{y}$:

$$\mathcal{L}_2 = \sum_{i=1}^{3} \mathcal{L}_{CE}(q_i, \mathbf{y}). \tag{13}$$

- The sum of the KL divergence loss of all student model predictions and the predictions of the corresponding teacher model:

$$\mathcal{L}_3 = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \mathcal{W}_{KL}^{ij} \mathcal{L}_{KL}(q_i, q_j), \tag{14}$$

- The sum of the ASM loss of all student model predictions and the predictions of the corresponding teacher model:

$$\mathcal{L}_4 = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \mathcal{W}_{ASM}^{ij} \mathcal{L}_{ASM}. \tag{15}$$

Here $\mathcal{W}_{KL}^{ij}$ and $\mathcal{W}_{KL}^{ij}$ are the weights of the KL divergence loss and the ASM loss, respectively, between the ith student model and its jth teacher model, as shown in Eq. (9).

Then the overall loss function can be expressed as:

$$\mathcal{L}_{total} = \underbrace{\alpha\mathcal{L}_1 + (1-\beta)\mathcal{L}_2}_{hard\ label\ loss} + \underbrace{\beta\mathcal{L}_3 + \gamma\mathcal{L}_4}_{distillation\ loss}, \tag{16}$$

Table 1
Evaluation metrics for compression efficiency.

Evaluation metric	Description
MFLOPs	Millions of floating-point operations in the model inference process
Parameters	Number of parameters of the model
Power consumption (PC)	Average energy consumption for model inference on Jetson Orin NX in Wh
Inference time (IT)	Time (ms) required by the model for a single inference of a 12-lead ECG of 10s length on Jetson Orin NX
Compression rate (CR)	Parameter ratios before and after model compression
Acceleration rate (AR)	Ratio of inference time before and after model compression

among them, α, β and γ are hyperparameters set. The overall training process is shown in Algorithm 1.

4. Experiments and results

4.1. Dataset and experimental settings

This study utilizes two arrhythmia databases, CPSC2018 and PTB-XL, to evaluate the effectiveness of the proposed method.

CPSC2018: This database comprises 12-lead ECG records from 6877 patients, featuring a balanced male-to-female ratio and sourced from 11 hospitals. Each record ranges from a few seconds to tens of seconds, sampled at 500 Hz. The database includes nine types of ECG records: atrial fibrillation (AF), first-degree atrioventricular block (IAVB), left bundle branch block (LBBB), right bundle branch block (RBBB), premature atrial contractions (PAC), premature ventricular contractions (PVC), ST segment depression (STD), and ST-segment elevation (STE). Each record is segmented into 10-s ECG signals with a 50% overlap, with the corresponding label assigned from the first label of the record. The database contains a total of 13,285 12-lead heart rhythms, each segment being 10 s long, with abnormal signals used for training and testing purposes.

PTB-XL: This database comprises 21,799 12-lead ECG records, each 10 s long, collected from 18,869 patients. It includes 71 tags covering various forms, rhythms, and diagnostic statements, categorized into five primary categories and 24 subcategories. For this study, we focus on the five primary categories (NORM, CD, MI, HYP, STTC) derived from diagnostic statements as tags. Each record may have one or multiple tags. Records without primary category labels were excluded, resulting in the utilization of 21,388 records.

G12EC: The database contains multi-label ECG records of 10 s in length and 500 Hz sampling rate from 10,344 patients (male: 5551, female: 4793) from Georgia in the southeastern United States. The database contains 27 types of disease labels including atrial fibrillation and other diseases. After excluding disease labels that only appear once, a total of 9407 records are used for training and testing.

The following are the settings for the optimizer and training hyperparameters, as well as the selection of evaluation metrics.

4.1.1. Training setting

In this study, parameters were set to $\alpha = 3$, $\beta = 0.3$ and $\gamma = 3000$. Training encompassed 200 epochs with a batch size of 64. Optimization utilized the SGD optimizer with an initial learning rate of 0.01 and employed a step learning rate scheduler. The dataset was randomly split into 80% for training and 20% for validation.

4.1.2. Evaluation metrics

For classification evaluation, metrics including Accuracy (Acc), and the Area Under the ROC Curve (AUC-score) are employed. These metrics are computed based on four indices: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), determined by the label's positive or negative classification. Where,

$$Acc = (TP + TN)/(TP + TN + FP + FN). \tag{17}$$

The $AUC - score$ represents the area under the ROC curve, which is plotted using various cut-off values. In this curve, Se (Sensitivity) is plotted on the vertical axis and Sp (Specificity) on the horizontal axis.

Six metrics are introduced to comprehensively evaluate compression efficiency in model compression: MFLOPs, parameters, power consumption (PC), inference time (IT), compression rate (CR), and acceleration rate (AR). Detailed descriptions of these metrics can be found in Table 1.

4.2. Performance of self-distillation

In this study, we validate the effectiveness of the proposed method on three arrhythmia databases, CPSC2018, PTB-XL and G12EC, using accuracy and AUC scores to evaluate model performance on their respective test sets. We conduct a comparative analysis across three scenarios: (i) full supervision using cross entropy alone, (ii) incorporation of KL divergence in an SD framework where each branch mimics only the deepest layer, and (iii) addition of feature map MSE loss in the SD setup. To demonstrate the method's applicability, we validate it on ResNet, xResNet, MobileNet V2, and ShuffleNet V2 architectures. Details are provided in Table 2.

As shown in Table 2, all distilled models outperformed the model trained solely with $\mathcal{L}_{CE}$. However, upon adding the feature map imitation loss $\mathcal{L}_{fea}$ to the SD framework, some model branches exhibited weaker performance compared to those using only $\mathcal{L}_{KL}$. This suggests that directly aligning the feature maps of teacher–student models during online distillation may lead to instability in the optimization process. Subsequently, employing the multi-teacher adaptive fusion SD framework resulted in improved performance across all model branches. Specifically, for the CPSC2018 database, the B_3 branch of ResNet34 achieved an accuracy of 87.05%. For the PTB-XL database, the B_3 branch of MobileNet V2 achieved an AUC score of 0.922. For the G12EC database, the B_4 branch of ResNet34 achieved an AUC score of 0.908.

Furthermore, it is noteworthy that shallow branches tend to exhibit greater performance gains compared to deep branches. For instance, in the CPSC2018 database, ResNet18 shows a performance improvement of 0.41% for the B_4 branch, whereas the B_1 branch improves by 2.1%. Similarly, xResNet18 demonstrates a 1.21% improvement for the B_4 branch and a 2.22% improvement for the B_1 branch, while ShuffleNet V2 shows a 1.74% improvement for the B_4 branch and a notable 3.84% improvement for the B_1 branch. This observation suggests that shallow branches, characterized by simpler representations, are more effective in learning relevant information and are easier to optimize compared to deeper branches with more complex representations.

4.3. Comparison with state-of-the-arts

In this section, we compare the proposed method with other state-of-the-art algorithms. Table 3 presents the performance differences between various methods and the proposed SD framework across different teacher–student model architectures.

In Table 3, we employed ResNet152 as the teacher model to assess multiple student models on the CPSC2018, PTB-XL and G12EC databases. Compared to offline distillation methods such as OKD [1], FitNet [11], and AT [28], as well as online distillation methods like DML [6] and BYOT [10], our proposed multi-teacher adaptive fusion SD approach consistently outperformed on both datasets. Notably, ResNet34 achieved the highest accuracy of 86.98% on CPSC2018,

Table 2
Comparison of performance of each classifier in self-distillation on *CPSC2018*, *PTB-XL* and *G12EC*.

Net	$\mathcal{L}_{CE}$	$\mathcal{L}_{KL}$	$\mathcal{L}_{fea}$	$\mathcal{L}_{asm}$	*CPSC2018*				*PTB-XL*				*G12EC*			
					B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4
RN18	✓				84.23	85.28	85.92	86.11	0.894	0.904	0.906	0.905	0.871	0.883	0.885	0.886
	✓	✓			85.20	85.92	86.07	86.03	0.909	0.916	0.918	0.918	0.883	0.888	0.892	0.892
	✓	✓	✓		85.99	86.37	86.22	86.22	0.908	0.914	0.917	0.916	0.877	0.881	0.886	0.884
	✓	✓		✓	**86.33**	**86.69**	**86.52**	**86.52**	**0.912**	**0.918**	**0.921**	**0.921**	**0.891**	**0.897**	**0.901**	**0.904**
RN34	✓				85.24	85.58	86.48	85.96	0.897	0.906	0.907	0.909	0.874	0.883	0.894	0.882
	✓	✓			86.11	86.30	86.63	86.37	0.909	0.915	0.918	0.917	0.881	0.886	0.890	0.890
	✓	✓	✓		86.07	86.45	86.56	86.60	0.908	0.915	0.916	0.916	0.871	0.879	0.886	0.885
	✓	✓		✓	**86.48**	**86.86**	**87.05**	**86.98**	**0.914**	**0.920**	**0.920**	**0.921**	**0.894**	**0.899**	**0.902**	**0.908**
xRN18	✓				83.74	84.83	85.32	85.50	0.887	0.889	0.897	0.896	0.856	0.876	0.883	0.887
	✓	✓			85.81	85.92	86.11	85.99	0.905	0.909	0.911	0.911	0.881	0.886	0.892	0.892
	✓	✓	✓		85.05	85.39	85.43	85.32	0.899	0.904	0.906	0.906	0.871	0.874	0.878	0.879
	✓	✓		✓	**85.96**	**86.14**	**86.48**	**86.71**	**0.913**	**0.917**	**0.919**	**0.919**	**0.885**	**0.890**	**0.895**	**0.899**
xRN34	✓				84.38	85.54	85.84	85.58	0.888	0.895	0.896	0.900	0.868	0.874	0.892	0.896
	✓	✓			85.35	85.62	85.73	85.54	0.906	0.911	0.915	0.915	0.875	0.883	0.889	0.887
	✓	✓	✓		85.36	85.62	85.81	85.66	0.898	0.902	0.905	0.904	0.875	0.878	0.881	0.881
	✓	✓		✓	**85.73**	**86.33**	**86.78**	**86.82**	**0.917**	**0.920**	**0.920**	**0.921**	**0.889**	**0.896**	**0.902**	**0.905**
MNV2	✓				82.57	83.66	84.68	84.87	0.892	0.899	0.899	0.904	0.845	0.849	0.851	0.885
	✓	✓			83.28	84.79	84.94	84.78	0.898	0.905	0.907	0.906	0.849	0.856	0.859	0.861
	✓	✓	✓		82.91	84.41	84.34	84.26	0.895	0.904	0.905	0.904	0.847	0.863	0.863	0.862
	✓	✓		✓	**83.89**	**85.43**	**85.92**	**86.41**	**0.911**	**0.921**	**0.922**	**0.920**	**0.864**	**0.879**	**0.887**	**0.898**
SNV2	✓				79.71	82.12	83.92	83.28	0.889	0.892	0.895	0.893	0.862	0.860	0.867	0.877
	✓	✓			80.73	83.43	84.19	84.11	0.897	0.905	0.908	0.906	0.868	0.872	0.873	0.871
	✓	✓	✓		80.69	82.53	83.85	83.25	0.899	0.905	0.908	0.906	0.867	0.869	0.874	0.872
	✓	✓		✓	**83.55**	**85.05**	**85.13**	**85.02**	**0.911**	**0.914**	**0.912**	**0.912**	**0.873**	**0.886**	**0.891**	**0.888**

RN: ResNet, xRN: xResNet, MNV2: MobileNet V2, SNV2: ShuffleNet V2.

Table 3
Comparison of performance on different teacher–student model architectures on *CPSC2018*, *PTB-XL* and *G12EC*.

Database	Teacher	Student	Baseline	OKD [1]	FitNet [11]	AT [28]	DML [6]	BYOT [10]	This work
CPSC2018	ResNet152	ResNet18	86.11	86.33	86.33	86.26	85.13	86.22	**86.52**
		ResNet18_S	84.23	85.99	84.64	84.94	86.26	85.99	**86.33**
		ResNet34	85.96	86.56	86.03	86.30	86.41	86.60	**86.98**
		ResNet34_S	85.24	86.30	85.54	85.77	**86.71**	86.07	86.48
		MobileNet V2	84.87	85.77	85.58	85.88	85.54	84.26	**86.41**
		ShuffleNet V2	83.28	84.49	83.66	84.19	84.38	83.25	**85.02**
PTB-XL		ResNet18	0.905	0.913	0.913	0.913	0.910	0.916	**0.921**
		ResNet18_S	0.894	0.910	0.911	0.909	0.913	0.908	**0.912**
		ResNet34	0.909	0.913	0.912	0.913	0.917	0.916	**0.921**
		ResNet34_S	0.897	0.912	0.908	0.911	**0.916**	0.908	0.914
		MobileNet V2	0.904	0.908	0.912	0.909	0.909	0.904	**0.920**
		ShuffleNet V2	0.893	0.906	0.905	0.904	0.905	0.906	**0.912**
G12EC		ResNet18	0.886	0.899	0.893	0.899	0.903	0.884	**0.904**
		ResNet18_S	0.871	0.888	0.879	0.889	0.874	0.877	**0.892**
		ResNet34	0.882	0.899	0.891	0.901	0.906	0.885	**0.908**
		ResNet34_S	0.874	0.894	0.889	0.893	0.884	0.871	**0.895**
		MobileNet V2	0.885	0.890	0.896	0.895	0.887	0.861	**0.898**
		ShuffleNet V2	0.877	0.891	0.884	**0.891**	0.889	0.872	0.889

ResNet18 and ResNet34 achieved the highest AUC score of 0.921 on PTB-XL, and ResNet34 achieved the highest AUC score of 0.908 on G12EC.

Our method exhibits significant advantages when the student model shares architectural similarities with the teacher model, such as ResNet18 or ResNet34. Even when the student model differs in architecture from ResNet152, as with MobileNet V2 and ShuffleNet V2, our approach still demonstrated superior performance. Unlike approaches relying on larger-scale ResNet152 as a teacher model, our method achieves representation transfer solely by leveraging knowledge across different levels of its hierarchy, eliminating the need for additional teacher models. Furthermore, by utilizing activation similarity graphs as knowledge representations and employing a multi-teacher fusion strategy, our approach addresses knowledge disparity across model levels, leading to more effective knowledge migration and model compression.

4.4. Compression efficiency evaluation

In this subsection, we evaluate the model efficiency post-compression using the NVIDIA Jetson Orin NX edge computing chip. We use several metrics to comprehensively assess compression efficiency, including MFLOPs, parameters, PC, IT, CR, and AR.

Initially, we compare the PC of the uncompressed model with that of the minimal model B_1, compressed using our proposed method, on an inference dataset comprising 13,285 12-lead ECG records from the CPSC2018 database. The specific PC comparison is illustrated in Fig. 6, where green denotes the inference PC of the compressed model (denoted as Model_S), and red represents the inference PC of the

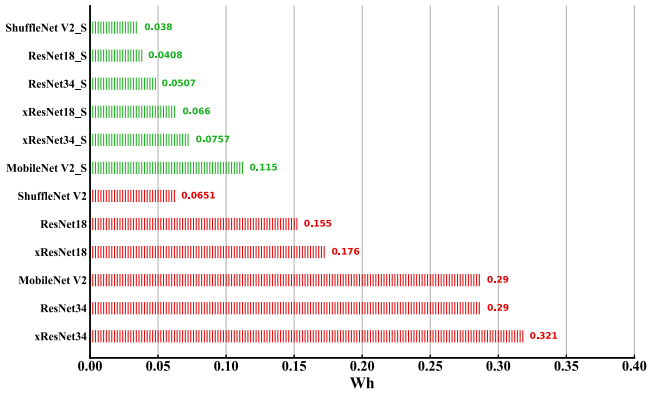

Fig. 6. Comparison of inference PC on Jetson Orin NX before and after model compression, with the PC of the model compressed to B_1 in green and the PC of the uncompressed model in red.

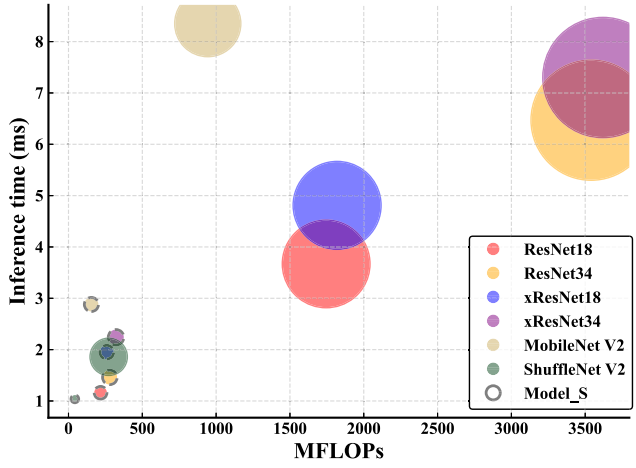

Fig. 7. Comparison of MFLOPs, IT, and parameters before and after model compression (B_1). Larger scatters indicate larger parameter counts.

uncompressed model. The PC of the compressed models is significantly lower than that of the original models, with ShuffleNet V2_S achieving the lowest PC of 0.038 Wh. ResNet34 shows the highest PC compression rate, reducing from 0.29 Wh before compression to 0.0507 Wh after compression, a compression ratio of 5.71 times. This demonstrates the effective reduction in inference PC achieved by our proposed compression method.

Subsequently, we compared the MFLOPs, IT, and parameters before and after model compression (B_1). A visual comparison of the combined three metrics is shown in Fig. 7. A larger scatter indicates a larger number of parameters in the model. As can be seen from Fig. 7, all the compression models have significant effect improvement in parametric quantity, IT, and MFLOPs metrics. Among them, ShuffleNet V2_S has the shortest single IT, the least MFLOPs, and the number of parameters. Also, Table 4 lists the specific MFLOPs, model parameters, IT, CR, and AR. As can be seen from the table, the MFLOPs were all reduced significantly after model compression, and the number of parameters were all below 150K, among which the ShuffleNet V2_S had the lowest number of parameters of 29.23K, and the ResNet34_S had the highest CR of 68.75 times. Meanwhile, the inference efficiency of the compressed models was significantly improved, in which ShuffleNet V2_S took only 1.04 ms for a single inference of 12-lead ECG with a length of 10 s, and the AR of the remaining models reached 2–3 times. It is further shown that the proposed method can effectively reduce the computational complexity and inference PC of the model.

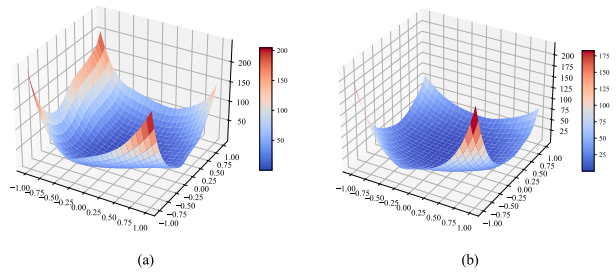

Fig. 8. Loss landscape around local minima. (a) Direct alignment of feature maps for single-teacher original SD. (b) Proposed MTSD.

4.5. Distillation stability and generalization

In this subsection, we assess the stability of our proposed method during the distillation process using the loss landscape visualization technique [53]. Fig. 8 illustrates the landscape of losses around the local minima of the model. Observing the figure, the region surrounding the local minima appears flatter with our proposed method, indicating less sensitivity of loss values to parameter changes. This robustness suggests reduced susceptibility to outliers during model training and a more stable optimization process. Furthermore, flatter local minima are indicative of improved generalization performance.

Activation similarity maps and loss weighting based on model gaps effectively contribute to enhancing the robustness of the distillation process. This approach mitigates to some extent the training instability typically associated with the direct alignment of feature maps in online distillation.

5. Discussion

5.1. Performance analysis

In this subsection, we analyze the performance of the proposed method and compare it with other methods.

Table 2 shows the comparison of the proposed multi-teacher adaptive fusion strategy with the three cases of direct training, distillation using only $\mathcal{L}_{KL}$ within a multi-branch SD framework, and distillation using both $\mathcal{L}_{KL}$ and $\mathcal{L}_{fea}$ within a multi-branch SD framework. When $\mathcal{L}_{fea}$ loss is introduced to the distillation process, the performance of most of the model branches shows a decreasing trend. This indicates that the pixel point changes of the feature map are more sensitive to the online distillation process and there is information redundancy, and the direct alignment of a large number of pixel points will lead to the instability of the distillation process and reduce the robustness of the model. The proposed method expands the single-teacher paradigm to multi-teacher based on SD and introduces a supervised loss adaptive weighting strategy based on information correctness and acceptability, which effectively facilitates information sharing at different levels in the backbone network. Meanwhile, the ECA module is utilized to enhance the mutual information and dependency relationship between channels and categories, and achieve more efficient and stable representation migration through channel activation similarity maps instead of feature maps, which achieves a better distillation effect.

In Table 3 the proposed method is compared with several state-of-the-art distillation methods under different teacher–student model architectures. The proposed method achieved relatively favorable results on CPSC2018, PTB-XL and G12EC databases. We believe that there is a large discrepancy between the deep and shallow features of the model, as shown in Fig. 5(a), (b), and that the discrepancy hinders the propagation of knowledge from the deep model to the shallow model. Even though ResNet152 possesses stronger knowledge representations as a large-scale model, it is difficult to fully migrate its knowledge

Table 4
Specific details of MFLOPs, Parameters, CR, IT, and AR before and after model compression.

Network	MFLOPs	Params	CR	IT	AR
ResNet18	1745.24	3.85M	–	3.67	–
ResNet18_S	**217.28**	**80.33K**	×47.92	**1.16**	×3.16
ResNet34	3540	7.23M	–	6.47	–
ResNet34_S	**280**	**105.16K**	×68.75	**1.46**	×4.43
xResNet18	1819.48	3.87M	–	4.81	–
xResNet18_S	**259.2**	**94.72K**	×40.85	**1.95**	×2.46
xResNet34	3620	7.24M	–	7.30	–
xResNet34_S	**321.92**	**119.56K**	×60.55	**2.24**	×3.25
MobileNet V2	943.22	2.19M	–	8.35	–
MobileNet V2_S	**155.06**	**98.66K**	×22.19	**2.88**	×2.89
ShuffleNet V2	272.22	697.53K	–	1.86	–
ShuffleNet V2_S	**43.04**	**29.23K**	×23.86	**1.04**	×1.78

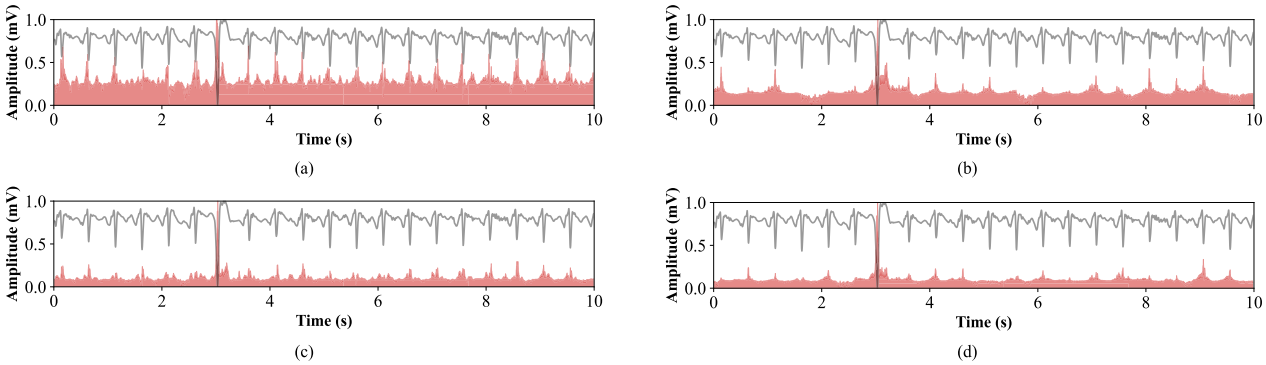

Fig. 9. The Relevance-CAM attention distribution of the model for PVC ECG signals. (a) Vanilla ResNet18 B_1 branch, (b) vanilla ResNet18, (c) the proposed MTSD ResNet18 B_1 branch, (d) the proposed MTSD ResNet18.

representations to the student model for compact student models such as ResNet18 and ResNet18_S. In contrast, the proposed method does not require additional large-scale teacher models and can achieve knowledge sharing at different levels by itself. Meanwhile, the activation similarity maps (Fig. 5(c), (d)) with small differences between the deep and shallow layers of the model are used for the representation migration, which makes the knowledge easier to disseminate to some extent.

5.2. Model efficiency analysis

The purpose of the model compression proposed in this work is to deploy the model in wearable ECG devices with computational resource constraints; thus, evaluating the practical usability of the compressed model from multiple perspectives is crucial.

Fig. 6 depicts the computational power consumption before and after model compression, measured on the NVIDIA Jetson Orin NX edge computing chip. Following compression via the proposed multi-teacher adaptive weighting SD framework, all student models exhibit a notable reduction in power consumption, averaging 4–6 times less compared to the original teacher model. This reduction is suitable for integration into small wearable ECG devices.

Furthermore, we conducted a detailed comparison of MFLOPs, parameters, CR, IT, and AR before and after model compression, as illustrated in Fig. 7 and detailed in Table 4. Fig. 7 visually contrasts the efficiency gains achieved by the student models compared to their respective teacher models. All student models exhibit significantly reduced parameters, MFLOPs, and IT, highlighting the effectiveness of the compression process. Notably, ShuffleNet V2, inherently designed for lightweight applications, demonstrates superior computational efficiency post-compression. With fewer than 30K parameters, it is well-suited for deployment in miniaturized wearable devices, achieving inference times close to 1 ms, and meeting the real-time demands of arrhythmia monitoring.

Moreover, it is observed that the CR and AR of the student models compressed using the proposed multi-teacher adaptive fusion SD framework tend to increase with the number of parameters in the original teacher models and their corresponding inference times. This trend arises from the exponential increase in convolution channels as model depth increases to enhance representations, resulting in larger gaps between parameter volumes in shallow and deeper layers, thereby boosting compression and acceleration rates.

5.3. Case study

In this subsection, we employ Relevance-CAM [54] to visualize the model's attention distribution towards input signals, assessing whether the student model achieves comparable representational capabilities to the teacher model. Fig. 9 illustrates the attention distribution for PVC ECG signals. It is observed that the model's attention predominantly focuses on the R-wave. Specifically, the PVC disease feature is most pronounced during the 7th heartbeat, with deeper layers showing heightened attention concentration towards this beat, while other parts of the R-wave receive weaker attention, as depicted in Fig. 9(a), (c). Following optimization using the MTSD framework, Fig. 9(d) reveals that ResNet18's attention distribution is notably concentrated in the onset region. Moreover, the attention distribution of the ResNet18 B_1 branch closely aligns with that of the teacher model, surpassing vanilla ResNet18 performance, as shown in Fig. 9(b). This highlights the efficacy of the proposed SD framework in transferring high-level representations from the teacher model.

5.4. Limitation

Although our framework has achieved remarkable results in arrhythmia monitoring, it still has some limitations. First, due to the complex model structure and multiple gradient extractions during training,

the computational cost is high and the training time is long, which may limit its application in large-scale datasets or resource-constrained environments. Second, our method performs well on arrhythmias, but its generalization ability in other tasks or domains has not been fully verified. In cross-domain tasks where data distribution varies greatly, the performance of the model may degrade. In addition, the model is sensitive to hyperparameter settings and may require more parameter tuning. Future research directions may be to make the student model adaptively evaluate, review, and summarize the teacher's knowledge base and the student's self-reflection, so that the learning task is closer to the process from easy to difficult, and verify the applicability of the framework in other tasks.

6. Conclusion

This paper introduces a multi-teacher self-distillation framework for compressing large-scale arrhythmia monitoring models intended for deployment in wearable ECG devices. The framework systematically integrates considerations of teacher supervision correctness and acceptability, leveraging activation similarity maps to effectively guide student model optimization. This approach enhances distillation process stability and model generalization, presenting a novel approach to multi-teacher online distillation.

Overall, the learning process of knowledge distillation requires two-way communication, and its effectiveness depends not only on how the teacher model extracts the minimum sufficient representation but also on the student's ability to transfer representations. The student model needs to effectively improve its performance from the teacher's distribution and self-reflection. Therefore, in future work, introducing curriculum learning to gradually train the model from simple to complex in an adaptive manner, allowing students to review and evaluate the knowledge they have learned, thereby improving learning efficiency and final performance is a possible development direction.

CRediT authorship contribution statement

Zhongyu Wang: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Caiyun Ma:** Writing – review & editing, Methodology, Data curation. **Shuo Zhang:** Methodology, Conceptualization. **Minghui Zhao:** Investigation, Data curation. **Yu Liu:** Methodology, Conceptualization. **Lina Zhao:** Data curation, Conceptualization. **Tingting Zhu:** Supervision, Methodology, Conceptualization. **Jianqing Li:** Supervision, Resources, Project administration, Funding acquisition. **Chengyu Liu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition.

Funding

This work was supported in part by the National Natural Science Foundation of China (62171123, 62071241, 62201144, and 62211530112), the China Postdoctoral Science Foundation, China (2024M760444) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX24_0382).

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.

Data availability

Data will be made available on request.

References

[1] Geoffrey E. Hinton, Oriol Vinyals, J. Michael Dean, Distilling the knowledge in a neural network, 2015, ArXiv: Machine Learning, ArXiv: Machine Learning.
[2] Jianping Gou, Baosheng Yu, Stephen J. Maybank, Dacheng Tao, Knowledge distillation: A survey, Int. J. Comput. Vis. (2021) 1789–1819.
[3] Yonglong Tian, Dilip Krishnan, Phillip Isola, Contrastive representation distillation, Learn. Learn. (2019).
[4] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, Dacheng Tao, Learning student networks via feature embedding, IEEE Trans. Neural Netw. Learn. Syst. (2021) 25–35.
[5] Peiye Liu, Wu Liu, Huadong Ma, Tao Mei, Mingoo Seok, KTAN: Knowledge transfer adversarial network, 2018, Cornell University - ArXiv, Cornell University - ArXiv.
[6] Ying Zhang, Tao Xiang, Timothy M. Hospedales, Huchuan Lu, Deep mutual learning, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[7] Inseop Chung, SeongUk Park, Jangho Kim, Nojun Kwak, Feature-map-level online adversarial knowledge distillation, 2020, Cornell University - ArXiv, Cornell University - ArXiv.
[8] Jianping Gou, Xiangshuo Xiong, Baosheng Yu, Lan Du, Yibing Zhan, Dacheng Tao, Multi-target knowledge distillation via student self-reflection, Int. J. Comput. Vis. (2023) 1857–1874.
[9] Zhilu Zhang, Mert R. Sabuncu, Self-distillation as instance-specific label smoothing, 2020, ArXiv: Learning, ArXiv: Learning.
[10] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, Kaisheng Ma, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019.
[11] Adriana Romero, Nicolas Ballas, SamiraEbrahimi Kahou, Antoine Chassang, Carlo Gatta, Yoshua Bengio, FitNets: Hints for thin deep nets, 2014, ArXiv: Learning, ArXiv: Learning.
[12] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, Hassan Ghasemzadeh, Improved knowledge distillation via teacher assistant, Proc. AAAI Conf. Artif. Intell. (2020) 5191–5198.
[13] Luyao Shi, Ning Jiang, Jialiang Tang, Xinlei Huang, Correlation guided multi-teacher knowledge distillation, in: Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, Chaojie Li (Eds.), Neural Information Processing, Springer Nature Singapore, Singapore, 2024, pp. 562–574.
[14] Zhiyuan Li, Huawei Liang, Hanqi Wang, Mingzhuo Zhao, Jian Wang, Xiaokun Zheng, MKD-cooper: Cooperative 3D object detection for autonomous driving via multi-teacher knowledge distillation, IEEE Trans. Intell. Veh. 9 (1) (2024) 1490–1500.
[15] Hailin Zhang, Defang Chen, Can Wang, Confidence-aware multi-teacher knowledge distillation, 2022.
[16] Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, Deng Cai, Reducing the teacher-student gap via spherical knowledge disitllation, 2021.
[17] Jun Rao, Xv Meng, Liang Ding, Shuhan Qi, Dacheng Tao, Parameter-efficient and student-friendly knowledge distillation, 2022.
[18] Rui Hu, Jie Chen, Li Zhou, A transformer-based deep neural network for arrhythmia detection using continuous ECG signals, Comput. Biol. Med. 144 (2022).
[19] Qihang Yao, Ruxin Wang, Xiaomao Fan, Jikui Liu, Ye Li, Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network, Inf. Fusion (2020) 174–182.
[20] Georgios Petmezas, Kostas Haris, Leandros Stefanopoulos, Vassilis Kilintzis, Andreas Tzavelis, John A Rogers, Aggelos K Katsaggelos, Nicos Maglaveras, Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets, Biomed. Signal Process. Control. (2021) 102194.
[21] Sahar Soltanieh, Javad Hashemi, Ali Etemad, In-distribution and out-of-distribution self-supervised ECG representation learning for arrhythmia detection, IEEE J. Biomed. Health Inform. 28 (2) (2024) 789–800.
[22] Wissal Midani, Wael Ouarda, Mounir Ben Ayed, DeepArr: An investigative tool for arrhythmia detection using a contextual deep neural network from electrocardiograms (ECG) signals, Biomed. Signal Process. Control 85 (2023).
[23] Chun Ding, Shenglun Wang, Xiaopeng Jin, Zhaoze Wang, Junsong Wang, A novel transformer-based ECG dimensionality reduction stacked auto-encoders for arrhythmia beat detection, Med. Phys. 50 (9) (2023) 5897–5912.
[24] Yaqoob Ansari, Omar Mourad, Khalid Qaraqe, Erchin Serpedin, Deep learning for ECG arrhythmia detection and classification: an overview of progress for period 2017–2023, Front. Physiol. 14 (2023).
[25] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M.M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P.S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira, Thomas B. Schön, Antonio Luiz P. Ribeiro, Automatic diagnosis of the 12-lead ECG using a deep neural network, Nat. Commun. (2020).
[26] Nitish Katal, Saurav Gupta, Pankaj Verma, Bhisham Sharma, Deep-learning-based arrhythmia detection using ECG signals: A comparative study and performance evaluation, Diagnostics 13 (24) (2023).

[27] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. (2013) 1798–1828.
[28] Sergey Zagoruyko, Nikos Komodakis, Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer, Cent. Pour Commun. Sci. Directe - HAL - Univ. Paris Descartes, Cent. Pour Commun. Sci. Directe - HAL - Univ. Paris Descartes (2016).
[29] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, Qun Liu, ALP-KD: Attention-based layer projection for knowledge distillation, Proc. AAAI Conf. Artif. Intell. (2022) 13657–13665.
[30] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, Chun Chen, Cross-layer distillation with semantic calibration, Proc. AAAI Conf. Artif. Intell. (2022) 7028–7036.
[31] Yuzhu Wang, Lechao Cheng, Manni Duan, Yongheng Wang, Zunlei Feng, Shu Kong, Improving knowledge distillation via regularizing feature direction and norm, in: Aleˇ s Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, Gül Varol (Eds.), Computer Vision – ECCV 2024, Springer Nature Switzerland, Cham, 2025, pp. 20–37.
[32] Wonpyo Park, Dongju Kim, Yan Lu, Minsu Cho, Relational knowledge distillation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019.
[33] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, Yu Liu, Correlation congruence for knowledge distillation, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019.
[34] Zhiqiang Shen, Marios Savvides, MEAL V2: Boosting vanilla ResNet-50 to 80 without tricks, 2020.
[35] Zhiyuan Li, Huawei Liang, Hanqi Wang, Mingzhuo Zhao, Jian Wang, Xiaokun Zheng, H Li, X Wang, MKD-Cooper: Cooperative 3D Object Detection for Autonomous Driving via Multi-teacher Knowledge Distillation.
[36] Lechao Cheng, Kaifeng Chen, Jiyang Li, Shengeng Tang, Shufei Zhang, Meng Wang, Dataset distillers are good label denoisers in the wild, 2024.
[37] Tianli Zhang, Mengqi Xue, Jiangtao Zhang, Haofei Zhang, Yu Wang, Lechao Cheng, Jie Song, Mingli Song, Generalization matters: Loss minima flattening via parameter hybridization for efficient online knowledge distillation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 20176–20185.
[38] Chaowei Fang, Qian Wang, Lechao Cheng, Zhifan Gao, Chengwei Pan, Zhen Cao, Zhaohui Zheng, Dingwen Zhang, Reliable mutual distillation for medical image segmentation under imperfect annotations, IEEE Trans. Med. Imaging 42 (6) (2023) 1720–1734.
[39] Ke Feng, J.C. Ji, Yongchao Zhang, Qing Ni, Zheng Liu, Michael Beer, Digital twin-driven intelligent assessment of gear surface degradation, Mech. Syst. Signal Process. 186 (2023) 109896.
[40] Sheng Li, Ke Feng, Yadong Xu, Yongbo Li, Qing Ni, Ke Zhang, Yulin Wang, Weiping Ding, Cross-modal zero-sample diagnosis framework utilizing non-contact sensing data fusion, Inf. Fusion 110 (2024) 102453.
[41] Sheng Li, Jinchen Ji, Ke Feng, Ke Zhang, Qing Ni, Yadong Xu, Composite neuro-fuzzy system-guided cross-modal zero-sample diagnostic framework using multisource heterogeneous noncontact sensing data, IEEE Trans. Fuzzy Syst. 33 (1) (2025) 302–313.
[42] Saman Taheri, Amirhossein Ahmadi, Behnam Mohammadi-Ivatloo, Somayeh Asadi, Fault detection diagnostic for HVAC systems via deep learning algorithms, Energy Build. 250 (2021) 111275.
[43] Xiaokang Zhou, Xuzhe Zheng, Xuesong Cui, Jiashuai Shi, Wei Liang, Zheng Yan, Laurence T. Yang, Shohei Shimizu, Kevin I-Kai Wang, Digital twin enhanced federated reinforcement learning with lightweight knowledge distillation in mobile networks, IEEE J. Sel. Areas Commun. 41 (10) (2023) 3191–3211.
[44] Xiucheng Wang, Nan Cheng, Longfei Ma, Ruijin Sun, Rong Chai, Ning Lu, Digital twin-assisted knowledge distillation framework for heterogeneous federated learning, China Commun. 20 (2) (2023) 61–78.
[45] Saurabh Gupta, Judy Hoffman, Jitendra Malik, Cross modal distillation for supervision transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[46] Yufeng Jin, Guosheng Hu, Haonan Chen, Duoqian Miao, Liang Hu, Cairong Zhao, Cross-modal distillation for speaker recognition, 2023, pp. 12977–12985, 37.
[47] Yadong Xu, Ke Feng, Xiaoan Yan, Xin Sheng, Beibei Sun, Zheng Liu, Ruqiang Yan, Cross-modal fusion convolutional neural networks with online soft-label training strategy for mechanical fault diagnosis, IEEE Trans. Ind. Inform. 20 (1) (2024) 73–84.
[48] Pengguang Chen, Shu Liu, Hengshuang Zhao, Jiaya Jia, Distilling knowledge via knowledge review, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[49] Yichen Zhu, Yi Wang, Student customized knowledge distillation: Bridging the gap between student and teacher, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021.
[50] Jang Hyun Cho, Bharath Hariharan, On the efficacy of knowledge distillation, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019.
[51] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, Qinghua Hu, ECA-net: Efficient channel attention for deep convolutional neural networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.
[52] Simon Kornblith, Mohammad Norouzi, Honglak Lee, GeoffreyE. Hinton, Similarity of neural network representations revisited, Int. Conf. Mach. Learn. Int. Conf. Mach. Learn. (2019).
[53] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, Tom Goldstein, Visualizing the loss landscape of neural nets, Neural Inf. Process. Syst. Neural Inf. Process. Syst. (2018).
[54] Jeong Ryong Lee, Sewon Kim, Inyong Park, Taejoon Eo, Dosik Hwang, Relevance-CAM: Your model already knows where to look, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021.