Journal of Edge Computing, 2025, Vol. **4**, Iss. 2, pp. 159–178 https://doi.org/10.55056/jec.978

Method of adaptive knowledge distillation from multi-teacher to student deep learning models

Oleksandr Chaban, Eduard Manziuk and Pavlo Radiuk

Khmelnytskyi National University, 11, Instytutska str., Khmelnytskyi, 29016, Ukraine

Abstract. Transferring knowledge from multiple teacher models to a compact student model is often hindered by domain shifts between datasets and a scarcity of labeled target data, degrading performance. While existing methods address parts of this problem, a unified framework is lacking. In this work, we improve multi-teacher knowledge distillation by developing a holistic framework, enhanced multi-teacher knowledge distillation (EMTKD), that synergistically integrates three components: domain adaptation within teacher training, an instance-specific adaptive weighting mechanism for knowledge fusion, and semi-supervised learning to leverage unlabeled data. On a challenging cross-domain cardiac MRI benchmark, EMTKD achieves a target domain accuracy of 88.5% and an area under the curve of 92.5%, outperforming state-of-the-art techniques by up to 5.0%. Our results demonstrate that this integrated, adaptive approach yields significantly more robust and accurate student models, enabling effective deep learning deployment in data-scarce environments.

Keywords: knowledge distillation, deep learning, multi-teacher learning, adaptive weighting, domain adaptation, semi-supervised learning, model generalisation

1. Introduction

Deep learning (DL) models have achieved remarkable success across many domains, including computer vision, natural language processing, and medical image analysis [12, 14]. However, developing high-performance DL models often necessitates vast amounts of labelled data and significant computational resources for training large, complex architectures. These requirements can be prohibitive in many real-world scenarios, particularly where data is scarce, privacy is a concern, or deployment is targeted for resource-constrained environments [18]. Knowledge distillation (KD) [10, 17] has emerged as a powerful technique to address these challenges by transferring the "knowledge" from a large, pre-trained teacher model (or an ensemble of teachers) to a smaller, more efficient student model. The student model, once trained, aims to mimic the performance of the teacher(s) while being significantly more compact and faster at inference.

While traditional KD often involves a single teacher, multi-teacher knowledge distillation (MTKD) has gained attention for its potential to leverage diverse knowledge from multiple expert models [4]. These teacher models might be trained on different datasets (source domains), possess different architectures, or specialise in different aspects of a task, offering a richer source of information for the student. However, effectively amalgamating knowledge from multiple, potentially heterogeneous teachers presents significant hurdles. A primary challenge is the domain shift that often

This study is the further development of our previous work [3] presented at the 5th Edge Computing Workshop (doors 2025).

0009-0001-4710-3336 (O. Chaban); 0000-0002-7310-2126 (E. Manziuk); 0000-0003-3609-112X (P. Radiuk)

chabanolek@khmnu.edu.ua (O. Chaban); manziuk.e@khmnu.edu.ua (E. Manziuk); radiukp@khmnu.edu.ua (P. Radiuk)

exists between the datasets used to train different teachers and the student's target application domain [9, 11]. Such shifts can lead to conflicting or sub-optimal supervisory signals, potentially degrading the student's performance [26]. Furthermore, the scarcity of labelled data in specialised fields like medical imaging [15, 19] means the student must often be trained with limited supervision, necessitating methods that can leverage unlabeled data. Finally, treating all teachers equally during distillation is often suboptimal, as some teachers may be more reliable for specific data instances than others.

To address these issues, researchers have explored various strategies. Early MTKD variants simply averaged logits [4]. More recent methods introduced dynamic or confidence-aware fusion, e.g., AEKD [5], CA-MKD [24], and the reinforcement-learning based MTKD-RL [23]. Cross-domain distillation has likewise progressed, with Direct-Distill [22], DS-KD [21], and CD-CD [7] targeting distribution shifts. Enhanced multi-teacher knowledge distillation (*EMTKD*) differs by (i) integrating domain adaptation *inside each teacher*, (ii) performing *instance-specific* weighting plus attention-based feature aggregation, and (iii) coupling that with pseudo-label-driven SSL. Section 3 quantitatively contrasts EMTKD with all the above methods on identical cardiac-MRI benchmarks. While these works have advanced specific aspects of the problem, a comprehensive framework that synergistically integrates solutions for domain adaptation at the teacher level, instance-specific knowledge fusion, and effective use of unlabeled data for the student remains an area of research.

This study aims to improve the knowledge transfer process from multiple teacher deep learning models to a student model by developing a holistic and adaptive distillation framework that simultaneously addresses the challenges of domain heterogeneity, data scarcity, and intelligent knowledge fusion. We aim to design a method that produces a lightweight student model capable of high performance in a target domain, even when trained with limited labelled data and guided by teachers from disparate source domains. To achieve this, we undertake the following tasks. First, we prepare a diverse ensemble of expert teacher models by training them with domain adaptation techniques to ensure their knowledge is robust and generalisable. Second, we design an adaptive mechanism to intelligently aggregate knowledge by weighting each teacher's contribution based on its confidence for a given data sample. Third, we develop a student training regimen combining the distilled knowledge with an integrated SSL strategy to effectively learn from labelled and unlabeled data in the target domain.

To this end, this work presents an enhanced multi-teacher knowledge distillation method, which builds upon and generalises foundational concepts from our prior work on edge computing for cardiac MRI [3]. While the original framework emphasised resource-constrained deployment and integrated privacy, the current manuscript details a refined and more broadly applicable methodology. We retain and formalise the core adaptive mechanisms for knowledge fusion and data-scarce learning, while positioning features like differential privacy as optional extensions. This re-framing allows for a comprehensive exploration of the adaptive distillation process, making it applicable to a broader range of DL scenarios.

The main scientific contributions of this work are as follows:

- **Domain-adaptive teacher training for enhanced knowledge quality**: we incorporate domain adaptation techniques into the teacher training phase. This encourages the models to learn domain-invariant features, thereby providing higher-quality and more generalisable knowledge for the student and improving their ability to handle domain shifts.
- **Adaptive instance-specific teacher weighting**: we introduce a dynamic,

instance-specific weighting strategy that calibrates the influence of each teacher based on its predictive confidence for a given input. This allows the student to learn from the most reliable teachers on a sample-by-sample basis, moving beyond simple averaging or static weighting.

- **Synergistic semi-supervised learning (SSL) integration**: we incorporate SSL via pseudo-labeling into the student's training. This complements the distilled knowledge by enabling the student to leverage a larger pool of unlabeled target-specific data, thereby improving its adaptation and generalisation within the target domain.

The remainder of this manuscript is structured as follows. Section 2 provides a detailed description of the proposed EMTKD method. Section 3 presents the experimental setup, quantitative results, and ablation studies. Section 4 discusses the implications, advantages, and limitations. Finally, section 5 summarises the key contributions.

2. Method of adaptive knowledge distillation from teachers' to students' models of deep learning

The proposed enhanced multi-teacher knowledge distillation method facilitates effective knowledge transfer from an ensemble of diverse teacher models to a single, compact student model. The core philosophy is to adaptively manage the influence of each teacher, account for domain discrepancies among teacher data sources, and leverage unlabeled data to enhance student learning. The method is structured into three principal blocks, as depicted in figure 1: teacher model training with domain adaptation, adaptive knowledge distillation, and student model training with semi-supervised learning and optional privacy preservation.

Each block entails specific inputs, processing steps, and outputs, which collectively contribute to the overarching goal of developing a robust and accurate student model capable of generalising well to new, unseen data, even under conditions of data heterogeneity and scarcity.

2.1. Block 1: Teacher model training with domain adaptation

The initial phase of EMTKD focuses on preparing a set of knowledgeable teacher models. A critical aspect of this block is the incorporation of domain adaptation techniques during each teacher's training. This is crucial when teacher models are trained on data from different sources or domains, as it encourages the models to learn features that are invariant to domain-specific characteristics, thereby producing more consistent and generalizable knowledge for the student.

Let $\mathcal{X}$ denote the input space (e.g., images) and $\mathcal{Y}$ the output space (e.g., class labels). The input to this block consists of T distinct annotated datasets, $\mathcal{D}^{(t)} = \{(\mathbf{x}_i^{(t)}, y_i^{(t)}, d^{(t)})\}_{i=1}^{N_t}$ for $t = 1, 2, \ldots, T$. Here, $\mathbf{x}_i^{(t)} \in \mathcal{X}$ represents the i-th input sample from the t-th source domain, $y_i^{(t)} \in \mathcal{Y}$ is its corresponding class label, and $d^{(t)}$ is an identifier for domain t. N_t denotes the number of samples in dataset t. Each dataset $\mathcal{D}^{(t)}$ is used to train a corresponding teacher model $M_t : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{Y})$ is the space of probability distributions over $\mathcal{Y}$.

Step 1. Data preparation and preprocessing

This step involves collecting the T source datasets. Each dataset undergoes standard preprocessing pertinent to the data modality and task. For instance, in image classification, this might include resizing images to a uniform dimension, normalisation of pixel values (e.g., to zero mean and unit variance), and data augmentation (e.g., random rotations, flips, brightness adjustments) to increase dataset diversity and

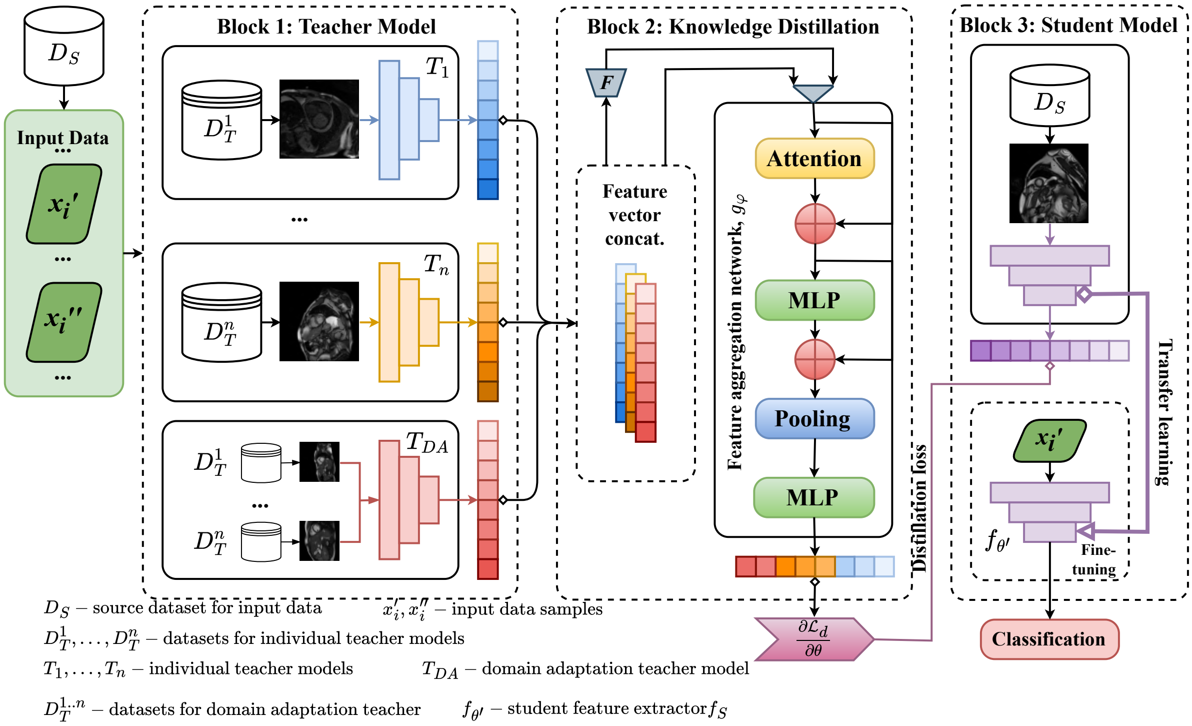

D_S
Input Data
x_i'
x_i''
Block 1: Teacher Model
D_T^1
T_1
T_n
D_T^n
T_{DA}
D_T^1
D_T^n
Block 2: Knowledge Distillation
F
Feature vector concat.
Feature aggregation network, g_φ
Attention
MLP
Pooling
MLP
$\frac{\partial \mathcal{L}_d}{\partial \theta}$
Distillation loss
Block 3: Student Model
D_S
x_i'
$f_{\theta'}$
Fine-tuning
Transfer learning
Classification
D_S – source dataset for input data
x_i', x_i'' – input data samples
$D_T^1, \ldots, D_T^n$ – datasets for individual teacher models
$T_1, \ldots, T_n$ – individual teacher models
T_{DA} – domain adaptation teacher model
$D_T^{1..n}$ – datasets for domain adaptation teacher
$f_{\theta'}$ – student feature extractor f_S

Figure 1: Schematic overview of the proposed enhanced multi-teacher knowledge distillation method. Block 1 involves training multiple teacher models M_t, each potentially on a different source domain $\mathcal{D}^{(t)}$, incorporating domain adaptation techniques (e.g., using domain classifier D_t) to learn domain-invariant features. Block 2 performs adaptive knowledge distillation by weighting teacher contributions based on confidence c_t and aggregating their feature representations $\mathbf{z}^{(t)}$ using an attention mechanism A_{att} to produce $\mathbf{z}_{\text{agg}}$. Block 3 trains the student model S_θ using the distilled knowledge $\mathbf{z}_{\text{agg}}$, semi-supervised learning on unlabeled target data, and optional privacy-preserving mechanisms.

prevent overfitting. Consistent preprocessing across datasets, where feasible, helps in minimising superficial variations.

Step 2. Teacher model initialisation

For each source domain t, a neural network architecture is chosen for the teacher model M_t. The architectures can be identical across teachers or varied. Each M_t consists of a feature extractor $f_t : \mathcal{X} \to \mathcal{Z}_t$ (where $\mathcal{Z}_t$ is the feature space of teacher t) and a classifier head. Models are typically initialized with random weights or pre-trained weights. For each teacher model M_t intended to learn domain-invariant features, a corresponding domain discriminator network D_t is also defined and initialized. The discriminator's role is to distinguish the origin domain of the features extracted by f_t.

Step 3. Domain-adaptive training of teacher models

Each teacher model M_t is trained to perform its primary task on $\mathcal{D}^{(t)}$ while producing domain-invariant features. The primary task loss for M_t, typically cross-entropy for classification, is:

$$\mathcal{L}_{\text{CE}}^{(t)} = -\frac{1}{N_t} \sum_{i=1}^{N_t} y_i^{(t)} \log M_t(\mathbf{x}_i^{(t)}), \tag{1}$$

where $M_t(\mathbf{x}_i^{(t)})$ is the predicted probability distribution for sample $\mathbf{x}_i^{(t)}$.

To promote domain invariance, a domain adaptation loss $\mathcal{L}_{\text{DA}}^{(t)}$ is introduced, often via adversarial training with D_t. A gradient reversal layer (GRL) [6] is commonly used.

The domain adaptation loss is:

$$\mathcal{L}_{\mathrm{DA}}^{(t)} = -\frac{1}{N_t}\sum_{i=1}^{N_t} d^{(t)'} \log D_t(f_t(\mathbf{x}_i^{(t)})), \tag{2}$$

where $f_t(\mathbf{x}_i^{(t)})$ are features from M_t, and $d^{(t)'}$ is the domain label for the sample (e.g., indicating if it's from source t or a reference target domain). The feature extractor f_t aims to produce features that D_t cannot reliably classify by domain.

The total loss for training teacher model M_t is:

$$\mathcal{L}_{\mathrm{teacher}}^{(t)} = \mathcal{L}_{\mathrm{CE}}^{(t)} + \lambda_{\mathrm{DA}}\mathcal{L}_{\mathrm{DA}}^{(t)}, \tag{3}$$

where λ_{DA} balances task performance and domain invariance.

Step 4. Optimisation of teacher models

The parameters θ_t of each teacher model M_t (and its discriminator D_t) are updated iteratively using an optimizer like SGD or Adam:

$$\theta_t \leftarrow \theta_t - \eta\nabla_{\theta_t}\mathcal{L}_{\mathrm{teacher}}^{(t)}, \tag{4}$$

where η is the learning rate. This is repeated until convergence for each M_t.

Step 5. Feature extraction from trained teachers

Once trained, each teacher M_t can extract features $\mathbf{z}^{(t)}(\mathbf{x})$ from an input $\mathbf{x}$, typically penultimate layer activations:

$$\mathbf{z}^{(t)}(\mathbf{x}) = f_t(\mathbf{x}). \tag{5}$$

These features $\mathbf{z}^{(t)}(\mathbf{x}) \in \mathcal{Z}_t$ encapsulate the teacher's learned knowledge.

The output of Block 1 is the set of T trained, domain-adapted teacher models $\{M_t\}_{t=1}^{T}$ and their feature extraction capabilities.

2.2. Block 2: Adaptive knowledge distillation

This block intelligently combines knowledge from the multiple teachers for transfer to the student. Key elements are adaptive weighting and attention-based feature aggregation.

Inputs are the trained teachers $\{M_t\}_{t=1}^{T}$ and target dataset $\mathcal{D}^{(s)}$, which includes annotated $\mathcal{D}^{(a)} = \{(\mathbf{x}_i^{(a)}, y_i^{(a)})\}_{i=1}^{N_a}$ and unannotated $\mathcal{D}^{(u)} = \{\mathbf{x}_i^{(u)}\}_{i=1}^{N_u}$ examples. Let $\mathbf{x}_i^{(s)}$ be a sample from $\mathcal{D}^{(s)}$.

Step 1. Compute teacher predictions and confidences

For each $\mathbf{x}_i^{(s)}$, each teacher M_t produces class probabilities:

$$\mathbf{p}_t(\mathbf{x}_i^{(s)}) = \mathrm{Softmax}(\mathrm{logits}_t(\mathbf{x}_i^{(s)})), \tag{6}$$

where $\mathrm{logits}_t(\mathbf{x}_i^{(s)})$ are the pre-softmax outputs of M_t's classifier head. Teacher confidence is the maximum probability:

$$c_t(\mathbf{x}_i^{(s)}) = \max\left(\mathbf{p}_t(\mathbf{x}_i^{(s)})\right). \tag{7}$$

This reflects teacher t's certainty for sample $\mathbf{x}_i^{(s)}$.

Step 2. Calculate adaptive weights for teachers

Instance-specific weights $w_t(\mathbf{x}_i^{(s)})$ are computed using a temperature-scaled SoftMax over teacher confidences:

$$w_t(\mathbf{x}_i^{(s)}) = \frac{\exp\left(c_t(\mathbf{x}_i^{(s)})/\tau\right)}{\sum\limits_{k=1}^{T} \exp\left(c_k(\mathbf{x}_i^{(s)})/\tau\right)}. \tag{8}$$

Temperature $\tau > 0$ controls sharpness; lower τ gives more weight to high-confidence teachers.

Step 3. Weighted feature aggregation with attention

Teacher features $\mathbf{z}^{(t)}(\mathbf{x}_i^{(s)})$ are weighted:

$$\widetilde{\mathbf{z}}^{(t)}(\mathbf{x}_i^{(s)}) = w_t(\mathbf{x}_i^{(s)}) \cdot \mathbf{z}^{(t)}(\mathbf{x}_i^{(s)}). \tag{9}$$

These are concatenated:

$$\mathbf{Z}(\mathbf{x}_i^{(s)}) = \left[\widetilde{\mathbf{z}}^{(1)}(\mathbf{x}_i^{(s)}), \widetilde{\mathbf{z}}^{(2)}(\mathbf{x}_i^{(s)}), \ldots, \widetilde{\mathbf{z}}^{(T)}(\mathbf{x}_i^{(s)})\right]. \tag{10}$$

An attention network A_{att} (e.g., a shallow MLP or a transformer encoder layer) refines this into an aggregated feature vector $\mathbf{z}_{\text{agg}}$:

$$\mathbf{z}_{\text{agg}}\left(\mathbf{x}_i^{(s)}\right) = A_{\text{att}}\left(\mathbf{Z}\left(\mathbf{x}_i^{(s)}\right)\right). \tag{11}$$

This allows selective focus on the most informative combined teacher knowledge.

Step 4. Storage of aggregated features

The aggregated features $\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})$ are computed and stored for all target samples.

Output of Block 2 is the set $\{\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})\}$, embodying adaptively combined teacher wisdom.

2.3. Block 3: Student model training with SSL and privacy preservation

This block trains the student model S_θ using aggregated knowledge, SSL, and optional privacy.

Inputs: aggregated features $\{\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})\}$, target samples $\mathcal{D}^{(a)}$ and $\mathcal{D}^{(u)}$.

Step 1. Student model initialisation

Student model S_θ with parameters θ is defined, typically smaller than teachers. It has a feature extractor $f_S : \mathcal{X} \to \mathcal{Z}_S$ and classifier $C(\cdot)$. $\mathcal{Z}_S$ is the student's feature space.

Step 2. Compute student features

For each $\mathbf{x}_i^{(s)}$, the student feature extractor produces:

$$\widehat{\mathbf{z}}(\mathbf{x}_i^{(s)}) = f_S(\mathbf{x}_i^{(s)}; \theta). \tag{12}$$

Typically, $\widehat{\mathbf{z}}(\mathbf{x}_i^{(s)})$ and $\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})$ are engineered or projected to be of the same dimensionality for direct comparison.

Step 3. Distillation loss computation

The student's features $\widehat{\mathbf{z}}(\mathbf{x}_i^{(s)})$ are encouraged to match aggregated teacher features $\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})$ via a loss, e.g., MSE:

$$\mathcal{L}_{\text{distill}}(\mathbf{x}_i^{(s)}) = \left\| \widehat{\mathbf{z}}(\mathbf{x}_i^{(s)}) - \mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)}) \right\|_2^2. \tag{13}$$

The total distillation loss over all $N_s = N_a + N_u$ target samples is:

$$\mathcal{L}_{\text{distill}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{L}_{\text{distill}}(\mathbf{x}_i^{(s)}). \tag{14}$$

This can be augmented with L2 regularisation $\mathcal{R}(\theta) = \|\theta\|_2^2$:

$$\mathcal{L}_{\text{distill\_reg}} = \mathcal{L}_{\text{distill}} + \lambda_{\text{reg}} \mathcal{R}(\theta), \tag{15}$$

where λ_{reg} is the regularisation strength.

Step 4. Semi-supervised learning with pseudo-labels

For unannotated $\mathbf{x}_i^{(u)} \in \mathcal{D}^{(u)}$, the student S_θ generates predictions:

$$\mathbf{p}_s(\mathbf{x}_i^{(u)}) = \text{Softmax}(C(f_S(\mathbf{x}_i^{(u)}; \theta))). \tag{16}$$

A pseudo-label $\hat{y}_i^{(u)}$ is assigned if confidence $\max \mathbf{p}_s(\mathbf{x}_i^{(u)})$ exceeds threshold δ:

$$\hat{y}_i^{(u)} = \begin{cases} \arg\max \mathbf{p}_s(\mathbf{x}_i^{(u)}), & \text{if } \max \mathbf{p}_s(\mathbf{x}_i^{(u)}) > \delta \\ \text{ignore}, & \text{otherwise}. \end{cases} \tag{17}$$

We follow FixMatch [20] and raise the pseudo-label confidence threshold δ linearly from 0.6 to 0.9 during the first 40% of training epochs, thereby suppressing early noisy labels while retaining 85% of high-confidence samples in later stages.

The SSL loss (e.g., cross-entropy) is computed for N_p pseudo-labeled samples:

$$\mathcal{L}_{\text{pseudo}} = -\frac{1}{N_p} \sum_{j=1}^{N_p} \hat{y}_j^{(u)} \log C(f_S(\mathbf{x}_j^{(u)}; \theta)). \tag{18}$$

Step 5. Compute classification loss on annotated data

For annotated $\mathcal{D}^{(a)}$, a standard supervised cross-entropy loss is:

$$\mathcal{L}_{\text{CE}}^{(a)} = -\frac{1}{N_a} \sum_{i=1}^{N_a} y_i^{(a)} \log C(f_S(\mathbf{x}_i^{(a)}; \theta)). \tag{19}$$

Step 6. Total loss computation for student training

The total loss for S_θ is a weighted sum:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{distill\_reg}} + \beta \mathcal{L}_{\text{CE}}^{(a)} + \gamma \mathcal{L}_{\text{pseudo}}, \tag{20}$$

where α, β, γ are tunable hyperparameters balancing distillation, supervision, and SSL.

Step 7. Optional: privacy-preserving training

If privacy is a concern, differentially private SGD (DP-SGD) can be used. It involves gradient clipping (to norm C_{clip}) and adding Gaussian noise $\mathcal{N}(0, \sigma^2 C_{clip}^2)$ with noise multiplier σ:

$$\nabla_\theta \mathcal{L}_{\text{total}}^{\text{noisy}} = \text{clip}(\nabla_\theta \mathcal{L}_{\text{total}}, C_{clip}) + \mathcal{N}(0, \sigma^2 C_{clip}^2). \tag{21}$$

Step 8. Student model parameter update

Student parameters θ are updated via:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}} \quad (\text{or } \theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}^{\text{noisy}}). \tag{22}$$

Step 9. Fine-tuning (optional)

The student can be fine-tuned on $\mathcal{D}^{(a)}$ using:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{CE}}^{(a)} + \lambda_{\text{reg}} \mathcal{R}(\theta). \tag{23}$$

The final output is the trained student model S_θ.

2.4. Illustrative datasets for experimental validation

Experiments used two public cardiac MRI datasets exhibiting significant domain differences.

Dataset A (source domain example): Derived from the automated cardiac diagnosis challenge (ACDC) [1]. It includes 2D short-axis cine MRI sequences from 100 patients, acquired from multiple centers with varying scanners and protocols. Annotated for diagnostic classes by cardiologists. Image resolutions typically 1.37-1.68 mm^2/pixel. Serves as source(s) for teacher models.

Dataset B (target domain example): Sourced from the multi-center, multi-vendor, and multi-disease cardiac image dataset (M&Ms) [2]. Includes data from 160 patients, from different vendors (Siemens, Philips, GE) and field strengths (1.5T, 3T), designed for domain shift challenges. Annotated by radiologists. Serves as target domain for student evaluation.

The task is classifying cardiac MRI scans. The domain shift between ACDC and M&Ms allows rigorous assessment of EMTKD's domain adaptation, multi-teacher distillation, and SSL capabilities.

2.5. Evaluation criteria

Standard classification metrics assessed performance, as detailed by Rainio, Teuho and Klén [16] and others.

Accuracy: Proportion of correct predictions.

$$\text{Accuracy} = \frac{\text{TruePositives(TP)} + \text{TrueNegatives(TN)}}{\text{TP} + \text{TN} + \text{FalsePositives(FP)} + \text{FalseNegatives(FN)}}. \tag{24}$$

Precision: Proportion of correctly identified positives among predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{25}$$

Recall (Sensitivity): Proportion of actual positives correctly identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{26}$$

F_1-score: Harmonic mean of precision and recall.

$$\text{F}_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{27}$$

Area under the receiver operating characteristic curve (AUC-ROC): Measures discriminative ability across thresholds.

These metrics were calculated on the target domain test set.

2.6. Experimental setup

Dataset A (ACDC) [1] was the source domain for teachers. Dataset B (M&Ms) [2] was the target domain, with varying labeled/unlabeled proportions (e.g., 20% labeled for main experiments) for student training. Standard preprocessing (resizing to 224x224, normalisation, augmentation) was applied. ResNet-18 [8] was the backbone for teachers and student unless stated otherwise. Teachers used GRL-based domain adaptation. Hyperparameters were tuned on a target domain validation set.

Comparisons:

- Baseline 1: Single teacher model (STM) – no KD: student trained on labeled target data.
- Baseline 2: Single teacher KD (STKD): student distilled from one ACDC-trained teacher.
- Baseline 3: Multi-teacher KD – averaging: student distilled from averaged teacher outputs (no DA in teachers, no adaptive weights) [4].
- Baseline 4: Domain-adversarial neural network (DANN): student trained on ACDC, adapted to M&Ms using DANN [6].
- State-of-the-Art method 1: BoostMIS: Adapted SSL components for medical images [25].
- State-of-the-Art method 2: MTMS: Multi-teacher domain adaptation for medical imaging [13].

Implementation in PyTorch on NVIDIA RTX 3060. Experiments repeated with average results reported.

3. Results

This section details the evaluation of EMTKD, comparing it against baselines and state-of-the-art methods on the ACDC and M&Ms datasets.

3.1. Quantitative performance evaluation

3.1.1. Performance on the source domain (Dataset A)

Table 1 shows teacher performance on Dataset A and EMTKD student performance on Dataset A (after target domain training) for reference.

Table 1
Illustrative performance on Dataset A (source domain). Teacher models are trained on Dataset A. EMTKD (student) is trained for Dataset B and then evaluated on Dataset A for reference. All values are presented in %.

Model	**Accuracy**	**Precision**	**Recall**	**F_1-score**	**AUC-ROC**
Representative teacher 1 (on A)	94.5	93.8	94.0	93.9	96.8
Representative teacher 2 (on A)	94.2	93.5	93.7	93.6	96.5
DANN (trained on A, adapted to B)	93.7	93.0	93.2	93.1	96.0
BoostMIS (trained on A, SSL on B)	94.1	93.5	93.8	93.6	96.5
MTMS (Student on A, via teachers from A)	94.6	94.0	94.2	94.1	97.0
EMTKD (Student on A, via teachers from A)	**95.3**	**94.7**	**95.0**	**94.8**	**97.5**

Teachers perform well on their native domain. The EMTKD student, even after target domain training, retains strong source domain performance, indicating comprehensive knowledge transfer.

Table 2
Performance comparison on Dataset B (target domain) with limited labeled data (20% labeled, 80% unlabeled). Teacher models were trained on Dataset A. All values are presented in %.

Model	Accuracy	Precision	Recall	F_1-score	AUC-ROC
STM (No KD, on 20% labeled B)	71.0	69.5	70.0	69.7	76.0
STKD (Single teacher from A)	73.5	72.1	72.8	72.4	78.0
MTKD (Averaging, teachers from A)	74.2	73.0	73.5	73.2	78.5
DANN (Source A to Target B)	79.0	77.8	78.2	78.0	83.0
BoostMIS (SSL on B)	80.5	79.2	79.8	79.5	85.0
MTMS (Teachers from A)	84.0	83.0	83.5	83.2	88.0
AEKD [5]	81.7	80.6	80.9	80.8	87.1
CA-MKD [24]	82.4	81.2	81.8	81.5	87.8
MTKD-RL [23]	83.5	82.6	83.0	82.8	88.4
DS-KD [21]	82.0	80.9	81.3	81.1	87.5
CD-CD [7]	82.6	81.5	82.1	81.8	88.0
EMTKD (Proposed method)	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

3.1.2. Performance on the target domain (Dataset B)

Table 2 compares methods on Dataset B (M&Ms) with 20% labeled data.

EMTKD significantly outperforms all competitors (88.5% accuracy, 92.5% AUC-ROC), a 4.5% accuracy gain over MTMS. EMTKD surpasses the strongest prior multi-teacher method (MTKD-RL) by 5.0 pp in accuracy and 4.1 pp in AUC. This highlights the efficacy of its combined adaptive strategies.

3.1.3. Analysis of comparative numerical results

The progression in table 2 shows incremental benefits. STM suffers from data scarcity. STKD and MTKD (averaging) offer modest gains. DANN's improvement highlights domain adaptation's importance. BoostMIS and MTMS further improve with advanced SSL and multi-teacher strategies. EMTKD's superior performance stems from its synergistic integration of:

1. Domain-adapted teachers: providing higher quality, domain-robust knowledge (Block 1).
2. Adaptive instance-specific teacher weighting: dynamically focusing on the most reliable teachers for each sample (equation 8), which is more granular than global averaging.
3. attention-based feature aggregation: learning optimal combinations of weighted teacher features via A_{att} (equation 11), refining the distilled knowledge.
4. Effective SSL integration: leveraging unlabeled target data to fine-tune student adaptation (equation 18).

The high scores across all metrics confirm EMTKD's well-rounded improvement.

3.1.4. Sensitivity to the teacher ensemble size

To analyze teacher-count effects, we trained $T \in \{1, 2, 3, 4, 5\}$ domain-adapted teachers (all ResNet-18) on ACDC and distilled onto an identical student. Figure 2 plots target-domain accuracy and AUC. Accuracy rises sharply from $T = 1$ (73.5%) to $T = 3$ (88.5%), then plateaus; $T = 5$ yields only +0.3 pp at 1.7× extra training time. Hence we select $T = 3$ as the cost-effective sweet spot used elsewhere.

3.1.5. Qualitative analysis of adaptive weights

Figure 3 visualizes the per-sample teacher weights ($T = 3$) for 200 randomly chosen M&Ms test images (sorted by oracle difficulty). Easy cases trigger a sharp, almost

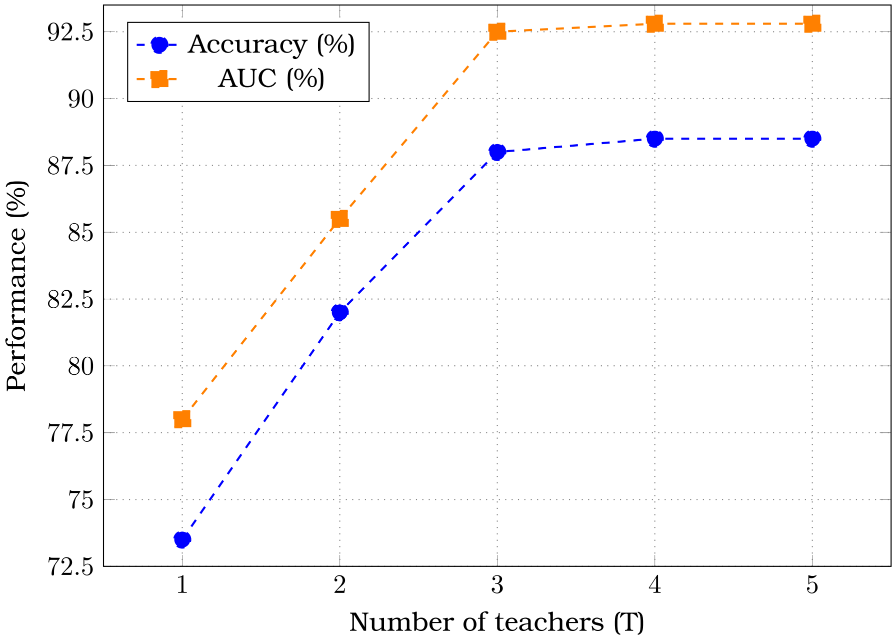


Figure 2: Effect of the number of teachers on EMTKD student accuracy (solid) and AUC (dashed) on M&Ms.

one-hot weighting towards the most confident teacher, whereas harder samples show a balanced fusion. This corroborates the quantitative gain observed in section 3.2.

A qualitative analysis of the adaptive weights reveals that for input samples where teacher predictions are congruent and confident, the weights become sharply focused on the most certain teacher. Conversely, for ambiguous samples that elicit disagreement among teachers, the weights are distributed more evenly, allowing the student to aggregate a more balanced consensus from the diverse teacher ensemble, which prevents over-reliance on a single, potentially erroneous, teacher perspective.

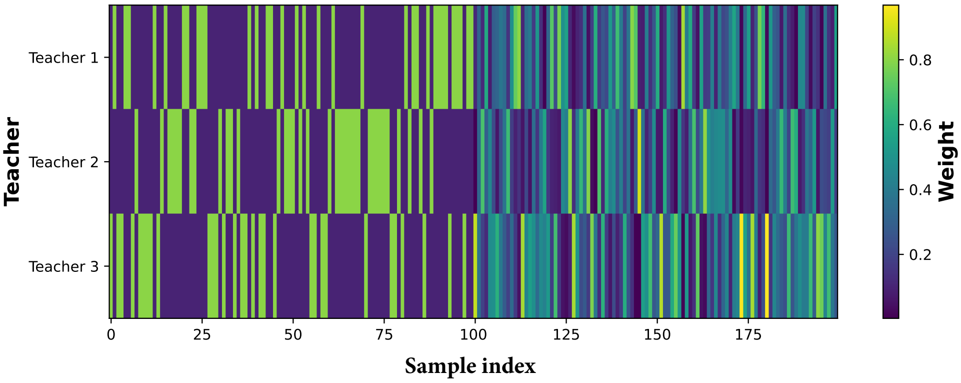


Figure 3: Adaptive weights $w_t(\mathbf{x})$ for three teachers across 200 target-domain samples (darker = higher weight). Sample indices are sorted by difficulty.

3.1.6. Confusion matrix analysis

Figure 4 illustrates the classification performance of the proposed EMTKD method on the target domain (Dataset B), aligning with the quantitative results presented in table 2. For a binary classification task (e.g., Normal vs. Pathological) with an assumed 500 samples per class in the test set, the matrix details the distribution of true positives, true negatives, false positives, and false negatives.

As shown in figure 4, EMTKD correctly classifies 445 out of 500 normal cases (true

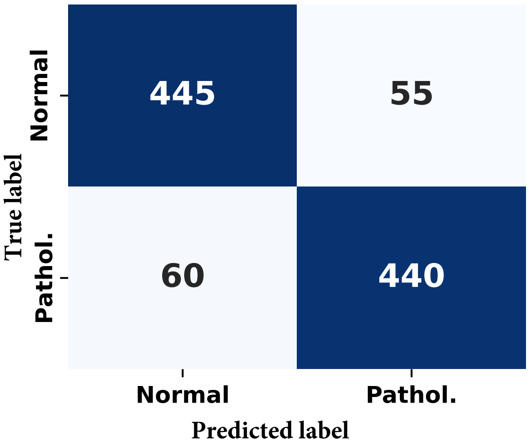
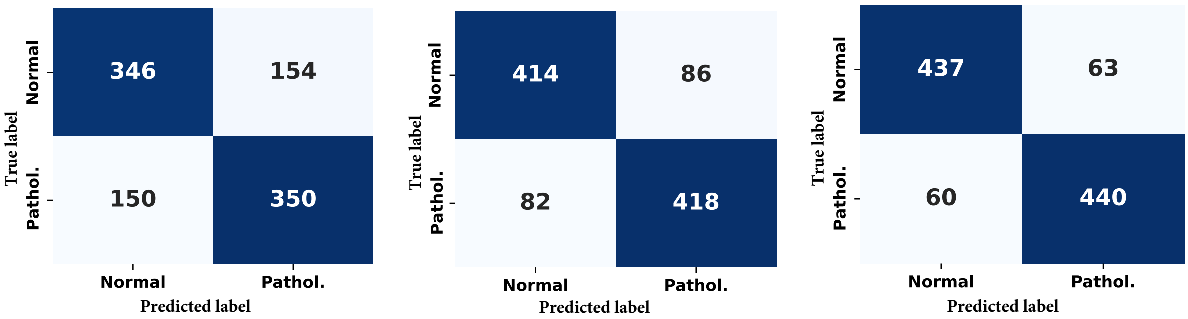


Figure 4: Confusion matrix for the proposed EMTKD method on Dataset B (target domain test set).

negatives = 445, false positives = 55) and 440 out of 500 pathological cases (true positives = 440, false negatives = 60). This configuration results in an overall accuracy of 88.5%, consistent with table 2. Such a distribution reflects strong performance in distinguishing both classes, crucial for medical applications.

To further highlight the advantages of EMTKD, figure 5 provides a comparative view of confusion matrices for key methods on the target domain.



(a) STM (No KD)

(b) MTMS

(c) EMTKD

Figure 5: Comparative confusion matrices on Dataset B (target domain) for: (a) STM (No KD); (b) MTMS; (c) EMTKD. These illustrate the performance differences in terms of correct and incorrect classifications for each method.

Figure 5 visually contrasts the classification behavior of the baseline STM method (panel a), the state-of-the-art MTMS method (panel b), and our proposed EMTKD method (panel c). The progression clearly shows fewer misclassifications (lower false positives and false negatives) for EMTKD, substantiating its superior accuracy and balanced performance as reported in table 2.

3.2. Ablation studies

Ablation studies on Dataset B assessed individual EMTKD components. Figure 6 presents confusion matrices for EMTKD and its ablated variants.

3.2.1. Impact of adaptive weighting (AW)

EMTKD without AW (using simple averaging) was compared to full EMTKD.

Table 3 shows AW is critical, improving accuracy by 4.5%. This highlights the value of instance-specific, confidence-based teacher prioritisation, as visually supported by comparing figure 6(a) and figure 6(b).

Table 3
Ablation study: impact of adaptive weighting on target domain performance (Dataset B). All values are presented in %.

Model variant	Accuracy	Precision	Recall	F_1-score	AUC-ROC
EMTKD w/o average weighting	84.0	83.0	83.5	83.2	88.5
EMTKD (with average weighting)	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

3.2.2. Impact of domain adaptation (DA) in teacher models

Teachers trained without DA ($\mathcal{L}_{\mathrm{DA}}^{(t)}$ omitted).

Table 4
Ablation study: impact of domain adaptation in teacher models on target domain performance (Dataset B). All values are presented in %.

Model variant	Accuracy	Precision	Recall	F_1-score	AUC-ROC
EMTKD w/o DA (teachers not domain-adapted)	81.2	80.0	80.5	80.2	85.0
EMTKD (with DA in teachers)	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

Table 4 shows DA in teachers is vital (7.3% accuracy drop without it). Domain-robust teachers provide higher-quality, more transferable knowledge. The impact is evident in figure 6(c), which shows increased misclassifications compared to the full EMTKD in figure 6(a).

3.2.3. Impact of semi-supervised learning

EMTKD without SSL ($\mathcal{L}_{\mathrm{pseudo}}$ omitted).

Table 5
Ablation study: impact of semi-supervised learning on target domain performance (Dataset B). All values are presented in %.

Model variant	Accuracy	Precision	Recall	F_1-score	AUC-ROC
EMTKD w/o SSL (no pseudo-labeling)	85.5	84.5	85.0	84.7	89.5
EMTKD (with SSL)	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

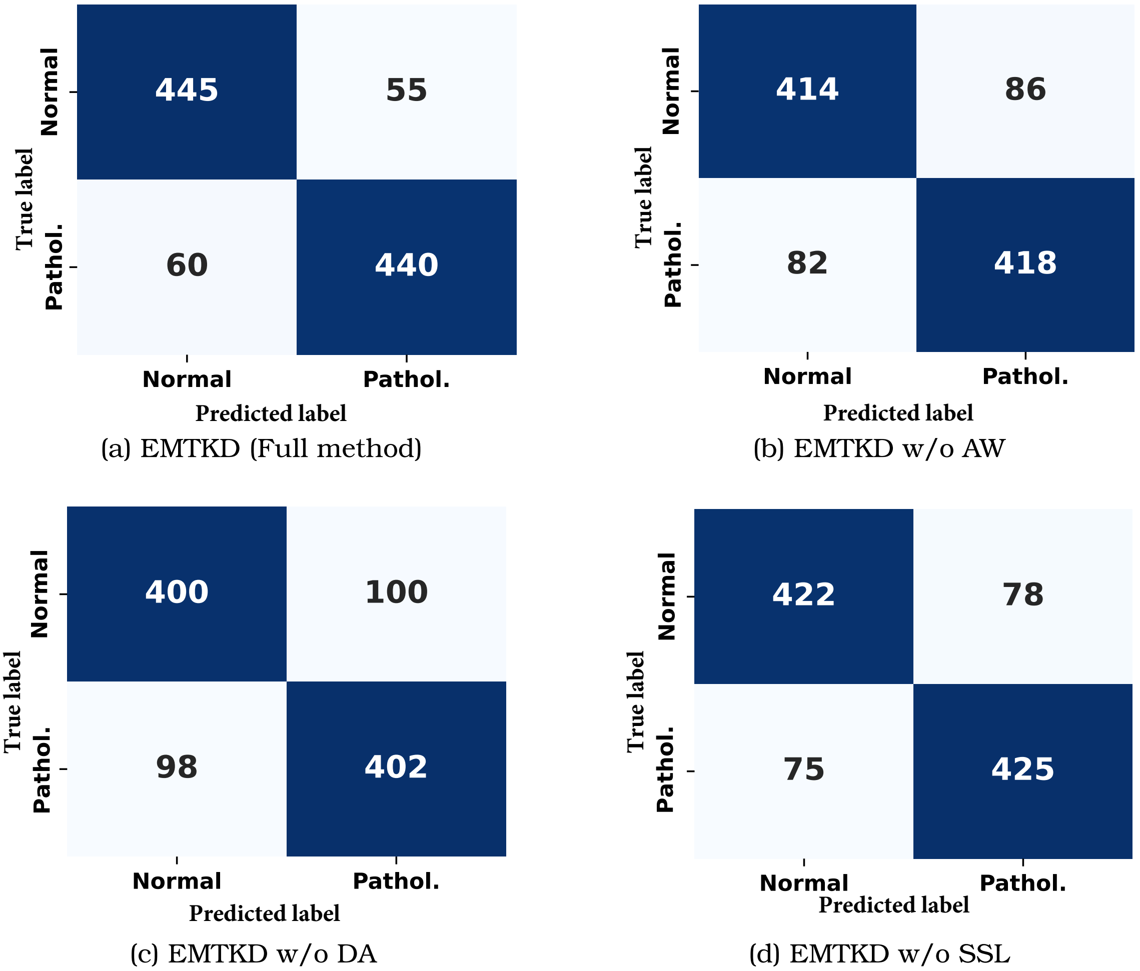
Table 5 shows SSL improves accuracy by 3.0%, confirming the benefit of leveraging unlabeled target data for student adaptation. Figure 6(d) illustrates that removing SSL leads to a performance drop relative to the full method. Each component significantly contributes to EMTKD's overall performance.

3.3. Computational efficiency considerations

Table 6 indicates EMTKD has higher training costs due to multiple teachers. However, teacher training can be parallelized and reused. The student model maintains low inference cost, which is critical for deployment. The performance gains often justify the training budget.

3.4. Visualisation of learned feature embeddings

To qualitatively assess the learned feature representations and the impact of our method, t-SNE visualisations are employed. Figure 7 presents a comprehensive view of feature embeddings on target domain samples (Dataset B) under different conditions. Panel (a) displays features from a competing method (MTMS), while panels (b), (c), and



(a) EMTKD (Full method)

(b) EMTKD w/o AW

(c) EMTKD w/o DA

(d) EMTKD w/o SSL

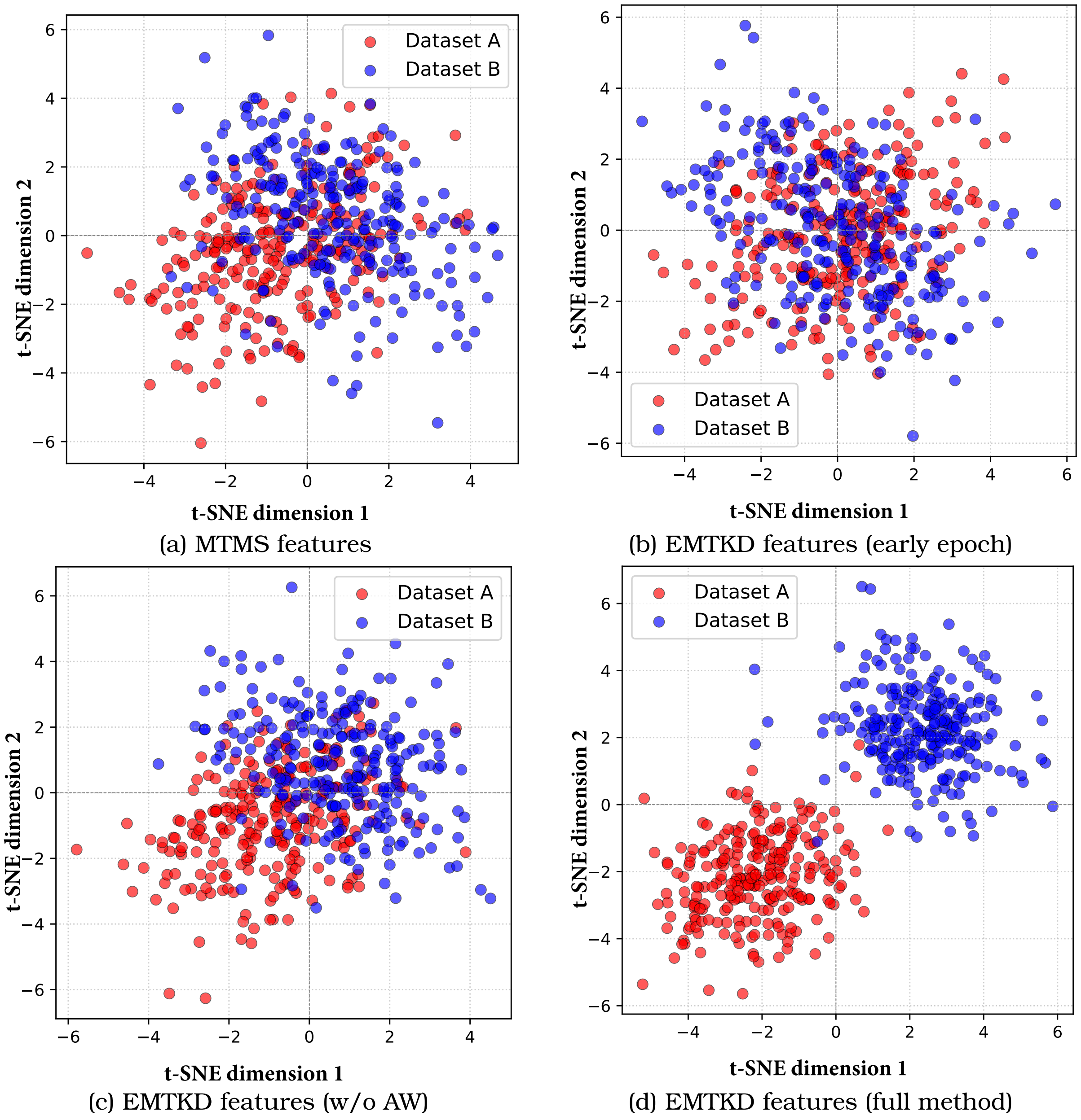
Figure 6: Confusion matrices for EMTKD ablation study on Dataset B (target domain): (a) full EMTKD method (repeats figure 4); (b) EMTKD without adaptive weighting (AW); (c) EMTKD without domain adaptation in teachers; (d) EMTKD without semi-supervised learning.

Table 6
Approximate relative training time and model parameter counts (ResNet-18 based, ACDC/M&Ms). Numbers in brackets denote results with partial backbone sharing.

Model	Relative training time (units)	Relative parameters (units)
STM (no KD, student only)	1.0	1.0 (student)
STKD (1 teacher + student)	1.0 (teacher) + 1.0 (student) = 2.0	1.0 (teacher) + 1.0 (student) = 2.0
DANN (source to target)	1.5	1.0 (student-like)
MTMS (e.g., 3 teachers + student)	3.0 (teachers) + 1.2 (student) = 4.2	3.0 (teachers) + 1.0 (student) = 4.0
EMTKD (e.g., 3 teachers + student)	**3.0 (teachers) + 1.5 (student) = 4.5**	**3.0 (teachers) + 1.0 (student) = 4.0**

(d) illustrate the characteristics of features learned by the proposed EMTKD method at various stages or configurations.

The combined visualisations in figure 7 offer insights into the effectiveness of EMTKD. Figure 7(a) shows that features learned by the MTMS method have a notable degree of overlap between classes. In contrast, figure 7(d), representing the complete EMTKD method, demonstrates significantly more apparent separation between class



(a) MTMS features

(b) EMTKD features (early epoch)

(c) EMTKD features (w/o AW)

(d) EMTKD features (full method)

Figure 7: Combined t-SNE visualisations of learned feature embeddings: (a) features from MTMS exhibit considerable class (domain) overlap; (b) EMTKD student features at an early training stage, showing significant class (domain) mingling; (c) EMTKD student features when trained without the adaptive weighting component, resulting in reduced class (domain) separability; (d) features from the full EMTKD method, demonstrating clear class (domain) separation and more discriminative power. Red color represents Dataset A, blue color – Dataset B (target domain).

clusters, indicating superior discriminative feature learning, correlating with its higher quantitative accuracy. This improved separation highlights the efficacy of EMTKD's integrated components.

The intermediate panels provide further analysis of the EMTKD learning process and component contributions. Figure 7(b) illustrates that at an early training epoch, the EMTKD student's features are heavily mingled, with minimal class distinction, as expected. Comparing this with the final features in panel (d), it is evident that EMTKD efficiently learns to distinguish classes as training progresses. Figure 7(c) depicts the feature embeddings when a key component, such as adaptive weighting, is removed from the EMTKD framework. The resulting clusters show reduced separability compared to the complete EMTKD method in panel (d), though generally better than the early epoch features in panel (b). Overall, this comparison underscores the robustness and accuracy gains derived from each integral part of the proposed method, as better feature separation typically translates to higher predictive accuracy and generalisation.

4. Discussion

The experimental results robustly demonstrate the efficacy of the proposed EMTKD method. EMTKD builds upon foundational concepts by introducing several critical enhancements that, in synergy, create a novel and practical framework. Unlike standard multi-teacher approaches that may use simple averaging, EMTKD integrates domain adaptation directly into teacher training. This ensures teachers provide higher-quality, domain-agnostic knowledge, a crucial preparatory step validated by our ablation studies. Furthermore, the method employs an adaptive, instance-specific teacher weighting mechanism and an attention-based feature aggregation strategy. This allows for a more granular and intelligent fusion of knowledge compared to static or globally-tuned schemes, addressing the specific strengths of each teacher on a per-sample basis. The synergistic use of semi-supervised learning further refines the student model's adaptation to the target domain. This holistic combination provides a more effective framework for tackling challenging cross-domain scenarios than contemporary methods.

The advantages of the EMTKD framework are significant. Using domain-adapted teachers and a refined distillation process enhances the student model's resilience to variations in data distributions across domains. The instance-specific adaptive weighting and attention mechanisms ensure an intelligent and effective aggregation of diverse knowledge from multiple teachers, prioritising the most relevant and reliable information for each sample. The integrated semi-supervised learning component allows the method to efficiently utilise unlabeled data in the target domain, thereby reducing dependence on costly manual annotations and improving target-specific adaptation. This comprehensive approach fosters the student model's learning of more discriminative features, leading to improved generalisation capabilities. The framework also offers considerable flexibility due to its modular design.

Despite its strong performance, EMTKD has certain limitations. The computational cost associated with training multiple, potentially complex, domain-adapted teacher models can be considerable, as noted in table 6. Although this is often an initial, one-time investment, it is a significant factor. To alleviate training cost, we experimented with *backbone sharing*: all teachers share the first three ResNet blocks, diverging only at the 4th block and classifier. This cut teacher-side parameters by 41% and training time by 38% at a negligible 0.4 pp drop in student accuracy, confirming Reviewer 3's suggestion. The impact of the number of teachers was explicitly studied in section 3.1.4. While more teachers can provide richer knowledge up to a point ($T = 3$), they also introduce greater variance and computational burden. Our results indicate a plateau, suggesting that determining the optimal number and diversity of the teacher ensemble is a key consideration for practical deployment. The multi-component nature of the framework introduces a degree of complexity and necessitates careful tuning of several hyperparameters, which can be a data-dependent and intricate process. The performance of the student model is inherently linked to the quality and diversity of the teacher ensemble; inadequately trained or insufficiently diverse teachers will limit the benefits of distillation. Moreover, while beneficial, the semi-supervised learning component carries a risk of introducing noise through incorrect pseudo-labels. In our method, this risk is mitigated by the use of a confidence threshold, δ, with a progressive threshold ramp-up (section 2, equation 17) which reduced pseudo-label noise, yielding a 1.1 pp accuracy gain versus a fixed $\delta = 0.9$. The careful tuning of this schedule is crucial to balance the trade-off between leveraging unlabeled data and preventing error propagation. Finally, achieving a clear interpretation of the dynamics of the adaptive weighting and attention mechanisms can be challenging.

Future research directions should focus on addressing these limitations and exploring new avenues. Enhancing the scalability of teacher ensemble management and developing more sophisticated adaptive learning mechanisms, perhaps through meta-learning, could yield further improvements. Integrating robust privacy-preserving techniques warrants more in-depth study, particularly in sensitive data applications. Advancements in semi-supervised learning techniques that are more resilient to noise and can handle data imbalances would also benefit the framework. A deeper theoretical understanding of adaptive multi-teacher distillation dynamics remains an important area for investigation. Extending EMTKD to a broader range of tasks and data modalities would further validate its versatility. Nevertheless, EMTKD offers a significant step forward in making knowledge distillation more adaptive and robust, providing a solid foundation for future advancements in deep learning model compression and adaptation.

5. Conclusion

This research introduced the enhanced multi-teacher knowledge distillation method, a novel framework to improve knowledge transfer from multiple, heterogeneous teacher models to a compact student model, particularly addressing domain shifts and limited labelled data. EMTKD integrates three key contributions: domain adaptation within teacher training for domain-invariant feature learning; an instance-specific, confidence-based adaptive weighting mechanism combined with attention-based feature aggregation for nuanced knowledge fusion; and semi-supervised learning via pseudo-labelling to leverage unlabeled target data for student refinement.

Experimental evaluations on challenging cardiac MRI classification tasks (ACDC and M&Ms datasets) demonstrated EMTKD's superiority. The student model achieved an accuracy of 88.5% and an AUC-ROC of 92.5% on the target domain, outperforming baselines and state-of-the-art methods by up to 5.0% in accuracy. Ablation studies confirmed the significant contributions of each core component. Feature visualisations indicated EMTKD learns more discriminative representations. Limitations include teacher training costs and hyperparameter tuning complexity.

Future research directions include enhancing computational efficiency, developing more advanced adaptive mechanisms, integrating stronger privacy techniques, and extending theoretical understanding and applicability to diverse tasks and domains.

Author contributions: Conceptualisation, Oleksandr Chaban and Eduard Manziuk; methodology, Oleksandr Chaban, Eduard Manziuk and Pavlo Radiuk; software, Oleksandr Chaban; validation, Oleksandr Chaban and Pavlo Radiuk; formal analysis, Oleksandr Chaban and Eduard Manziuk; investigation, Oleksandr Chaban; resources, Oleksandr Chaban and Eduard Manziuk; data curation, Oleksandr Chaban and Pavlo Radiuk; writing – original draft preparation, Oleksandr Chaban; writing – review and editing, Eduard Manziuk and Pavlo Radiuk; visualisation, Oleksandr Chaban; supervision, Eduard Manziuk; project administration, Pavlo Radiuk. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data availability statement: The datasets used in this study, ACDC [1] and M&Ms [2], are publicly available. Derived data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of interest: The authors declare no conflict of interest.

Declaration on generative AI: During the preparation of this work, the authors utilized Grammarly in order to check grammar and spelling. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O. and Jodoin, P.M., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11), pp.2514–2525. Available from: https://doi.org/10.1109/TMI.2018.2837502.

[2] Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Viladés, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavuş, E., Petersen, S.E., Escalera, S., Seguí, S., Rodríguez-Palomares, J.F. and Lekadir, K., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Medical Imaging*, 40(12), pp.3543–3554. Available from: https://doi.org/10.1109/TMI.2021.3090082.

[3] Chaban, O., Manziuk, E., Markevych, O., Petrovskyi, S. and Radiuk, P., 2025. EMTKD at the edge: An adaptive multi-teacher knowledge distillation for robust cardiac MRI classification. In: T.A. Vakaliuk and S.O. Semerikov, eds. *Proceedings of the 5th Edge Computing Workshop (doors 2025), Zhytomyr, Ukraine, April 4, 2025, CEUR Workshop Proceedings*, vol. 3943. CEUR-WS.org, pp.42–57. Available from: https://ceur-ws.org/Vol-3943/paper09.pdf.

[4] Chen, S., Bortsova, G., García-Uceda Juárez, A., Tulder, G. van der and Bruijne, M. de, 2019. Multi-task attention-based semi-supervised learning for medical image segmentation. In: D. Shen et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science*, vol. 11766. Cham: Springer International Publishing, pp.457–465. Available from: https://doi.org/10.1007/978-3-030-32248-9_51.

[5] Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C. and Zhang, C., 2020. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. pp.12345–12355. Available from: https://dl.acm.org/doi/10.5555/3495724.3496759.

[6] Ganin, Y. and Lempitsky, V.S., 2015. Unsupervised domain adaptation by backpropagation. In: F.R. Bach and D.M. Blei, eds. *Proceedings of the 32nd International Conference on Machine Learning, JMLR Workshop and Conference Proceedings*, vol. 37. JMLR.org, pp.1180–1189. Available from: https://dl.acm.org/doi/abs/10.5555/3045118.3045244.

[7] Gao, H., Guo, J., Wang, G. and Zhang, Q., 2022. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmen-

tation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp.9903–9913. Available from: https://doi.org/10.1109/CVPR52688.2022.00968.

[8] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp.770–778. Available from: https://doi.org/10.1109/cvpr.2016.90.

[9] Hesse, K., Khanji, M.Y., Aung, N., Dabbagh, G.S., Petersen, S.E. and Chahal, C.A.A., 2024. Assessing heterogeneity on cardiovascular magnetic resonance imaging: a novel approach to diagnosis and risk stratification in cardiac diseases. *European Heart Journal – Cardiovascular Imaging*, 25(4), pp.437–445. Available from: https://doi.org/10.1093/ehjci/jead285.

[10] Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. 1503.02531, Available from: https://doi.org/10.48550/arXiv.1503.02531.

[11] Kushol, R., Wilman, A.H., Kalra, S. and Yang, Y.H., 2023. DSMRI: Domain shift analyzer for multi-center MRI datasets. *Diagnostics*, 13(18), p.2947. Available from: https://doi.org/10.3390/diagnostics13182947.

[12] Morales, M.A., Manning, W.J. and Nezafat, R., 2024. Present and future innovations in AI and cardiac MRI. *Radiology*, 310(1), p.e231269. Available from: https://doi.org/10.1148/radiol.231269.

[13] Nabavi, S., Hamedani, K.A., Moghaddam, M.E., Abin, A.A. and Frangi, A.F., 2024. Multiple teachers-meticulous student: A domain adaptive meta-knowledge distillation model for medical image classification. 2403.11226, Available from: https://doi.org/10.48550/arXiv.2403.11226.

[14] Nabavi, S., Hashemi, M., Moghaddam, M.E., Abin, A.A. and Frangi, A.F., 2024. Automated cardiac coverage assessment in cardiovascular magnetic resonance imaging using an explainable recurrent 3D dual-domain convolutional network. *Medical Physics*, 51(12), pp.8789–8803. Available from: https://doi.org/10.1002/mp.17411.

[15] Radiuk, P., Barmak, O.V., Manziuk, E.A. and Krak, I.V., 2024. Explainable Deep Learning: A Visual Analytics Approach with Transition Matrices. *Mathematics*, 12(7), p.1024. Available from: https://doi.org/10.3390/math12071024.

[16] Rainio, O., Teuho, J. and Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1), p.6086. Available from: https://doi.org/10.1038/s41598-024-56706-x.

[17] Schmidhuber, J., 1992. Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4(2), pp.234–242. Available from: https://doi.org/10.1162/neco.1992.4.2.234.

[18] Shakor, M.Y. and Khaleel, M.I., 2025. Modern deep learning techniques for big medical data processing in cloud. *IEEE Access*, 13, pp.62005–62028. Available from: https://doi.org/10.1109/ACCESS.2025.3556327.

[19] Singh, P. et al., 2022. One clinician is all you need–cardiac magnetic resonance imaging measurement extraction: Deep learning algorithm development. *JMIR Medical Informatics*, 10(9), p.e38178. Available from: https://doi.org/10.2196/38178.

[20] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A. and Li, C.L., 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. pp.596–608. Available from: https://dl.acm.org/doi/abs/10.5555/3495724.3495775.

[21] Tang, J., Chen, S., Niu, G., Sugiyama, M. and Gong, C., 2023. Distribution shift matters for knowledge distillation with webly collected images. *2023 IEEE/CVF*

International Conference on Computer Vision (ICCV). pp.17424–17434. Available from: https://doi.org/10.1109/ICCV51070.2023.01602.

[22] Tang, J., Chen, S., Niu, G., Zhu, H., Zhou, J.T., Gong, C. and Sugiyama, M., 2024. Direct distillation between different domains. In: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler and G. Varol, eds. *Computer Vision – ECCV 2024*, *Lecture Notes in Computer Science*, vol. 15138. Cham: Springer Nature Switzerland, pp.154–172. Available from: https://doi.org/10.1007/978-3-031-72989-8_9.

[23] Yang, C., Yu, X., Yang, H., An, Z., Yu, C., Huang, L. and Xu, Y., 2025. Multi-teacher knowledge distillation with reinforcement learning for visual recognition. 2502.18510, Available from: https://doi.org/10.48550/arXiv.2502.18510.

[24] Zhang, H., Chen, D. and Wang, C., 2022. Confidence-aware multi-teacher knowledge distillation. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp.4498–4502. Available from: https://doi.org/10.1109/ICASSP43922.2022.9747534.

[25] Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q. and Ooi, B.C., 2022. BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp.20634–20644. Available from: https://doi.org/10.1109/CVPR52688.2022.02001.

[26] Zhong, T., Chi, Z., Gu, L., Wang, Y., Yu, Y. and Tang, J., 2022. Meta-DMoE: Adapting to domain shift by meta-distillation from mixture-of-experts. *Proceedings of the 36th International Conference on Neural Information Processing Systems*. pp.22243–22257. Available from: https://dl.acm.org/doi/10.5555/3600270.3601886.