Pattern Recognition 134 (2023) 109076

Contents lists available at ScienceDirect

Pattern Recognition

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

ML-DSVM+: A meta-learning based deep SVM+ for computer-aided diagnosis

Xiangmin Han [a,b], Jun Wang [a,b], Shihui Ying [c], Jun Shi [a,b,*], Dinggang Shen [d,e,**]

[a] Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, School of Communication and Information Engineering, Shanghai University, China
[b] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, China
[c] Department of Mathematics, School of Science, Shanghai University, Shanghai, China
[d] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[e] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

ARTICLE INFO

Article history:
Received 23 March 2021
Revised 10 June 2022
Accepted 25 September 2022
Available online 27 September 2022

Keywords:
Deep neural network
Support vector machine plus
Learning using privileged information
Meta-learning

ABSTRACT

Transfer learning (TL) can improve the performance of a single-modal medical imaging-based computer-aided diagnosis (CAD) by transferring knowledge from related imaging modalities. Support vector machine plus (SVM+) is a supervised TL classifier specially designed for TL between the paired data in the source and target domains with shared labels. In this work, a novel deep neural network (DNN) based SVM+ (DSVM+) algorithm is proposed for single-modal imaging-based CAD. DSVM+ integrates the bi-channel DNNs and SVM+ classifier into a unified framework to improve the performance of both feature representation and classification. In particular, a new coupled hinge loss function is developed to conduct bidirectional TL between the source and target domains, which further promotes knowledge transfer together with the feature representation under the guidance of shared labels. To alleviate the overfitting caused by the increased parameters in DNNs for limited training samples, the meta-learning based DSVM+ (ML-DSVM+) is further developed, which designs randomly selecting samples from the training data instead of other CAD tasks for meta-tasks. This sampling strategy also can avoid the issue of class imbalance. ML-DSVM+ is evaluated on three medical imaging datasets. It achieves the best results of 88.26±1.40%, 90.45±5.00%, and 87.63±5.56% on accuracy, sensitivity and specificity, respectively, on the Bimodal Breast Ultrasound Image dataset, 90.00±1.05%, 72.55±3.87%, and 96.40±2.26% of the corresponding indices on the Alzheimer's Disease Neuroimaging Initiative dataset, and 85.76±3.12% of classification accuracy, 88.73±7.22% of sensitivity, and 82.60±1.56% of specificity for the Autism Brain Imaging Data Exchange dataset.

1. Introduction

Medical imaging-based computer-aided diagnosis (CAD) has attracted great attention in recent years, which helps radiologists to improve diagnosis accuracy together with good consistency and reproducibility [1].

It is well known that multi-modal medical imaging can provide more comprehensive information for clinical diagnosis. Therefore, a multi-modal imaging-based CAD generally outperforms the corresponding single-modal approach. However, in clinical practice, the diagnosis costs increase with more modalities, and advanced multi-modal imaging devices are always scarce in most hospitals, which limits the wide application of multi-modal imaging-based CAD [2]. In contrast, single-modal imaging-based CAD has wider and more flexible applications, especially in rural areas, but single-modal imaging provides only one-fold type of information for diagnosis, which limits the performance of its corresponding CAD. Therefore, it is necessary to further improve single-modal imaging-based CAD with advanced artificial intelligence techniques.

Transfer learning (TL) aims to transfer knowledge from the related source domain (SD) to the target domain (TD) to improve the model performance in TD [3]. TL has been successfully applied in the analysis and processing of medical images. Some recent studies indicate that TL can effectively promote the diagnosis accuracy of single-modal imaging-based CAD by transferring knowledge from related imaging modalities or diseases [4]. TL is also an effective way to alleviate the small sample size (SSS) problem [3],

* Corresponding author at: School of Communication and Information Engineering, Shanghai University, No. 99, Shangda Road, Shanghai, China.
** Corresponding author at: School of Biomedical Engineering, ShanghaiTech University, No. 393, Middle Huaxia Road, Shanghai, China.
E-mail addresses: junshi@shu.edu.cn (J. Shi), Dinggang.Shen@gmail.com (D. Shen).

https://doi.org/10.1016/j.patcog.2022.109076
0031-3203/

which commonly exists in the field of CAD because of the time-consuming and expensive procedure of collection and annotation of medical images[1]. However, these current works mainly adopt conventional TL methods for medical imaging data, and there is still room for improvement by making full use of the properties of multi-modal imaging, such as label information.

In fact, multi-modal imaging data generally share the same label, because they scan the same organs or tissues of a subject. The shared label information is expected to help explore more transferable knowledge and improve knowledge transfer in a supervised manner. However, conventional TL algorithms generally have no constraints on data labels of the SD and the TD, and cannot make full use of label information to guide knowledge transfer [3].

Learning using privileged information (LUPI) is a supervised TL paradigm, which is specially designed for the paired data in the SD (also called privileged information in LUPI) and TD with shared labels [5]. Support vector machine plus (SVM+) is a typical LUPI-based classifier that adopts an additional SD to help construct an optimal separating hyperplane [5]. SVM+ generally outperforms the conventional TL classifiers because the shared labels can effectively guide and improve knowledge transfer. Therefore, SVM+ and its variants have been successfully applied to TL-based CAD [6]. However, its performance is usually limited by a small number of paired training samples with shared labels.

On the other hand, deep learning (DL) has shown its effectiveness for feature representation, which is critical in a CAD system [1]. In addition to end-to-end DL methods (such as convolutional neural networks (CNN)) for classification tasks, it is also an alternative and popular approach to further learn and improve feature representation from hand-crafted features by using other deep neural network (DNN) algorithms, such as stacked auto-encoder (SAE) [7] and deep belief networks (DBN), especially on limited samples [8]. This strategy has also been widely used in the field of CAD because of the SSS problem [1]. However, most of these DNN algorithms perform unsupervised learning, and they are independent of the classifier module in a CAD system. This suggests that DNN-based feature learning can be further improved with the guidance of labels in a supervised manner [9]. Therefore, integrating both DNN and SVM+ classifier construction into a unified framework, named DSVM+ in this work, is expected to improve a CAD system. Furthermore, because DSVM+ contains a bi-channel DNNs, the bidirectional transfer strategy can be used to further improve its performance, which has shown its superior performance to the unidirectional TL [10].

It is worth noting that the introduction of DNNs into DSVM+ will significantly increase the parameters in the model, which may result in an overfitting problem when the labeled training data are not provided sufficiently [11]. Meta-learning is one of the approaches for few-shot learning [12]. Its main idea is to train a meta-model to capture common characteristics across a set of different tasks as prior information, which is then effectively adapted to a new task with only a few labeled training data [12]. Meta-learning has been successfully used for various few-shot learning tasks [13], such as classification, regression, and reinforcement learning. These studies indicate that meta-learning has the potential to improve the performance of DNN-based TL algorithms.

In this work, we propose a novel meta-learning based DSVM+ (ML-DSVM+) algorithm to improve single-modal imaging-based CAD by transferring knowledge from other related imaging modalities. ML-DSVM+ contains three modules, namely the meta-sampling module, bi-channel DNNs, and an SVM+ classifier. A random sampling strategy is used to generate meta-tasks for meta-training and meta-validation. Then, we propose a coupled hinge loss to optimize the bi-channel DNNs and the SVM+ classifier in a unified framework. The proposed ML-DSVM+ is evaluated on three medical imaging datasets, and all the experimental results show its effectiveness on the problems of SSS and class imbalance.

The main contributions of this work are following three-fold:

1) We propose a novel DSVM+ classifier to improve the performance of single-modal imaging-based CAD with transferable knowledge. DSVM+ integrates the bi-channel DNNs and SVM+ classifier into a unified framework for jointly learning, which can effectively improve the performance of both feature representation of DNNs and SVM+ classifier with the guidance of shared labels in a supervised manner.
2) We develop a new coupled hinge loss function in DSVM+ to perform bidirectional knowledge transfer between SD and TD instead of the unidirectional approach in the original SVM+, which can promote the feature learning in both domains and also extract more transferable knowledge simultaneously.
3) We further propose a meta-learning based DSVM+ (ML-DSVM+) algorithm with a new meta-learning strategy. ML-DSVM+ self-generates the meta-tasks by randomly selecting samples from the bimodal training data. This strategy not only improves the performance of DSVM+ with a small number of training samples, but also alleviates the issue of class imbalance to a certain extent.

2. Related work

The existing LUPI works mainly focus on developing LUPI-based classifiers. SVM+ is a classical approach under the LUPI paradigm [5], and many improved SVM+ algorithms have been proposed, such as robust SVM+ [14], fast SVM+ [15], adaptive SVM+ [16], and multi-view SVM+ [17]. These algorithms have been successfully applied for different classification tasks, indicating the effectiveness of the LUPI paradigm.

Recently, several pioneering works have shown the effectiveness of LUPI in CADs. For example, Duan *et al.* used the single nucleic polymorphisms as SD for the fundus image based glaucoma detection with SVM+ [6]; Zheng *et al.* developed a boosted LUPI classification framework by ensembling multiple SVM+ classifiers for magnetic resonance imaging (MRI) based diagnosis of brain disorders [18]; Alahmadi *et al.* utilized generalized matrix learning vector quantization (GMLVQ) classifier to diagnose mild cognitive impairment (MCI), which adopted cognitive data as the diagnostic data with fMRI data as SD [19]. All these works suggest that the LUPI paradigm is suitable for promoting single-modal imaging-based CAD by adopting the corresponding modality as SD with shared labels. However, these LUPI classifiers still have room to be improved. For example, the current feature representation and SVM+ are separated modules, and the weak representation will degrade the classification performance. It is expected to improve their performance by integrating both modules into a unified framework for jointly learning.

In fact, SVM has been successfully embedded in the DL-based classification models followed after a DNN or CNN [9][20]. Tang replaced the softmax classifier in a DNN-based classification model with the linear SVM, and used the objective function of L2-SVM to train DNN for the classification [9]. Although its optimization procedure becomes more complex than the models with a softmax classifier, the classification performance generally improves. Li and Zhang proposed a deep neural mapping SVM (DNMSVM) algorithm by taking the DNN as a kernel mapping from the original input space into a feature space [20]. It is worth noting that the DNN in SVM can approximate kernel mapping to avoid the issue of kernel selection, because according to the universal approximation theorem, any multivariate continuous function can be well approximated by a neural network under certain conditions [21].

Therefore, we propose a DNN-based SVM+ (DSVM+) to improve the performance of the original SVM+.

On the other hand, bidirectional learning is another effective approach to improve the transfer performance of TL. It not only improves the performance of both domains, but also reduces the dependency on a large number of training data compared with unidirectional learning [10]. This strategy has attracted considerable attention in recent years. For example, Chen *et al.* proposed an unsupervised bidirectional cross-modality algorithm to conduct bidirectional domain adaptation between MRI and computed tomography (CT) images, which improved the performance of the segmentation network in the TD [22]; Li *et al.* developed a bidirectional learning framework for semantic image segmentation task, in which the image translation model and the segmentation adaptation model could be trained alternatively to promote each other [10]. Inspired by the effectiveness of this strategy, we propose a coupled hinge loss to perform bidirectional knowledge transfer, which alternatively makes SD and TD serve as privileged information for each other to learn and extract more implicit transferable knowledge between the bi-channel DNNs in DSVM+.

Although DSVM+ can improve TL performance, it has also significantly increased model parameters, resulting in an overfitting problem for limited medical imaging. Meta-learning provides an effective way to alleviate the SSS problem in DNN by learning a specific base-learner for a new task by training on a set of different tasks [1]. Existing meta-learning methods can be divided mainly into three categories: 1) model-based approaches that aim to use a hand-designed meta-learner to optimize the model with limited training data; 2) metric-based methods that focus on learning an effective metric or distance function for samples in the same class; 3) optimization-based methods that try to learn a parameterized base-learner for a new task with only a few gradient descent steps. The first and second approaches require complicated and manually designed optimization models, while recent studies indicate that optimization-based meta-learning methods are more suitable for DNN-based architectures [23]. Therefore, we propose an ML-DSVM+ algorithm based on the latter to alleviate the problems of SSS or class imbalance.

Conventional meta-learning methods require a large number of different meta-tasks to train a robust meta-model. However, it is generally difficult to collect and generate many different classification meta-tasks in most medical applications. To this end, we propose a new meta-learning strategy to improve DSVM+, where multiple meta-tasks are generated by direct random selecting from the bimodal training samples. Meanwhile, because we randomly select an equal number of samples from positive and negative classes to build a meta-task each time, the class imbalance problem can be avoided in the training dataset.

3. Method

3.1. Preliminary of SVM+

For a binary classification task, given a training set $\mathcal{D} = \{(x_i^t, x_i^s, y_i)\}_{i=1}^N$, where x_i^t and x_i^s are a pair of feature vectors for the i-th sample from TD and SD, respectively, y_i is the shared label by x_i^t and x_i^s. SVM+ solves the constrained optimization problem as follows:

$$\min \frac{1}{2}\left(\left\|w^t\right\|^2 + \lambda\|w^s\|^2\right) + C\sum_{i=1}^{N}\left(\xi\left(w^s, b^s, \psi\left(x_i^s\right)\right)\right) \tag{1}$$

s.t. $y_i(w^t\phi(x_i^t) + b^t) \geq 1 - \xi(w^s, b^s, \psi(x_i^s))$ and $\xi(w^s, b^s, \psi(x_i^s)) \geq 0,\ i = 1, \ldots, N$where (w^t, b^t) and (w^s, b^s) are the weight vector and bias parameters in TD and SD, respectively, $\phi(\cdot)$ and $\psi(\cdot)$ denote the feature mapping function induced by the kernel in TD and SD, respectively, $\xi(w^s, b^s, \psi(x_i^s)) = w^s x_i^s + b^s$ is a slack function (or correcting function) defined in the SD, $\lambda > 0$ is the trade-off parameter, and $C > 0$ is the penalty parameter.

3.2. DSVM+

To improve the performance of SVM+, a DNN-based SVM+ is proposed in this study. As shown in Fig. 1, DSVM+ consists of two modules: the bi-channel DNNs and the SVM+ classifier. The former contains two DNNs with the same architecture corresponding to the fed features in SD and TD. The latter utilizes the knowledge learned from SD to guide the learning of classification hyperplane and then train a powerful classifier in TD. DSVM+ integrates the bi-channel DNNs and SVM+ classifier into a unified framework and then simultaneously improves the performance of feature representation and classification with the guidance of shared labels. In particular, we propose a coupled hinge loss to perform bidirectional knowledge transfer between SD and TD, which can learn and extract more transferable knowledge as well as simultaneously promote feature learning in both domains.

(1) **Training stage of DSVM+**

As shown in Fig. 1, the bi-channel DNNs in both TD and SD have the same network structure. The TD network can be computed by:

$$\begin{aligned} a_0^t &= x^t \\ \{a_l^t &= \mathrm{relu}\left(W_l^t a_{l-1}^t + b_l^t\right),\ l = 1, \ldots, L \\ z^t &= W_L^t a_{L-1}^t + b_L^t \end{aligned} \tag{5}$$

where a_l^t, W_l^t and b_l^t are the output, weight matrix and bias vector of the l-th layer in the TD network, respectively, and z^t denotes the output of the TD network. The widely used activation function $\mathrm{relu}(x) = \max(x,\ 0)$ is used here.

Similarly, the DNN in the SD network can be computed as:

$$\begin{aligned} a_0^s &= x^s \\ \{a_l^s &= relu\left(W_l^s a_{l-1}^s + b_l^s\right),\ l = 1, \ldots, L \\ z^s &= W_L^s a_{L-1}^s + b_L^s \end{aligned} \tag{6}$$

where (a_l^s, W_l^s, b_l^s) are the output, weight matrix and bias vector of the l-th layer in the SD network, respectively, and z^s is the output of the SD network.

We develop the following coupled hinge loss for the DSVM+:

$$\begin{aligned} \mathcal{L}\left(W_L^t, b_L^t, W_L^s, b_L^s\right) &= \mathcal{L}^t\left(W_L^t, W_L^s, b_L^t\right) + \mathcal{L}^s\left(W_L^t, W_L^s, b_L^s\right) \\ &\quad + \mathcal{L}^{coupled}\left(W_L^t, W_L^s, b_L^t, b_L^s\right) \end{aligned} \tag{7}$$

where $\mathcal{L}^t(W_L^t, W_L^s, b_L^t)$ and $\mathcal{L}^s(W_L^t, W_L^s, b_L^s)$ are the classification terms in TD and SD, respectively, and $\mathcal{L}^{coupled}(W_L^t, W_L^s, b_L^t, b_L^s)$ is the coupled term. These three components are defined by:

$$\begin{aligned} &\mathcal{L}^t\left(W_L^t, W_L^s, b_L^t\right) \\ &= \min \frac{\lambda_2}{2}\|W_L^s\|^2 + C_1\sum \max\left(0,\ 1 - y\left(W_L^t a_{L-1}^t + b_L^t\right)\right)^2 \end{aligned} \tag{8}$$

$$\begin{aligned} &\mathcal{L}^s\left(W_L^t, W_L^s, b_L^s\right) \\ &= \min \frac{\lambda_1}{2}\left\|W_L^t\right\|^2 + C_2\sum \max\left(0,\ 1 - y\left(W_L^s a_{L-1}^s + b_L^s\right)\right)^2 \end{aligned} \tag{9}$$

$$\begin{aligned} &\mathcal{L}^{coupled}\left(W_L^t, W_L^s, b_L^t, b_L^s\right) \\ &= min\sum \max\left(0,\ 1 - y\left(W_L^t a_{L-1}^t + b_L^t + W_L^s a_{L-1}^s + b_L^s\right)\right)^2 \end{aligned} \tag{10}$$

where the first terms in Eq. (8) and (9) are the regularization terms in the bi-channel DNNs to avoid the overfitting, $\lambda_1 > 0$ and $\lambda_2 > 0$

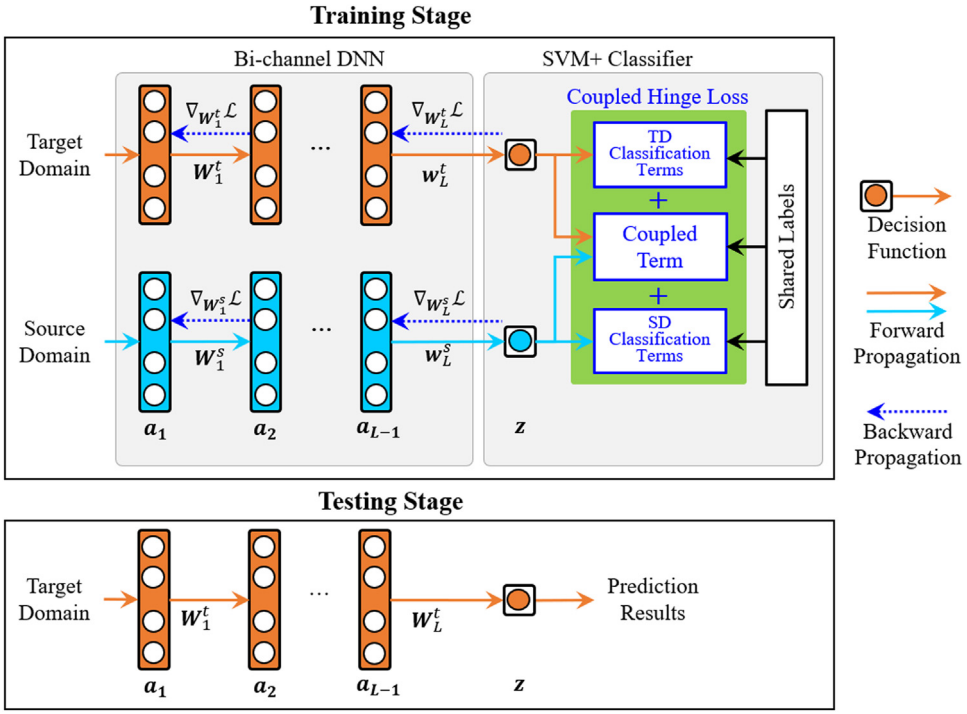


Fig. 1. Architecture of the proposed DSVM+ algorithm and flowchart of its training and testing stages. During the training stage, both TD and SD data are used to train the bi-channel DNNs and the SVM+ classifier, and the coupled hinge loss makes full use of the shared labels to perform bidirectional knowledge transfer between DNNs. In the testing stage, only TD data is available to predict the classification results.

are trade-off parameters to balance the relationship between the two domains, and $C_1 > 0$ and $C_2 > 0$ are the parameters to balance the hinge loss and regularizers in the TD and SD, respectively.

Comparted with the original SVM+ that includes only one hinge loss to learn a decision relu in TD, we adopt two classification terms $\mathcal{L}^t(W_L^t, W_L^s, b_L^t)$ and $\mathcal{L}^s(W_L^t, W_L^s, b_L^s)$ for the corresponding bi-channel DNNs. Each classification term can not only train an independent classifier, but also optimize the corresponding DNN in a unified framework. Since the bimodal imaging data provide complementary information, they can learn additional knowledge from each other. Thus, we develop a new coupled term $\mathcal{L}^{coupled}(W_L^t, W_L^s, b_L^t, b_L^s)$ in hinge loss to perform bidirectional knowledge transfer between bimodal data under the guidance of shared label y. In this unified framework, the coupled hinge loss can simultaneously improve the performance of both feature representation of the bi-channel DNNs and the corresponding two classifiers. Moreover, it can capture more transferable knowledge from two domains. It is worth noting that the regularization term of the hinge loss is applied on the last layer of DNN in DNMSVM [20], which is considered as the feature representation of DNN for the followed SVM. Therefore, we also introduce two regularization terms only on the last layer in DSVM+.

To optimize DSVM+, the stochastic gradient descent (SGD) algorithm is adopted with random initialization for training, such that the parameters are updated by:

$$\begin{aligned} W^t &= W^t - \alpha \nabla_{W^t} \mathcal{L} \\ b^t &= b^t - \alpha \nabla_{b^t} \mathcal{L} \\ W^s &= W^s - \alpha \nabla_{W^s} \mathcal{L} \\ b^s &= b^s - \alpha \nabla_{b^s} \mathcal{L} \end{aligned} \tag{11}$$

After several iteration steps, both the classification performance and feature representation of DSVM+ can be improved progressively and simultaneously.

The procedure for jointly optimizing the network in both TD and SD is summarized in Algorithm 1.

Algorithm 1 DSVM+

Input: Training set $\mathcal{D} = \{(x_i^t, x_i^s, y_i)\}_{i=1}^N$
Initialization: Set learning rate α, randomly initialize $W_l^t, b_l^t, W_l^s, b_l^s$, $l = 1, \ldots, L$
1: **while** not convergence **do**
2: **Bidirectional transfer strategy:**
3: Forward propagation in TD: Compute a_l^t, $l = 1, \ldots, L-1$, and z^t by (5), and compute $\mathcal{L}^t$ by (8);
4: Forward propagation in SD: Compute a_l^s, $l = 1, \ldots, L-1$, and z^s by (6), and compute $\mathcal{L}^s$ by (9);
5: Compute the coupled loss values by (10) and calculate the total loss;
6: Backward propagation in TD: Compute $\{\nabla_{W_l^t}\mathcal{L}\}_{l=1}^L$ and $\{\nabla_{b_l^t}\mathcal{L}\}_{l=1}^L$, and update $\{W_l^t\}_{l=1}^L$, $\{b_l^t\}_{l=1}^L$ as follows:$W_l^t = W_l^t - \alpha\nabla_{W_l^t}\mathcal{L}$, $b_l^t = b_l^t - \alpha\nabla_{b_l^t}\mathcal{L}$
7: Backward propagation in SD: Compute $\{\nabla_{W_l^s}\mathcal{L}\}_{l=1}^L$ and $\{\nabla_{b_l^s}\mathcal{L}\}_{l=1}^L$, and update $\{W_l^s\}_{l=1}^L$, $\{b_l^s\}_{l=1}^L$ as follows:$W_l^s = W_l^s - \alpha\nabla_{W_l^s}\mathcal{L}$), $b_l^s = b_l^s - \alpha\nabla_{b_l^s}\mathcal{L}$
8: **end**

(1) Testing stage of DSVM+

During the testing stage, the SD modality is not available, and only the TD modality is fed to the trained DSVM+. The final prediction is given by:

$$y_{pred} = \text{sign}\left(W_L^t a_{L-1}^t + b_L^t\right) \tag{12}$$

where W_L^t and b_L^t are the optimized parameters in the training stage, and a_{L-1}^t is the output of the (L-1)-th layer in TD.

3.3. ML-DSVM+

Although DSVM+ can improve feature representation and then explore more transferable knowledge from both domains, the bi-channel DNNs significantly introduce more parameters into this model, easily resulting in overfitting when the training data are limited bimodal. Meta-learning effectively alleviates this issue in a DL model by learning knowledge from other related tasks [13]. The conventional meta-learning methods required a lot of similar tasks

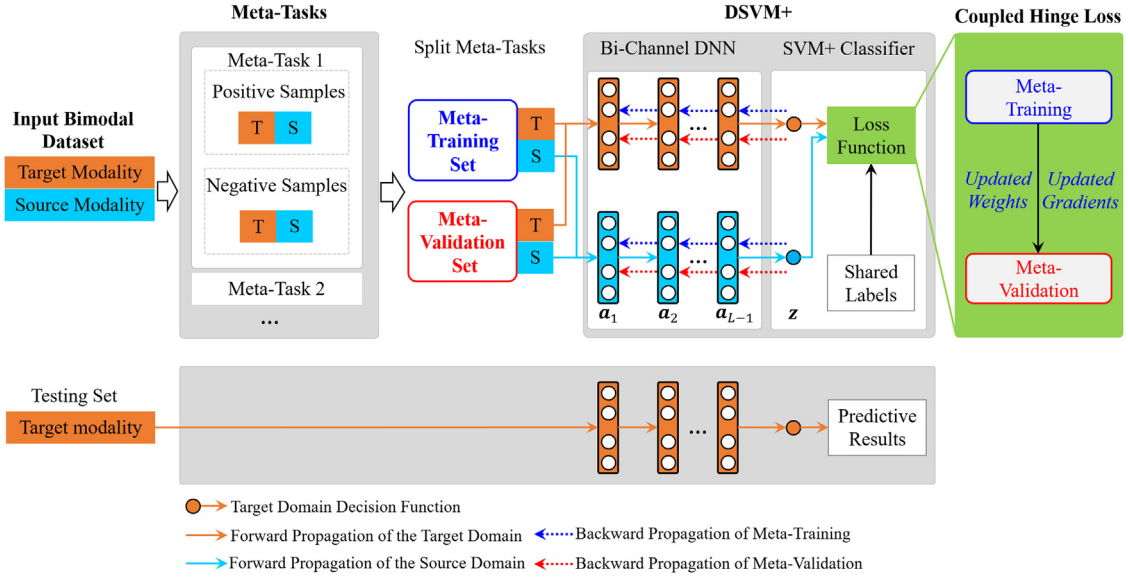


Fig. 2. Flowchart of the proposed ML-DSVM+ algorithm. A random sampling strategy is adopted to generate meta-tasks. The training procedure in meta-learning includes two stages, the meta-training stage (also called the inner loop) and the meta-validation stage (also called the outer loop).

to learn a robust model [13]. However, it is generally difficult to collect lots of medical imaging data from different modalities and diseases to design related CAD tasks as meta-tasks, because both data collection and annotation are extremely time-consuming and expensive [1]. To this end, we propose to adopt a random sampling strategy to self-generate meta-tasks only from the training samples for ML-DSVM+ in this work, which also can alleviate the problem of class imbalance.

As shown in Fig. 2, we propose a meta-learning based DSVM+, *i.e.*, ML-DSVM+. A random sampling strategy is adopted on the bimodal data to build multiple meta-tasks, which are then utilized to train the base-learner (DSVM+). Each meta-task contains an equal number of positive and negative bimodal samples. The training procedure of ML-DSVM+ includes two stages: the meta-training stage (also named the inner loop) and the meta-validation stage (also called the outer loop). After training the DSVM+ model, only an image of the target modality is fed to DSVM+ to conduct the disease prediction.

Suppose that the data distribution over the dataset is $p(\mathcal{T})$, and the proposed DSVM+ in Section 3.2 is considered as the base-learner $f_{(W^t,b^t,W^s,b^s)}$. A random sampling procedure is used to generate multiple meta-tasks $\{\mathcal{T}_j\}_{j=1}^{T}$ from the data distribution $p(\mathcal{T})$. In particular, *K-positive* and *K-negative* samples are sampled into a meta-task. Each meta-task is divided into a training set $\mathcal{D}^{tr}$ for meta-training and a validation set $\mathcal{D}^{val}$ for meta-validation. Both $\mathcal{D}^{tr}$ and $\mathcal{D}^{val}$ contain $K/2$ positive samples and $K/2$ negative samples, respectively (i.e., *K*-shot). In addition, $\mathcal{D}^{val}$ is used to compute the second derivative in the meta-validation stage, which should not be confused with the commonly used validation set for model selection.

In the meta-training stage, prior knowledge across the meta-training set $\mathcal{D}^{tr}$ can be learned by gradient-based learning. The model parameters W_l^t, b_l^t, W_l^s and b_l^s are updated iteratively using gradient descent, and the update rule can be formulated as:

$$\begin{aligned} W_l^{t,new} &= W_l^t - \alpha \nabla_{W_l^t} \mathcal{L}_{\mathcal{T}_j} \\ b_l^{t,new} &= b_l^t - \alpha \nabla_{b_l^t} \mathcal{L}_{\mathcal{T}_j} \\ W_l^{s,new} &= W_l^s - \alpha \nabla_{W_l^s} \mathcal{L}_{\mathcal{T}_j} \\ b_l^{s,new} &= b_l^s - \alpha \nabla_{b_l^s} \mathcal{L}_{\mathcal{T}_j} \end{aligned} \tag{13}$$

where $W_l^{t,new}$, $b_l^{t,new}$, $W_l^{s,new}$ and $b_l^{s,new}$ are the updated parameters during the meta-training stage, and α denotes the meta-training step size. According to the work done by Finn et al. (2017), multiple gradient descent steps can be applied in this stage, and we consider one gradient update step in this work.

During the meta-validation stage, the learned task-level knowledge is transferred to the base-learner through the second derivatives across the meta-validation set $\mathcal{D}^{val}$. The updated parameters $W_l^{t,new}$, $b_l^{t,new}$, $W_l^{s,new}$ and $b_l^{s,new}$ in Eq. (13) are considered as the initialization in the meta-validation stage, and then the object of meta-validation is formulated as:

$$\min_{W_l^t, b_l^t, W_l^s, b_l^s} \sum_{\mathcal{D}^{val}} \mathcal{L}_{\mathcal{T}_j} \left(f_{W_l^{t,new}, b_l^{t,new}, W_l^{s,new}, b_l^{s,new}} \right) \tag{14}$$

As observed in Eq. (14), the meta-validation involves second derivatives, which require the computation of Hessian-vector products. The resulting update for the model parameters $W_l^t, b_l^t, W_l^s, b_l^s$ can be expressed as

$$\begin{aligned} W_l^t &= W_l^{t,new} - \beta \nabla_{W_l^{t,new}} \sum_{\mathcal{D}^{val}} \mathcal{L}_{\mathcal{T}_j} \\ b_l^t &= b_l^{t,new} - \beta \nabla_{b_l^{t,new}} \sum_{\mathcal{D}^{val}} \mathcal{L}_{\mathcal{T}_j} \\ W_l^s &= W_l^{s,new} - \beta \nabla_{W_l^{s,new}} \sum_{\mathcal{D}^{val}} \mathcal{L}_{\mathcal{T}_j} \\ b_l^s &= b_l^{s,new} - \beta \nabla_{b_l^{s,new}} \sum_{\mathcal{D}^{val}} \mathcal{L}_{\mathcal{T}_j} \end{aligned} \tag{15}$$

where β is the meta-validation step size.

Algorithm 2 summarizes the training process of ML-DSVM+.

It is worth noting that in the existing meta-learning methods, the meta-tasks and the target task are generally developed based on different datasets or modalities, and the fine-tuning is then conducted on the trained model by meta-learning with limited training data of the target task. However, in this work, there are no additional training samples from other modalities or other diseases that are used for building the meta-tasks during the meta-training and meta-validation stages, and the diagnosis modality in the target domain is the same as the diagnosis modality in the source domain. Therefore, the fine-tuning stage is not necessary in this work.

Algorithm 2 ML-DSVM+

Input: Meta-tasks $\mathcal{T}_j = (\mathcal{D}^{tr}, \mathcal{D}^{val})$
Initialization: Set step size α, β, randomly initialize W, b
1: **while** not convergence **do**
2: **Meta-training stage:**
3: **for** each $\mathcal{D}^{tr}$, **do**
4: Forward propagation based on $\mathcal{D}^{tr}$:
5: Compute a_l^t, $l = 1, \ldots, L-1$, and z^t
6: Compute a_l^s, $l = 1, \ldots, L-1$, and z^s
7: Calculate $\mathcal{L}(W_l^t, b_l^t, W_l^s, b_l^s)$ by (7)
8: Compute $\nabla_{W_l^t}\mathcal{L}_{\mathcal{T}_j}, \nabla_{b_l^t}\mathcal{L}_{\mathcal{T}_j}, \nabla_{W_l^s}\mathcal{L}_{\mathcal{T}_j}, \nabla_{b_l^s}\mathcal{L}_{\mathcal{T}_j}$, and update $W_l^t, b_l^t, W_l^s, b_l^s$ by (13)
9: **end for**
10: **Meta-validation stage:**
11: Forward propagation using $\mathcal{D}^{val}$, and calculate $\mathcal{L}(W_l^{t,new}, b_l^{t,new}, W_l^{s,new}, b_l^{s,new})$
12: Evaluate $\nabla_{W_l^{t,new}}\mathcal{L}_{\mathcal{T}_j}, \nabla_{b_l^{t,new}}\mathcal{L}_{\mathcal{T}_j}, \nabla_{W_l\ s,new}\mathcal{L}_{\mathcal{T}_j}, \nabla_{b_l^{s,new}}\mathcal{L}_{\mathcal{T}_j}$, and update $W_l^t, b_l^t, W_l^s, b_l^s$ by (15)
13: **end**

4. Experiments and results

4.1. Datasets and data processing

The proposed ML-DSVM+ algorithm was evaluated on three multi-modal medical imaging datasets, namely the Bimodal Breast Ultrasound Image (BBUI) dataset, the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [24][1], and the Autism Brain Imaging Data Exchange (ABIDE) dataset [25][2].

BBUI Dataset: This dataset was collected from the Nanjing Drum Tower Hospital, including 264 pairs of B-mode ultrasound (BUS) and elastography ultrasound (EUS) images (129 patients with benign tumors and 135 patients with malignant cancers). All lesions were underwent biopsy and pathologically proven. Approval from the ethics committee of the hospital was obtained, and all patients provided signed informed consent.

The images in the BBUI were scanned using a Mindray Resona7 ultrasound scanner with the L11-3U probe by an experienced sonologist. All the acquired BUS and EUS images had the same size of 550×420 pixels. A region of interest (ROI), including the tumor region, was selected by an experienced sinologist from each ultrasound image. The statistical features, texture features, and features of Hu's invariant moments were then directly extracted from the ROI without image resizing [26]. The statistical features were calculated from the intensities of all pixels, including the mean, standard deviation, coefficient of variance, skewness, kurtosis, entropy of histogram, area ratio, combined area ratio, several percentiles, etc. The texture features were extracted from the gray-level cooccurrence matrix (GLCM), including the energy, contrast, homogeneity, and entropy of GLCM. Moreover, the Hu moment invariants were also extracted as features. A total of 71 features were thus generated from each ROI in both the BUS and EUS images. Please refer to [26] for more details about feature extraction.

In this experiment, we selected the BUS images as TD and the EUS images as SD, because BUS has been widely used for the diagnosis of breast cancers in clinical practice, while EUS is yet to be used as a routine diagnostic tool, especially in rural hospitals.

TADPOLE Challenge Dataset: This dataset was acquired for the Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE) challenge[3], which is a subset of ADNI [24]. In this work, we selected bimodal imaging data, namely T1-weighted MRI and FluoroDeoxyGlucose positron emission tomography (FDG-PET), captured on the same subject at the baseline period (i.e., no reused subjects). After filtering the meaningless data and removing the missing data from these candidates, 190 subjects (51 AD subjects and 139 MCI subjects) were then obtained in this experiment. Note that this dataset was selected to evaluate the performance of ML-DSVM+ for the classification of imbalanced classes.

The MRI scans were preprocessed with standard ADNI pipelines[4], including correction for gradient non-linearity and B1 non-uniformity, and sharpening of the peak. The regional features, including ROI volume, ROI cortical thickness, and ROI surface areas, were then extracted using the FreeSurfer cross-sectional and longitudinal pipelines. After removing the missing values, 343 features were obtained. On the other hand, each FDG-PET image has been processed by the standard ADNI protocol[5], including co-registration, averaging, standardization, and smoothing. After the FDG-PET images were registered with the corresponding MR images, standardized uptake value ratio (SUVR) measures were extracted for the relevant ROI using the SPM5 software [27]. The missing values and meaningless data were filtered, and 158 features were finally obtained for each image. Please refer to the work by [28] for more details on the feature extraction of MRI and FDG-PET images, respectively.

PET devices are known to be very expensive and equipped only in a few hospitals, while MRI devices are more accessible and widely used in clinical practice. Therefore, we considered MRI and PET as TD and SD, respectively.

ABIDE dataset: The ABIDE dataset is a collection of 16 international imaging centers, aiming to study the neural mechanism of autism spectrum disorder (ASD) [25]. In our experiment, we used both T1-weighted MRI and resting-state functional magnetic resonance imaging (rs-fMRI) acquired from the New York University (NYU) Langone Medical Center. After removing the missing and meaningless data, we obtained a dataset including 112 subjects (54 ASD patients and 58 normal controls (NC)), who were all between 6 and 15 years of age [29].

The primary steps of the preprocessing protocol for rs-fMRI images are as follows[6]: discarding the first 10 volumes for magnetization equilibrium, performing slice timing and head motion correction, normalizing the images to the MNI space of the resolution 3×3×3 mm, conducting the Nuisance variable regression, parcellating the images into 116 ROIs, applying band-pass filtering, scrubbing the unmatched volumes, and computing the Pearson correlation coefficients. Finally, a 116×116 correlation matrix was obtained for each rs-fMRI scan. Functional connectivity (FC) features from the gray matter (GM) regions were extracted from the rs-fMRI images [29]. We performed the *t*-test to select the most relevant 303 features, and then a 303-dimensional feature vector was obtained for each rs-fMRI scan in our experiment. By contrast, the regional morphological features were extracted from MRI scans using the FreeSurfer pipeline. In particular, multiple atlases containing different ROIs were used to extract the following features: cerebral cortical gray matter (GM) volumes, subcortical white matter (WM) volumes, mean cortical thickness measures, subcortical structure volumes, and the volumes and thickness measures of the Brodmann areas. A total of 303 features were extracted for each MRI scan in our experiments. Please refer to the works by [29] for more details on the feature extraction of MRI scans.

MRI and rs-fMRI are two widely applied imaging tools for ASD diagnosis. Because rs-fMRI generally achieves relatively higher accuracy than MRI [30], we selected rs-fMRI as TD and MRI as SD in this experiment.

[1] http://adni.loni.usc.edu.
[2] http://preprocessed-connectomes-project.org/abide.
[3] https://tadpole.grand-challenge.org/.
[4] See MRI analysis on ADNI website: http://adni.loni.usc.edu/methods/mri-analysis/ mri-pre-processing.
[5] See PET analysis on ADNI website: http://adni.loni.usc.edu/methods/pet-analysis/pre-processing.
[6] Standard pipeline provided by ABIDE with AFNI: https://afni.nimh.nih.gov/afni/.

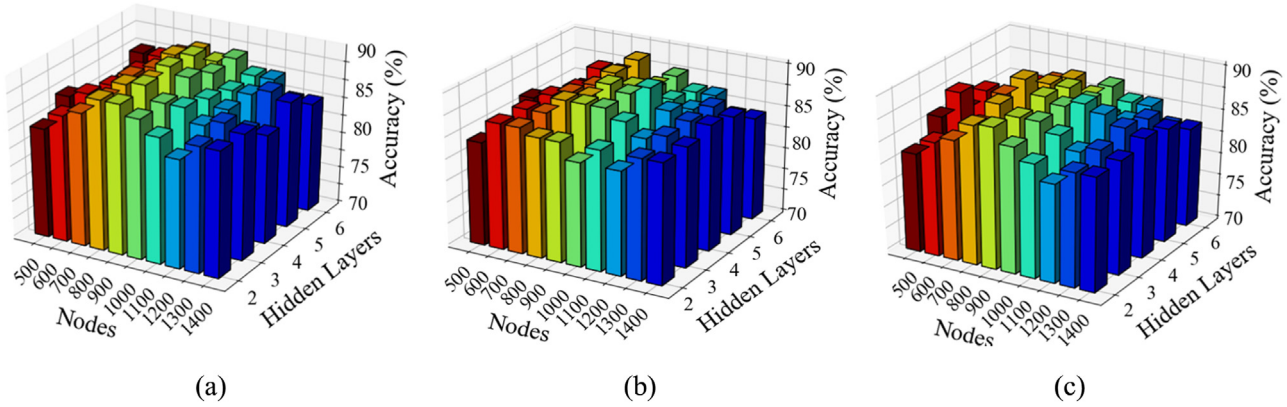


Fig. 3. The classification accuracy on the BBUI dataset with different combinations of hidden layers and nodes in the cases of (a) 5-shot, (b) 10-shot, and (c) 15-shot by the proposed ML-DSVM+. Each bar shows the average accuracy of 5-fold cross-validation, and the color of each bar is allocated only for better visualization.

4.2. Experimental settings

The proposed ML-DSVM+ was compared to the following related or state-of-the-art TL algorithms.

1) **SVM** [31]: The commonly used SVM was directly performed on the TD data without TL.
2) **DNMSVM** [20]: This typical DNN-based SVM algorithm was performed on the TD data without TL.
3) **DCCA-SVM** [32]: The deep canonical correlation analysis (DCCA) is a DNN-based CCA algorithm. It can be used for feature-level TL, because it learns bimodal representation and then handles missing modalities during the testing stage [32]. In this experiment, DCCA was conducted on the hand-crafted bimodal features, and the learned features corresponding to the TD modality were then fed to SVM for classification.
4) **MRBM-SVM** [33]: This multi-modal restricted Boltzmann machine (MRBM) algorithm can be used for feature-level TL, since it could handle missing modality [33]. Similar to DCCA, the finally learned features corresponding to the TD modality were fed to SVM.
5) **MDL-CW** [34]: MDL-CW is a bi-channel DNN-based multi-modal learning algorithm that can learn the cross weights between two DNNs. It can perform multi-modal learning with shared labels in a supervised manner and handle the missing modality in the testing stage. Therefore, it can be considered as a deep LUPI algorithm to perform feature-level transfer learning. The finally learned features corresponding to the TD modality were fed to SVM.
6) **DPL-HD** [35]: DPL-HD is a CNN-based deep LUPI algorithm for feature-level TL, which utilizes privileged information to control the variance of the dropout. To conduct a fair comparison on the hand-crafted features in this work, we modified this algorithm by replacing the CNN in DPL-HD with DNN.
7) **HFA** [36]: The heterogeneous feature augmentation (HFA) is a widely used TL classifier algorithm. It was used as classifier-level TL.
8) **KRVFL+** [37]: The kernel random vector functional link plus is a newly proposed LUPI classifier. It was used as classifier-level TL.
9) **FSVM+** [15]: The fast implementation of SVM+ was selected as the baseline for classifier-level TL in this work.
10) **U-DSVM+:** The unidirectional TL version of DSVM+ was proposed for classifier-level TL. U-DSVM+ only conducted unidirectional knowledge transfer from SD to TD using Eq. (11) and Eq. (13). The SD network was optimized using an independent hinge loss.
11) **DSVM+**: The proposed DSVM+ algorithm was used for classifier-level TL.

We fed the same hand-crafted features to the proposed DSVM+ and ML-DSVM+ together with all the compared algorithms for a fair comparison. The classification performance will be analyzed in the following perspectives: (1) U-DSVM+ and FSVM+ were compared to evaluate the performance of DNN in U-DSVM+ in improving the feature representation of hand-crafted features; (2) DSVM+ and U-DSVM+ were compared to assess the performance of bidirectional and unidirectional transfer learning; (3) The proposed DSVM+ was compared with other transfer learning algorithms to indicate its effectiveness; (4) ML-DSVM+ was finally compared with DSVM+ to evaluate the performance of the introduced meta-learning strategy.

All the algorithms were conducted with the five-fold cross-validation strategy to avoid the sampling bias introduced by the randomly splitting dataset. The accuracy (ACC), sensitivity (SEN), specificity (SPE) and Youden index (YI) were selected as evaluation indices. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) value were also used to evaluate the performance. The final results were given in the mean ± standard deviation format.

4.3. Experimental results on BBUI dataset

Fig. 3 shows the classification accuracies in 5-shot, 10-shot, and 15-shot cases of ML-DSVM+, which are performed with different hidden layers and layer nodes. It can be observed that hyperparameter combinations of DNN can affect classification accuracy. The best accuracy in the 15-shot is 88.26±1.40% with the hidden layers as 3, nodes in each hidden layer as 600, and the parameters C_1, C_2, λ_1 and λ_2 as 1, 1, 0.05 and 0.05, respectively. The best accuracies in 5-shot and 10-shot are 87.50±1.54% and 87.87±1.89%, respectively. We then use the accuracy in 15-shot to compare with other algorithms with optimized parameters.

Table 1 shows the classification results of different algorithms on the BBUI dataset. It can be seen that the proposed ML-DSVM+ achieves the best mean classification accuracy of 88.26±1.40%, sensitivity of 90.45±5.00%, specificity of 87.63±5.56%, and YI of 78.08±4.45%.

In particular, compared with FSVM+, U-DSVM+ improves the accuracy by 1.51%, sensitivity by 3.03% and YI by 3.02%. It indicates that DNN can effectively improve the feature representation of the hand-crafted features. DSVM+ improves 1.07%, 0.71%, 3.41%,

Table 1
Classification Results of different algorithms on BBUI dataset (Unit: %)

Algorithms	Transfer Learning	Modality	ACC	SEN	SPE	YI
SVM	no TL	BUS	82.21±2.46	81.43±3.92	83.05±6.30	64.48±5.24
DNMSVM			84.11±3.36	87.34±3.69	80.73±5.66	68.07±6.67
DCCA-SVM	Feature-level TL	BUS with	83.34±2.32	87.39±5.39	79.16±7.70	66.56±4.48
MRBM-SVM		EUS as SD	83.35±3.55	84.32±5.30	82.24±5.61	66.56±7.10
MDL-CW			84.48±4.35	85.18±0.35	83.92±8.77	69.10±9.05
DPL-HD			85.23±4.35	85.92±1.53	84.66±7.69	70.58±8.76
HFA	Classifier-level TL	BUS with	84.85±4.62	86.69±4.96	83.03±6.41	69.72±9.20
KRVFL+		EUS as SD	85.21±5.31	85.95±6.33	84.60±8.41	70.54±10.54
FSVM+	Classifier-level TL	BUS with	84.11±2.78	84.37±3.08	83.77±3.73	68.15±5.87
U-DSVM+		EUS as SD	85.62±2.51	87.40±2.97	83.77±2.57	71.17±4.97
DSVM+			86.69±3.78	88.11±4.42	86.91±5.94	75.02±8.86
ML-DSVM+			**88.26±1.40**	**90.45±5.00**	**87.63±5.56**	**78.08±4.45**

and 3.85% in accuracy, sensitivity, specificity, and YI, respectively, over U-DSVM+, indicating the effectiveness of our proposed bidirectional transfer strategy. DSVM+ also archives superior classification results compared to other TL algorithms and improves at least 1.46%, 2.25%, 0.72%, and 4.44% in terms of accuracy, sensitivity, specificity, and YI, respectively. By further introducing the meta-learning into DSVM+, ML-DSVM+ improves 1.57%, 2.34%, 0.72%, and 3.06%, respectively, in the corresponding indices, over DSVM+.

Fig. 4 gives the ROC curves and AUC values of all algorithms in our experiments. The proposed ML-DSVM+ achieves the best AUC value of 0.921, suggesting its effectiveness on the BBUI dataset.

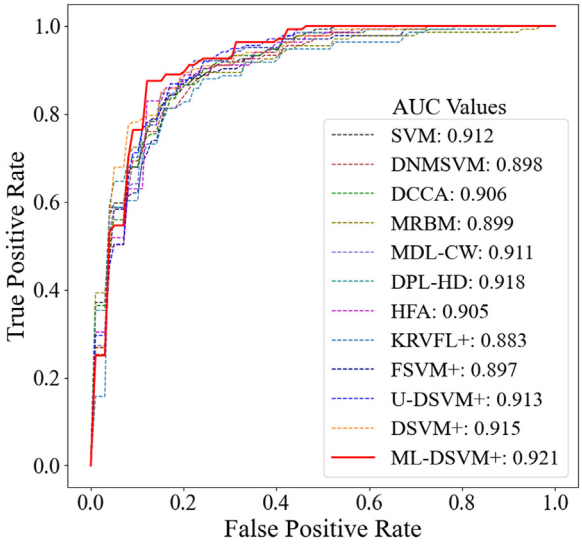


Fig. 4. ROC curves and corresponding AUC values of different algorithms on the BBUI dataset.

4.4. Experimental results on TADPOLE challenge dataset

The subset of TADPOLE challenge dataset used in this work is an imbalanced dataset that includes 139 MCI subjects and 51 AD subjects. Fig. 5 shows the classification accuracies of the 5-shot, 10-shot, and 15-shot cases in ML-DSVM+ with different numbers of hidden layers and nodes.

The best accuracy in 5-shot and 10-shot cases are 89.47±2.88% and 90.00±2.57%, while the best accuracy in 15-shot case is 90.00±1.05% with the hidden layers as 3, nodes in each hidden layer as 500, and the parameters C_1, C_2, λ_1 and λ_2 as 1, 1, 0.05 and 0.05, respectively. Obviously, the performance of the 15-shot case is better than the other two. Therefore, the 15-shot case is then compared with other algorithms with optimized parameters.

Table 2 gives the classification results of different algorithms on the TADPOLE challenge dataset, which have a similar trend to those in Table 1. ML-DSVM+ again exhibits superior performance compared with all the other algorithms. The best mean classification accuracy, sensitivity, specificity, and YI are 90.00±1.05%, 72.55±3.87%, 96.40±2.26%, and 68.95±2.41%, respectively. U-DSVM+ improves 1.58%, 4.00%, 0.74%, and 4.74% in accuracy, sensitivity, specificity, and YI, respectively, over FSVM+, which indicates that the feature representation has been improved by DNN. The proposed bidirectional transfer algorithm DSVM+ improves 3.82% and 3.10% in sensitivity, and YI, respectively, over unidirectional transfer algorithms U-DSVM+. It indicates the superior performance of the proposed bidirectional knowledge transfer strategy. DSVM+ also achieves superior performance to the other

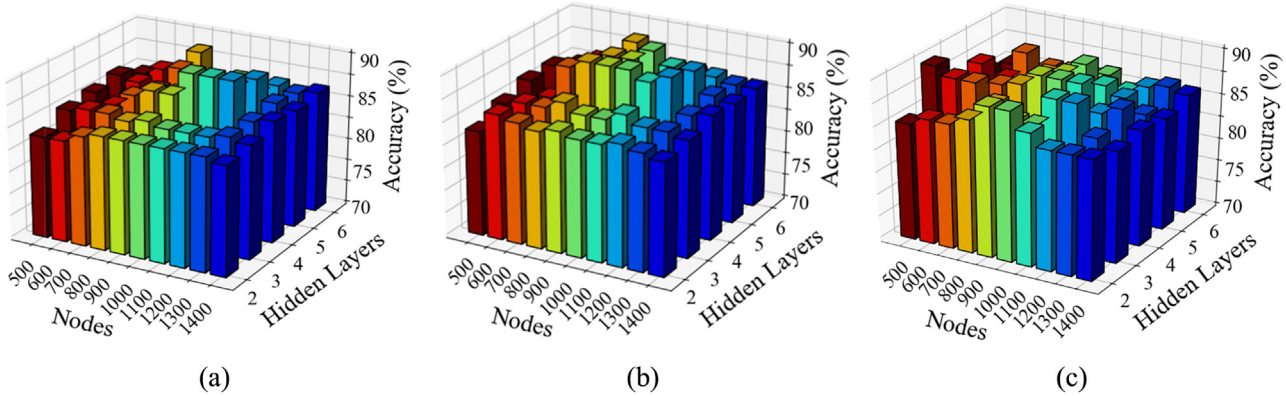


Fig. 5. The classification accuracy on the TADPOLE challenge dataset with different combinations of hidden layers and nodes in the cases of (a) 5-shot, (b) 10-shot, and (c) 15-shot by the proposed ML-DSVM+.

Table 2
Classification results of different algorithms on TADPOLE challenge dataset (Unit: %)

Algorithms	Transfer Learning	Modality	ACC	SEN	SPE	YI
SVM	no TL	MRI	83.16±4.88	52.91±11.65	94.23±2.89	47.14±13.88
DNMSVM			86.32±3.07	64.73±4.51	94.23±3.66	58.96±5.88
DCCA-SVM	Feature-level TL	MRI with	85.79±2.11	58.91±6.69	95.69±3.49	54.60±5.20
MRBM-SVM		PET as SD	85.26±3.57	56.91±9.82	95.63±4.29	52.54±9.14
MDL-CW			87.37±1.97	62.91±8.80	96.40±2.26	59.31±7.38
DPL-HD			87.89±1.29	66.73±4.50	96.40±0.05	61.13±4.51
HFA	Classifier-level TL	MRI with	84.21±2.35	62.91±6.12	92.09±2.66	55.00±5.94
KRVFL+		PET as SD	84.74±4.21	60.91±8.07	93.52±3.51	54.43±10.27
FSVM+	Classifier-level TL	MRI with	86.84±3.72	62.73±7.43	95.66±3.54	58.39±8.98
U-DSVM+		PET as SD	88.42±1.29	66.73±4.17	96.40±0.05	63.13±4.19
DSVM+			88.95±1.05	70.55±6.42	95.69±1.42	66.23±5.44
ML-DSVM+			**90.00±1.05**	**72.55±3.87**	**96.40±2.26**	**68.95±2.41**

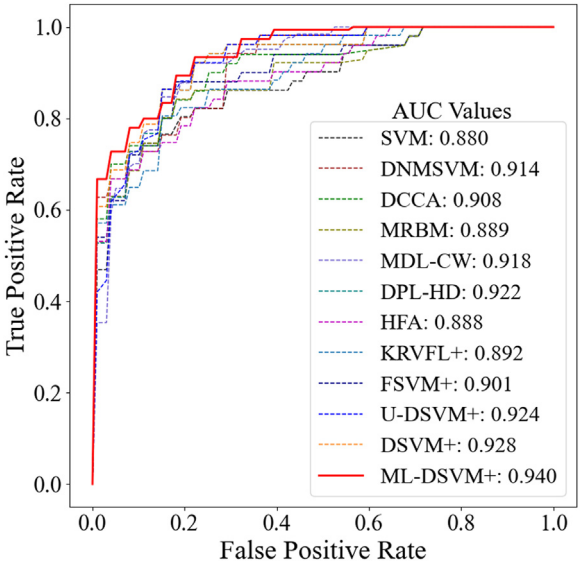


Fig. 6. ROC curves and corresponding AUC values of different algorithms on the TADPOLE challenge dataset.

TL algorithms with improvements by at least 1.06%, 3.82%, and 5.1% in accuracy, sensitivity, and YI, respectively. Higher sensitivity means a higher true positive rate, which is more important for the medical imaging classification task with imbalanced data. By introducing meta-learning, ML-DSVM+ improves 1.05%, 2.00%, 0.71%, and 2.72% in the corresponding indices, over DSVM+.

As shown in Fig. 6, ML-DSVM+ again outperforms other compared approaches with the highest true positive rate and the least false positive rate with the best AUC value of 0.940.

4.5. Experimental results on ABIDE dataset

The comparison results of 5-shot, 10-shot and 15-shot in ML-DSVM+ are shown in Fig. 7. The best accuracies in the 5-shot and 10-shot cases are 84.77±2.45% and 84.85±1.49%, respectively. The best accuracy in the 15-shot case is 85.76±3.12%, when the ML-DSVM+ has 3 hidden layers with 500 nodes, and the parameters C_1, C_2, λ_1 and λ_2 are 1, 1, 0.05 and 0.05, respectively. It is then selected for comparison with other algorithms with optimized parameters.

The classification results of different algorithms on the ABIDE dataset are shown in Table 3. ML-DSVM+ again outperforms all the compared algorithms with the best mean classification accuracy of 85.76±3.12%, sensitivity of 88.73±7.22%, specificity of 82.60±1.56%, and YI of 71.32±6.40%, respectively.

As shown in Table 3, DSVM+ obtains superior accuracy than FSVM+, indicating that the feature representation can be effectively improved by DNN. DSVM+ also achieves better accuracy than the unidirectional TL algorithm U-DSVM+, demonstrating the effectiveness of the proposed bidirectional TL strategy. The proposed ML-DSVM+ achieves improvements of 3.64%, 1.82%, 5.46%, and 7.27% on accuracy, specificity, and YI, respectively, over DSVM+, respectively. It can also be found that ML-DSVM+ achieves the best AUC value of 0.820 in Fig. 8.

4.6. Convergence and stability

Fig. 9 shows the loss curves of DNMSVM, DSVM+, and ML-DSVM+ on the three datasets. To implement a fair comparison, we set the identical learning rates for each algorithm, so do the other parameters. The loss values are sampled every ten steps to visualize the trends clearly. It can be observed that the training loss

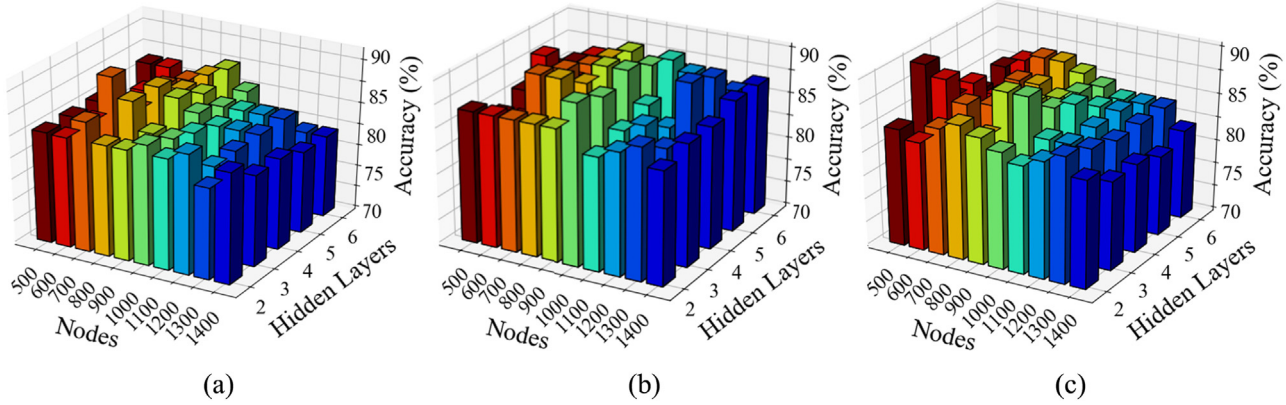


Fig. 7. The classification accuracy on the ABIDE dataset with different combinations of hidden layers and nodes in the cases of (a) 5-shot, (b) 10-shot, and (c) 15-shot by the proposed ML-DSVM+. Each bar shows the average accuracy of 5-fold cross-validation.

Table 3
Classification results of different algorithms on ABIDE dataset (Unit: %)

Algorithms	Transfer Learning	Modality	ACC	SEN	SPE	YI
SVM	no TL	rs-fMRI	73.26±8.89	70.55±10.25	76.49±13.49	47.04±17.66
DNMSVM			81.21±2.06	85.09±4.80	77.14±5.55	62.23±3.93
DCCA-SVM	Feature-level TL	rs-fMRI	81.21±3.53	83.45±5.77	79.62±8.08	63.07±6.78
MRBM-SVM		with MRI as SD	80.38±1.86	78.20±3.62	82.95±4.84	61.16±3.17
MDL-CW			80.30±2.54	81.45±5.80	78.96±5.28	60.42±4.85
DPL-HD			81.29±1.06	83.27±3.88	79.35±3.54	62.63±2.03
HFA	Classifier-level TL	rs-fMRI	77.65±4.14	79.23±7.69	75.93±5.80	55.16±8.44
KRVFL+		with MRI as SD	78.56±3.38	75.19±9.87	80.95±3.35	56.14±8.60
FSVM+	Classifier-level TL	rs-fMRI	80.23±9.52	79.09±7.36	81.15±14.73	60.24±18.37
U-DSVM+		with MRI as SD	81.36±4.64	81.45±5.80	81.56±8.43	60.01±8.89
DSVM+			82.12±2.94	86.91±4.93	77.14±5.55	64.05±5.81
ML-DSVM+			**85.76±3.12**	**88.73±7.22**	**82.60±1.56**	**71.32±6.40**

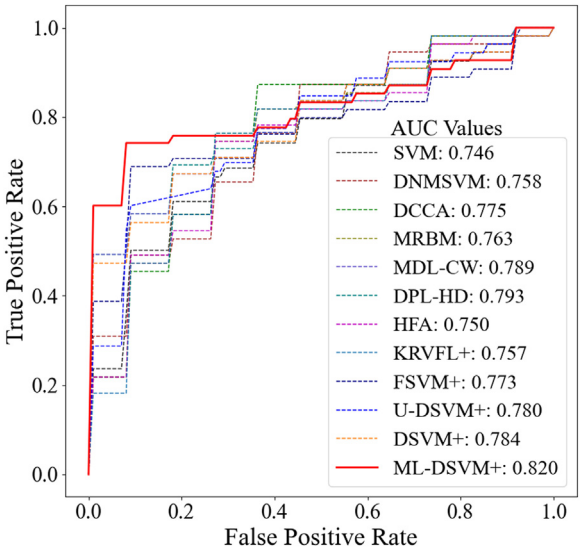


Fig. 8. ROC curves and corresponding AUC values of different algorithms on the ABIDE dataset.

of ML-DSVM+ within 50 steps and always converges within 100-150 steps, which indicates the effectiveness of the proposed ML-DSVM+. By contrast, when the training losses of DNMSVM and DSVM+ stagnate early, the loss of ML-DSVM+ continues to decrease. Fig. 9 shows that the loss values of DNMSVM and DSVM+ vibrate after 100 steps, while the ML-DSVM+ is already convergent.

4.7. Sensitivity analysis

It is worth noting that there are four critical hyperparameters in the proposed ML-DSVM+, *i.e.*, C_1, C_2, λ_1, and λ_2. Thus, the TADPOLE dataset is selected as the example for the sensitivity analysis about these hyperparameters, because it is a famous public dataset. As shown in Fig. 10, we fix three hyperparameters and then conduct the training procedure with different values of another hyperparameter.

The candidate of C_1 and C_2 are set as {0.001, 0.01, 0.1, 1, 10, 100, 1000}. The candidate of λ_1 and λ_2 are set as {0.0005, 0.005, 0.05, 0.5, 0.1, 1, 5}. As shown in Fig. 10, it can be found that the best accuracy is obtained when C_1 and C_2 are close to 1. Moreover, λ_1 and λ_2 can be set to 0.05 for optimal accuracy.

4.8. Implementation efficiency

In order to evaluate the efficiency of ML-DSVM+, we compared both averaged training and testing time of different algorithms on the TADPOLE dataset as an example. The averaged training time was calculated using the values of all the 5-fold experiments, and the averaged testing time was calculated on each testing sample. As shown in Table 4, ML-DSVM+ takes more training time to converge. It is because meta-learning involves second-order computation. However, the average testing time of ML-DSVM+ is on the same order of magnitude as other algorithms.

5. Discussion

In this work, a novel ML-DSVM+ algorithm is proposed to improve the performance of single-modal imaging-based CAD with transferred knowledge from the related imaging modality. We first

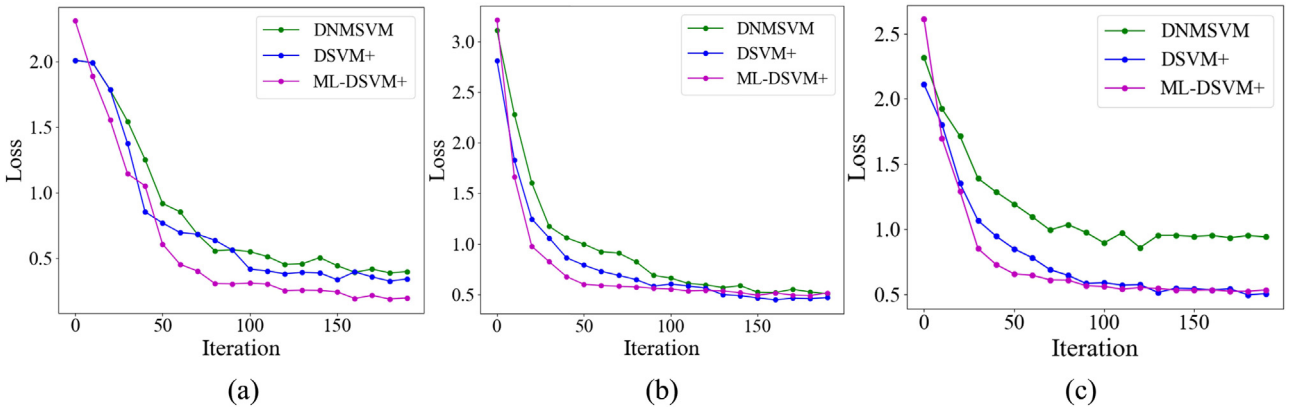


Fig. 9. Loss curves of ML-DSVM+, DSVM+ and DNMSVM on (a) BBUI dataset, (b) TADPOLE challenge dataset and (c) ABIDE dataset.

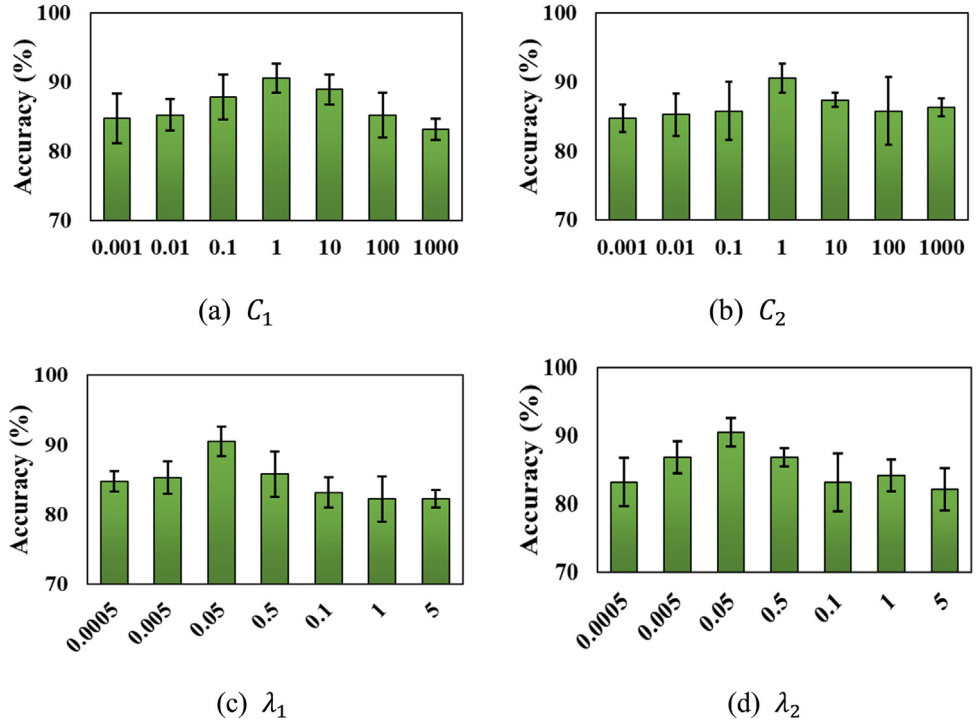


Fig. 10. Sensitivity analysis results for (a) C_1, (b) C_2, (c) λ_1, and (d) λ_2 on the TADPOLE dataset.

Table 4
Implementation efficiency of different algorithms. $1K$ denotes the batch size.

Algorithms	Training time (Unit: Sec)	Testing time (Unit: Sec)
DSVM	142.33	0.0048
U-DSVM+	178.15	0.0051
DSVM+	180.43	0.0062
ML-DSVM+	208.26	0.0049

integrate the bi-channel DNNs and SVM+ classifier into a uniformed framework, and then propose a coupled hinge loss for bidirectional TL to improve the transfer performance. Moreover, the meta-learning strategy is further adopted into DSVM+ to alleviate the overfitting caused by the significantly increased parameters in DNNs on small-size training samples. The experimental results on three medical imaging datasets indicate the effectiveness of the proposed ML-DSVM+.

In clinical practice, paired multi-modal medical imaging data generally share the same label. It is expected that the supervised TL can achieve superior transfer performance for this type of paired data with the guidance of label information. Therefore, the LUPI paradigm is suitable for knowledge transfer between paired data with shared labels. It is known that feature representation is another critical module except for the classifier in a CAD system. Because DL can improve the representation of hand-crafted features, we propose the DSVM+ algorithm by integrating the bi-channel DNNs and SVM+ into a unified framework. The introduction of DNNs can also avoid the issue of kernel selection in the original SVM+ algorithm [20]. Moreover, we design a new coupled hinge loss function to conduct bidirectional transfer in DSVM+ instead of unidirectional transfer in the original SVM+. The proposed coupled hinge loss has several advantages: 1) The bidirectional transfer strategy can simultaneously improve the feature representation of both the TD and SD networks with more knowledge interaction, and extract more transferable implicit knowledge from these two domains; 2) The three terms in the coupled hinge loss can make full use of the shared labels to explore and guide knowledge transfer in a supervised manner, and simultaneously improve both feature representation and classification performance in a unified framework; 3) Although we do not add the transfer module in both DNNs, they intrinsically realize knowledge transfer by updating the network parameters guided by the new coupled loss function. In addition, DSVM+ simultaneously trains two independent classifiers, which makes it more flexible for applications.

Although the bi-channel DNNs improve feature representation, they also significantly increase the parameters, which will result in the overfitting problem with limited medical imaging training samples. To mitigate this issue, we propose introducing meta-learning to train the DSVM+ model. In contrast to existing meta-learning methods that require a large number of different tasks to train a model, the meta-tasks in ML-DSVM+ are self-generated by randomly sampling instances from the training dataset. This strategy to generate meta-tasks is suitable for CAD tasks with improved performance, because it is generally difficult to collect many different medical images with different diseases to build a large number of meta-tasks. Moreover, our sampling strategy controls each meta-task to have equal numbers of positive and negative samples, and therefore, ML-DSVM+ also overcomes the problem of class imbalance, which is very common in the field of CAD. It is worth noting that the proposed ML-DSVM+ can also be applied to the conventional meta-learning scenario, which adopts a lot of related tasks to train the base learner.

6. Conclusion

In summary, we propose a novel LUPI-based algorithm, namely ML-DSVM+, for the single-modal imaging-based CAD. ML-DSVM+ integrates the bi-channel DNNs and SVM+ into a unified framework, which simultaneously improves both feature representation and classification under the guidance of shared labels. The experimental results on three datasets show that ML-DSVM+ effectively improves the performance of single-modal imaging-based CAD with the transferred knowledge from SD. It suggests the potential of ML-DSVM+ for CAD, even with a small and class-imbalanced training set.

It still has room to improve the proposed ML-DSVM+. Due to the limited number of training samples, the shallow DNNs is adopted in ML-DSVM+ to further improve the representation of hand-crafted features. In our future work, the proposed ML-DSVM+ can be easily extended to a CNN version by replacing the DNN backbone with CNN, which will have more applications. Besides, to alleviate the overfitting issue on limited training samples, we will develop new optimization strategies to effectively reduce the complexity and also improve stability with small standard deviation. Moreover, the feature-level TL strategy will be integrated into the current ML-DSVM+ to further promote its performance.

Declaration of Competing Interest

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

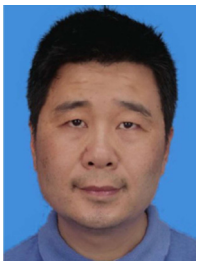
Acknowledgments

This work is supported by the National Natural Science Foundation of China (81830058, 11971296) and the 111 Project (D20031). The authors would like to thank Mrs. Weijun Zhou with Nanjing Drum Tower Hospital and the First Affiliated Hospital of Nanjing University Medical School to provide the BBUI dataset in our experiments.

References

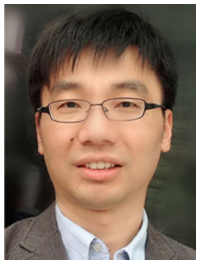
[1] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
[2] X. Fei, L. Shen, S. Ying, Y. Cai, Q. Zhang, W. Kong, W. Zhou, J. Shi, Parameter transfer deep neural network for single-modal B-Mode ultrasound-based computer-aided diagnosis, Cogn. Comput. 12 (2020) 1252–1264.
[3] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: International Conference on Artificial Neural Networks (ICANN), 2018, pp. 270–279.
[4] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: a survey of semi–supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296.
[5] V. Vapnik, R. Izmailov, Learning using privileged information: similarity control and knowledge transfer, J. Mach. Learn. Res. 16 (1) (2015) 2023–2049.
[6] L. Duan, Y. Xu, W. Li, L. Chen, D.W.K. Wong, T.Y. Wong, J. Liu, Incorporating privileged genetic information for fundus image based glaucoma detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2014, pp. 204–211.
[7] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[8] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neur. Comput. 18 (2006) 1527–1554.
[9] Y. Tang, Deep learning using linear support vector machines, International Conference on Machine Learning (ICML) Challenges in Representation Learning Workshop, 2013.
[10] Y. Li, L. Yuan, N. Vasconcelos, Bidirectional learning for domain adaptation of semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6936–6945.
[11] L. Fang, X. Wang, Brain tumor segmentation based on the dual-path network of multi-modal MRI images, Pattern Recognit. 124 (2022) 108434.
[12] M. Huisman, J.N. Van Rijn, A. Plaat, A survey of deep meta-learning, Artif. Intell. Rev. 54 (6) (2021) 4483–4541.
[13] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1126–1135.
[14] X. Li, B. Du, C. Xu, Y. Zhang, L. Zhang, D. Tao, R-SVM+: Robust learning with privileged information, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 2411–2417.
[15] W. Li, D. Dai, M. Tan, D. Xu, L.V. Gool, Fast algorithms for linear and kernel SVM+, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2258–2266.
[16] N. Sarafianos, M. Vrigkas, I.A. Kakadiaris, Adaptive SVM+: Learning with privileged information for domain adaptation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2637–2644.
[17] J. Tang, Y. Tian, P. Zhang, X. Liu, Multiview privileged support vector machines, IEEE Trans. Neural Netw. Learn. Syst. 29 (2018) 3463–3477.
[18] X. Zheng, J. Shi, S. Ying, Q. Zhang, Y. Li, Improving single-modal neuroimaging based diagnosis of brain disorders via boosted privileged information learning framework, in: International Workshop on Machine Learning in Medical Imaging (MLMI), 2016, pp. 95–103.
[19] H.H. Alahmadi, Y. Shen, S. Fouad, C.D.B. Luft, P. Bentham, Z. Kourtzi, P. Tino, Classifying cognitive profiles using machine learning with privileged information in mild cognitive impairment, Front. Comput. Neurosc. 10 (2016) 117.
[20] Y. Li, T. Zhang, Deep neural mapping support vector machines, Neur. Netw 93 (2017) 185–194.
[21] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neur. Netw. 4 (1991) 251–257.
[22] C. Chen, Q. Dou, H. Chen, J. Qin, P.A. Heng, Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation, IEEE Trans. Med. Imaging 39 (7) (2020) 2494–2505.
[23] R. Singh, V. Bharti, V. Purohit, A. Kumar, A.K. Singh, S.K. Singh, MetaMed: Few-shot medical image classification using gradient-based meta-learning, Pattern Recognit. 120 (2021) 108111.
[24] C.R. Jack, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reason. Imaging 27 (2008) 685–691.
[25] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, J.S. Anderson, M. Assaf, S.Y. Bookheimer, M. Dapretto, The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatry 19 (2014) 659–667.
[26] Q. Zhang, Y. Xiao, J. Suo, J. Shi, J. Yu, Y. Guo, Y. Wang, H. Zheng, Sonoelastomics for breast tumor classification: A radiomics approach with clustering-based feature selection on sonoelastography, Ultrasound Med. Biol. 43 (2017) 1058–1069.
[27] J. Ashburner, Computational anatomy with the SPM software, Magn. Reson. Imaging 27 (8) (2009) 1163–1174.
[28] M. Reuter, N.J. Schmansky, H.D. Rosas, B. Fischl, Within-subject template estimation for unbiased longitudinal image analysis, Neuroimage 61 (4) (2012) 1402–1418.
[29] J. Wang, Q. Wang, J. Peng, D. Nie, F. Zhao, M. Kim, H. Zhang, C.-Y. Wee, S. Wang, D. Shen, Multi-task diagnosis for autism spectrum disorders using multi-modality features: a multi-center study, Hum. Brain Mapp. 38 (2017) 3081–3097.
[30] F. Zhao, L. Qiao, F. Shi, P.T. Yap, D. Shen, Feature fusion via hierarchsical supervised local CCA for diagnosis of autism spectrum disorder, Brain Imaging Behav. 11 (4) (2017) 1050–1060.
[31] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[32] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the 30th International Conference on Machine Learning (ICML), 2013, pp. 1247–1255.
[33] K. Sohn, W. Shang, H. Lee, Improved multi-modal deep learning with variation of information, in: Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 2141–2149.
[34] S. Rastegar, M. Soleymani, H.R. Rabiee, S.M. Shojaee, MDL-CW: A multi–modal deep learning framework with cross weights, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2601–2609.
[35] J. Lambert, O. Sener, S. Savarese, Deep learning under privileged information using heteroscedastic dropout, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8886–8895.
[36] W. Li, L. Duan, D. Xu, I.W. Tsang, Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1134–1148.
[37] P.B. Zhang, Z.X. Yang, A new learning paradigm for random vector functional-link network: RVFL+, Neural Netw. 122 (2020) 94–105.

Xiangmin Han received his B.S. degree and M.S. degree in software engineering from Beijing Union University, China, in 2014 and 2018, respectively. He is currently a Ph. D. candidate at the School of Communication and Information Engineering, Shanghai University. His research interests include machine learning in medical image computing.



Jun Wang received his Ph.D. degree in pattern recognition and intelligence systems from the School of Computer Science and Technology in Nanjing University of Science and Technology, China, in 2011. He has been a Research Assistant in the Department of Computing, Hong Kong Polytechnic University, China, and a postdoc research fellow in the Department of Radiology and BRIC, School of Medicine, University of North Carolina at Chapel Hill, USA, respectively. He is currently an Associate Professor with Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, China. His research interests include machine learning, fuzzy systems and medical image classification.

Shihui Ying received the B.Eng. degree in Mechanical Engineering and the Ph.D. degree in Applied Mathematics from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2008, respectively. He held a postdoctoral position with the Biomedical Research Imaging Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, from 2012 to 2013. He is currently a Professor with the Department of Mathematics, School of Science, Shanghai University, Shanghai, China. His current research interests include geometric theory and methods for medical image processing, and machine learning.



Jun Shi received the B.S. degree and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China in 2000 and 2005, respectively. In 2005, he joined the School of Communication and Information Engineering, Shanghai University, Shanghai, China, where he has been a Professor since 2015. From 2011 to 2012, he was a visiting scholar with the University of North Carolina at Chapel Hill, Chapel Hill, USA. His current research interests include machine learning in medical imaging.

Dinggang Shen, Professor, IEEE Fellow, AIMBE Fellow, IAPR Fellow. His research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 1000 papers in the international journals and conference proceedings, with h-index of 122. He serves as an editorial board member for eight international journals, and was General Chair for MICCAI 2019.