Semi-Supervised Distillation Network Toward Noise-Resistant Medical Image Classification

Yifan Peng[1,#], Yong Hu[2,#], Shen Zhang[3,#], Zongsheng Li[4],
Weifan Xu[4] and Xiangtong Du[3,*]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China

[2] School of Electrical and Electronic Engineering, Anhui Institute of Information Technology, Wuhu, 241000, China

[3] School of Medical Imaging, Xuzhou Medical University, Xuzhou, 221004, China

[4] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China

[#] These authors contributed equally to this work

Revista Internacional
Métodos numéricos
para cálculo y diseño en ingeniería
RIMNI
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH
In cooperation with
CIMNE

INFORMATION

Keywords:
Semi-supervised learning
adaptive distillation
noisy-robust classification

DOI: 10.23967/j.rimni.2025.10.73661

SCIPEDIA

Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH In cooperation with CIMNE

Semi-Supervised Distillation Network Toward Noise-Resistant Medical Image Classification

Yifan Peng[1,#], Yong Hu[2,#], Shen Zhang[3,#], Zongsheng Li[4], Weifan Xu[4] and Xiangtong Du[3,*]

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China

[2]School of Electrical and Electronic Engineering, Anhui Institute of Information Technology, Wuhu, 241000, China

[3]School of Medical Imaging, Xuzhou Medical University, Xuzhou, 221004, China

[4]School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China

[#]These authors contributed equally to this work

ABSTRACT

Deep learning (DL)-based models have demonstrated significant advancements in medical image classification. However, the scarcity of accurately labeled training data and the prevalence of label noise remain critical obstacles to further performance improvements. Although semi-supervised learning and learning with noisy labels (LNL) methods each offer partial remedies, their independent application often leads to suboptimal outcomes. To address this, we propose a unified framework termed the Semi-supervised Adaptive Distillation Network (SAD-Net), which synergistically integrates semi-supervised training with noise-robust distillation. SAD-Net consists of three core components. First, a semi-supervised learning framework is employed to generate pseudo-labels from unlabeled data, thereby augmenting the training set. Subsequently, a Noise Filtering Module (NF-Module) is introduced, which combines a Convolutional Neural Network (CNN) with an Improved Fuzzy C-Means (IFCM) algorithm using a weighted average distance metric. This module produces weighted soft labels from both models and filters out noisy samples based on a confidence threshold. Finally, an Adaptive Weighted Distillation Module (AWD-Module) is designed, incorporating the IFCM along with two CNN architectures. It processes the high-confidence samples selected by the NF-Module and performs classification via dynamically weighted soft labels derived from all three models. Extensive experiments on two medical image datasets show that SAD-Net achieves superior performance compared to state-of-the-art semi-supervised methods, attaining the highest scores in accuracy, sensitivity, specificity, and F1-score. Moreover, it outperforms leading LNL approaches across all evaluated metrics. These results validate the efficacy of the proposed SAD-Net in simultaneously mitigating the problems of limited labeled data and noisy labels in medical image classification.

OPEN ACCESS

Received: 23/09/2025

Accepted: 03/12/2025

Published: 20/03/2026

DOI
10.23967/j.rimni.2025.10.73661

Keywords:
Semi-supervised learning
adaptive distillation
noisy-robust classification

*Correspondence: Xiangtong Du (duxiangtong@xzhmu.edu.cn).

1 Introduction

Since the emergence of deep learning (DL), convolutional neural networks (CNNs) and related architectures have transformed medical image recognition and diagnosis. DL models have improved diverse tasks such as cancer lesion grading [1], disease detection [2], cell classification [3], and other classification challenges [4]. These advances utilise a range of imaging modalities, including CT for structural scans [5], Ultrasound (US) for tissue visualisation [6], X-rays for radiographic assessment [7], and MRI for anatomical and functional analysis [8], among others [9].

Despite these advances, CNN-based classifiers can overfit when trained on limited annotated data. High-quality, labelled medical images are scarce, which makes generalisation difficult [10]. Acquiring medical images needs specialised equipment, and many conditions are rare [11]. Privacy regulations can also restrict data access [12]. Even available datasets may contain inconsistent or unreliable labels. Furthermore, diseases with similar appearances can lead to variability, even among experts.

Limited data and label noise remain major challenges in medical imaging. Semi-supervised learning (SSL) methods, such as regularisation consistency [13,14] and entropy minimisation [15–17], use unlabeled samples to generate pseudo-labels. These methods help with limited annotations but assume clean labels, making them vulnerable to noise, confirmation bias, and degraded performance.

Learning with noisy labels (LNL) employs error regularisation [18–20], label correction [21,22], and hybrid strategies [23,24] to manage mislabeled data. However, most LNL methods underutilise unlabeled data and perform poorly with little annotated data. A unified strategy addressing both limited labels and label noise is needed.

We propose the Semi-supervised Adaptive Distillation Network (SAD-Net) to address these challenges in noise-robust classification. SAD-Net first uses a semi-supervised pipeline to generate pseudo-labels from unlabeled data, expanding the training set. To handle potential noise from pseudo-labels, we introduce a Noise Filtering Module (NF-Module) that combines a CNN with an Improved Fuzzy C-Means (IFCM) clustering algorithm using a weighted distance metric. FCM is robust to label noise and provides soft cluster assignments via membership values. The enhanced IFCM further improves reliability. The NF-Module builds consensus using soft labels from both the CNN and the IFCM, filtering samples based on a confidence threshold.

Samples with confidence scores above the threshold—whether originally labelled or pseudo-labelled—are kept to create a refined data set. Those below the threshold are either discarded or reused in additional training cycles. This curated data set is then used by the Adaptive Weighted Distillation Module (AWD-Module), which integrates the IFCM, a deep CNN, and a lightweight CNN to generate dynamically weighted soft labels for final classification.

We evaluated SAD-Net using two public medical image datasets, BloodMnist and OrganaMnist. Results show SAD-Net successfully addresses both annotation scarcity and label noise. SAD-Net outperformed leading SSL and LNL baselines on multiple metrics. The main contributions of our work are:

1) We design a Noise Filtering Module (NF-Module) that hybridises a CNN and an improved fuzzy C-means (IFCM) algorithm with weighted distance metrics to filter out noisy samples via confidence-based thresholding.

2) We develop an Adaptive Weighted Distillation Module (AWD-Module) that leverages IFCM and two complementary CNNs to perform distillation with dynamic soft labelling for enhanced classification robustness.

3) We empirically validate our approach on two medical image datasets, BloodMnist and OrganaMnist [25], showing that SAD-Net outperforms existing pseudo-label-based semi-supervised and LNL methods in noisy medical image classification.

2 Related Works

Scarcely labelled data and pervasive label noise impede medical image classification. This section overviews semi-supervised pseudo-label methods and approaches to learning with noisy labels.

Methods based on Pseudo-labelling aim to augment the labelled training set by generating artificial labels, which are then used in conjunction with the original annotations to train classifiers. These techniques can be broadly categorised into regularisation consistency [13,14] and entropy minimisation [15–17]. Regularisation consistency methods leverage data augmentations to produce high-confidence pseudo-labels from unlabeled samples while enforcing prediction consistency across different augmented views of the same image—as implemented in methods such as Mixmatch [13] and Fixmatch [14]. Although these approaches yield a large number of pseudo-labels, their reliability is often uncertain. Entropy minimisation methods, conversely, construct low-entropy pseudo-labels for unlabeled data and incorporate them into the cross-entropy loss. For example, Zhen et al. [15] introduced a pseudo-label estimation framework for semi-supervised classification. Zhen et al. [16] proposed Anti-Curse Pseudo Labelling (ACPL), and Yang et al. [17] developed Adaptive Pseudo Labelling (AdaPL) to dynamically adjust pseudo-labels during training. However, erroneous pseudo-labels can introduce significant noise and misguide the model, particularly during early training, thereby increasing the model's susceptibility to label inaccuracies.

Despite these efforts, effectively mitigating label noise remains a persistent challenge in medical image analysis. Common strategies can be broadly categorised into error avoidance and label correction. For instance, Xue et al. [18] developed a collaborative training framework with an integrated noise filter to distinguish clean from noisy samples, thereby reducing the impact of mislabeled data. However, this approach relies heavily on assumptions about the underlying noise distribution. As an alternative, Zhu et al. [21] proposed a label-correction method to directly identify and correct noisy labels. Given the frequent scarcity of annotated medical images, several methods aim to address both limited samples and label noise simultaneously. Co-Correcting [23] applied a dual-network architecture combined with curriculum learning and label update modules, though its effectiveness under extremely small labelled datasets remains unclear, and it cannot fully leverage unlabeled data. Du et al. [24] introduced a noise-resistant distillation framework (NRD-Net), which integrates segmentation and distillation to alleviate and correct label noise. Nevertheless, its dependency on segmentation limits broader applicability across diverse imaging modalities. In parallel, Bhalaji et al. [26] proposed DevideMix++ and Monte-Carlo MixMatch to enhance robustness against model memorisation and improve handling of label uncertainty within the learning-with-noisy-labels (LNL) framework, respectively. Still, the practicality of these methods in data-scarce medical scenarios requires further validation.

Several existing methods aim to address the challenges posed by insufficient and noisy labels simultaneously. Cai et al. [27] proposed SSS-Net, a semi-supervised model that integrates a loss-similarity-based clustering method (LSCM) with shadowed-sets theory to enhance noise-resistant classification. Cordeiro et al. [28] introduced LongReMix, which applied a two-stage learning process to improve the precision of unsupervised clean-noisy sample separation and to strengthen SSL robustness when clean-labelled data is scarce. Despite these advances, the practical efficacy of these methods in real-world medical imaging scenarios remains to be clinically validated.

Therefore, the development of methods capable of effectively handling label noise in settings with limited annotated samples remains an urgent need in medical image classification. Furthermore, improving the utilisation of unlabeled data to enhance both model training and classification performance represents a critical direction for future research.

3 New Methodology

This paper introduces the Semi-supervised Adaptive Distillation Network (SAD-Net), a novel framework that addresses two major challenges in medical image classification: the scarcity of labelled training samples and the prevalence of label noise. Building upon a pseudo-label-based semi-supervised learning approach, the proposed method integrates two key components to enhance robustness and performance. As shown in Fig. 1, these include: (1) a Noise Filtering Module (NF-Module), which combines an Improved Fuzzy C-Means (IFCM) model with a CNN classifier to improve resilience to label noise; and (2) an Adaptive Weighted Distillation Module (AWD-Module), which incorporates the IFCM, a lightweight CNN (ResNet50) [29], and a deep CNN (DenseNet121) [30] to strengthen classification accuracy.

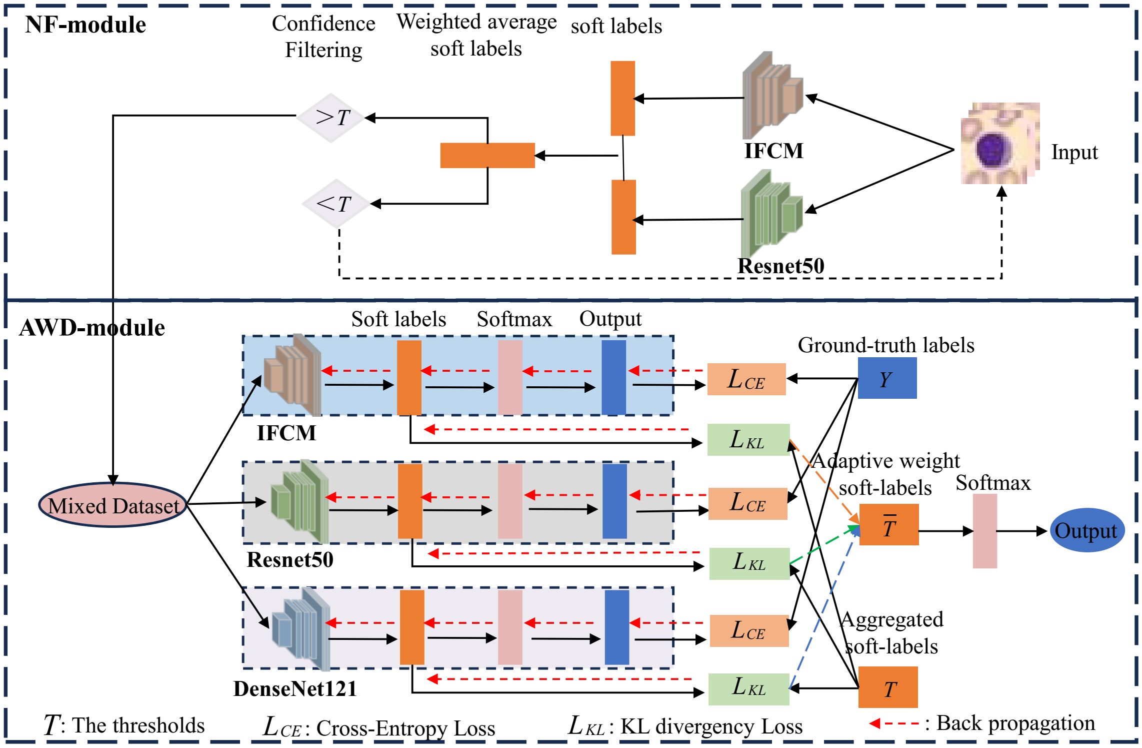

Figure 1: The semi-supervised adaptive distillation network (SAD-Net) framework consists of two main components: the Noise Filtering Module (NF-Module), shown in the upper part of the figure, and the Adaptive Weighted Distillation Module (AWD-Module) in the lower section. In the NF-Module, labelled samples with confidence scores exceeding predefined thresholds are included in a hybrid dataset constructed for subsequent training, while those under the thresholds are recycled into the input stream. Within the AWD-Module, weighted-average soft labels—produced by three student models (IFCM, ResNet50, and DenseNet121) with complementary architectures—guide the training of each model. The entire network is optimised using a combination of cross-entropy (CE) and Kullback–Leibler (KL) divergence losses. Final predictions are obtained from an adaptively weighted fusion of the soft labels derived through the three student models

As shown in Fig. 1, the input data is processed by computing a weighted average derived from the soft labels generated by both the IFCM and CNN models. This integrated value is compared against a predefined confidence threshold. Samples that exceed this threshold are retained in a hybrid dataset for additional training, while low-confidence samples are recycled as input for further learning. High-confidence unlabeled instances are assigned pseudo-labels and incorporated into the mixed dataset. The mixed data are then subjected to knowledge distillation, during which dynamically weighted soft targets generated by several student models are utilised to mutually supervise and refine their learning. The final classification result is derived from an adaptive weighted aggregation of these soft labels.

3.1 Noise Filtering Module (NF-Module)

In the proposed framework, the NF-Module is incorporated to filter the training dataset based on confidence estimates derived from soft labels, thereby reducing the influence of noisy data during training. An Improved Fuzzy C-Means (IFCM) model is first established using a label assignment strategy, enabling it to produce soft labels that reflect class affiliation degrees. The NF-Module is then formed by integrating this IFCM with a convolutional neural network (CNN) model.

3.1.1 An Improved Fuzzy c-Means (IFCM)

FCM [31], as a representative unsupervised clustering approach, organises data according to feature similarity and inherently alleviates the adverse effects of label noise. Given the input $\{x_1, x_2, \ldots, x_c\}$, the objective function of the improved fuzzy C-means (IFCM) is:

$$J = \sum\nolimits_{i=1}^{C} \sum\nolimits_{k=1}^{N} u_{ik}^{m} \overline{\Phi}_{d_{kc}^{2}} + \sum\nolimits_{i=1}^{c} \pi_A(x_i) e^{[1-\pi_A(x_i)]} \tag{1}$$

where c is the number of the clustering centres. The term u_{ik} is defined as the membership degree that quantifies how strongly sample x_j belongs to class k, with the constraint $u_{ij} \geq 0$. Here, m functions as a weighting exponent, also referred to as the fuzzification parameter and $1 \leq m \leq +\infty$, $\overline{\Phi}_{d_{kc}^{2}} = \frac{\|\Phi(x_k)-\Phi(v_i)\|^2}{\Phi_i} = \frac{\Phi_{kc}^2}{\Phi_i}$, and Φ_i is the weighted average distance from i to the k-th cluster center and can be calculated by:

$$\Phi_i = \left(\frac{\sum_{k=1}^{K} u_{ik}^{m} \Phi(d_{kc}^{2})}{\sum_{k=1}^{K} u_{ik}^{m}} \right)^{1/2} \tag{2}$$

The Fuzzy C-Means (FCM) algorithm achieves satisfactory performance in numerous applications; however, its efficacy is frequently compromised in the presence of noisy data, where outliers can significantly perturb the estimation of cluster centres. To mitigate this limitation, the Improved Fuzzy C-Means (IFCM) algorithm incorporates a possibility measure. This allows each data point to be described not only by a membership degree but also by a possibility value, thereby improving the model's capacity to manage uncertainty and noise.

Similar to FCM, IFCM operates through an iterative optimisation procedure that successively updates the possibility memberships and the cluster centres. A fundamental distinction lies in their theoretical foundations: while FCM relies on probabilistic assignments that enforce a partition of unity, IFCM is grounded in possibility theory. This approach evaluates the absolute, non-mutually-exclusive potential of a data point's affiliation with a cluster.

As a result, IFCM does not necessitate that every data point be assigned to a cluster. The affiliation is determined individually for each cluster centre based solely on spatial proximity. This formulation enhances robustness in noisy environments by relaxing the constraint that membership degrees must sum to one. Outliers can thus be assigned low-probability values and may remain unassigned, effectively mitigating a critical drawback of conventional FCM: noise is often forcibly incorporated into clusters. The step-by-step procedure of the IFCM algorithm is summarised as follows:

1) Initialisation: At the beginning, set the number of clusters c, initialise the cluster centres $\{v_1, v_2, \ldots, v_c\}$, select the fuzzy factor m and the maximum number of iterations.

2) Calculation of distance: For each data point x_i and each cluster center v_i, calculate the Euclidean distance between them:

$$d_{ij} = \left\| x_j - v_i \right\|, \tag{3}$$

where such a distance is used to measure the similarity between data points and cluster centres.

3) Update of the membership degree: Calculate the membership degree u_{ij} of each data point x_j to the cluster centre v_i through the following formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}}, \tag{4}$$

where the fuzziness index m determines the degree of fuzziness of the clusters.

4) Update of the possibility metric: IFCM introduces a possibility metric p_{ij}, which is used to represent the possibility that data points x_j belong to a cluster i. The calculation formula for the possibility measurement is as follows:

$$p_{ij} = \frac{u_{ij}^m}{\sum_{k=1}^{C} u_{ik}^m}, \tag{5}$$

5) Update of the cluster centres: The cluster centre v_i is updated based on membership degree and likelihood metrics, with the formula being:

$$v_i = \frac{\sum_{j=1}^{n} p_{ij}^m x_j}{\sum_{j=1}^{n} p_{ij}^m}, \tag{6}$$

Compared to standard FCM, the IFCM algorithm offers notable advantages in handling noisy datasets. By incorporating a possibility measure, IFCM can more effectively discriminate between valid data points and outliers. The combined use of fuzzy membership degrees and possibility values yields a more robust clustering framework, thereby enhancing its ability to handle real-world data characterised by uncertainty and noise.

3.1.2 The Construction of NF-Module

Owing to the distinctive nature of medical datasets, data features within the same category often exhibit high similarity, making it essential to preprocess features before clustering to enhance inter-cluster separation. To improve clustering outcomes, Principal Component Analysis (PCA) [32] is employed to enhance feature discriminability.

PCA reduces the original features to a limited set of the most representative features, which are mutually independent, thereby increasing the model's generalizability. The method works by mapping the original data to a new coordinate system via a linear transformation that aligns the axes with the directions of maximum variance, thereby achieving effective dimensionality reduction.

The central concept of PCA is to project high-dimensional data onto a new set of orthogonal axes—the principal components—ordered by the amount of variance they capture. The first few components typically account for most of the data's variance. Thus, by retaining these key components, PCA preserves essential information from the original dataset while significantly reducing its dimensionality.

For example, to reduce the dimension of a dataset $x = \{x_1, x_2, ..., x_N\}$ to a dimension k, The procedure is structured in the following steps:

1) Perform de-averaging (centralisation) by subtracting the mean value of each feature.

2) Derive the covariance matrix as $\frac{1}{n}XX^T$.

3) Apply eigenvalue decomposition to $\frac{1}{n}XX^T$ to extract eigenvalues and eigenvectors.

4) Order the eigenvalues in decreasing magnitude, retain the top k, and construct the eigenvector matrix P from their associated eigenvectors.

(5) Map the data into the reduced subspace defined by these k eigenvectors, i.e., $Y = PX$.

Subsequently, the confidence of the weighted soft labels derived from IFCM and the classification network is assessed to preliminarily remove noisy samples. While high-confidence instances do not guarantee correctness, low-confidence instances are highly indicative of label noise.

Taking x as the input, the soft labels calculated by the CNN model are x is $\{s_1, s_2, \ldots, s_c\}$ and $\sum_1^C s_c = 1, C$ represents the overall count of categories. The soft labels calculated by IFCM is $\{s'_1, s'_2, \ldots, s'_c\}$, and $\sum_1^C s'_c = 1$. The resulting soft labels are formulated as:

$$S(x) = \alpha \cdot S_c + (1 - \alpha) \cdot S'_c \tag{7}$$

where α functions as the trade-off parameter.

Through thresholding, the NF-module can filter the noise labels of all input data.

3.2 The Contraction of the AWD-Module

To achieve more precise classification, we introduce a three-branch distillation framework constructed from student models of diverse architectures. The IFCM component reduces the influence of noisy annotations by generating soft labels with high confidence. DenseNet121, acting as a more complex convolutional network, captures complex semantic features and contributes semantically rich soft labels, albeit with a higher tendency to overfit. To mitigate this risk, ResNet50 is incorporated as an additional student model, providing stronger generalisation capability. The learning process is supervised by adaptively weighted integration of soft labels from all three branches, which collectively serve as teacher knowledge to guide the student models.

3.2.1 Independent Training on Three Student Models

Firstly, each of the three branches undergoes independent training. The classification branch is optimised using the standard Cross-Entropy loss (CE loss) L_{CE}. It should be noted that the IFCM also operates in a supervised manner during cluster updates by incorporating L_{CE} into its objective function, as shown in Eq. (1). The definition of L_{CE} is provided below:

$$L_{\mathrm{CE}} = \frac{1}{Q}\sum_{q=1}^{Q} y_q \log(p_q), \tag{8}$$

Specifically, Q stands for the total number of training samples. For the q-th sample, y_q denotes its ground-truth label, whereas p_q represents the probability output by each classification branch.

3.2.2 Online Distillation of Three Student Models

After aggregation, the soft labels produced by the three models are computed and adopted as teacher knowledge, denoted as $\overline{T}$, to supervise the training of the student models through online distillation. This teacher signal provides initial guidance for the learning process and is defined as follows:

$$\overline{T} = \sum_{m=1}^{M} w_m \cdot \bar{s}_m \tag{9}$$

Formally, $\bar{s}m$ designates the soft label of the m-th student, M indicates the total number of students, and w_m is the weight assigned to the m-th student, computed as $\frac{vm}{\sum m{=}1^{M}\, vm}$. The term v_m represents the accuracy achieved by that student model.

The KL divergence loss between the teacher signal $\overline{T}$, and the soft output $\bar{s}_m$ from the m-th student model is used to encourage each student to learn from the ensemble teacher. The teacher is formed by an adaptively weighted aggregation of the outputs from the three student models. We define the KL divergence loss for the m-th student and the overall classification loss as:

$$L_{KL}(\overline{S}_m, \overline{T}) = \sum_{q=1}^{Q}\sum_{m=1}^{M} S_m^q * log\frac{S_m^q}{\overline{S}_m^q}. \tag{10}$$

Here, Q denotes the overall sample size, while M represents the number of student models involved.

3.2.3 Adaptive Weighted Distillation Based on Three Student Models

To enhance classification robustness, the weights assigned to the student models are adaptively updated based on their individual contributions to the final prediction. By calculating the KL-loss of the soft labels of the three models and the soft labels obtained by weighted averaging, the results of L_1, L_2, and L_3 can be obtained, then the final weights w_m of the three students can be calculated as:

To achieve more robust classification outcomes, the weight assigned to each student model is adaptively updated based on its respective contribution to the final prediction. The KL divergence between the soft labels produced by the three models and their weighted average is calculated, yielding losses L_1, L_2, and L_3. These values are then used to determine the final weights w_m for the three student models as follows:

$$w_m = \frac{L_{KL}}{\sum_1^M log\left(\frac{max\,(L_{KL}\,|m \in \{1, 2, ..., m\})}{L_{KL}}\right) + 1}. \tag{11}$$

Finally, the aggregated soft labels used to obtain the final classification result can be calculated using Eq. (5).

4 Results of Experiments

We perform three separate experiments on two publicly available datasets to assess the effectiveness and generalization ability of the proposed approach. The first experiment compares the performance of our approach against state-of-the-art pseudo-label-based semi-supervised approaches to evaluate its efficacy in leveraging unlabeled data. The second experiment contrasts our model with leading LNL (Learning with Noisy Labels) methods to assess its robustness under label noise. The third experiment comprises a series of ablation studies designed to systematically examine the contribution of each key component within the proposed framework.

4.1 Experiment Settings

Medical image Datasets. The BloodMnist dataset (BD) [25] is considered as the first dataset in our evaluation., comprising 15,362 images labeled into 7 types. The second dataset, OrganaMnist (OA) [25], consists of 52,335 images labelled into 11 types. The third dataset, BloodCells (BC) [25], consists of 12,500 images labelled into 4 types. To mitigate the issue of inconsistent parameters due to varying image sizes, all images are resized to $128 * 128$.

Models. The two student models employed in this work are ResNet50 [29] and DenseNet121 [30].

Experimental Parameter Settings. The optimiser is configured with a learning rate of 0.0005, along with the momentum fixed at 0.9 and the weight decay configured as 0.0005. The scheduler configuration sets $T_{\max}$ to 100. Pre-training proceeds on the IFCM model until convergence, after which the model undergoes 100 training epochs. Within the NF-Module, the label allocation and confidence filtering thresholds are established at 0.9 and 0.8, respectively. All experimental results are averages across three runs.

Implementation Details. Networks referenced in this section were carried out in PyTorch. The experiments were run on a system equipped with two NVIDIA GeForce RTX 3090 Ti GPUs.

Evaluation metrics. To evaluate performance on the two datasets, we adopt Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), and F1-score (F1) as evaluation metrics.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{12}$$

$$Sen = \frac{TP}{TP + FN}. \tag{13}$$

$$Spe = \frac{TN}{TN + FP}. \tag{14}$$

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{15}$$

4.2 Semi-Supervised Experiments with BD Dataset

Two semi-supervised experiments are conducted on the BD dataset in this section to evaluate the proposed model's performance. The comparison involves self-training, FixMatch [14], FlexMatch [33], and FaxMatch [15]. It is important to note that all entries presented in bold font correspond to the optimal results within their respective columns or rows of the tables.

Table 1 presents the semi-supervised classification results on dataset BD. For a fair comparison, DenseNet121 [30] serves as the backbone for the other three models. With a labelled data proportion of 30%, SAD-Net achieves the highest evaluation scores: Acc 91.07%, Spe 98.72%, Sen 89.14%, and

F1 89.68%, respectively. The classification performance of SAD-Net surpasses that of other semi-supervised models.

Table 1: Comparison of semi-supervised models on BD dataset (%)

Model	Labeled	Unlabeled	Acc	Spe	Sen	F1
Baseline	30%	0	85.55	95.79	88.85	87.16
Self-training	30%	70%	89.72	98.53	89.01	88.08
FixMatch [14]	30%	70%	89.45	98.49	86.30	88.32
FlexMatch [33]	30%	70%	90.07	98.58	88.54	88.50
SAD-Net	30%	70%	**91.07**	**98.72**	**89.14**	**89.68**

To provide a deeper evaluation, the proposed method is compared with other semi-supervised models. We visualise per-class accuracy in Fig. 2, where results on the OA dataset show that SAD-Net achieves higher accuracy across the majority of classes. Although certain categories show reduced performance—attributable to imbalanced classes—The strong generalisation capacity of the method in semi-supervised classification is demonstrated by the overall results.

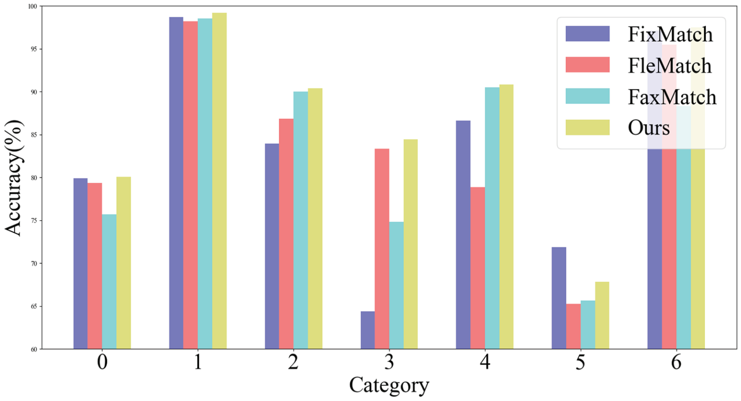

Figure 2: Visualisation of Acc values per class across semi-supervised models on OA dataset

4.3 The LNL Experiments on Datasets BD and OA

We further evaluate the proposed framework's ability to handle learning with noisy labels (LNL). Two experiments were carried out on the BD and OA datasets. The comparative methods selected for this analysis include DivideMix++ [26], Co-Training [18], Co-Teaching [21], and Co-Correcting [23].

To assess LNL performance under varying levels of label noise, we introduce label noise at 10%, 20%, 30%, and 40%. Notably, SAD-Net was trained using only 50% of the labelled data. The comparison results are presented in the table below, with the highest accuracy values highlighted in red and the second-highest in blue.

As shown in Table 2, the proposed SAD-Net, as a semi-supervised approach, achieves higher accuracy (Acc) across varying noise ratios than supervised methods such as DivideMix++, Co-Training, and Co-Teaching. Although SAD-Net's accuracy is slightly lower than that of the fully supervised Co-Correcting method at higher noise levels, it still demonstrates noise robustness comparable to such supervised baselines, highlighting its effectiveness as a semi-supervised model.

Table 2: LNL results under different noise ratios in dataset BD (%)

Methods	Clean	0.1	0.2	0.3	0.4
DivideMix++ [26]	90.78	89.04	87.12	84.43	81.32
Co-Training [18]	89.95	88.92	86.75	83.72	80.62
Co-Teaching [21]	89.74	88.82	86.90	83.87	80.85
Co-Correcting [23]	90.85	89.12	87.63	84.85	82.10
SAD-Net	89.83	89.10	87.25	84.52	81.76

For the AWD-Module, the experiment is conducted using a labelled data ratio of 50%. Fig. 3 illustrates that under noise-free conditions, SAD-Net delivers classification accuracy on par with the three supervised comparison models. As noise levels rise, the robustness of SAD-Net becomes increasingly pronounced. While its accuracy remains marginally below that of the fully supervised Co-Correcting method at every noise level, SAD-Net exhibits notable noise resistance—particularly noteworthy given its semi-supervised framework.

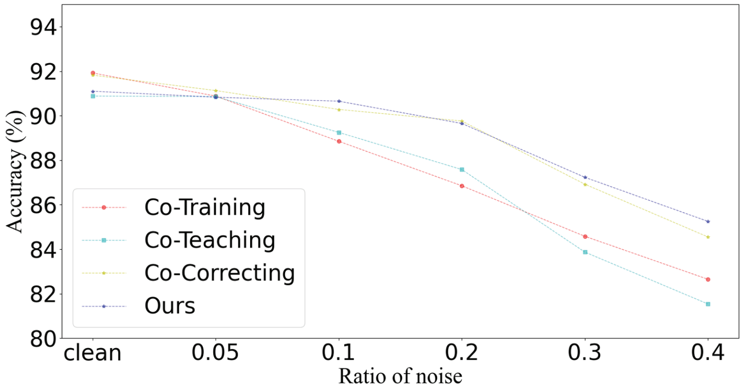

Figure 3: Accuracy comparison under different noise ratios on OA dataset

4.4 Ablation Studies

Three ablation experiments are performed to validate the roles of individual modules within the proposed approach. These experiments evaluate SAD-Net with different amounts of labelled data, as well as the effectiveness of both the NF-Module and the AWD-Module.

4.4.1 Verification of SAD-Net under Different Ratios of Labelled Samples

Initially, we validate the semi-supervised classification experiments with labelled data ratios of 20%, 30%, and 40%, respectively. For comparison, FixMatch [14], FlexMatch [33], and self-training are selected as the benchmark models.

To ensure a fair comparison, ResNet50 is adopted as the backbone network for all three baseline methods. As presented in Table 3, SAD-Net demonstrates superior classification performance across various labelled data ratios, outperforming other semi-supervised models in all evaluation metrics. Notably, with only 40% of labelled data, SAD-Net nearly matches the performance of fully supervised approaches, achieving an Acc of 92.45%, Sen of 91.78%, Spe of 99.32%, and F1 of 92.23%.

Table 3: Evaluation of semi-supervised classification on dataset OA (%)

Model	Labeled	Unlabeled	Acc	Sen	Spe	F1
Upper Bound	100%	0	95.02	93.13	99.53	93.12
Baseline	20%	0	77.47	73.25	92.36	73.33
FixMatch [14]	20%	80%	79.32	72.55	96.73	75.36
FlexMatch [33]	20%	80%	83.43	80.19	98.34	81.40
Self-training	20%	80%	90.39	90.27	99.04	89.97
SAD-Net	20%	80%	**91.33**	**90.98**	**99.13**	**90.89**
Baseline	30%	0	79.21	75.68	97.92	77.82
FixMatch [14]	30%	70%	80.44	73.32	98.04	76.01
FlexMatch [33]	30%	70%	86.09	82.45	98.61	84.27
Self-training	30%	70%	90.21	91.42	99.22	91.60
SAD-Net	30%	70%	**92.13**	**91.44**	**99.21**	**91.85**
Baseline	40%	0	80.21	76.44	95.32	78.16
FixMatch [14]	40%	60%	83.21	78.44	98.32	80.16
FlexMatch [33]	40%	60%	86.55	82.79	98.85	85.24
Self-training	40%	60%	91.01	90.72	99.20	90.95
SAD-Net	40%	60%	**92.45**	**91.78**	**99.32**	**92.23**

4.4.2 Effectiveness Verification of the NF-Module

Under the SAD-Net framework, the NF-Module's primary function is confidence filtering, which improves the detection of noisy data. To assess the effectiveness of the semi-supervised distillation network without this module (denoted as SAD-Net1) across varying noise levels, the proportion of labelled data is set at 30%.

As presented in Table 4, classification performance drops considerably when the NF-Module is removed. underscoring its essential role in mitigating the outcome of label noise. The beneficial effect of noise filtering in SAD-Net becomes increasingly pronounced as the noise ratio rises.

Table 4: The verification of the effectiveness of the NF-module across varying noise ratios on dataset BD (%)

Model	Noisy data	Clean data	Acc	Spe	Sen	F1
SAD-Net	10%	90%	**89.10**	**98.72**	**89.23**	**89.39**
SAD-Net1	10%	90%	88.64	96.43	87.68	87.21
SAD-Net	20%	80%	**88.56**	**95.53**	**87.14**	**86.86**
SAD-Net1	20%	80%	85.25	92.73	85.54	84.59

Note: It is important to note that all values presented in bold font correspond to the highest valuesoptimal results within their respective columns or rows of the tables.

4.4.3 The Verification of the Effectiveness of AWD-Module

Then, we conduct an evaluation with 30% labelled data to examine the classification performance of the semi-supervised network in the absence of the Adaptive Weighted Distillation Module (AWD-Module) (referred to as SAD-Net 2) under various noise proportions.

Table 5 demonstrates that the AWD-Module effectively mitigates, to some extent, the adverse effects of noisy labels. In addition, the distillation framework achieves superior classification performance with increased confidence levels.

Table 5: The verification of the effectiveness of AWD-module across varying noise ratios on dataset OA (%)

Model	Noisy data	Clean data	Acc	Spe	Sen	F1
SAD-Net	10%	90%	**91.53**	**97.56**	**91.35**	**91.65**
SAD-Net2	10%	90%	88.24	94.23	88.08	88.21
SAD-Net	20%	80%	**87.96**	**93.23**	**87.26**	**87.86**
SAD-Net2	20%	80%	83.25	89.73	83.74	83.49

Note: It is important to note that all values presented in bold font correspond to the highest valuesoptimal results within their respective columns or rows of the tables.

4.4.4 The Verification of the Training Time and Inference Speed

Finally, we compare the running time and memory consumption of different models. The experiment is conducted under a semi-supervised setting using the BD dataset, with 30% of the data designated as labelled.

As shown in Table 6, the three-model distillation SAD-Net achieves higher accuracy at the cost of longer training time than a single model. Since diagnostic accuracy is paramount in clinical decision-making, this trade-off is justified. For future work, we will focus on designing lightweight model architectures and efficient fusion strategies to mitigate this computational overhead, thereby enhancing both diagnostic efficiency and practicality for medical settings.

Table 6: The verification of the training time and inference speed from different models

Models	Dataset	Acc	Training time	Memory consumption
FlexMatch	BD	0.901	450.7 s	1.72 GB
FixMatch	BD	0.895	544.5 s	1.97 GB
Self-training	BD	0.897	734.1 s	1.98 GB
SAD-Net	BD	0.911	1013.3 s	2.56 GB

5 Discussions and Conclusions

This paper presents a Semi-supervised Adaptive Distillation Network (SAD-Net), a novel framework designed to concurrently address the challenges of limited labelled data and label noise in medical image classification. SAD-Net integrates two core components: a Noise Filtering Module (NF-Module), which combines an Improved Fuzzy C-Means (IFCM) algorithm with a convolutional neural network (CNN) to identify and filter noisy labels, and an Adaptive Weighted Distillation Module (AWD-Module) that leverages the IFCM alongside ResNet50 and DenseNet121 models to facilitate robust knowledge distillation. Extensive experimental evaluations demonstrate the model's efficacy in handling both sparse annotations and label noise simultaneously. The following section reviews the benefits and drawbacks of our approach, highlighting its contributions and providing insights for subsequent studies.

Advantages: 1) SAD-Net presents a novel semi-supervised approach that simultaneously alleviates the problems of scarce samples and labelling noise. 2) By employing a label assignment strategy, the unsupervised IFCM model produces soft labels based on membership degrees, broadening the applicability of deep unsupervised learning. 3) The AWD-Module leverages adaptive distillation and integrates IFCM-derived soft labels to yield more robust classification outcomes.

Limitations: 1) Experimental results indicate that class imbalance adversely affects both semi-supervised learning performance and noise robustness. Future work will focus on integrating advanced resampling or cost-sensitive strategies to effectively handle class-imbalanced datasets. 2) The training process inherits certain limitations of the IFCM algorithm, such as high computational complexity, extended training time, and sensitivity to initial cluster centroids. Developing more efficient clustering alternatives or initialisation strategies will be a critical direction for future research. 3) Current hyperparameters and thresholds for noise filtering and label assignment are configured manually. We intend to investigate dynamic, optimisation-based parameter adjustment mechanisms to automate this process and improve experimental reproducibility. 4) SAD-Net requires more training time and memory usage, and it does not address the challenges associated with online learning and continuous learning. In subsequent work, it is essential to optimise the model to improve training efficiency and strengthen its cross-domain continual learning capabilities.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm the contribution to the paper as follows: Conceptualization, Yifan Peng and Yong Hu; methodology, Yifan Peng, Shen Zhang, Zongsheng Li, and Xiangtong Du; software, Yong Hu, Zongsheng Li, and Weifan Xu; validation, Yifan Peng, Yong Hu, and Shen Zhang; formal analysis, Weifan Xu and Xiangtong Du; investigation, Yifan Peng and Xiangtong Du; resources, Xiangtong Du; writing—original draft preparation, Yifan Peng, Yong Hu, Shen Zhang, and Zongsheng Li; writing—review and editing, Xiangtong Du. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Chen XY, Liu X, Wu YK, Wang ZL, Wang SH. Research related to the diagnosis of prostate cancer based on machine learning medical images: a review. Int J Med Inform. 2023;181(3):105279. doi:10.1016/j.ijmedinf.2023.105279.

2. Li YY, Zhang HL, Sun Y, Fan Q, Wang L, Ji C, et al. Deep learning-based platform performs high detection sensitivity of intracranial aneurysms in 3D brain TOF-MRA: an external clinical validation study. Int J Med Inform. 2024;188(7):105487. doi:10.1016/j.ijmedinf.2024.105487.
3. Dimauro G, Ciprandi G, Deperte F, Girardi F, Ladisa E, Latrofa S, et al. Nasal cytology with deep learning techniques. Int J Med Inform. 2019;122(7553):13–9. doi:10.1016/j.ijmedinf.2018.11.010.
4. Michael L, Jonas S, Martin PN, Christian JG. Deep learning-based classification of organs at risk and delineation guideline in pelvic cancer radiation therapy. J Appl Clin Med Phys. 2023;24(9):e14022. doi:10.1002/acm2.14022.
5. Fan Y, Gong H. An improved COVID-19 classification model on chest radiography by dual-ended multiple attention learning. IEEE J Biomed Health Inform. 2024;28(1):145–56. doi:10.1109/jbhi.2023.3324333.
6. Katherine M, Katia FB, Sonia LH, Ahmed OS, Patel MV, Bader KB. Development of convolutional neural network to segment ultrasound images of histotripsy ablation. IEEE Trans Biomed Eng. 2024;71(6):1789–97. doi:10.1109/tbme.2024.3352538.
7. Liu K, Bao C, Liu S. Semi-supervised medical image classification based on sample intrinsic similarity using canonical correlation analysis. Comput Mater Contin. 2025;82(3):4451–68. doi:10.32604/cmc.2024.059053.
8. Sepehri K, Song X, Proulx R, Hajra SG, Dobberthien B, Liu CC, et al. Towards effective machine learning in medical imaging analysis: a novel approach and expert evaluation of high-grade glioma 'ground truth' simulation on MRI. Int J Med Inform. 2021;146(sup 6):104348. doi:10.1016/j.ijmedinf.2020.104348.
9. Du XT, Liu ZD, Feng ZL, Deng H. DataMap: dataset transferability map for medical image classification. Pattern Recognit. 2024;146(8):110044. doi:10.1016/j.patcog.2023.110044.
10. Wang J, Gao M, Li Q, Kim H, Jeon G. A survey on supervised, unsupervised, and semi-supervised approaches in crowd counting. Comput Mater Contin. 2024;81(3):3561–82. doi:10.32604/cmc.2024.058637.
11. Kamal H, Fatemeh D, Maryam R, Haghighat M, Jalili M. A comprehensive survey on multi-view classification: methods, applications, and challenges. ACM Trans Intell Syst Technol. 2025;16(6):1–34. doi:10.1145/3767728.
12. Mohammadi M, Berahmand K, Azizi S, Sheikhpour R, Khosravi H. Semi-supervised adaptive symmetric nonnegative matrix factorization for multi-view clustering. IEEE Trans Netw Sci Eng. 2025;12(6):4967–81. doi:10.1109/tnse.2025.3578315.
13. Berthelot D, Carlini N, Goodfellow I, Oliver A, Papernot N, Raffel C. MixMatch: a holistic approach to semi-supervised learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Red Hook, NY, USA: Curran Associates Inc.; 2019. p. 5049–59.
14. Kihyuk S, David B, Chun-Liang L. FixMatch: simplifying semi-supervised learning with consistency and confidence. arXiv:2001.07685. 2020.
15. Zhen P, Dezhi Z, Shenwei T, Wu W, Yu L, Zhou S, et al. FaxMatch: multi-Curriculum Pseudo-Labeling for semi-supervised medical image classification. Med Phys. 2023;50(5):3210–22. doi:10.1002/mp.16312.
16. Zhen P, Shenwei T, Dezhi Z. Semi-supervised medical image classification with adaptive threshold pseudo-labeling and unreliable sample contrastive loss. Biomed Signal Process Control. 2023;79(9):104142. doi:10.1016/j.bspc.2022.104142.
17. Jiawei M, Xuesong Y, Guodao Z, Chen B, Chang Y, Chen W, et al. Pseudo-labeling generative adversarial networks for medical image classification. Comput Biol Med. 2022;147:105729. doi:10.1016/j.compbiomed.2022.105729.
18. Xue C, Yu L, Chen P, Dou Q, Heng P. Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE Trans Med Imaging. 2022;41(6):1371–82. doi:10.1109/tmi.2021.3140140.
19. Yi K, Wu JX. Probabilistic end-to-end noise correction for learning with noisy labels. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Piscataway, NJ, USA: IEEE; 2019. p. 7017–25.

20. Li JN, Socher R, Hoi SCH. DivideMix: learning with noisy labels as semi-supervised learning. arXiv:2002.07394. 2020.
21. Zhu M, Zhang L, Wang L, Li D, Zhang J, Yi Z. Robust co-teaching learning with consistency-based noisy label correction for medical image classification. Int J Comput Assist Radiol Surg. 2023;18(4):675–83. doi:10.1007/s11548-022-02799-6.
22. Ren M, Zeng W, Yang B, Urtasun R. Learning to reweight examples for robust deep learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML); London, UK: PMLR; 2018. p. 4334–43.
23. Liu J, Li R, Sun C. Noise-tolerant medical image classification via mutual label correction. IEEE Trans Med Imaging. 2021;40(12):3580–92. doi:10.1109/tmi.2021.3091178.
24. Du XT, Shen A, Wang XM, Feng ZL, Deng H. NRD-Net: a noise-resistant distillation network for accurate diagnosis of prostate cancer with bi-parametric MRI images. Multim Tools Appl. 2024;83(11):33597–614. doi:10.1007/s11042-023-16712-z.
25. Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. MedMNIST v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data. 2023;10(1):41. doi:10.1038/s41597-022-01721-8.
26. Bhalaji N, Marques R, Eduardo A, Petia R. Bayesian DivideMix++ for enhanced learning with noisy labels. Neural Netw. 2024;172(2):106122. doi:10.1016/j.neunet.2024.106122.
27. Cai K, Zhang H, Pedrycz W, Miao D. SSS-Net: a shadowed-sets-based semi-supervised sample selection network for classification on noise labeled images. Knowl Based Syst. 2023;276(C):110732. doi:10.1016/j.knosys.2023.110732.
28. Cordeiro FR, Sachdeva R, Belagiannis V, Reid I, Carneiro G. Longremix: robust learning with high confidence samples in a noisy label environment. Pattern Recognit. 2023;133(5):109013. doi:10.1016/j.patcog.2022.109013.
29. Anju S, Rajnish K, Prabha G. Deep learning-based prediction model for diagnosing gastrointestinal diseases using endoscopy images. Int J Med Inform. 2023;177(5):105142. doi:10.1016/j.ijmedinf.2023.105142.
30. Li Z, Jiang J, Zhou H, Zheng Q, Liu X, Chen K, et al. Development of a deep learning-based image eligibility verification system for detecting and filtering out ineligible fundus images: a multicentre study. Int J Med Inform. 2021;147(6468):104363. doi:10.1016/j.ijmedinf.2020.104363.
31. Salar A. Fuzzy C-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers. Rev Develop Expert Syst Appl. 2021;165(5):113856. doi:10.1016/j.eswa.2020.113856.
32. Lever J, Krzywinski M, Altman N. Principal component analysis. Nat Methods. 2017;14(7):641–2. doi:10.1038/nmeth.4346.
33. Zhang BW, Wang YD, Hou WX, Wu H, Wang J, Okumura M, et al. FlexMatch: boosting semi-supervised learning with curriculum pseudo labeling. In: NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2021. p. 18408–19.