---

**Article:**
Yang, Y., Guo, J., Ye, Q. et al. (2021) A weighted multi-feature transfer learning framework for intelligent medical decision making. Applied Soft Computing, 105. 107242. ISSN: 1568-4946

https://doi.org/10.1016/j.asoc.2021.107242

---

# A Weighted Multi-Feature Transfer Learning Framework for Intelligent Medical Decision Making

Yun Yang[a], Jing Guo[b], Qiongwei Ye[c], Yuelong Xia[b], Po Yang[d], Amin Ullah[e], Khan Muhammad[e]

[a]*School of Software, Yunnan University, Kunming, China*

[b]*School of Information Science and Engineering, Yunnan University, Kunming, China*

[c]*School of Business, Yunnan University of Finance and Economics, Kunming, China*

[d]*Department of Computer Science, Sheffield University, Sheffield, UK*

[e]*Department of Software, Sejong University, Seoul 143-747, Republic of Korea*

## ABSTRACT

Transformative computing provides an emerging technology to data analysis and information processing, but how to effectively connect the data derived from different domains has aroused much of concern. Especially on medical areas, the scarcity of annotated medical data makes it hard to build a robust classification model, thus, the utilization of medical resources from different sources is particularly important. Transfer learning leverages the knowledge gained from the related domain to enhance the computational effectivity on the target domain. In this work, we extend transfer learning with ensemble learning to present a novel Weighted Multi-Feature Hybrid Transfer Learning Framework (W-MHTL) that builds a transformative approach to connect different domains and applies it to medical decision making. Our approach lessens the distribution variances from multiple perspectives by applying variant types of feature-based transfer learning methods. In each feature space, we construct the transfer model by evaluating the corelations and obtain the predicting result from each model. Finally, a feasible ensemble strategy is used to jointly consider each result. We evaluate our approach on synthetic datasets and UCI medical benchmarks, and a cerebral stroke dataset collected from local hospital. The experiment results reveal our method achieves superior performances with the currently available alternatives.

***Keywords:*** *medical decision making; transfer learning; ensemble learning; distribution variances; transformative computing*

## 1. Introduction

With the continues expansion of Internet and the evolution of technology, the data scale grows exponentially, which lights the enthusiasm for researchers to apply more intelligent learning methods and mine additional useful information in huge number of datasets [1-3]. The Wise Information Technology of 120 (WIT120) aims to build a regional medical information platform for health records and utilize the intelligent technologies to help doctors and researchers improve the efficiency and quality of patients' medical treatments [4]. Due to the efficient data analysis and information mining abilities,

Corresponding author: Qiongwei Ye, E-mail: yeqiongwei@163.com; Yun Yang and Jing Guo contributed equally to this work.

machine learning has become a mainstream approach in WIT120, including healthcare systems [5-7], early interventions [8, 9], disease diagnosis support [10-12], and others [13-15]. In real-world applications, data analysis is usually run based on large information sets gathered, obtained from varied resources, which are frequently independent of each other. Especially in the medical fields, the uneven development of medical condition leads to the problems, such as lacking medical resources, limited diagnosis information and high diagnostic error rates, which will cause insufficient or even no reliable medical data to build a robust classification model. Compared to remote areas, developed areas tend to possess sufficient medical resources, including clinical records, medical images, and so on. However, these independent medical resources between different areas exist distribution variances caused by discrepancies in medical facilities, age groups, geography and other factors [16]. The direct use of existing medical models for medical decision making in remote area often fails to reach a satisfactory result [17]. Transformative computing aims to provide a persuasive method to execute computations or analysis on the data obtained from various resources [18, 19]. Through this way, we can make a rational use of multiple resources, to improve the processing ability of target tasks. As illustrated in Fig. 1, the main objective of this paper is to construct a transformative method to fix the chasm between different medical domains and utilize the medical resources from developed areas as auxiliary supplement to offer better medical treatments for medical decision making in remote areas.

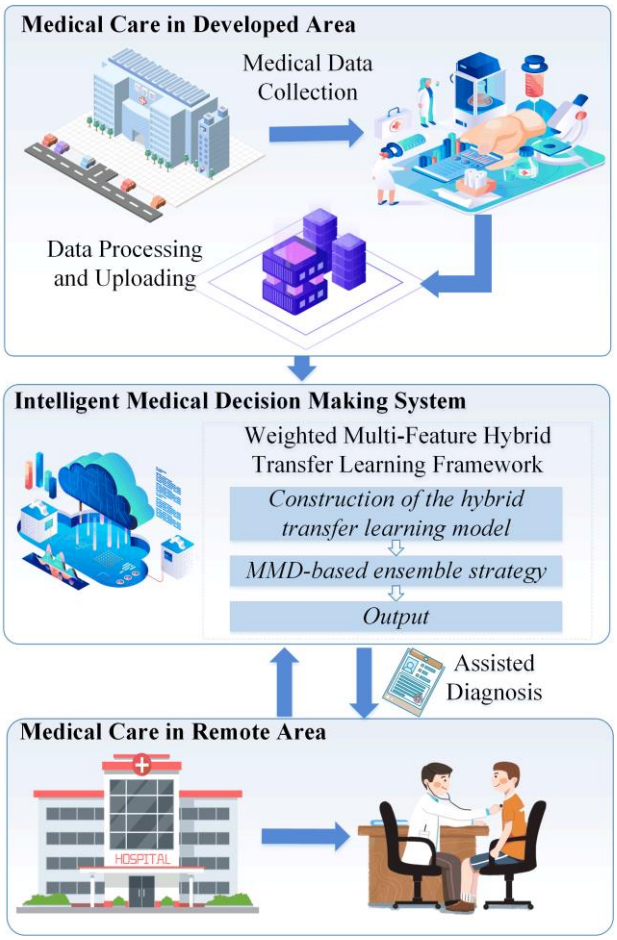



**Fig. 1.** Flow chart of the intelligent medical decision making.

Transfer learning [20] breaks the hypothesis of traditional machine learning that the distribution of training data and testing data must be consistent. It can improve the ability to solve target tasks by decreasing the distribution variances and building the connection between domains such that the knowledge obtained from the related resources can be effectively used [21]. Following the terminology of transfer learning, we denote the one that contains sufficient annotated data as the source domain and the one with only limited amounts of data, or none at all, as the target domain [22]. By utilizing massive, similar and high-quality medical resources collected in developed areas, transfer learning can be used as a transformative way to improve the medical decision making in remote areas. There are four categories of transfer learning approaches [20]: instance-based transfer learning [23-25], feature-based transfer learning [26-29], model-based transfer learning [30, 31] and relation-based transfer learning [17, 32]. In our framework, we mainly consider feature-based and instance-based transfer methods. The feature-based transfer learning aims to reduce the discrepancy between source and target domains through feature transformation. By rationally transforming the origin source domain and target domain into a new feature space, the distribution variance can be effectively reduced. The instance-based transfer learning method reuses data based on generating weighting rules and has better theoretical foundations. However, some problems still exist in the aforementioned transfer learning approaches in practical applications. Feature-based transfer learning methods can only reduce the distribution variance from a single perspective, so it is difficult to select an optimal method for a given data set. Moreover, when mapped into the new feature space, the traditional methods directly use the source domain to train the classifier for target domain, which will cause low generalization by irrelevant instances. Instance-based transfer learning is effective in domain distributions with small discrepancies, but it has problems when addressing complex distributions. Considering the defects of traditional transfer learning methods in medical decision making, we apply various feature mapping methods to transform the source domain and target domain into several new feature spaces, in which the discrepancy between domains can be effectively reduced from different views. After that, the instance-based transfer learning method is used to weight the training sets on account of the different contributions for target domains. The result of each single-feature transfer model can be obtained after applying our hybrid strategy.

To aggregate each single-feature transfer model and integrate each output for a better prediction result, we extend ensemble approach and propose a feasible ensemble strategy for our framework. Ensemble learning completes classification tasks by constructing and combining multiple weak classifiers to achieve better results [33]. The advantages of the combination of ensemble learning and transfer learning can be summarized from three main aspects: (1) Accuracy: A combined result achieves a better average performance than does a single transfer result; (2) Transferability: A combined solution can extract more domain knowledge from the source domain; (3) Robustness: A combined solution can reduce sensitivity to distribution variances. In this regard, researchers have conducted different studies. For instance, Acharya et al. [34] described an optimization framework that utilized existing classifiers trained on source data and a similarity matrix from a cluster ensemble operating on the target data that yielded a consensus labelling of the target data. Kandaswamy et al. [35] proposed a methodology to reduce the impact of selective layer-based transference and provided an optimized framework that works for many transfer learning cases. Bhatt et al. [36] combined transfer learning and co-training paradigms

and applied this co-transfer learning framework to perform cross-resolution face matching. Our previous studies illustrated the advantages of using ensemble learning algorithms [37-41].

In this paper, we present a transformative approach which extends transfer learning and ensemble learning to solve problems in medical decision making. Our framework has two parts: First, we take a combination of feature-based and instance-based transfer learning approaches to form the hybrid transfer model. Specifically, considering the complicated distribution variance, we apply multiple feature-based methods that decrease the discrepancy from different views. Then the instance-based transfer learning method can be used to construct the hybrid models and enhance the connection between different domains. Second, we propose an ensemble strategy to make further analysis for each single-feature hybrid transfer model and reconcile the predictions to obtain a final transfer result. The main contributions of our work are summarized as follows:

- We leverage the transfer learning and ensemble learning to propose a transformative approach that decreases the distribution variance among different medical resources and provides an effective solution for target classification tasks on medical decision making.
- Compared with other solutions, our approach utilizes different feature-based transfer learning methods to reduce the distribution discrepancy from various perspectives, which solves the difficulty of searching the optimal feature transformation in the conditions of the complex distributions.
- By combining feature-based transfer learning and instance-based transfer learning to form a hybrid transfer model, our approach overcomes the fundamental weakness of reducing the distribution discrepancy and eliminates the negative effects by offering a proper weighting strategy for irrelevant samples.
- Based on the different distribution discrepancies after feature mapping, we measure the correspondence between source domain and target domain in the transformed feature space and propose a novel weighted ensemble strategy to obtain the final results.

The following parts of this paper are organized as follows. Section 2 presents related works. Section 3 describes the proposed framework and provides mathematical analysis. Section 4 reports the simulation results on a collection of medical datasets. Section 5 provides further discussion of the experimental results and possible future work. The conclusions are drawn in Section 6.

## 2. Related work

Domain adaptation is an important branch of feature-based transfer learning that maps source and target data into a Reproducing Kernel Hilbert Space (RKHS) [42] in which the distributions of the source and target domains should be as similar as possible. Maximum Mean Difference (MMD) [43] is a commonly used measurement to compute the differences among various domains [26-29]. By minimizing the MMD, the optimal mapping space can be found. The distribution issues that occur between the source and target domains can largely be summarized in two different aspects: marginal distribution, namely $P(X_S)=P(X_T)$ and conditional distribution, namely $P(Y_S|\ X_S) = P(Y_T|\ X_T)$ [21]. Many researchers focus on how to provide a proper way to find the optimal transformation based on the above variances. But without performing a qualitative analysis of the data distribution, it is quite difficult

to choose the optimal method to adapt the distribution variances between domains. The instance-based transfer learning methods reuse data based on generating weighting rules and has better theoretical foundations [23, 24, 44]. For example, TrAdaBoost [23] proposed a new mechanism to automatically adjust the weights of training samples. In the source domain, the weights of instances reduced when the instances were misclassified, while in the target domain, as in AdaBoost [45], the weights were reduced for correctly classified instances and increased for incorrectly classified instances. When assigned reasonable weights, the training instances can efficiently transfer the source domain knowledge to solve the target task. However, this kind of approach is effective in domain distributions with small discrepancies, it will lose its efficiency when addressing complex distributions.

Recently, lots of researchers have applied transfer learning methods to medical decision making. Samala et al. [46] presented a multi-stage transfer learning framework for the classification of malignant and benign masses in digital mammography tomography synthesis. By conducting a two-step fine-tuning strategy, a well-trained neural network can be firstly fine-tuned with the mammography dataset and utilize the annotated digital breast dataset to make a further training. The simulation results demonstrate the advantageous of transfer learning in improving the performance of model in target tasks when training data is limited. Due to effects of contrast, brightness and artifacts in medical images and the time and labour consuming to examine and evaluate, Martinez et al. [47] leveraged transfer learning to present an aided diagnosis tool to solve the classification tasks on retinography. They prove its ability in distinguishing among different grades of diabetic retinopathy. Wang et al. [48] proposed a transfer learning approach with least squares support vector machine, which can leverage the limited training sets to maintain the robustness of the model. The proposed approach is applied to a real-world clinical dataset to predict overall and cancer-specific mortality in patients with bladder cancer at 5 years after radical cystectomy.

## 3. Weighted Multi-Feature Hybrid Transfer Learning Framework

In this section, we propose the W-MHTL. This framework can be mainly divided into 2 parts: the *construction of the hybrid transfer learning model* and *MMD-based weighted ensemble*.

### *3.1. Problem Statement*

Let $\mathcal{D}_s = \{x_i, y_i\} (i = 1,...,n)$ be the source domain. To enable our transfer learning model, we use part of the labelled data as training data in the target domain and denote as $\mathcal{D}_T = \mathcal{D}_T^{train} \cup \mathcal{D}_T^{test} = \{x_j^{train}, y_j^{train}\} \cup \{x_j^{test}\} (j = 1,...,m)$. We assume the feature space is $\mathcal{X}_S = \mathcal{X}_T$, and the label space is $\mathcal{Y}_S = \mathcal{Y}_T$. The data distribution between the different domains exists as variances in either the conditional probability distribution $P(Y_S | X_S) = P(Y_T | X_T)$ or the marginal probability distribution $P(X_S) = P(X_T)$. Therefore, our transfer learning model aims to learn the labels of $\mathcal{D}_T^{test}$ using the source domain $\mathcal{D}_S$ and the labelled target domain $\mathcal{D}_T^{train}$.

*3.2. Construction of the Hybrid Transfer Learning Model*

Regarding the different types of domain distributions, it is hard to find a high-performing domain adaptation method for all given datasets. Therefore, we adopt different types of methods to narrow the distribution variance by mapping them into multiple RKHSs. In this research, we adopt four classical domain adaptation methods to map the source domain and target domain into new feature spaces and we define the mapping function as $\varphi_c\left(c=1,...,C\right)$. We denote $\mathcal{H}$ as RKHS, $l\in\{1,2,...,L\}$ as the distinct class label. The instances belonging to class $l$ in the source and target domains can be denoted as $\mathcal{D}_s^{(l)}$ and $\mathcal{D}_T^{(l)}$.

- Transfer Component Analysis (TCA) [26] is a classical domain adaptation algorithm that reduces the distribution variance between the source and target domains mainly from the marginal distribution perspective, as shown in Eq. (1).

$$\min_{\varphi_c} Dist_c^{\,2}(\mathcal{D}_S,\mathcal{D}_T)=\min_{\varphi_c}\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi_c(x_i)-\frac{1}{m}\sum_{j=1}^{m}\varphi_c(x_j)\right\|_{\mathcal{H}}^2 \tag{1}$$

- Transfer Joint Matching (TJM) [27] lessens the discrepancy through the perspective of marginal distribution adaptation and the instances are weighted by their importance, as shown in Eq. (2). To solve this problem, TJM weights the source domain instances by introducing $\ell_{2,1}$ [49] to reflect the instance correlations. It should be noted that the way of reweighting the instance in TJM occurs in feature mapping steps and it is different from the weighting strategy when constructing our transfer model.

$$\min_{\varphi_c} Dist_c^{\,2}(\mathcal{D}_S,\mathcal{D}_T)=\min_{\varphi_c}\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi_c(x_i)-\frac{1}{m}\sum_{j=1}^{m}\varphi_c(x_j)\right\|_{\mathcal{H}}^2+\lambda\left\|\varphi_c\right\|_{2,1} \tag{2}$$

- Joint Distribution Adaptation (JDA) [28] reduces the distance between the source and target domain through the perspective of the joint probability distribution, which takes both the marginal and conditional distributions into consideration. For the labels in the target domain, using Eq. (3), the adopted method generates fake labels by training a simple classifier with a source domain $\mathcal{D}_s$ and constantly reducing the differences over the iterations.

$$\min_{\varphi_c} Dist_c^{\,2}(\mathcal{D}_S,\mathcal{D}_T)=\min_{\varphi_c}\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi_c(x_i)-\frac{1}{m}\sum_{j=1}^{m}\varphi_c(x_j)\right\|_{\mathcal{H}}^2+\sum_{l=1}^{L}\left\|\frac{1}{n_c}\sum_{x_i\in\mathcal{D}_S^{(l)}}\varphi_c(x_i)-\frac{1}{m_c}\sum_{x_j\in\mathcal{D}_T^{(l)}}\varphi_c(x_j)\right\|_{\mathcal{H}}^2 \tag{3}$$

- Balanced Distribution Adaptation (BDA) [29] assumes that marginal distribution adaptation and conditional distribution adaptation are unequally important. It adaptively adjusts the importance of the marginal and conditional distributions in domain adaptation by dynamically adjusting the distance between two distributions using a balance factor $\mu$ where $\mu\in[0,1]$. The balance factor in Eq. (4) defines the importance of distributions. BDA considers that marginal distribution adaption

is more important when $\mu \to 0$, and considers the conditional distribution adaption more important when $\mu \to 1$:

$$\min_{\varphi_c} Dist_c^{\,2}(\mathcal{D}_S, \mathcal{D}_T) = \min_{\varphi_c} (1-\mu) \left\| \frac{1}{n} \sum_{i=1}^{n} \varphi_c(x_i) - \frac{1}{m} \sum_{j=1}^{m} \varphi_c(x_j) \right\|_{\mathcal{H}}^2 + \mu \sum_{l=1}^{L} \left\| \frac{1}{n_c} \sum_{x_i \in \mathcal{D}_S^{(l)}} \varphi_c(x_i) - \frac{1}{m_c} \sum_{x_j \in \mathcal{D}_T^{(l)}} \varphi_c(x_j) \right\|_{\mathcal{H}}^2 \tag{4}$$

These four feature mapping approaches adopt the idea of MMD to measure the distribution discrepancy. However, they solve the transfer learning problems from different perspectives or different constraints. The analysis is based on the conceptual framework of Maximum Mean Discrepancy Embedding (MMDE) [31], which first takes advantages of kernel tricks to change the way of learning the mapping function $\varphi_c$ into learning an adaptation matrix $A_c$. Thus, the source and target domains are transformed into RKHS, which can be notated as $\mathcal{D}_{s_c} = \left\{ A_c^{\top} x_i, y_i \right\} (i = 1,...,n)$ in the source domain and $\mathcal{D}_{T_c} = \left\{ A_c^{\top} x_j^{Train}, y_j^{Train} \right\} \mathrm{U} \left\{ A_c^{\top} x_j^{Test} \right\} (j = 1,...,m)$ in the target domain. The new data distribution after feature mapping has been well proven $P_c\left(\mathcal{D}_{S_c}\right) \approx P_c\left(\mathcal{D}_{T_c}\right)$ [26-29].

$$\varepsilon_t = \sum_{j=1}^{m'} \frac{\mu_j^t \left| f\left(A_c^{\top} x_j\right) - y_j \right|}{\sum_{j=1}^{m'} \mu_j^t} \tag{5}$$

where $\mu_j^t$ represents the weights of target training sets in the $t^{\text{th}} \in \{1,...,N\}$ iteration, and the classifier $f$ is trained by all the input training datasets. By denoting $\beta = 1/\left(1+\sqrt{2\ln n / N}\right)$ and $\beta_t = \varepsilon_t / (1-\varepsilon_t)$, where N is the iteration number and n denotes source instances, the update of new training weights $\mu^{t+1}$ can be summarized as follows:

$$\mu^{t+1} = \begin{cases} \mu_i^t \beta^{\left| f(A_c^{\top} x_i) - y_i \right|}, & 1 \le i \le n \\ \mu_j^t \beta_t^{-\left| f(A_c^{\top} x_j) - y_j \right|}, & 1 \le j \le m' \end{cases} \tag{6}$$

The model iterates until the error of the training sets is 0 or reach the iteration number *N*. By giving such a hybrid transfer learning strategy, a robust model $h_c(x)$ can be acquired to predict the testing datasets and obtain the transfer results from each model.

### *3.3. MMD-based Weighted Ensemble*

For the complex data distribution in practical problems, different feature-based transfer learning methods contribute unequally to the final results, so it is crucial to design a reasonable combination. The results of the ensemble owe to the prediction result from each single-feature model, so our assessment criterion is to use MMD as a distance metric to evaluate the similarity in each new feature space and we denote $Dist_c$ as

$$Dist_c = \left\| \frac{1}{n}\sum_{i=1}^{n}\varphi_c\left(x_i\right) - \frac{1}{m}\sum_{j=1}^{m}\varphi_c\left(x_j\right) \right\|_{\mathcal{H}} \tag{7}$$

To ensure each single-feature transfer learning model makes a constructive contribution to the consensus solution, we introduce an efficient weighting scheme. Based on the assumption that a larger MMD distance has less similarity, we define the weight $w_c$ as follows

$$w_c = \frac{exp\left(-Dist_c\right)}{\sum_{c=1}^{C} exp\left(-Dist_c\right)} \tag{8}$$

where $w_c > 0$ and $\sum_{c=1}^{C} w_c = 1$. We use an inverse function to simulate the purpose of our ensemble strategy. Lastly, the weighted hybrid transfer learning framework $H\left(x\right)$ is constructed by a linear combination of each model $h_c$ with its weight $w_c$. The whole process of W-MHTL is summarized in Algorithm 1.

$$H\left(x\right) = \sum_{c=1}^{C} w_c h_c\left(x\right) \tag{9}$$

**Algorithm 1:** W-MHTL

**Input:** training sets $\mathcal{D}_s$ and $\mathcal{D}_T^{train}$ with labels and testing set $\mathcal{D}_T^{test}$ without labels; parameters for domain adaptation

**Output:** Ensemble classifier $H$

**for** $c$=1 to $C$ **do**

Find the adaptation matrix $A_c$ of the feature-based transfer learning method.

Transform all original data into the new pace $\mathcal{H}_c$.

Calculate the ambiguity $Dist_c$ by Eq. (7).

Use TrAdaBoost strategy to weight the training sets $\mathcal{D}_{S_c}$ and $\mathcal{D}_{T_c}^{train}$.

Construct a single-feature hybrid transfer model $h_c$.

**end for**

Get the weights $w_c$ for every transfer model using Eq. (8).

$H \leftarrow$ Multi-feature hybrid model ensemble classification by Eq. (9)

*3.4. Algorithm Analysis*

In our proposed framework, we assume that the classifier from each single-feature model is $h_c = h_1, h_2, ..., h_C$, and $H$ denotes the final ensemble output. Similar to the analysis in [50], we define

the ambiguity of an input instance $x$ between each individual model and the ensemble framework as follows:

$$V\left(h_c|x\right)=\left\|h_c\left(x\right)-H\left(x\right)\right\|^2 \tag{10}$$

Thus, the weighted mean of the ensemble ambiguity can be illustrated as:

$$\begin{aligned}\overline{V}\left(h\,|\,x\right)&=\sum_{c=1}^{C}w_cV\left(h_c\,|\,x\right)\\&=\sum_{c=1}^{C}w_c\left\|h_c\left(x\right)-H\left(x\right)\right\|^2\end{aligned} \tag{11}$$

This function illustrates the ambiguity of the weighted mean ensemble output by each model. In other words, it measures the disagreement among single-feature models on input $x$. We denote $y$ as the ground truth label; consequently, the loss of the individual model and the ensemble are adjusted as follows:

$$L\left(h_c\,|\,x\right)=\left\|h_c\left(x\right)-y\right\|^2 \tag{12}$$

$$L\left(H\,|\,x\right)=\left\|H\left(x\right)-y\right\|^2 \tag{13}$$

Let $\overline{L}\left(h|\,x\right)=\sum_{c=1}^{C}w_cL\left(h_c\,|\,x\right)$ represents the mean of the weighted loss for all individual models. The formula in Eq. (11) can be transformed into:

$$\begin{aligned}\overline{V}\left(h\,|\,x\right)&=\sum_{c=1}^{C}w_c\left\|h_c\left(x\right)-H\left(x\right)\right\|^2\\&=h_c\left(x\right)\cdot H\left(x\right)-\sum_{c=1}^{C}w_cH^2\left(x\right)\\&=\sum_{c=1}^{C}w_c\left\|h_c\left(x\right)-y\right\|^2-\left\|H\left(x\right)-y\right\|^2\\&=\sum_{c=1}^{C}w_cL\left(h_c\,|\,x\right)-L\left(H\,|\,x\right)\\&=\overline{L}\left(h|\,x\right)-L\left(H\,|\,x\right)\end{aligned} \tag{14}$$

In our framework, the distribution variances between different domains are reduced after the feature mapping. Under the ideal circumstance, we consider the distribution $P_c\left(\mathcal{D}_{S_c}\right)=P_c\left(\mathcal{D}_{T_c}\right)$ after feature mapping. Therefore, the ensemble ambiguity for all the input instances can be denoted by:

$$\sum_{c=1}^{C}w_c\int V\left(h_c\,|\,x\right)P\left(x\right)\mathrm{d}x=\sum_{c=1}^{C}w_c\int L\left(h_c\,|\,x\right)P\left(x\right)dx-\int L\left(H\,|\,x\right)P\left(x\right)dx \tag{15}$$

Similarly, the loss value and the ambiguity terms of an individual model for all input instances can be denoted as follows:

$$L_c = \int L(h_c \mid x) P(x) \mathrm{d}x \tag{16}$$

$$V_c = \int V(h_c \mid x) P(x) \mathrm{d}x \tag{17}$$

and the loss function of the ensemble is:

$$L = \int L(H \mid x) P(x) \mathrm{d}x \tag{18}$$

Here, $\overline{L} = \sum_{c=1}^{C} w_c L_c$ represents the weighted average of the loss value, and $\overline{V} = \sum_{c=1}^{C} w_c V_c$ denotes the weighted average of the ambiguity. So, the Eq. (15) can be rewritten as follows:

$$\underbrace{\int L(H \mid x) P(x) \mathrm{d}x}_{L} = \underbrace{\sum_{c=1}^{C} w_c \int L(h_c \mid x) P(x) \mathrm{d}x}_{\overline{L}} - \underbrace{\sum_{c=1}^{C} w_c \int V(h_c \mid x) P(x) \mathrm{d}x}_{\overline{V}} \tag{19}$$

Equation (19) gives a brief interpretation of the ensemble loss $L = \overline{L} - \overline{V}$, in which the former depends on the mean loss values of each hybrid models and latter one includes the variances of each model with the ensemble. The greater the accuracy and diversity of each individual model is, the better the ensemble results will be. To further analyse Eq. (19), $\overline{L}$ is influenced by the loss in each model. In our framework, we construct the hybrid transfer learning model by combination of the feature-based and instance-based transfer learning, which was sufficiently proven in [23] to be able to reduce the training loss both in the source data and target data. The second term in Eq. (19), $\overline{V}$ represents the weighted average of the ambiguity. The differences among the individual models mainly stems from the results of each domain adaptation method. As we mentioned before, it is rare to find a method that performs well for all given datasets. Although the distribution variance may be reduced after each feature mapping method, some discrepancies will still exist in most of the feature spaces. We expect that such properties can be measured by MMD so that the ambiguity of each model can be well reflected. In practical medical applications, domain adaptation methods are generally not available for finding the best mapping function for a given dataset. Therefore, instead of finding a specific feature mapping function, we measure the capacities of the adopted methods from different perspectives by giving them a proper weighting strategy to integrate our framework.

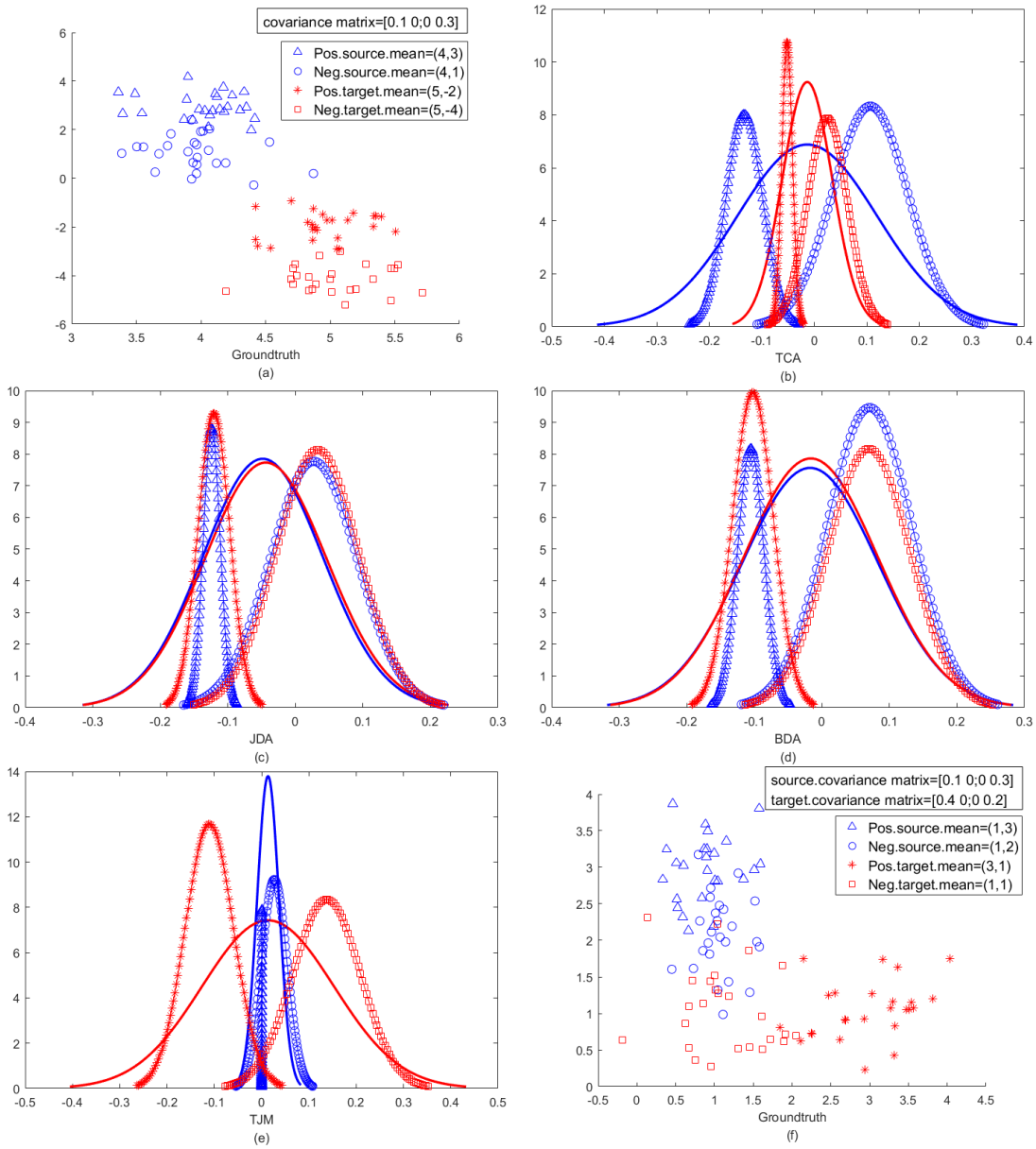
## 4. Simulations

In this section, the extensive evaluations and experiments are conducted to validate the effectiveness and performance of W-MHTL. We design the experiments from three types of datasets: synthetic datasets, UCI medical benchmarks and a cerebral stroke dataset collected from *The First People's Hospital of Yunnan Province*[1], China.

[1] http://www.ypfph.com

*4.1. Synthetic Datasets*

It is known that domain adaptation methods can effectively reduce the distribution variances, nonetheless different methods come from different perspectives, they convey discrepant transferability for the given datasets. For illustration, we use two 2-D synthetic datasets to simulate the marginal distribution variance and conditional distribution variance in practical medical applications. To demonstrate the discrepancies among the four adopted methods, we compare the differences in the Gaussian distribution curve between the source domain and target domain after dimension reduction. The datasets are produced by a Gaussian distribution function with different means and covariance matrixes as illustrated in Fig. 2.



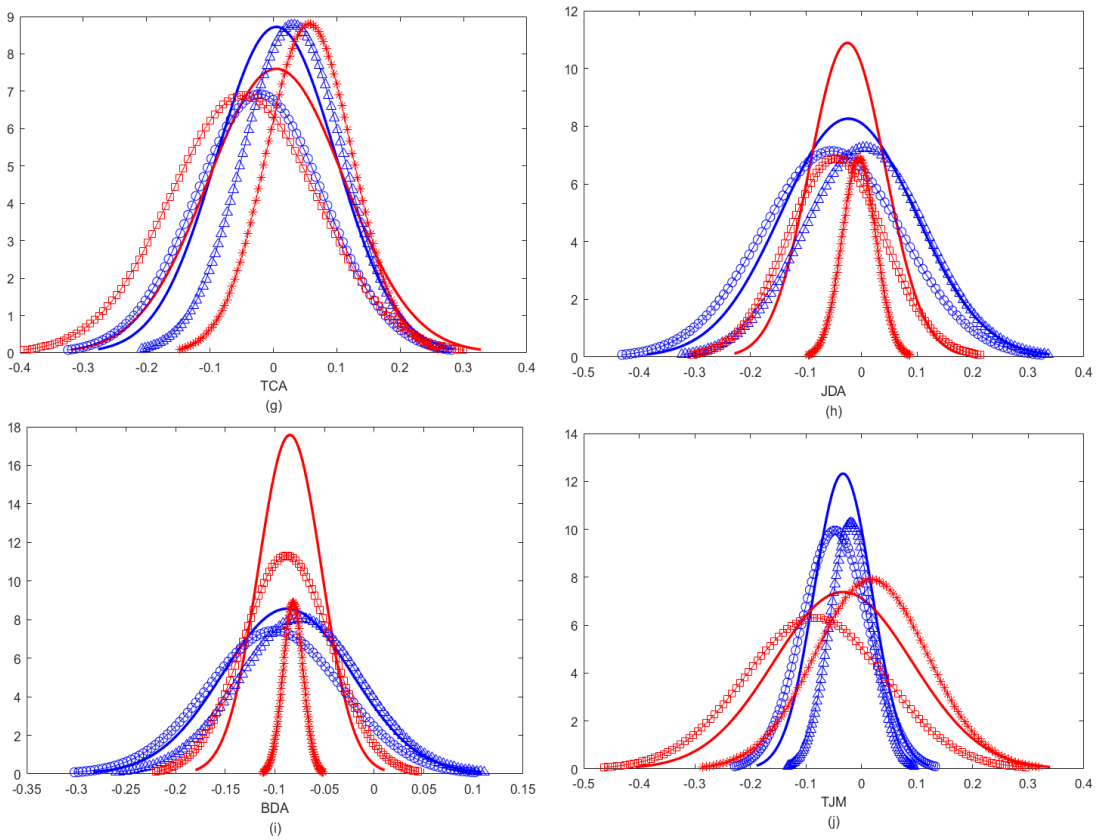



**Fig. 2.** Synthetic data sets demonstrate the different impacts when using different methods: (a) and (f) represent the 2-D datasets of the ground-truth; (b)–(e) and (g)–(j) illustrate the Gaussian distribution curves from different types of domain adaptation methods.

***1) Conditional distribution variance***: Fig. 2. (a) shows the simulated conditional distribution variance situation in which the data are similar overall but not specific to each class. By mapping them into 1-D space, Fig. 2. (b)–(e) show the distribution variances of different domains. By adding the source label information into the domain adaptation process, the distribution curves of JDA and BDA become more similar, while TCA and TJM perform poorly in this situation. For further illustration, the Gaussian distributions of JDA fit perfectly not only in the whole view of datasets but also specific to the curve of each class. BDA can reach a better performance like JDA but remains some discrepancy in each class that may come from some irrelevant data. While TCA and TJM achieve unsatisfactory results, they narrow the distribution only from the perspective of the marginal distribution variance with different constraints; thus, they do not achieve good performance in this situation.

***2) Marginal distribution variance***: To simulate the marginal distribution variance, we adopt different covariance matrices in the source domain and target domain, as shown in Fig. 2. (f). For this kind of complex distribution, we can observe from the curves of the source domain and target domain that the mathematical expectations for each method are basically similar. On the other hand, from the perspective of the standard deviations, the amplitudes of the distributions exhibit a huge discrepancy, which means that simply narrowing the distance on holistic measurements of datasets is insufficient because some irrelevant instances still exist. In Fig. 2. (e), we see that TCA achieves a better performance because the curves show the best matches between domains, while the other three adopted methods show unsatisfactory performance.

In summary, without performing a qualitative analysis of the data distribution, it is quite difficult to select an optimal feature mapping method to adapt the distribution variances between the source and target domains. Furthermore, as shown in Fig. 2, while the distributions after the domain adaptation are similar overall, uncorrelated instances still exist that may cause classifier deviations. Thus, the above intuitive demonstration strongly advocates the necessity for the joint use of diverse perspectives and the need to reweight the instances to achieve better classification.

### *4.2. UCI Medical Benchmarks*

To evaluate the properties of our framework, 18 representative medical datasets from the UCI repository [51] were selected to conduct the experiments. By utilizing the methods in [23], we apply binary feature values in the datasets, such as age group, area, etc. to split the datasets and simulate inconsistent data distributions. In the last 4 groups of datasets, the attributes are all continuous variables. One way to separate the source domain and target domain is to select a given attribute and use K-means to cluster it into two partitions. Intuitively, these two partitions will have different data distribution and we use MMD to further evaluate the variances between different domains. The information of concerning datasets is shown in Table 1 in detail.

**Table 1**
The descriptions of the UCI medical data sets.

| **Datasets** | ***Feature*** | ***Sample*** | ***MMD*** | ***Source Domain*** | | ***Target Domain*** | |
|---|---|---|---|---|---|---|---|
| | | | | ***Pos.*** | ***Neg.*** | ***Pos.*** | ***Neg.*** |
| ***Autism*** | 20 | 346 | 0.082 | 126 | 122 | 62 | 36 |
| ***Heart Disease*** | 12 | 270 | 0.013 | 83 | 100 | 67 | 20 |
| ***Brest Cancer*** | 8 | 277 | 0.211 | 32 | 30 | 164 | 51 |
| ***Diabetic*** | 18 | 1151 | 3.264 | 194 | 193 | 417 | 347 |
| ***ILPD*** | 9 | 579 | 0.457 | 91 | 49 | 323 | 116 |
| ***Sani*** | 54 | 303 | 0.184 | 129 | 40 | 87 | 47 |
| ***Thoracic*** | 15 | 470 | 0.035 | 55 | 268 | 15 | 132 |
| ***Lung*** | 7 | 365 | 0.018 | 149 | 100 | 56 | 60 |
| ***Colic*** | 16 | 368 | 0.184 | 160 | 110 | 72 | 28 |
| ***Cervical*** | 32 | 668 | 0.385 | 37 | 535 | 8 | 88 |
| ***Cleve*** | 12 | 296 | 0.008 | 71 | 24 | 89 | 112 |
| ***Sick*** | 27 | 2643 | 0.048 | 1633 | 131 | 798 | 81 |
| ***Hepatitis*** | 18 | 80 | 0.176 | 19 | 9 | 28 | 24 |
| ***Bupa*** | 6 | 345 | 0.315 | 58 | 61 | 87 | 139 |
| ***Parkinson*** | 22 | 195 | 0.032 | 31 | 19 | 116 | 29 |
| ***Mammography*** | 5 | 830 | 0.072 | 23 | 44 | 380 | 383 |
| ***Pima*** | 8 | 768 | 1.189 | 94 | 192 | 174 | 308 |
| ***WDBC*** | 14 | 569 | 0.222 | 185 | 93 | 172 | 119 |

For the main parameters of the domain adaptation methods, the adopted experiments settings are showed as follows. We simulate the parameter adjustment methods in [27, 28] and seek the optimal parameters through an empirical approach. The iterations of the domain adaptation methods (TJM, JDA, BDA) are set to 20. The optimal $\lambda$ is obtained from $\lambda \in \{0.01, 0.1, 1, 10\}$. The domain adaptation methods involve dimensionality reduction, we select $\{1/4,\ 1/2 \text{ and } 3/4\}$ of the initial data dimensions for dimensionality reduction and choose the best parameter for each model. In practical applications, there is no uniform conclusion on the selection of kernel function, thus empirical adjustments are usually employed according to a given dataset. In general, linear kernel is used for linearly separable cases, and RBF is mainly used for linearly inseparable cases. We find the optimal kernel function for each data set, the RBF kernel is chosen for the datasets *Autism*, *Lung* and *Cervical*,

while the linear kernel is used for the rest of the datasets. For all the selection of parameters, we adopt the grid search to find the optimal parameter sets. The base classifier of this experiment is mainly used to validate the effectiveness of W-MHTL; thus, we chose C4.5 as the base classifier for TrAdaBoost with T=10 iterations. We randomly select 10% target samples as target training sets in each experiment and set the same random seed to guarantee the fairness. The results are compared with the average accuracy and standard deviations of the 10-fold cross validation results.

***1) The Effectiveness of W-MHTL***: To further demonstrate the effectiveness of W-MHTL, we compare the results of each single-feature hybrid model and different ensemble strategies with our approach. Table 2 collectively lists all the experimental results from two aspects. Regarding transference on single features, no single-feature model can produce optimal results on all the datasets; the winning performances spread across the four models on different datasets. These results illustrate the difficulty of choosing the most effective domain adaptation method for a given dataset. To compare with the MHTL which considers the importance of each model equally, our weighting strategy also shows its superiority. As shown in Table 2, our approach achieves expressively better performance in terms of both classification and standard deviations. The objective of our W-MHTL framework is to obtain a more universal way that considers complex data distribution from different perspectives and weighting methods. Our approach achieves the best performances on 16 out of 18 datasets, which provides a strong indication of its effectiveness.

**Table 2**
Classification accuracy (%) of different transfer learning models on medical benchmarks.

| Datasets | *Single Feature Mapping* | | | | *Multi Feature Mapping* | |
|---|---|---|---|---|---|---|
| | TCA [26] | JDA [28] | BDA [29] | TJM [27] | MHTL | W-MHTL |
| ***Autism*** | 75.3±1.6 | 75.7±6.6 | 74.1±3.6 | 71.5±3.2 | **80.1±3.9** | 78.1±3.4 |
| ***Heart Disease*** | 78.6±2.8 | 79.2±4.1 | 76.4±3.1 | 77.4±5.3 | 79.6±2.9 | **80.2±2.6** |
| ***Brest Cancer*** | 72.5±1.7 | 70.5±2.8 | 68.8±3.2 | 71.3±2.2 | 72.7±1.8 | **74.4±1.2** |
| ***Diabetic*** | 66.3±1.3 | 66.1±1.3 | 66.9±1.4 | 66.7±1.2 | 67.3±1.2 | **68.4±0.8** |
| ***ILPD*** | 67.4±2.1 | 65.4±1.1 | 67.1±1.6 | 66.7±1.5 | 67.5±2.1 | **69.9±0.9** |
| ***Sani*** | 70.3±2.4 | 69.4±2.5 | 69.1±3.1 | 68.2±2.9 | 68.7±2.8 | **72.3±2.8** |
| ***Thoracic*** | 84.8±1.9 | 84.8±1.3 | 85.4±2.2 | 84.3±1.1 | 86.6±1.9 | **87.1±1.7** |
| ***Lung*** | 68.4±2.2 | 66.2±3.0 | 68.1±3.1 | 67.7±1.2 | 68.8±3.2 | **70.4±2.3** |
| ***Colic*** | 72.1±1.5 | 70.5±1.5 | 71.9±2.8 | 69.9±2.7 | 74.3±2.7 | **76.7±2.2** |
| ***Cervical*** | 86.4±1.7 | 88.6±2.9 | 88.1±4.3 | 87.8±3.4 | 90.6±2.2 | **93.6±1.8** |
| ***Cleve*** | 68.9±2.1 | 69.8±2.9 | 70.1±2.6 | 69.1±3.2 | 70.5±2.7 | **72.1±2.3** |
| ***Sick*** | 93.1±0.7 | 93.4±0.4 | 92.4±0.7 | 93.2±1.2 | 93.9±0.8 | **95.1±0.4** |
| ***Hepatitis*** | 63.2±3.9 | 66.2±4.1 | 65.5±3.5 | 67.1±3.8 | 67.4±3.3 | **68.9±3.2** |
| ***Bupa*** | 65.4±1.3 | 67.8±2.2 | 66.9±3.2 | 66.1±1.9 | 69.1±1.2 | **72.9±1.7** |
| ***Parkinson*** | 73.2±1.7 | 75.6±3.7 | 76.0±2.3 | 75.3±1.4 | 77.2±3.8 | **79.4±2.4** |
| ***Mammography*** | 69.5±1.4 | 71.3±2.4 | 70.7±2.6 | 71.1±2.0 | 71.3±2.3 | **73.8±1.8** |
| ***Pima*** | 61.1±1.4 | 62.4±2.7 | 63.1±2.1 | 61.5±2.3 | 64.1±1.3 | **66.8±1.8** |
| ***WDBC*** | 93.3±0.8 | 92.5±0.8 | **94.5±1.0** | 91.7±1.8 | 93.4±0.9 | 93.8±0.4 |

***2) The Performance of W-MHTL***: In this part, we compare our W-MHTL approach with other three state-of-the-art transfer learning algorithms for medical classification problems, i.e., TrAdaBoost [23], MTLF [52] and CODA [53]. All the compared algorithms require data with class notation in target

domain, and we use C4.5 as the base learner for all adopted methods. Table 3 shows that TrAdaBoost and the MTLF algorithm each win on two datasets, and the CODA wins on one dataset, while our W-MHTL approach achieves the best results on the other 13 datasets. For further analysis, compared with the feature selection and the matrix processing, the direct sample weighting retains the original data properties thus, the TrAdaBoost algorithm has advantages for processing data with small distribution differences. However, as the distribution variances become larger between the source and target domains, TrAdaBoost will lose its efficacy. The MTLF and CODA belong to feature-based transfer learning methods. From Table 3, we can easily see their effectiveness when the instance-based methods loss its power in datasets *Sani, Lung* and *Sick*. The MTLF takes advantage of the Mahalanobis distance [54] to measure the distribution discrepancy between source and target domains instead of MMD. The CODA conducts a feature selection strategy based on the Pearson correlation coefficient [55], then co-training is used to improve the classifier. However, we can also see that it has limitations when faced with complex distribution variances.

**Table 3**
Classification accuracy of different transfer learning methods on benchmarks.

| *Datasets* | *TrAdaBoost* [23] | *MTLF* [52] | *CODA* [53] | *W-MHTL* |
|---|---|---|---|---|
| ***Autism*** | **83.6±3.3** | 87.6±2.6 | 80.7±2.3 | 78.1±3.4 |
| ***Heart Disease*** | 70.8±1.3 | 78.9±1.1 | 75.4±1.2 | **80.2±2.6** |
| ***Brest Cancer*** | 69.1±3.2 | 71.4±3.6 | 73.1±2.3 | **74.4±1.2** |
| ***Diabetic*** | 62.8±1.3 | 64.4±1.2 | 65.7±1.3 | **68.4±0.8** |
| ***ILPD*** | 68.9±1.5 | 68.1±1.3 | 66.1±2.5 | **69.9±0.9** |
| ***Sani*** | 72.5±2.8 | **74.8±1.7** | 73.4±2.9 | 72.3±2.8 |
| ***Thoracic*** | 86.1±1.8 | 80.2±2.4 | 82.1±1.7 | **87.1±1.7** |
| ***Lung*** | 67.2±2.9 | 68.8±1.7 | **77.4±2.1** | 70.4±2.3 |
| ***Colic*** | 74.7±3.4 | 76.1±1.0 | 75.7±1.6 | **76.7±2.2** |
| ***Cervical*** | 91.4±0.9 | 93.3±0.5 | 90.3±0.7 | **93.6±1.8** |
| ***Cleve*** | 71.2±3.0 | 70.6±1.4 | 69.5±1.2 | **72.1±2.3** |
| ***Sick*** | 92.5±0.7 | **95.9±0.1** | 90.6±0.2 | 95.1±0.4 |
| ***hepatitis*** | 61.9±2.4 | 65.7±1.9 | 65.3±2.8 | **68.9±3.2** |
| ***Bupa*** | 64.3±2.3 | 65.1±2.3 | 68.1±3.2 | **72.9±1.7** |
| ***Parkinson*** | **83.4±1.7** | 81.5±0.5 | 81.2±1.2 | 79.4±2.4 |
| ***Mammography*** | 70.4±1.3 | 71.2±1.6 | 69.8±2.3 | **73.8±1.8** |
| ***Pima*** | 64.0±1.6 | 61.8±1.8 | 60.9±3.5 | **66.8±1.8** |
| ***WDBC*** | 89.8±1.5 | 90.7±0.9 | 90.5±1.5 | **93.8±0.4** |

In Fig. 3, we compare our W-MHTL framework with other algorithms regarding the limited training sample problems. We focus on the datasets *Breast Cancer* and *Diabetic*. In Fig. 3, the X-axis respects the percentages of the training data in the target domain: 5%, 10%, 20%, 30%, 40% and 50%, while the Y-axis represents the accuracy. As shown in Fig. 3, we see the advantages of transfer learning when addressing limited training sample problems. By utilizing knowledge from the source domain, the task of the target domain can be solved well. The W-MHTL yields better performances than any of the other algorithms when there is limited training data on target domain. The accuracy curves improve sharply when the ratio is less than 0.1, especially for the *Breast Cancer*, but gradually flatten out after the ratio is larger than 0.1. Fig. 3 shows that traditional machine learning may be able to learn a suitable classifier when the training data ratio is larger than 0.3. To further demonstrate the performance of W-MHTL, we select two representative unbalanced datasets and use the Receiver Operating Characteristic (ROC) and

Area Under Curve (AUC) for comparison. The experiment results are reported in Fig. 4. It can be observed that W-MHTL achieves better AUC which means our method is more robust than other compared algorithms and has better resistance to the processing of unbalanced datasets.

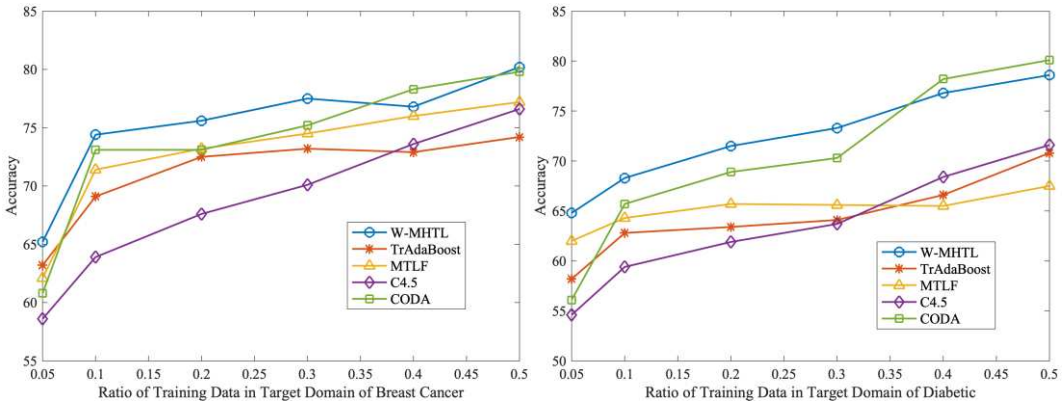



**Fig. 3.** Experimental results on different ratios of target training data.

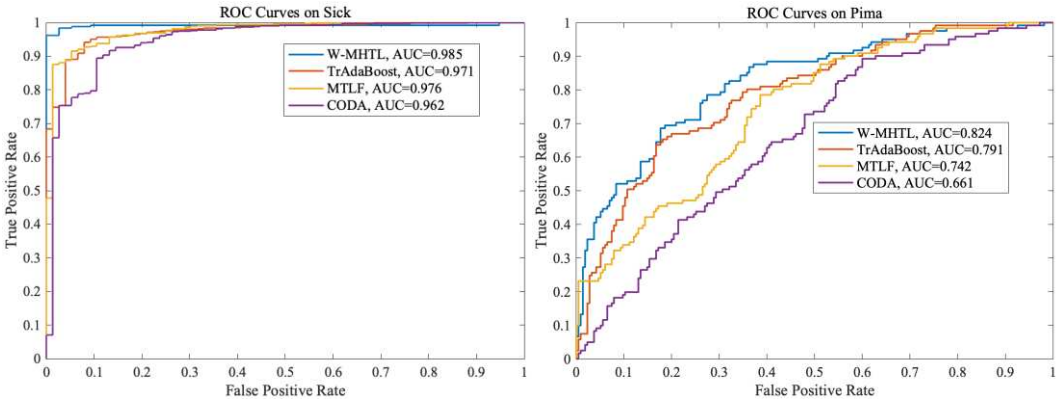



**Fig. 4.** The ROC curves on unbalanced datasets.

### *4.3. Application on Actual Medical Data*

In this section, we verify our W-MHTL framework using actual medical data obtained from *The First People's Hospital of Yunnan Province*, China. Cerebral stroke is an acute cerebral vascular disease, a type of brain tissue injury caused by sudden blood vessel ruptures in the brain or blood flow interruptions to the brain because of vascular obstruction, including both ischaemic and haemorrhagic strokes [56]. Due to lack of effective medical treatment, early prevention is the best approach, among which hypertension is considered as an important controllable stroke risk factor. Therefore, we aim to simulate an experiment in this section to diagnose cerebral strokes in people with hypertension and people without hypertension. As described above, we separate the datasets on the hypertension feature to form the source and target domains and the partition results are reported in Table 4. Our task is to use the knowledge from the data of patients with hypertension to assist in the final task of diagnosing the risk of cerebral stroke in people without hypertension.

**Table 4**
The description of real dataset.

| ***Data Sets*** | ***Feature*** | ***Sample*** | ***MMD*** | ***Source Domain*** | | ***Target Domain*** | |
|---|---|---|---|---|---|---|---|
| | | | | ***Pos.*** | ***Neg.*** | ***Pos.*** | ***Neg.*** |
| ***Cerebral*** | 31 | 3438 | 0.235 | 2246 | 49 | 1083 | 61 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ***Stroke*** | | | | | | | |

Considering the unbalanced nature of origin dataset, we use SMOTE [57] to make the number of positive and negative instances balanced before applying the transfer learning methods for them. By adopting the experimental setups described in Section 4.3, we expect to demonstrate the performance of our approach in practical applications involving limited training samples as well as its effectiveness on resource-sharing problems. As shown in Fig. 5, our framework achieves the best performance among all the compared methods when the training data ratio is below 0.2. This result demonstrates the feasibility of transfer learning methods when training data in the target domain are rare. As the number of training data increases, it is sufficient to simply train a classifier for the final task, and there is no need to use transfer learning. However, considering the scarce training data in practical problems, especially in medical problems, it is necessary to construct a robust classifier by transfer learning methods and our methods provide a satisfactory solution in such cases.

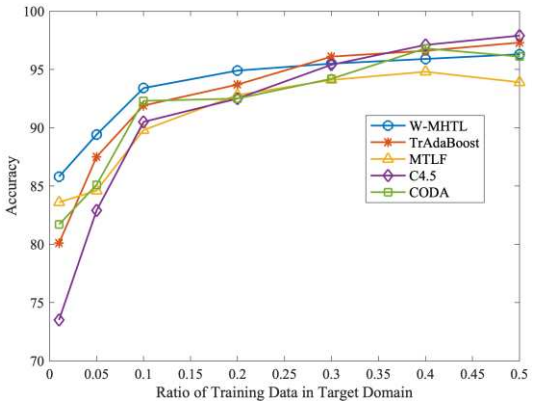



**Fig. 5.** Experimental results on different ratios of real datasets.

## 5. Discussion

As illustrated in Section 4, W-MHTL obtains high-quality transfer learning results even with extremely limited target training data. Therefore, it provides a promising yet easy-to-use technique for addressing transfer learning problems. After analysing from the mathematical formulas and the overall experimental results, the novelties and the deficiencies of our approach are summarized as follows.

Firstly, the usage of different feature perspectives in our framework plays an important character in mitigating the complex distribution variances in medical applications. As shown in Fig. 2, it is difficult to find a high-performing method for all given datasets in complex practical applications. The simulation results in Table 2 and Table 3 consistently indicate that our framework not only provides better classification accuracy than any other compared methods, but also has the stability for the given datasets. Furthermore, by applying multiple feature-based transfer learning methods, W-MHTL can improve the generalization especially on unbalanced datasets.

Secondly, our approach combines the strengths of both feature-based and instance-based transfer learning approaches. As shown in Fig. 2 the data distribution after domain adaptation is similar in the

mass but some irrelevant instances still exist that may cause negative transfer. Thus, a hybrid strategy that combines the feature-based and instance-based transfer learning approaches can solve both data variance and instance reuse problems. Table 2 lists persuasive experimental results that such a combination substantially improves the approach to solving transfer learning problems. Nevertheless, we are convinced that when there are few differences in the data distributions, instance-based transfer learning forms a better solution for representing the information loss during feature-based transfer learning methods. We list the MMD values of the collected datasets in Table 1 and find an interesting phenomenon that data with small variances in distribution tend to get higher accuracy. However, TrAdaBoost does not achieve excellent classification results in all datasets with small distribution variances which reveals that feature-based transfer learning methods can reduce the divergences. Furthermore, they enhance the separability of data for a better classification.

Thirdly, the proposed ensemble strategy produces a consolidated solution for combining each single-feature hybrid transfer learning model through a quantitative analysis of the transfer ability in each feature space. The objective function derived in (15) declare that the performance of our framework depends on both the quality and the diversity of each model, which means that the transfer problem must be solved from different viewpoints. The performance of TrAdaBoost was well-proven in its proposing paper; the diversity of each model stems mainly from the distance between the source domain and the target domain, which we measure using MMD. On a collection of benchmark datasets, the experimental results shown in Table 2 demonstrate the effectiveness of the proposed ensemble strategy in comparison with its prototype. Compared with other classical transfer learning methods, Table 3 shows the superiority of our approach, while Fig. 3 illustrates its effectiveness in dealing with limited training sample problems. Finally, in our tests, the W-MHTL also illustrates its feasibility on real-world problems and achieves satisfactory results.

Although the experimental results demonstrate that our proposed method achieves a certain level of superiority on benchmark datasets and real-world dataset, some constrains still exist for further study. The computational overhead of domain adaptation is quite expensive; thus, our framework requires relatively high computation to find a suitable feature space to reduce the distribution variances. Moreover, this method requires many parameters, but we can seek the optimal parameter values only through an empirical approach, which makes finding the best parameter values a bit difficult. Furthermore, while our approach is compatible with any other domain adaptation algorithms for initial feature mapping, how to select appropriate methods for the target domain, which has different characteristics, remains an interesting topic for further research. Finally, the measurement of transferability is crucial in transfer learning problems—not only to evaluate the data variances between domains but also to consider the target task's accuracy. Such measurement can effectively avoid negative transfer problems, which means that we need to define a more reasonable way to select the optimal module in our framework.

## 6. Conclusion

In this paper, we presented a transformative method named W-MHTL which constructed the connection between different domains and provided a feasible solution for the fundamental problems in

medical decision making. Under the condition of limited training data on target domain, our model combined both feature-based and instance-based transfer learning to make a sufficient leverage of the related domain as auxiliary to improve the performance on target domain. Considering the complicated distribution variance in practical application, we utilized multiple feature-based transfer learning methods to decrease the gap between domains from different perspectives and weighted the instances by their correlations. Finally, we designed a consensus function to further enhance the connection between the source domain and target domain. The simulation results showed that our transfer learning framework yields better results when solving transfer learning problems in medical fields. The experiments to test our framework on real medical data sets demonstrated its robustness.

**Acknowledgment**

The authors are grateful to Eamonn Keogh for providing the benchmark dataset, the University of California for publishing the UCI Machine Learning Repository and the People's Hospital of Yunnan Province for providing the real cerebral stroke dataset. We also thank Jingdong Wang, who provided the transfer learning MATLAB code used in our simulation. We would like to acknowledge the financial support provided by the Chinese Natural Science Foundation, under Grant: 61876166 and 61663046; Yunnan provincial major science and technology special plan projects: digitization research and application demonstration of Yunnan characteristic industry, under Grant: 202002AD080001.

**References**

[1] H. Hu, K. Wang, C. Lv, J. Wu, and Z. Yang, "Semi-supervised metric learning-based anchor graph hashing for large-scale image retrieval," *IEEE Transactions on Image Processing,* vol. 28, no. 2, pp. 739-754, 2018.

[2] C. Xu, K. Wang, P. Li, R. Xia, S. Guo, and M. Guo, "Renewable energy-aware big data analytics in geo-distributed data centers with reinforcement learning," *IEEE Transactions on Network Science and Engineering,* vol. 7, no. 1, pp. 205-215, 2018.

[3] X. He, K. Wang, and W. Xu, "QoE-driven content-centric caching with deep reinforcement learning in edge-enabled IoT," *IEEE Computational Intelligence Magazine,* vol. 14, no. 4, pp. 12-20, 2019.

[4] G.-J. Cheng, L.-T. Liu, X.-J. Qiang, and Y. Liu, "Industry 4.0 development and application of intelligent manufacturing," in *2016 international conference on information system and artificial intelligence (ISAI)*, 2016: IEEE, pp. 407-410.

[5] I. You, J. Choi, C. Choi, and P. Kim, "Intelligent healthcare service based on context inference using smart device," *Soft Computing,* vol. 18, no. 12, pp. 2577-2586, 2014.

[6] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access,* vol. 3, pp. 678-708, 2015.

[7] J. Choi, C. Choi, H. Ko, and P. Kim, "Intelligent healthcare service using health lifelog analysis," *Journal of medical systems,* vol. 40, no. 8, pp. 1-10, 2016.

[8] O. Abuzaghleh, B. D. Barkana, and M. Faezipour, "Noninvasive Real-Time Automated Skin Lesion Analysis System for Melanoma Early Detection and Prevention," *IEEE Journal of Translational Engineering in Health & Medicine,* vol. 3, no. 4300212, pp. 1-12, 2015.

[9] M. Abdollahian and T. Das, "A MDP model for breast and ovarian cancer intervention strategies for BRCA1/2 mutation carriers," *IEEE J Biomed Health Inform,* vol. 19, no. 2, pp. 720-727, 2014.

[10] P. Annangi, S. Thiruvenkadam, A. Raja, H. Xu, X. Sun, and L. Mao, "A region based active contour method for x-ray lung segmentation using prior shape and low level features," in *2010 IEEE international symposium on biomedical imaging: from nano to macro*, 2010: IEEE, pp. 892-895.

[11] M. Abdel-Basset, G. Manogaran, A. Gamal, and V. Chang, "A novel intelligent medical decision support model based on soft computing and IoT," *IEEE Internet of Things Journal,* vol. 7, no. 5, pp. 4160-4170, 2019.

[12] L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. S. Goodenday, "Knowledge discovery approach to automated cardiac SPECT diagnosis," *Artificial intelligence in medicine,* vol. 23, no. 2, pp. 149-169, 2001.

[13] H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," *IEEE Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1153-1159, 2016.

[14] D. Erhan, P. A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training," *Immunology of Fungal Infections,* vol. 5, pp. 153-160, 2009.

[15] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans Med Imaging,* vol. 35, no. 5, pp. 1299-1312, 2016.

[16] B. Reed, S. Rhodes, P. Schofield, and K. Wylie, "Gender variance in the UK: Prevalence, incidence, growth and geographic distribution," *Retrieved June,* vol. 8, p. 2011, 2009.

[17] L. Mihalkova and R. J. Mooney, "Transfer Learning from Minimal Target Data by Mapping across Relational Domains," in *International Jont Conference on Artifical Intelligence*, 2009, pp. 1163-1168.

[18] L. Ogiela, "Transformative computing in advanced data analysis processes in the cloud," *Information Processing & Management,* vol. 57, no. 5, p. 102260, 2020.

[19] L. Ogiela, M. Takizawa, and U. Ogiela, "Transformative Computing for Distributed Services Management Protocols," in *International Conference on Advanced Information Networking and Applications*, 2020: Springer, pp. 470-475.

[20] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge & Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[21] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data,* vol. 3, no. 1, pp. 1-40, 2016.

[22] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 41, no. 4, pp. 801-814, 2018.

[23] W. Dai, Q. Yang, G. R. Xue, and Y. Yu, "Boosting for transfer learning," in *International Conference on Machine Learning*, 2007, pp. 193-200.

[24] B. Tan, Y. Song, E. Zhong, and Q. Yang, "Transitive Transfer Learning," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1155-1164.

[25] M. N. A. Khan and D. R. Heisterkamp, "Adapting instance weights for unsupervised domain adaptation using quadratic mutual information and subspace learning," in *International Conference on Pattern Recognition*, 2017, pp. 1560-1565.

[26] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks,* vol. 22, no. 2, pp. 199-210, 2010.

[27] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer Joint Matching for Unsupervised Domain Adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1410-1417.

[28] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer Feature Learning with Joint Distribution Adaptation," in *IEEE International Conference on Computer Vision*, 2014, pp. 2200-2207.

[29] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, "Balanced Distribution Adaptation for Transfer Learning," in *IEEE International Conference on Data Mining*, 2017, pp. 1129-1134.

[30] Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu, "Cross-People Mobile-Phone Based Activity Recognition," in *International Joint Conference on Artificial Intelligence*, 2011, pp. 2545-2550.

[31] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," vol. 2, pp. 677-682, 2008.

[32] J. Davis and P. Domingos, "Deep Transfer via Second-Order Markov Logic," in *International Conference on Machine Learning*, 2009, pp. 217-224.

[33] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Taylor & Francis, 2012, pp. 77-79.

[34] A. Acharya, E. R. Hruschka, J. Ghosh, and S. Acharyya, "An Optimization Framework for Combining Ensembles of Classifiers and Clusterers with Applications to Nontransductive Semisupervised Learning and Transfer Learning," *Acm Transactions on Knowledge Discovery from Data,* vol. 9, no. 1, pp. 1-35, 2014.

[35] C. Kandaswamy, L. M. Silva, L. A. Alexandre, and J. M. Santos, "Deep transfer learning ensemble for classification," in *International Work-Conference on Artificial Neural Networks*, 2015: Springer, pp. 335-348.

[36] H. S. Bhatt, S. Richa, V. Mayank, and N. K. Ratha, "Improving cross-resolution face matching using ensemble-based co-transfer learning," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society,* vol. 23, no. 12, pp. 5654-69, 2014.

[37] Y. Yang and J. Jiang, "Hybrid Sampling-Based Clustering Ensemble With Global and Local Constitutions," *IEEE Trans Neural Netw Learn Syst,* vol. 27, no. 5, pp. 952-965, 2017.

[38] Y. Yang and J. Jiang, "HMM-based hybrid meta-clustering ensemble for temporal data," *Knowledge-Based Systems,* vol. 56, no. C, pp. 299-310, 2014.

[39] Y. Yang and K. Chen, "Time Series Clustering Via RPCL Network Ensemble With Different Representations," *IEEE Transactions on Systems Man & Cybernetics Part C,* vol. 41, no. 2, pp. 190-199, 2011.

[40] Y. Yang and K. Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations," *IEEE Transactions on Knowledge and Data Engineering,* vol. 23, no. 2, pp. 307-320, 2010.

[41] Y. Yang, Z. Li, W. Wang, and D. Tao, "An adaptive semi-supervised clustering approach via multiple density-based information," *Neurocomputing,* vol. 257, pp. 193-205, 2017.

[42] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *International Conference on Neural Information Processing Systems*, 2006, pp. 601-608.

[43] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by Kernel Maximum Mean Discrepancy," *Bioinformatics,* vol. 22, no. 14, pp. 49-57, 2006.

[44] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3712-3722.

[45] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23-37.

[46] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, and K. H. Cha, "Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE transactions on medical imaging,* vol. 38, no. 3, pp. 686-696, 2018.

[47] F. J. Martinez-Murcia, A. Ortiz, J. Ramírez, J. M. Górriz, and R. Cruz, "Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy," *Neurocomputing,* 2020.

[48] G. Wang, G. Zhang, K.-S. Choi, K.-M. Lam, and J. Lu, "Output based transfer learning with least squares support vector machine and its application in bladder cancer prognosis," *Neurocomputing,* vol. 387, pp. 279-292, 2020.

[49] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint ℓ 2,1 -norms minimization," in *International Conference on Neural Information Processing Systems*, 2010, pp. 1813-1821.

[50] A. Krogh, and J. Vedelsby, "Neural Network Ensembles, Cross Validation, and Active Learning.," *International Conference on Neural Information Processing Systems MIT Press,* pp. 231-238, 1995.

[51] UCI Machine Learning Repository [Online] Available: http://archive.ics.uci.edu/ml/datasets.html

[52] Y. Xu *et al.*, "A Unified Framework for Metric Transfer Learning," *IEEE Transactions on Knowledge & Data Engineering,* vol. 29, no. 6, pp. 1158-1171, 2017.

[53] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Advances in neural information processing systems*, 2011, pp. 2456-2464.

[54] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems,* vol. 50, no. 1, pp. 1-18, 2000.

[55] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*: Springer, 2009, pp. 1-4.

[56] M. J. O'donnell *et al.*, "Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study," *The Lancet,* vol. 376, no. 9735, pp. 112-123, 2010.

[57] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, no. 1, pp. 321-357, 2002.