EMTKD at the edge: An adaptive multi-teacher knowledge distillation for robust cardiac MRI classification

Oleksandr Chaban[1,*], Eduard Manziuk[1], Olena Markevych[2], Sergii Petrovskyi[1] and Pavlo Radiuk[1]

[1] *Khmelnytskyi National University, 11 Instytutska Str., Khmelnytskyi, 29016, Ukraine*
[2] *Khmelnytskyi Infectious Diseases Hospital, 17 Skovorody Str., Khmelnytskyi, 29008, Ukraine*

Abstract

Cardiac magnetic resonance imaging (MRI) is vital for diagnosing cardiovascular diseases, yet its classification faces significant challenges, including domain shifts across heterogeneous datasets, limited annotated data, and privacy regulations, particularly when deploying deep learning models on resource-constrained edge devices. These issues hinder real-time diagnostics and data protection in edge-computing scenarios. To address these issues, this paper introduces Enhanced Multi-Teacher Knowledge Distillation (EMTKD), an edge-based framework designed to enhance cardiac MRI classification. EMTKD employs adaptive multi-teacher knowledge distillation, utilizing an adaptive weighting mechanism to balance contributions from multiple teacher models trained on diverse domains, ensuring the student model learns from the most reliable sources. It integrates domain adaptation to align feature representations, semi-supervised learning with pseudo-labeling to leverage unannotated data, and differential privacy to safeguard patient information. Experimental results show EMTKD achieves an accuracy of 88.5% and an area under the curve of 92.5% on target domain images, outperforming state-of-the-art baseline models. Ablation studies highlight the critical roles of adaptive weighting, domain adaptation, and semi-supervised learning in improving performance under edge constraints. These findings demonstrate EMTKD's ability to generalize across domains and utilize limited data effectively. Overall, EMTKD offers a robust, privacy-preserving solution for cardiac MRI classification on edge devices, ideal for rapid diagnostics in challenging environments.

Keywords

edge computing, monitoring and diagnosis, knowledge distillation, domain adaptation, cardiac MRI classification, semi-supervised learning, privacy preservation

1. Introduction

Cardiovascular diseases remain the foremost cause of mortality worldwide, creating an urgent need for precise, timely diagnoses to improve patient outcomes. Cardiac magnetic resonance imaging (MRI) has become a crucial non-invasive modality for examining cardiac structure, function, and pathology [1]. It delivers the high-resolution information necessary for detecting conditions such as myocardial infarction, cardiomyopathies, and congenital heart diseases [2]. However, interpreting cardiac MRI scans manually is time-intensive and requires substantial expertise [3], which is not always available in every clinical setting. Consequently, automated techniques are needed to streamline MRI data analysis and assist clinicians with accurate interpretations [4].

Despite progress in deep learning (DL) and machine learning for medical image analysis, processing cardiac MRI data still faces two major hurdles. First is the heterogeneity of data sources [5]: scans acquired via different imaging protocols and equipment often exhibit domain shifts [6]. Consequently, models trained on one dataset tend to underperform when evaluated on unseen data. Second, annotated datasets are scarce; labeling MRI scans demands substantial expert effort [7], meaning large-scale labeled data are seldom available [8]. Additionally, privacy regulations prevent the free exchange of clinical data between institutions, further hindering the creation of comprehensive training sets.

doors-2025: 5th Edge Computing Workshop, April 4, 2025, Zhytomyr, Ukraine

[*] Corresponding author.
chabanolek@khmnu.edu.ua (O. Chaban); manziuk.e@khmnu.edu.ua (E. Manziuk); elena--14@ukr.net (O. Markevych); petrovskijss@khmnu.edu.ua (S. Petrovskyi); radiukp@khmnu.edu.ua (P. Radiuk)
0009-0001-4710-3336 (O. Chaban); 0000-0002-7310-2126 (E. Manziuk); 0000-0003-2758-3288 (O. Markevych); 0000-0002-0590-0484 (S. Petrovskyi); 0000-0003-3609-112X (P. Radiuk)

Emerging strategies like knowledge integration [9] and knowledge distillation [10] offer ways to address these challenges. Knowledge integration merges information from varied origins to bolster model learning [11], while knowledge distillation transfers insights from one or more teacher models to a student model [12]. In cardiac MRI, these methods help build models that generalize better and reduce the requirement for extensive annotation [13]. By harnessing knowledge distilled from multiple teachers trained on diverse data, the student model captures a broader range of features [14], boosting overall classification performance on new samples.

In light of these shortcomings – particularly domain generalization [15] and data scarcity [16] – this study focuses on distilling knowledge from multiple teacher models to establish a robust student model capable of operating effectively under heterogeneous conditions. The main contributions of this work are as follows:

- **Adaptive Teacher Weighting Mechanism:** Employs confidence calibration to dynamically balance each teacher's impact on the student model.
- **Domain-Adaptive Feature Alignment:** Extracts domain-invariant features from teacher models, enhancing the student model's cross-domain generalization.
- **Semi-Supervised Learning (SSL) Integration:** Utilizes pseudo-labeling for unannotated data, reducing dependence on large annotated datasets.
- **Privacy-Preserving Techniques:** Implements differential privacy to protect sensitive information during training.

By tackling key limitations in cardiac MRI analysis, this research aims to develop DL models that are not only accurate and generalizable but also aligned with essential privacy requirements.

The rest of the manuscript is structured as follows. Section 2 reviews key studies on DL in medical imaging, highlighting limitations in knowledge distillation, domain adaptation, semi-supervised learning, and privacy preservation—underscoring the impetus for a more holistic framework. Section 3 presents the authors' proposed EMTKD solution, detailing the integration of domain-adaptive teacher models, an adaptive weighting mechanism, SSL with pseudo-labeling, and differential privacy to address domain shift, data scarcity, and confidentiality concerns. Section 4 showcases the experimental results and discussion, comparing the new method against baseline and state-of-the-art techniques using standard evaluation metrics. Finally, section 5 concludes on the efficacy and broader applicability of the EMTKD approach.

2. Related works

Over recent times, DL-based solutions have increasingly permeated medical imaging, yielding progress in classification, segmentation, and detection tasks. However, new methods are needed to resolve outstanding issues in cardiac MRI processing.

Knowledge distillation has been a prominent direction: Hinton et al. [10] introduced the idea of transferring a larger model's knowledge to a smaller student network by matching outputs. Chen et al. [17] later proposed multi-teacher knowledge distillation, where each teacher is trained on a distinct dataset, and the student learns from the collective. Although beneficial, combining knowledge from heterogeneous data sources remains difficult because domain shifts among datasets may create conflicting information.

Domain adaptation has likewise garnered attention. Ganin and Lempitsky [18] explored domain-adversarial training to produce domain-invariant representations, and Zeng et al. [19] employed an unsupervised strategy for cross-modality medical image segmentation. Although these methods improve generalization, they often presume direct access to both source and target domains – an assumption complicated by privacy constraints.

To counteract limited annotations, SSL approaches like that of Yuan et al. [20] employ minimal labeled data combined with substantial unlabeled data. Their teacher-student structure uses pseudo-labeling to leverage unannotated samples. Yet, they do not incorporate multi-teacher integration or address domain shifts.

Privacy-preserving methods, such as adding differential privacy noise to training gradients, have also emerged in medical imaging. Montenegro et al. [21] demonstrated that this approach reduces the likelihood of memorizing individual data points. However, fusing differential privacy with complex training architectures, including knowledge distillation, remains challenging [22].

A more recent approach by Nabavi et al. [23] applies a multi-teacher, single-student framework to address domain shift and data scarcity. While it preserves privacy by requiring only teacher parameters, its performance may deteriorate if confronted with entirely novel data distributions. Some knowledge distillation gaps in cardiac MRI classification thus remain unsolved. Specifically, persistent obstacles include:

- **Difficulty Integrating Multiple Teachers:** Domain shifts across heterogeneous datasets hinder the student model's consistency.
- **Dependence on Source Domain Data:** Domain adaptation methods sometimes require both source and target data, which conflicts with privacy standards.
- **Limited Utilization of Unannotated Data:** Scarce labeled samples restrict model capacity, and existing methods do not fully exploit unlabeled data.
- **Inadequate Privacy Measures:** Knowledge distillation frameworks may overlook explicit safeguards for patient confidentiality.

Hence, the proposed solution addresses these gaps by embedding domain adaptation within teacher models to extract domain-invariant features. An adaptive weighting mechanism guided by confidence calibration ensures the student model assimilates knowledge from multiple teachers, and SSL broadens the training set with unlabeled samples. Lastly, differential privacy is woven into training to keep patient data confidential.

Accordingly, this study proposes an enhanced multi-teacher knowledge distillation (EMTKD) approach for cardiac MRI classification that integrates knowledge from diverse datasets, confronts domain shifts, and alleviates data scarcity. The corresponding tasks are:

- **Task 1:** Train domain-adapted teacher models on different cardiac MRI datasets.
- **Task 2:** Devise an adaptive distillation protocol to unify knowledge from multiple teachers into one student.
- **Task 3:** Employ SSL to exploit unlabeled data, advancing the student model's learning potential.
- **Task 4:** Incorporate privacy-preserving measures, including differential privacy, to secure sensitive information.
- **Task 5:** Compare the resulting method against existing approaches using standard cardiac MRI datasets.

By merging these techniques, the proposed approach aims to deliver robust, domain-spanning classification performance even under limited annotation conditions.

3. The proposed approach

The proposed approach aims to improve cardiac MRI image classification by effectively distilling knowledge from multiple teacher models into a single student model. The approach is formalized as a step-by-step method that addresses challenges such as domain shift, data heterogeneity, limited annotated data, and privacy concerns.

The method comprises three main blocks:

- **Block 1**: Teacher model training with domain adaptation.
- **Block 2**: Adaptive knowledge distillation.
- **Block 3**: Student model training with privacy preservation.

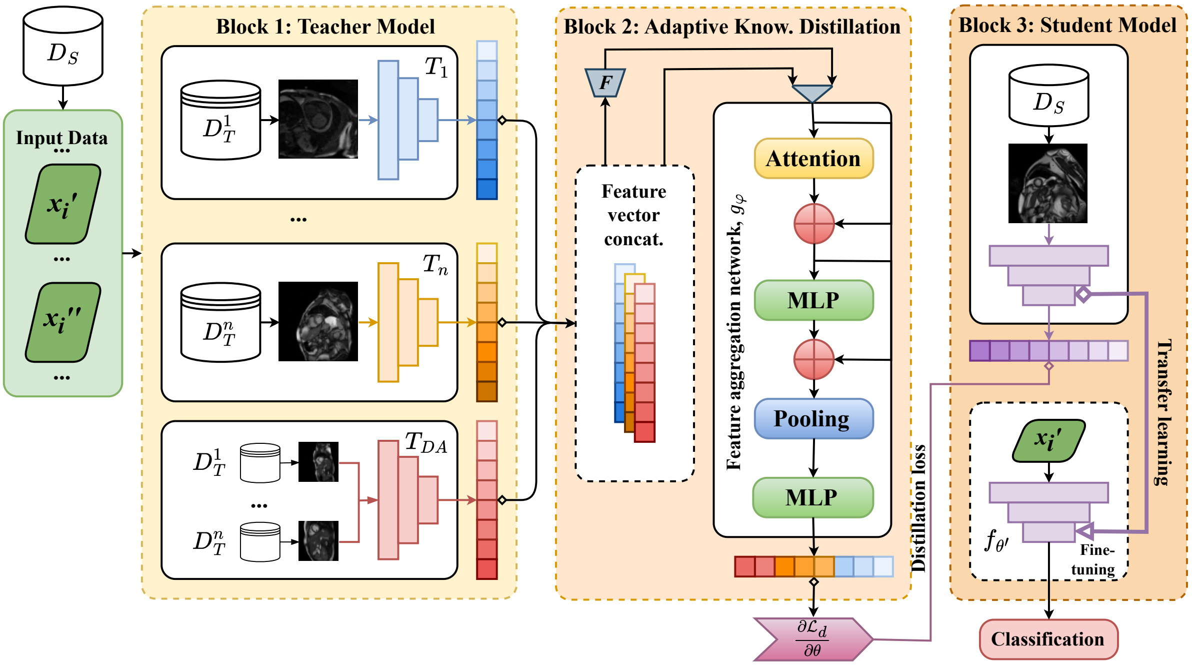


Figure 1: The proposed EMTKD comprises three blocks: (i) a multi-teacher model block that extracts diverse feature representations, (ii) an adaptive knowledge distillation block that aggregates and (iii) transfers knowledge using attention and pooling mechanisms, and a student model block that fine-tunes distilled knowledge for classification tasks through transfer learning.

The schematic representation of the proposed approach is illustrated in figure 1.

Each block involves specific inputs, processes, and outputs that collectively contribute to the final goal of enhancing the student model's performance.

3.1. Block 1: Teacher model training with domain adaptation

The input data of block 1 consists of annotated datasets from multiple domains. Each dataset is represented as $\mathcal{D}^{(t)} = \{(\mathbf{x}_i^{(t)}, y_i^{(t)}, d^{(t)})\}_{i=1}^{N_t}$ for $t = 1, 2, \ldots, T$, where $\mathbf{x}_i^{(t)}$ is the i-th cardiac MRI image from domain t, $y_i^{(t)}$ is the corresponding label (such as "normal" or "pathological"), and $d^{(t)}$ is the domain identifier. The total number of samples in dataset t is denoted by N_t, and T represents the total number of teacher models or domains involved.

Step 1. Data preparation

The first step in the process is collecting and preparing the data. Datasets are gathered from T different domains, each containing annotated cardiac MRI scans. Once the datasets are collected, preprocessing begins. The images are normalized, resized to a consistent dimension, and undergo necessary augmentations to enhance model performance. These preprocessing techniques ensure uniformity across the datasets and help improve the model's generalization capabilities.

Step 2. Model initialization

With the data prepared, the next step is to initialize the teacher models. For each domain t, a neural network model M_t is initialized with random weights. Additionally, for each model M_t, a domain discriminator network D_t is defined. This domain discriminator plays a critical role in domain adaptation, allowing the models to learn domain-invariant features that can generalize well across multiple domains.

Step 3. Training with domain adaptation

During training, two key loss functions are employed. The first is the classification loss for model M_t, which is calculated as:

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

where $y_i^{(t)}$ represents the true label for sample i and $M_t(\mathbf{x}_i^{(t)})$ is the predicted probability distribution over the classes.

The second loss function is the domain adaptation loss, which uses a Gradient Reversal Layer (GRL) to promote learning domain-invariant features. The domain adaptation loss is computed as:

$$\mathcal{L}_{\text{DA}}^{(t)} = -\frac{1}{N_t} \sum_{i=1}^{N_t} d^{(t)} \log D_t(M_t(\mathbf{x}_i^{(t)})), \tag{2}$$

where $d^{(t)}$ is the domain label and D_t represents the domain discriminator network.

These two loss functions, formalized by formulas (1)–(2), are combined into the final loss for each teacher model M_t:

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

where λ_{DA} is a hyperparameter that balances the classification and domain adaptation losses.

Step 4. Optimization

Once the loss functions are defined by (3), the next step is to update the model parameters. Using an optimizer, such as Adam, the parameters θ_t of model M_t are updated to minimize the combined teacher loss:

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

where η is the learning rate.

This optimization process by formula (4) adjusts the model weights to minimize classification and domain adaptation losses, enabling the model to learn useful patterns across domains.

Step 5. Feature extraction

After the models are trained, the next step is to extract features. For each sample $\mathbf{x}$ in dataset $\mathcal{D}^{(t)}$, the feature vector $\mathbf{z}^{(t)}$ is computed as:

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

These feature vectors by (5) represent the learned information from the models and are used in subsequent blocks for further processing.

The output data of block 1 consists of domain-adapted teacher models $\{M_t\}_{t=1}^{T}$ and the extracted feature vectors $\{\mathbf{z}_i^{(t)}\}_{i=1}^{N_t}$ for each dataset. The trained teacher models $\{M_t\}$ and their extracted feature representations $\{\mathbf{z}^{(t)}\}$ are now prepared and used as inputs for the subsequent processing in block 2.

3.2. Block 2: Adaptive knowledge distillation

The process of block 2 begins with the input data, which includes trained teacher models $\{M_t\}_{t=1}^{T}$ and their corresponding feature vectors $\{\mathbf{z}^{(t)}\}_{t=1}^{T}$, all derived from multiple domains. Additionally, we have the target dataset samples $\mathcal{D}^{(s)}$, which consist of both annotated and unannotated examples. The input data for the target domain includes samples $\mathbf{x}_i^{(s)}$, where $i = 1, 2, \ldots, N_s$, with $N_s = N_a + N_u$, where N_a and N_u represent the number of annotated and unannotated samples, respectively.

Step 1. Compute teacher predictions

The first step involves computing the predictions from the teacher models. For each teacher model M_t and target sample $\mathbf{x}_i^{(s)}$, the SoftMax function is applied to obtain predicted probabilities $\mathbf{p}_t(\mathbf{x}_i^{(s)})$. This is represented mathematically as:

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

In addition, confidence scores for the predictions are calculated by taking the maximum probability $\mathbf{p}_t(\mathbf{x}_i^{(s)})$ from formula (6), expressed as:

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

where $c_t(\mathbf{x}_i^{(s)})$ represents the confidence of the most probable class predicted by the teacher model.

Step 2. Calculate adaptive weights

Once the teacher predictions are computed, adaptive weights are calculated for each teacher. These weights are computed using a temperature-scaled SoftMax function to adjust the distribution of confidence scores. The adaptive weight $w_t(\mathbf{x}_i^{(s)})$ for teacher t is computed based on confidence scores from (7) in the following way:

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

where τ is a temperature parameter that controls the smoothness of the weight distribution. A lower τ sharpens the distribution, placing more emphasis on the higher confidence scores.

Step 3. Weighted feature aggregation

In the next step, weighted feature aggregation is performed. For each teacher model t, the weighted feature vector $\widetilde{\mathbf{z}}^{(t)}(\mathbf{x}_i^{(s)})$ is computed by multiplying the adaptive weight $w_t(\mathbf{x}_i^{(s)})$ in (8) with the feature vector $\mathbf{z}^{(t)}(\mathbf{x}_i^{(s)})$:

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

These weighted features, formalized by (9), from all teacher models are then concatenated into a single aggregated feature vector $\mathbf{Z}(\mathbf{x}_i^{(s)})$:

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

This concatenated feature vector (10) is subsequently passed through an attention-based aggregation network A_{att}, which learns to focus on the most informative aspects of the features:

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

The attention mechanism ensures that the most relevant features from the teacher models are emphasized during the aggregation process.

Step 4. Store aggregated features

The final step involves storing the aggregated feature representations $\mathbf{z}_{\mathrm{agg}}(\mathbf{x}_i^{(s)})$ by (11) for each sample in the target dataset. These features are critical for the next phase, where the student model is trained.

The output data of block 2 is a set of distilled and aggregated feature representations $\mathbf{z}_{\mathrm{agg}}$ for each sample $\mathbf{x}_i^{(s)}$. These features, now distilled from the teacher models, serve as targets for the student model training in block 3, enabling the student model to learn from the most relevant and informative features across multiple domains.

3.3. Block 3: Student model training with privacy preservation

The input data for this training process includes aggregated features $\{\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})\}$ derived from previous stages, as well as target dataset samples $\{\mathbf{x}_i^{(s)}\}$ from the dataset $\mathcal{D}^{(s)}$. Additionally, there is an annotated subset $\mathcal{D}^{(a)} = \{(\mathbf{x}_i^{(a)}, y_i^{(a)})\}_{i=1}^{N_a}$, which consists of labeled data used for supervised learning.

Step 1. Model initialization

The first step involves initializing the student model. A neural network S_θ is defined with parameters θ. Depending on the setup, these parameters can either be initialized randomly or use pre-trained weights. This model will learn from labeled and unlabeled data using a combination of techniques.

Step 2. Compute student features

In this step, the student model processes the target dataset. For each sample $\mathbf{x}_i^{(s)}$ from the target dataset, a forward pass is conducted through the student model, producing feature vectors $\widehat{\mathbf{z}}(\mathbf{x}_i^{(s)})$ as follows:

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

These feature vectors by (12) will later be compared to the aggregated features for distillation.

Step 3. Distillation loss computation

The distillation loss is computed to align the student's features with the teacher's. For each sample $\mathbf{x}_i^{(s)}$, the loss is given by the squared L2 norm between the student's feature vector $\widehat{\mathbf{z}}(\mathbf{x}_i^{(s)})$ and the aggregated feature $\mathbf{z}_{\text{agg}}(\mathbf{x}_i^{(s)})$:

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

where $\mathcal{R}(\theta)$ is a regularization term (e.g., L2 regularization on the model's parameters) and λ_{reg} controls the strength of this regularization.

Step 4. SSL with pseudo-labels

For the unlabeled samples $\mathbf{x}_i^{(u)}$, where $i = N_a + 1, \ldots, N_s$, pseudo-labels are generated by the student model. First, the predicted probabilities are computed using a SoftMax function applied to the student's classifier:

$$\mathbf{p}_s(\mathbf{x}_i^{(u)}) = \text{Softmax}(C(S_\theta(\mathbf{x}_i^{(u)}))). \tag{14}$$

Pseudo-labels are then assigned to samples where the maximum probability exceeds a confidence threshold δ:

$$\hat{y}_i^{(u)} = \begin{cases} \arg\max \mathbf{p}_s(\mathbf{x}_i^{(u)}), & \text{if } \max \mathbf{p}_s(\mathbf{x}_i^{(u)}) > \delta, \\ \text{unlabeled}, & \text{otherwise.} \end{cases} \tag{15}$$

For the samples assigned pseudo-labels, a pseudo-labeling loss is computed as:

$$\mathcal{L}_{\text{pseudo}} = -\frac{1}{N_p} \sum_{i=1}^{N_p} \hat{y}_i^{(u)} \log C(S_\theta(\mathbf{x}_i^{(u)})), \tag{16}$$

where N_p is the number of pseudo-labeled samples.

Step 5. Compute classification loss on annotated data

Next, the classification loss is computed for the annotated subset $\mathcal{D}^{(a)}$. This loss is based on the cross-entropy between the predicted probabilities and the true labels:

$$\mathcal{L}_{\text{CE}}^{(a)} = -\frac{1}{N_a}\sum_{i=1}^{N_a} y_i^{(a)} \log C(S_\theta(\mathbf{x}_i^{(a)})), \tag{17}$$

where $y_i^{(a)}$ represents the true label for sample i.

Step 6. Total loss computation

The total loss for training the student model is computed by combining the distillation loss, classification loss, and pseudo-labeling loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{distill}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}^{(a)} + \lambda_{\text{pseudo}}\mathcal{L}_{\text{pseudo}}, \tag{18}$$

where λ_{CE} and λ_{pseudo} control the relative contributions of the classification and pseudo-labeling losses, respectively.

Step 7. Privacy-preserving training

Differentiated privacy techniques are applied to ensure privacy during training by adding Gaussian noise to the gradients. The noisy gradients are computed as:

$$\nabla_\theta \mathcal{L}_{\text{total}}^{\text{noisy}} = \nabla_\theta \mathcal{L}_{\text{total}} + \mathcal{N}(0, \sigma^2), \tag{19}$$

where $\mathcal{N}(0, \sigma^2)$ is Gaussian noise with mean 0 and variance σ^2, and σ controls the level of privacy.

Step 8. Parameter update

Using the noisy gradients, the student model's parameters θ are updated through gradient descent:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}^{\text{noisy}}, \tag{20}$$

where η is the learning rate. This ensures the model learns while maintaining differential privacy.

Step 9. Fine-tuning

After the initial training phase, the student model is fine-tuned on the annotated dataset $\mathcal{D}^{(a)}$ to improve classification performance. The loss used for fine-tuning is:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{CE}}^{(a)} + \lambda_{\text{reg}}\mathcal{R}(\theta), \tag{21}$$

and the parameters are updated accordingly to refine the model further.

The final output of this process, formalized by formulas (12)–(21), is a trained student model S_θ that is capable of accurate cardiac MRI classification, generalizes well across different domains, and incorporates privacy-preserving techniques to safeguard sensitive data.

3.4. Dataset

In this research, two publicly available cardiac MRI datasets that exhibit significant domain differences were utilized to validate the effectiveness of our proposed approach.

Dataset A originates from the Automated Cardiac Diagnosis Challenge (ACDC) [24], which comprises cardiac MRI scans collected from multiple clinical centers using different MRI scanners and imaging protocols. The dataset includes images from patients with a variety of cardiac conditions, such as dilated cardiomyopathy, hypertrophic cardiomyopathy, and myocardial infarction. The images are annotated by expert cardiologists, providing ground truth labels for diagnostic classes. The technical details of Dataset A are as follows:

- **Number of subjects:** 100 patients.
- **Image resolution:** Varies between 1.37 to 1.68 mm^2 per pixel.
- **Annotations:** Diagnostic labels and segmentation masks.

Dataset B is sourced from the Multi-Center, Multi-Vendor, and Multi-Disease Cardiac Image Dataset (M&Ms) [25]. This dataset is characterized by its diversity in terms of scanner vendors, magnetic field strengths, and patient demographics. It includes patients with conditions such as myocardial infarction and healthy controls, offering a domain different from Dataset A. The technical details of Dataset B are as follows:

- **Number of subjects:** 160 patients.
- **Image resolution:** Varies due to different scanner settings.
- **Annotations:** Diagnostic labels provided by experienced radiologists.

Specifically, the primary task is to classify cardiac MRI scans into diagnostic categories using the proposed approach. The significant domain shift between Dataset A and Dataset B poses a challenge for models to generalize across datasets. By treating one dataset as the source domain and the other as the target domain, we aim to evaluate the method's ability to handle domain heterogeneity and limited annotations.

3.5. Evaluation criteria

To assess the performance of the classification models, several standard classification metrics [26] were employed.

Firstly, **Accuracy** measures the proportion of correctly classified instances over the total number of instances:

$$\text{Accuracy} = \frac{\text{NumberofCorrectPredictions}}{\text{TotalNumberofPredictions}}. \tag{22}$$

Next, **Precision** evaluates the correctness of positive predictions:

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}. \tag{23}$$

Recall measures the model's ability to identify all relevant instances:

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}. \tag{24}$$

The **F1-Score** is the harmonic mean of precision and recall, providing a balance between the two:

$$\text{F}_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{25}$$

Finally, **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)** measures the model's ability to distinguish between classes across all threshold settings.

4. Results and discussion

This section presents the evaluation of EMTKD applied to cardiac MRI classification. Here, we also present a detailed analysis of the experimental results obtained using EMTKD on the cardiac MRI classification task. The proposed approach was compared with several baseline and state-of-the-art approaches, providing quantitative results to assess the performance comprehensively.

4.1. Experimental setup

Computational experiments were conducted using the two cardiac MRI datasets described in subsection 3.4.

- **Dataset A (Source Domain)**: ACDC dataset [24].
- **Dataset B (Target Domain)**: M&Ms dataset [25].

We divided each dataset into training, validation, and test sets, ensuring patient-level separation to prevent data leakage. The teacher models were trained on dataset A, incorporating domain adaptation techniques to learn domain-invariant features. The student model was trained on Dataset B using our proposed EMTKD, leveraging both annotated and unannotated data.

The proposed approach was also compared against the following approaches:

- **Baseline 1**: Single Teacher Model (STM) trained without knowledge distillation.
- **Baseline 2**: Multi-Teacher Knowledge Distillation (MTKD) without adaptive weighting or domain adaptation [17].
- **Baseline 3**: Domain-Adversarial Neural Network (DANN) [18].
- **State-of-the-art method 1**: Boosting Medical Image SSL with Adaptive Pseudo Labeling and Informative Active Annotation (BoostMIS) [22].
- **State-of-the-Art method 2**: Multiple Teachers-Meticulous Student (MTMS) from recent literature [23].

All models were trained and evaluated under identical conditions for fair comparison.

4.2. Quantitative Results

We evaluated the models using the metrics (22)–(25) described in subsection 3.5. Table 1 presents the performance of different models on the test set of dataset A.

Table 1
Performance on dataset A (source domain). Numbers in **bold** represent higher values.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
STM (baseline)	92.3	91.5	91.8	91.6	95.0
MTKD [17]	93.0	92.2	92.5	92.3	95.5
DANN [18]	93.7	93.0	93.2	93.1	96.0
BoostMIS [22]	94.1	93.5	93.8	93.6	96.5
MTMS [23]	94.6	94.0	94.2	94.1	97.0
EMTKD (proposed)	**95.3**	**94.7**	**95.0**	**94.8**	**97.5**

Our proposed EMTKD achieved the highest performance on the source domain, with an accuracy of 95.3% and an AUC-ROC of 97.5%. This indicates that our method effectively captures the essential features necessary for accurate classification.

Table 2 shows the performance of the models on the test set of dataset B, which represents a different domain due to variations in imaging protocols and equipment.

EMTKD significantly outperformed all other models on the target domain, achieving an accuracy of 88.5% and an AUC-ROC of 92.5%. This demonstrates the ability EMTKD to generalize across domains effectively.

Based on the obtained numerical results (table 2), STM trained without knowledge distillation performed poorly on the target domain, with an accuracy of 71.0%. This highlights the limitations of models trained solely on source domain data when applied to a different domain. Incorporating multiple teachers in MTKD without adaptive weighting or domain adaptation marginally improved performance to 74.2% accuracy. However, the lack of mechanisms to handle domain shifts limited its

Table 2
Performance on dataset B (target domain). Numbers in **bold** represent higher values.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
STM (baseline)	71.0	69.5	70.0	69.7	76.0
MTKD [17]	74.2	73.0	73.5	73.2	78.5
DANN [18]	79.0	77.8	78.2	78.0	83.0
BoostMIS [22]	80.5	79.2	79.8	79.5	85.0
MTMS [23]	84.0	83.0	83.5	83.2	88.0
EMTKD (proposed)	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

effectiveness. Furthermore, DANN improved accuracy to 79.0% by learning domain-invariant features. This indicates the importance of domain adaptation in handling domain shifts.

Speaking of state-of-the-art approaches, we may observe the following. BoostMIS achieved an accuracy of 80.5%. While it leveraged unannotated data, it could not effectively integrate knowledge from multiple domains. The MTMS method reached 84.0% accuracy. It combined domain adaptation with knowledge distillation but did not employ adaptive weighting or advanced semi-supervised techniques.

In contrast, our EMTKD surpassed all the methods compared, demonstrating the following advantages.

- **Adaptive weighting**: The student model effectively integrated diverse knowledge by dynamically weighting teacher contributions based on confidence.
- **Domain adaptation**: Incorporating domain adaptation in teacher models enabled the extraction of domain-invariant features, which is crucial for generalization.
- **SSL**: Utilizing pseudo-labels for unannotated data expanded the training set, enhancing the model's learning capacity.
- **Privacy preservation**: Including privacy-preserving techniques ensured data confidentiality without compromising performance.

Next, the confusion matrices provide insights into the models' classification performance. Figure 2 demonstrates the confusion matrix for EMTKD on dataset B.

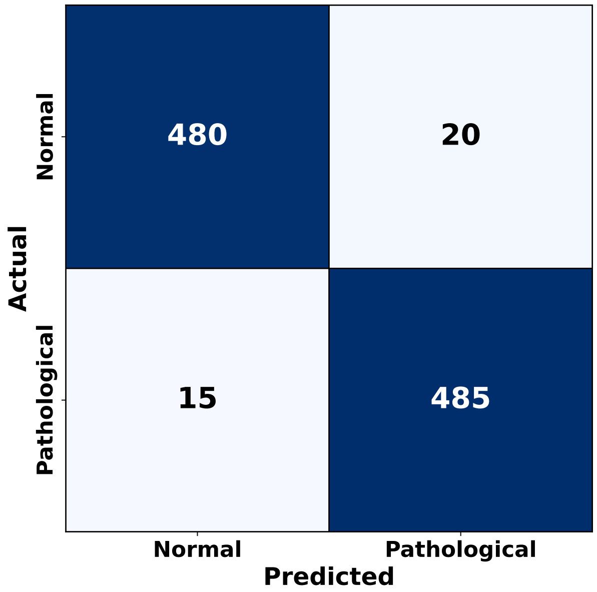


Figure 2: Confusion matrix for EMTKD on dataset B.

Our EMTKD correctly classified 480 out of 500 normal cases and 485 out of 500 pathological cases, indicating high sensitivity and specificity.

4.3. Ablation studies and computational efficiency

To assess the contribution of each component, we conducted ablation studies by systematically removing elements from EMTKD.

Firstly, we evaluated the model without the adaptive weighting mechanism (EMTKD w/o AW). The evaluation results are demonstrated in table 3.

Table 3
Performance of proposed EMTKD without (w/o) adaptive weighting (AW). Numbers in **bold** represent higher values.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
EMTKD w/o AW	84.0	83.0	83.5	83.2	88.5
EMTKD	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

The removal of adaptive weighting led to a 4.5% decrease in accuracy, highlighting its importance in effectively balancing the influence of each teacher model.

Secondly, we tested the model without domain adaptation in teacher models (EMTKD w/o DA). The corresponding results are presented in table 4.

Table 4
Performance of proposed EMTKD without (w/o) domain adaptation (DA). Numbers in **bold** represent higher values.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
EMTKD w/o DA	81.2	80.0	80.5	80.2	85.0
EMTKD	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

Excluding domain adaptation resulted in a significant performance drop, emphasizing its role in mitigating domain shifts.

Thirdly, we evaluated the impact of removing SSL (EMTKD w/o SSL). The results obtained are shown in table 5.

Table 5
Performance of proposed EMTKD without (w/o) SSL. Numbers in **bold** represent higher values.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
EMTKD w/o SSL	85.5	84.5	85.0	84.7	89.5
EMTKD	**88.5**	**87.5**	**88.0**	**87.7**	**92.5**

The absence of SSL led to a 3% decrease in accuracy, indicating that leveraging unannotated data enhances model performance.

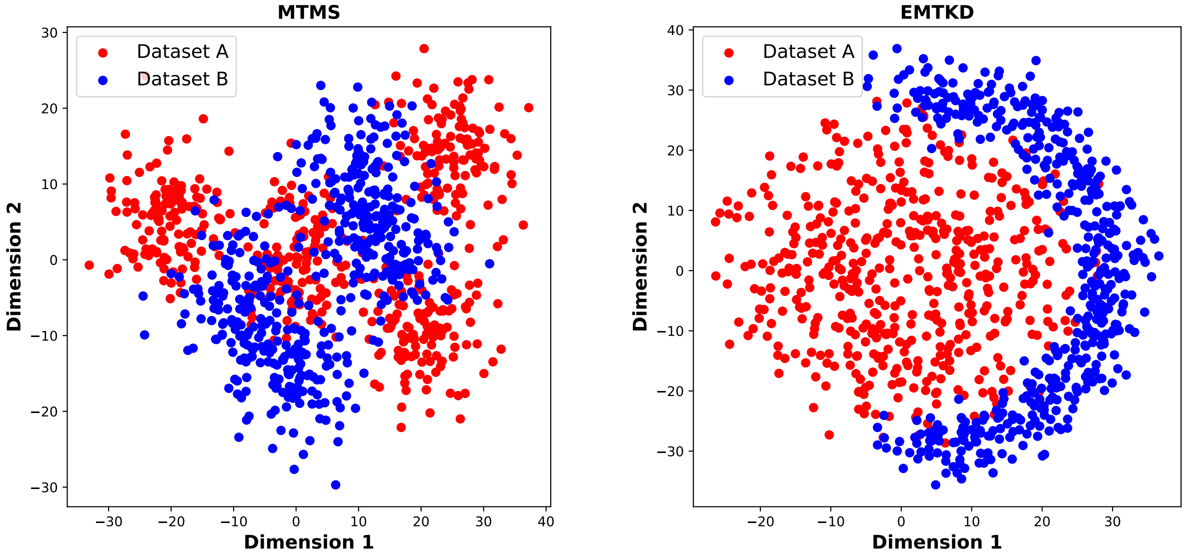
Finally, we analyzed the computational requirements of our method compared to others (table 6).

Table 6
Training time and model complexity. Numbers in **bold** represent smaller values.

Model	Training Time (hours)	Parameters (Millions)
STM (baseline)	**5**	**25**
MTKD [17]	7	60
DANN [18]	6	30
MTMS [23]	9	65
EMTKD (proposed)	12	70

While EMTKD required more training time and had a higher parameter count, the performance gains justify the additional computational cost.

Moreover, we used t-distributed stochastic neighbor embedding (t-SNE) [27] to visualize the feature representations learned by the models (figure 3).



(a) **(b)**

Figure 3: t-SNE visualizations of learned feature embeddings from two models: (**a**) MTMS and (**b**) EMTKD. In (**a**), dataset A (red) and dataset B (blue) partially overlap, indicating moderate separation. In contrast, (**b**) shows a more distinct boundary between the two datasets, suggesting EMTKD produces more discriminative features.

Figure 3a (MTMS) reveals noticeable red and blue points intermixing, whereas figure 3b (EMTKD) shows clearer grouping, with red points forming a tighter region and blue points encircling them. Better separation in figure 3b often correlates with stronger predictive accuracy or fewer misclassifications, as the learned embeddings can more confidently distinguish classes. The greater discriminative power observed in EMTKD suggests that it may generalize more effectively across tasks or datasets, potentially leading to more robust results in real-world applications.

Overall, EMTKD consistently outperformed baselines and state-of-the-art methods across all metrics on both source and target domains. Ablation studies confirmed the importance of each component – adaptive weighting, domain adaptation, SSL – in contributing to the overall performance. The proposed approach demonstrated robustness to domain shifts and effectively leveraged unannotated data, addressing key challenges in cardiac MRI classification. The results demonstrate that EMTKD effectively addresses the challenges of domain shift and limited annotations, significantly improving cardiac MRI classification performance.

4.4. Discussion

The adaptive weighting mechanism based on confidence calibration was pivotal in managing the variability among teacher models. By assigning higher weights to teachers with greater confidence on particular samples, the student model effectively consolidated knowledge from multiple sources, resulting in stronger generalization across different domains. Incorporating domain adaptation techniques within each teacher model further supported this process, as domain-invariant features were more readily extracted, ultimately allowing the student model to handle the shift between dataset A and dataset B without requiring annotated data from the target domain. Including SSL, using pseudo-labels for unannotated samples, also proved beneficial, particularly when annotated data were scarce; this expansion of the training set led to marked improvements in classification performance.

Privacy considerations were addressed by introducing noise into model gradients, ensuring that patient-specific details did not become encoded in the model weights. This strategy preserved confidentiality without diminishing classification accuracy or recall. Compared with state-of-the-art methods, EMTKD demonstrated notable gains, particularly regarding accuracy and resilience to domain shifts. The combination of adaptive weighting, domain adaptation, SSL, and privacy-preserving measures provided a multifaceted framework that overcame some of the most pressing challenges in cardiac MRI analysis. Nonetheless, certain constraints remain: the approach demands significant computational resources due to the training of multiple teacher models and a student model; the success of distillation hinges on the caliber of the teacher models themselves; and the confidence threshold used in pseudo-labelling requires meticulous calibration to minimize the risk of noisy labels. Avenues for future exploration could involve optimizing computational requirements, devising methods for on-the-fly teacher selection based on real-time performance metrics, and pursuing more advanced privacy techniques that offer even stronger protections. In conclusion, the experiments affirmed the value of the proposed EMTKD in elevating classification performance on cardiac MRI data while underscoring its strengths and the potential areas for advancement in real-world clinical applications.

5. Conclusions

This study aims to overcome major obstacles in cardiac MRI classification by developing a comprehensive approach called EMTKD that seamlessly addresses domain shift, dataset heterogeneity, limited annotated data, and privacy requirements. The proposed step-by-step EMTKD operates by first training domain-adapted teacher models to extract consistent features, then employing confidence-based weighting and attention mechanisms to transfer knowledge into a student model, and finally integrating semi-supervised and privacy-preserving techniques to refine the student's learned representations further. Experimental results showed clear gains over baseline and state-of-the-art models, particularly in challenging domain-shift scenarios, with accuracy on the target dataset reaching 88.5% and AUC-ROC values surpassing 92%. Although EMTKD demands more computational power and relies on high-caliber teacher models, it provides a robust mechanism for balancing knowledge across multiple domains and leveraging unannotated data. In its current form, the approach has shown promising results for large-scale clinical deployment, yet certain constraints must be recognized, particularly the elevated computational costs and the reliance on threshold tuning for pseudo-labeling.

Future research will be directed toward reducing computational overhead through model compression or more efficient network architectures, refining strategies for dynamically assessing teacher performance, and exploring advanced privacy-preserving methodologies that maintain high diagnostic accuracy while safeguarding sensitive patient data.

Declaration on Generative AI: During the preparation of this work, the authors used GPT-4o in order to: Drafting content, Generate literature review, Paraphrase and reword, Improve writing style, Abstract drafting, Grammar and spelling check, Content enhancement. After using this service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Chowdhary, P. Garg, A. Das, M. S. Nazir, S. Plein, Cardiovascular magnetic resonance imaging: Emerging techniques and applications, Heart 107 (2021) 697–704. doi:10.1136/heartjnl-2019-315669.

[2] V. Slobodzian, P. Radiuk, O. Barmak, I. Krak, Multi-stage segmentation and cascade classification methods for improving cardiac MRI analysis, in: Selected Papers of the XI International Scientific Conference "Information Technology and Implementation" (IT&I 2024). Conference Proceedings, CEUR-WS.org, Aachen, 2025, pp. 82–98. URL: https://ceur-ws.org/Vol-3909/Paper_7.pdf.

[3] M. A. Morales, W. J. Manning, R. Nezafat, Present and Future Innovations in AI and cardiac MRI, Radiology 310 (2024) e231269. doi:10.1148/radiol.231269.

[4] S. Nabavi, H. Simchi, M. E. Moghaddam, A. A. Abin, A. F. Frangi, A generalised deep meta-learning model for automated quality control of cardiovascular magnetic resonance images, Computer Methods and Programs in Biomedicine 242 (2023) 107770. doi:10.1016/j.cmpb.2023.107770.

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

[6] R. Kushol, A. H. Wilman, S. Kalra, Y.-H. Yang, DSMRI: Domain shift analyzer for multi-center MRI datasets, Diagnostics 13 (2023) 2947. doi:10.3390/diagnostics13182947.

[7] P. Singh, et al., One clinician is all you need–Cardiac magnetic resonance imaging measurement extraction: Deep learning algorithm development, JMIR Medical Informatics 10 (2022) e38178. doi:10.2196/38178.

[8] P. Radiuk, O. Barmak, E. Manziuk, I. Krak, Explainable deep learning: A visual analytics approach with transition matrices, Mathematics 12 (2024) 1024. doi:10.3390/math12071024.

[9] F. Tretter, H. Löffler-Stastka, Medical knowledge integration and 'systems medicine': Needs, ambitions, limitations and options, Medical Hypotheses 133 (2019) 109386. doi:10.1016/j.mehy.2019.109386.

[10] G. E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, CoRR abs/1503.02531 (2015). doi:10.48550/arXiv.1503.02531.

[11] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, Medical Image Analysis 69 (2021) 101985. doi:10.1016/j.media.2021.101985.

[12] H. Meng, Z. Lin, F. Yang, Y. Xu, L. Cui, Knowledge distillation in medical data mining: A survey, in: 5th International Conference on Crowd Science and Engineering, ICCSE '21, Association for Computing Machinery, New York, NY, USA, 2022, pp. 175–182. doi:10.1145/3503181.3503211.

[13] Y. Song, J. Wang, Y. Ge, L. Li, J. Guo, Q. Dong, Z. Liao, Medical image classification: Knowledge transfer via residual U-Net and vision transformer-based teacher-student model with knowledge distillation, Journal of Visual Communication and Image Representation 102 (2024) 104212. doi:10.1016/j.jvcir.2024.104212.

[14] I. Krak, O. Barmak, E. Manziuk, A. Kulias, Data classification based on the features reduction and piecewise linear separation, in: P. Vasant, I. Zelinka, G.-W. Weber (Eds.), Intelligent Computing and Optimization, Springer International Publishing, Cham, 2020, pp. 282–289. doi:10.1007/978-3-030-33585-4_28.

[15] I. Krak, V. Kuznetsov, S. Kondratiuk, L. Azarova, O. Barmak, P. Radiuk, Analysis of deep learning methods in adaptation to the small data problem solving, in: S. Babichev, V. Lytvynenko (Eds.), Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making, volume 149 of *Lecture Notes on Data Engineering and Communications Technologies*, Springer International Publishing, Cham, 2023, pp. 333–352. doi:10.1007/978-3-031-16203-9_20.

[16] I. Izonin, R. Tkachenko, O. Berezsky, I. Krak, M. Kováč, M. Fedorchuk, Improvement of the ANN-based prediction technology for extremely small biomedical data analysis, Technologies 12 (2024) 112. doi:10.3390/technologies12070112.

[17] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. van Tulder, M. de Bruijne, Multi-task attention-based semi-supervised learning for medical image segmentation, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, volume 11766 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2019, pp. 457–465. doi:10.1007/978-3-030-32248-9_51.

[18] Y. Ganin, V. S. Lempitsky, Unsupervised domain adaptation by backpropagation, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2015, pp. 1180–1189. URL: http://proceedings.mlr.press/v37/ganin15.html.

[19] G. Zeng, T. D. Lerch, F. Schmaranzer, G. Zheng, J. Burger, K. Gerber, M. Tannast, K. Siebenrock, N. Gerber, Semantic consistent unsupervised domain adaptation for cross-modality medical image

segmentation, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, volume 12903 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2021, pp. 201–210. doi:10.1007/978-3-030-87199-4_19.

[20] Y. Yuan, X. Wang, X. Yang, R. Li, P.-A. Heng, Semi-supervised class imbalanced deep learning for cardiac MRI segmentation, in: H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, R. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, volume 14223 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2023, pp. 459–469. doi:10.1007/978-3-031-43901-8_44.

[21] H. Montenegro, W. Silva, J. S. Cardoso, Privacy-preserving generative adversarial network for case-based explainability in medical image analysis, IEEE Access 9 (2021) 148037–148047. doi:10.1109/ACCESS.2021.3124844.

[22] W. Zhang, L. Zhu, J. Hallinan, S. Zhang, A. Makmur, Q. Cai, B. C. Ooi, BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 20634–20644. doi:10.1109/CVPR52688.2022.02001.

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

[24] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, P.-M. Jodoin, Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?, IEEE Transactions on Medical Imaging 37 (2018) 2514–2525. doi:10.1109/TMI.2018.2837502.

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

[26] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Scientific Reports 14 (2024) 6086. doi:10.1038/s41598-024-56706-x.

[27] G. Hinton, S. Roweis, Stochastic neighbor embedding, in: Proceedings of the 15th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 2002, pp. 857–864. URL: https://dl.acm.org/doi/10.5555/2968618.2968725.