Chinese Journal of Chemical Engineering 59 (2023) 231–239

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Full Length Article

Minimax entropy-based co-training for fault diagnosis of blast furnace

Dali Gao [1], Chunjie Yang [1,*], Bo Yang [2], Yu Chen [3], Ruilong Deng [1]

[1] The State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
[2] Jianwei Digital Sphere Co., Ltd, Lianyungang 222113, China
[3] Jiangsu Binxin Iron and Steel Group Co., Ltd, Chongqing 401220, China

ARTICLE INFO

Article history:
Received 21 September 2022
Received in revised form 21 December 2022
Accepted 22 December 2022
Available online 05 January 2023

Keywords:
Co-training
Fault diagnosis
Blast furnace
Minimax entropy
Transfer learning

ABSTRACT

Due to the problems of few fault samples and large data fluctuations in the blast furnace (BF) ironmaking process, some transfer learning-based fault diagnosis methods are proposed. The vast majority of such methods perform distribution adaptation by reducing the distance between data distributions and applying a classifier to generate pseudo-labels for self-training. However, since the training data is dominated by labeled source domain data, such classifiers tend to be weak classifiers in the target domain. In addition, the features generated after domain adaptation are likely to be at the decision boundary, resulting in a loss of classification performance. Hence, we propose a novel method called minimax entropy-based co-training (MMEC) that adversarially optimizes a transferable fault diagnosis model for the BF. The structure of MMEC includes a dual-view feature extractor, followed by two classifiers that compute the feature's cosine similarity to representative vector of each class. Knowledge transfer is achieved by alternately increasing and decreasing the entropy of unlabeled target samples with the classifier and the feature extractor, respectively. Transfer BF fault diagnosis experiments show that our method improves accuracy by about 5% over state-of-the-art methods.

1. Introduction

Steel manufacturing is the pillar industry of the economic system. The blast furnace (BF) is the critical equipment for steel manufacturing, consuming more than 70% of the total energy consumption [1,2]. In blast furnace ironmaking process, iron ore (sinter, pellets) and coke are charged alternately from the top according to the preset proportion, and hot air (1100–1200 °C) is blown into the blast furnace from the bottom tuyere. The blast furnace performs the reduction reaction to produce molten iron, which discharges from the iron tap. After dust removal, the gas is released from the top of BF and enters the gas pipe network.

The blast furnace is the key equipment in the ironmaking process. Once a fault occurs, it may cause incalculable economic losses and even casualties. Therefore, it is vital to conduct timely and accurate fault diagnoses for the BF. However, due to the complicated physical and chemical processes occurring inside the BF and the harsh environments where conventional instruments do not work properly, it is challenging for devices to directly and precisely assess the conditions inside the BF. Which lead to the fault diagnosis of the BF particularly difficult.

The expert system-based method is currently the most widely used approach in BF diagnosis. However, the existing prior knowledge about BF is not comprehensive due to the exceptionally complex reactions in BF. In addition, part of the expert knowledge comes from the personal experience of operators acquired during long practice, whose universality cannot be guaranteed and is difficult to express accurately and clearly in mathematical formulas. Therefore, it is difficult for the expert system to achieve the expected performance.

With the advancement of information technology, data-driven fault diagnosis methods have been greatly developed. Compared with the fault diagnosis methods based on expert systems, it has the characteristics of high accuracy and easy application, the representative methods mainly include principal component analysis (PCA), independent principal component analysis (ICA), partial least square (PLS), support vector machine (SVM), artificial neural network (ANN), and related improvements to these methods [3–7].

It is important to note that the conventional data-driven approaches outlined above require the following two presuppositions: 1) A lot of data with labels. 2) Training and test data obey the same distribution. However, because of the following issues,

* Corresponding author.
E-mail address: cjyang999@zju.edu.cn (C. Yang).

https://doi.org/10.1016/j.cjche.2022.12.007
1004-9541/

it is difficult to satisfy these hypotheses in the actual BF fault diagnosis scenario:

(1) The fault data are too few to meet the training set requirements. At first, the operator regulates the blast furnace timely to avoid serious accidents [8]. In this case, fault data are not available. Second, the BF data fluctuates significantly when ore quality and working mode change [9]. Since working mode switching is generally not recorded, it is difficult to isolate fault data from fluctuating data. In addition, manual labeling is extremely laborious in the BF ironmaking process, which makes it more difficult to obtain labeled data.
(2) Since the BF data exhibit multi-mode and time-varying properties [10]. After a period of blast furnace operation, the probability distribution of the BF data may vary over time, which could lead to inconsistent distributions of the training and test data. A model trained on labeled data may not be efficient when trying to classify unlabeled data that had been collected over a different time period.

A new intelligent fault diagnosis method is urgently needed to address the above issues. With such a method, a model that trained on labeled data from one period can be applied to unlabeled data from another period.

Transfer learning (TL) can transfer knowledge learned in source domain to other different but related target domain [11]. It helps alleviate the problem of small samples by introducing samples from auxiliary domains. In addition, transfer learning can effectively handle data obeying different distributions and has shown great performance in the related recognition or classification tasks. In order to improve the accuracy of fault diagnosis, provide more reliable data support and regulation basis for blast furnace operators, we introduce transfer learning into BF fault diagnosis to effectively deal with the problems of data probability drift and lack of labeled fault samples during BF production.

The method that combines with deep learning is a focus of TL because of its efficient feature extraction capabilities. Such methods generally apply adaptation to marginal or conditional distributions or to both. In general, there are two main types of implementations: 1) Using techniques like maximum mean discrepancy (MMD), Kullback-Leibler divergence to measure distribution differences and employing the outputs as a loss function to train the model, such as the methods deep domain confusion (DDC) [12], balanced distribution adaptation (BDA) [13], and domain adaptation network based on coral loss (Deep Coral) [14]. 2) Extracting domain invariant features by fooling domain discriminators in adversarial transfer learning, such as domain adaptive neural network (DANN) [15] and dynamic adversarial adaptation network (DAAN) [16].

Currently, some scholars have applied deep transfer learning methods to industrial fault diagnosis [17,18]. Guo *et al.* [19] added a domain discriminator to DDC to reduce the marginal distribution discrepancy, the performance of the proposed method was verified in the fault diagnosis experiment. Zhang *et al.* [20] proposed a collaborative transfer fault diagnosis model for machinery fault diagnosis, experiments conducted on two decentralized fault diagnosis datasets demonstrate the effectiveness of its approach. Qian *et al.* [21] proposed an improved joint distribution adaptation to align both the marginal and conditional distributions. Fault diagnosis experiments on vibration signal datasets of roller bearings verify its effectiveness and applicability. Gao *et al.* [22] proposed a fault diagnosis method for BF which aligns the prior and joint distributions. Transfer BF fault diagnosis experiments demonstrate the effectiveness of the proposed method. Zhang *et al.* [23] proposed a new method which can automatically recognize the unknown fault modes by using additional outlier identifier.

We can see that existing transfer learning-based methods generally achieve knowledge transfer by aligning distributions. However, as shown in Ref. [24], the classifier may not perform effectively on the target domain even if the distributions are matched with the non-discriminative representations. In addition, due to the absence of target labels, many methods, such as BDA, DAAN, *etc*., have proposed utilizing a classifier's predictions as pseudo-labels of unlabeled samples, and retraining it, a process called self-training. It's based on the premise that one's own output with high confidence is correct, and only in such a case can the pseudo-labels will further improve the performance of the classifier. However, since the training data is dominated by labeled source domain data, such classifiers tend to be weak in the target domain. As a result, it is difficult to verify the accuracy of the pseudo-labels, which could lead to a degradation of performance. In addition, the features generated after domain adaptation are likely to be at the decision boundary, resulting in a loss of classification performance.

To improve performance of weak classifiers in the target domain caused by self-training in traditional transfer learning and achieve low-density decision boundaries, we propose minimax entropy-based co-training (MMEC) approach that adversarially optimizes a transferable fault diagnosis model for the BF. To learn discriminative class boundaries, minimax entropy is introduced to the model. Inspired by Ref. [25], the direction of a classifier's weight vector often represents the normalized characteristics of the relevant class, the weight vectors can be seen as representative vectors or centers of classes. Furthermore, the entropy on the unlabeled target samples shows discrepancy between the estimated center and the target features. To estimate domain-invariant centers, we shift the centers towards the target features by increasing the entropy of the unlabeled target samples. Then, decrease the entropy of the unlabeled samples with the feature extractor so that they are better clustered around the center.

As for the problem of weak classifiers leading to low accuracy in pseudo-labels, the performance of weak classifiers can be effectively enhanced by co-training [26]. In co-training, two classifiers are trained with different views. Only when the outputs from the two classifiers are the same and at least one classifier has high confidence, the input sample can be moved to the training dataset. In this way, the classification performance of weak classifiers can be effectively improved. In our setup, we can split our features into two mutually exclusive views so that co-training is effective. In addition, co-training also contributes to transfer knowledge, as explained in detail in Section 2.

The structure of MMEC includes a dual-view feature extractor, followed by two classifiers that compute the feature's cosine similarity to representative vector of each class. Knowledge transfer is achieved by alternately increasing and decreasing the entropy of unlabeled target samples with the classifier and the feature extractor, respectively. Transfer BF fault diagnosis experiments show that our method improves accuracy by about 5% over state-of-the-art methods.

The main insights and contributions of this paper are summarized as follows.

(1) To address the problem of fluctuating blast furnace data distribution and low fault data, we propose a novel transfer learning-based fault diagnosis method called MMEC.
(2) Compared to traditional transfer learning methods, we extend minimax entropy from few-shot learning to the unsupervised transfer learning. In the case of unlabeled target data, knowledge transfer and co-training ensure that the feature vectors are always clustering towards the category centers with high confidence to learn discriminative representations.

(3) Unlike the usual self-training approach, co-training is used to solve the problem of low accuracy of pseudo-labeling caused by weak classifiers.

The remainder of this paper is structured as follows. The transfer learning issue is discussed in Section 2. The proposed method is detailed in Section 3. The transfer fault diagnosis experiments utilizing actual industrial BF data are carried out in Section 4. Finally, conclusions are drawn in Section 5.

2. Transfer Learning Problems

Our method focuses on the transfer learning-based diagnosis method. This study is generally conducted using the following definition. The domain is denoted as D={x,$P(x)$,$Q(y|x)$}, which generally consists of three parts: data space, marginal distribution and conditional distribution of data. Given a source domain D_s and corresponding learning task T_s, a target domain D_t and learning task T_t. Transfer learning is to use the knowledge in the D_s and learning task T_s to improve the generalization performance of the model in the D_t and T_t, where $D_s \neq D_t$ or $T_s \neq T_t$.

We consider labeled BF samples as the D_s, and unlabeled BF data to be monitored as the target domain D_t. The source domain and target domain data from different sampling time periods can be regarded to obey different data distributions under the premise of large fluctuation of BF data, *i.e.*, $D_s \neq D_t$. In this paper, BF data have the same variables and the same types of fault conditions, and only differ significantly in terms of data distribution, so the research in this paper could be classified as a feature-based homogeneous unsupervised transfer learning problem.

The main goal of transfer learning is to learn the knowledge in the D_s to improve the generalization performance of the model in the D_t. Ben-David *et al.* [24] showed that the upper bound on the generalization error in the target domain depends on following three items: 1) classification error when training on D_s, 2) divergence between the D_s and D_t, 3) error of an ideal joint hypothesis. Let H be the hypothesis class, the generalization error in the target domain is defined as:

$$\forall h \in H, R_t(h) \leq R_s(h) + \frac{1}{2} d_{H\Delta H}(s,t) + \varepsilon \tag{1}$$

where R denotes the expected error for each hypothesis and the divergence between the source and target domains is defined as $H\Delta H$-distance:

$$d_{H\Delta H}(s,t) = 2 \sup_{(h,h') \in H^2} \left| \underset{x \sim s}{E} \left[h(x) \neq h'(x) \right] - \underset{x \sim t}{E} \left[h(x) \neq h'(x) \right] \right| \tag{2}$$

The ideal joint hypothesis is defined as $h^* = \arg\min_{h \in H} (R_s(h) + R_t(h))$ and its corresponding error ε is defined as:

$$\varepsilon = R_s(h^*) + R_t(h^*) \tag{3}$$

Because ε is usually considered to be very low based on fixed features and cannot be evaluated in the absence of labeled target samples, most methods attempt to decrease $R_s(h)$ and $d_{H\Delta H}(s,t)$. MMD or the loss of a domain classifier is generally utilized to measure the divergence between domains.

However, ε is very important when training CNNs, since CNNs extract and classify representations simultaneously. When the representations in the D_t are non-discriminative, ε can get quite large. The third term is therefore taken into consideration while concentrating on how to learn discriminative representations in the target domain.

3. Proposed Method

The complete training process of MMEC is summarized in Algorithm 1.

Algorithm 1. *Minimax entropy-based co-training* (*MMEC*)

Input: Source dataset $X_s = \{x_s, y_s\}_{i=1}^m$, unlabeled target dataset $X_{tu} = \{x_t\}_{i=1}^n$, labeled target dataset $X_{tl} = \varnothing$, labeled dataset $X_l = X_s \cup X_{tl}$
Output: Adaptive classifier
Begin:
  According to the loss function Eq. (9), train the F, C_1, C_2 with (x_s, y_s) until convergence
  Repeat:
    According to the Eq. (9), calculate the classification loss L with X_l
    According to the Eq. (10), calculate the entropy H with X_{tu}
    According to the loss function Eq. (11) and Eq. (12), update θ_F and θ_C through the stochastic gradient descent (SGD) algorithm
    Assign pseudo-labels $\widehat{y_t}$ to unlabeled target samples when the predictions of C_1 and C_2 are the same, and at least one of them has high confidence
    Remove unlabeled target sample x_t from X_{tu} and add $(x_t, \widehat{y_t})$ to X_{tl}
  Until: $X_{tu} = \varnothing$
End: Return an adaptive classifier

3.1. Minimax entropy-based co-training (MMEC)

As shown in Fig. 1, the structure of MMEC includes a dual-view feature extractor, followed by two classifiers that compute the features' cosine similarity to weight vectors of classifier. Due to the direction of a classifier's weight vector often represents the normalized characteristics of the relevant class. the weight vectors can be seen as representative vectors or centers of classes. Knowledge transfer is achieved by alternately increasing and decreasing the entropy of unlabeled target samples with the classifier and the feature extractor, respectively.

(1) Feature extractor: The feature extractor is constructed by a convolutional neural network (CNN) with 10 layers. The corresponding detailed parameters are shown in Table 1.

Considering the disadvantages that individual samples are more sensitive to noise and may contain insufficient valid information for fault diagnosis. We introduce a moving window to improve the accuracy of fault diagnosis by exploiting the inter-sample dependence. We take a matrix consisting of 35 BF observed variables at 35 consecutive moments as a BF sample. The convolution kernel extracts the response features

$$h_{i,j}^l = \mathrm{Relu}\left(W^l h_{i,j}^{l-1} + b^l \right) \tag{4}$$

where $h_{i,j}^l$ is the corresponding feature representation of $x_{i,j}$ in the lth layer. W^l and b^l are the weights and bias of the lth layer, Relu $(\cdot)$ denotes a nonlinear activation function.

We use the max-pooling layer to reduce the computational cost and increase the convolutional kernel receptive field to prevent overfitting:

$$p_m = \max(0, h_{i,j}) \quad i,j \in (m \times k, (m+1) \times k) \tag{5}$$

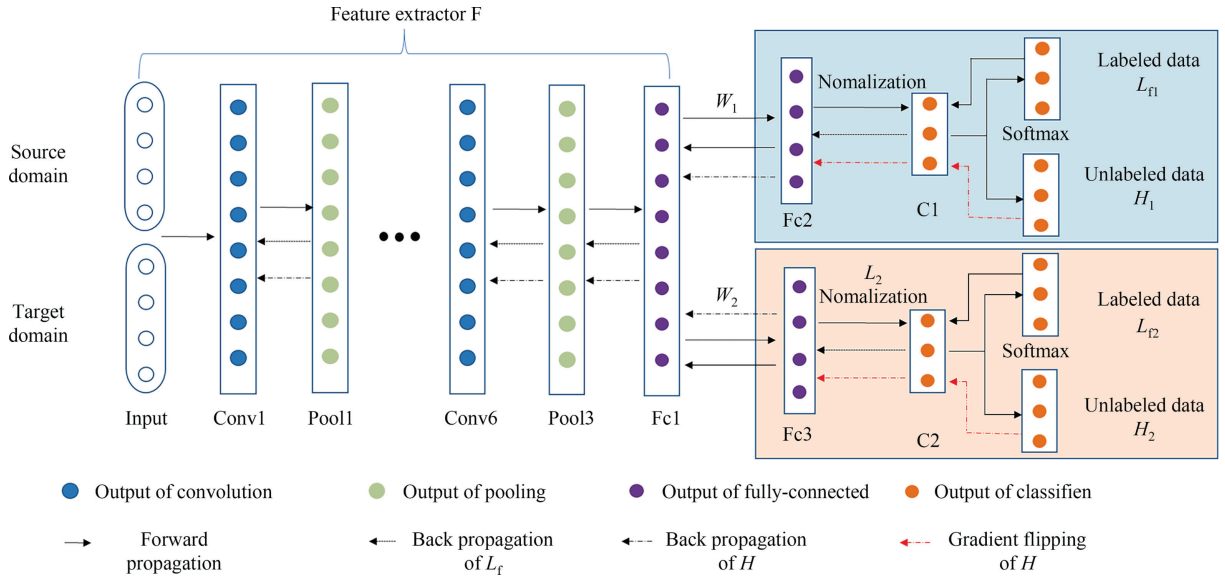

Fig. 1. Structure diagram of the MMEC.

Table 1
Parameters of feature extractor

Layer	Symbol	Operator	Kernel size/stride/padding/channel
1	Input	Input samples	35 × 35
2	Conv1	Convolution	11 × 11/2/2/64
3	Pool1	Max pooling	3 × 3/2
4	Conv2	Convolution	5 × 5/1/2/192
5	Pool2	Max pooling	3 × 3/2
6	Conv3	Convolution	3 × 3/1/1/384
7	Conv4	Convolution	3 × 3/1/1/256
8	Conv5	Convolution	3 × 3/1/1/64
9	Pool3	Max pooling	2 × 2/1
10	Fc1	Fully-connected	/

where k and p_m denote the size and the mth output of pooling layer, respectively.

After the convolution and pooling operation of the neural network on the BF samples, we use the vectors in Fc1 as the output features of the feature extractor, Fc1 is obtained by flattening the vectors in Pool3, and other fully connected layers are obtained by.

$$f = \varphi\left(\boldsymbol{W}_{\mathrm{f}}^{\mathrm{T}} I_{\mathrm{f}} + b_{\mathrm{f}}\right) \tag{6}$$

where I_{f} and f mean the input and output, respectively. $\boldsymbol{W}_{\mathrm{f}}$ is the weight matrix, and b_{f} is the bias.

(2) Classifiers based on Co-training: Since our goal is to label the target samples with high accuracy through co-training, we expect C1 and C2 to classify samples according to different feature spaces. Therefore, we constrain the weights of Fc2 and Fc3 to split features into two mutually exclusive views.

$$\sum_{i=1}^{d} w_{1_i}^2 w_{2_i}^2 = 0 \tag{7}$$

Where d means dimensions of weight vectors. In order to facilitate parameter optimization, we use a L1 norm $\left|\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_2\right|$ to make this constraint less strict.

Due to the direction of a classifier's weight vector often represents the normalized characteristics of the relevant class, we perform L2 normalization on the output of layer Fc2 and Fc3. For the classifier C1, which consist of weight vectors $\boldsymbol{W} = [w_1, w_2 \cdots w_k]$ where k represents the number of the furnace conditions of BF. C1 takes the normalized feature vector as input and outputs $\frac{\boldsymbol{W}^{\mathrm{T}} f_2}{\|f_2\|}$, where f_2 is the output of Fc2. Then, the softmax layer utilizes the output of C1 to get probabilistic output as

$$p(x) = \sigma\left(\frac{\boldsymbol{W}^{\mathrm{T}} f_2}{\|f_2\|}\right) \tag{8}$$

where σ denotes a softmax function. The working process of C2 is basically the same as that of C1, and will not be repeated here.

At last, we assign pseudo-labels $\widehat{y}_t$ to unlabeled target samples when the predictions of C_1 and C_2 are the same, and at least one of them has high confidence, in this paper, this confidence-threshold is set to 0.8.

3.2. Optimization objective

(1) Object 1: MMEC needs to correctly classify the labeled data under the condition that the input features of the two classifiers are different. Therefore, MMEC have to minimize the sum of classification loss on the labeled data and the L1 norm $\left|\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_2\right|$. The first objective function can be written as follows:

$$L = E_{(x,y)\in D_l} L_{\mathrm{ce}}(p(x), y) + \left|\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{W}_2\right| \tag{9}$$

where L_{ce} is the standard cross-entropy loss function, D_l is the labeled dataset.

(2) Object 2: Since the labeled dataset is dominated by source domain data, the classifier's weight vector will be near source distributions. To improving the generalization performance of the model in the target domain, minimax entropy is combined with transfer learning. To obtain the domain-invariant centers of classes, we increase the similarity between classifier's weight vectors $\boldsymbol{W}$ and unlabeled target features by maximizing entropy so that the weight vector $\boldsymbol{W}$ of the k-way linear classifier is shifted to the target distribution. To obtain the discriminative feature representation in the target domain, we should cluster the target features toward the estimated centers. We propose to generate the desired discriminative features by reducing the entropy of the feature extractor F on unlabeled target samples. Entropy is calculated as follows,

$$H = -E_{(x,y)\in D_u} \sum_{i=1}^{k} p(y=i|x) \lg p(y=i|x) \tag{10}$$

Combining above optimization objects, our approach can be understood as adversarial learning between classifiers and feature extractor. Classifiers and feature extractor are trained by alternately increasing and decreasing the entropy of unlabeled target samples respectively. During training, we set a gradient reversal layer to achieve the inversion of the gradient of the entropy loss on the unlabeled target samples. Both classifiers and feature extractor are trained to correctly classify the labeled samples. The overall optimization object is shown below

$$\hat{\theta}_{\mathrm{F}} = \arg\min_{\theta_{\mathrm{F}}} (L + \mu H) \tag{11}$$

$$\hat{\theta}_{\mathrm{C}} = \arg\min_{\theta_{\mathrm{C}}} (L - \mu H) \tag{12}$$

where θ_{F} and θ_{C} denote parameters of feature extractor F and weight vectors of classifiers respectively, $\mu \in [0,1]$ is a hyperparameter which leverages the importance between the minimax entropy and classification error.

3.3. Theoretical analysis

As shown in DAAN, domain divergence can be measured by domain classifier, Eq. (2) can be rewritten as,

$$d_{H\Delta H}(s,t) = 2 \sup_{h\in H} \left| \underset{x\sim s}{E} [h(f_s) = 1] - \underset{x\sim t}{E} [h(f_t) = 1] \right| \tag{13}$$

where f_s and f_t are obtained from the last fully connected layer of the deep model.

An upper bound on the error of the minmax entropy in semi-supervised transfer learning was demonstrated by Saito *et al.* [25]. While in this paper, we extend minimax entropy to unsupervised transfer learning and ensure its effectiveness by combining it with co-training. It is promising that the theoretical error upper bound of the minmax method is also applicable to unsupervised transfer learning with the guarantee of high confidence in the pseudo label. Unlike semi-supervised transfer learning, fault diagnosis is different from fields like image recognition, where the data to be monitored, *i.e.*, the target domain data is unlabeled. It means that we do not have labeled samples as the initial feature transfer direction, which causes difficulties in the application of the minmax entropy in BF fault diagnosis. However, co-training can effectively improve the weak classifier performance in the target domain, which makes the pseudo-labels in the process closer to the true labels. Therefore, unlike the self-training approach in traditional transfer learning, we combine the co-training method with minimax entropy to ensure that the feature vectors are clustered towards the correct category centers in the target domain with high confidence to realize the improvement of diagnosis performance.

In our case, take classifier C1 as an example, the features are outputs of Fc2. Although domain classifiers are not used in our model to obtain domain-invariant features, our approach can be seen as minimizing domain divergence by decreasing entropy on unlabeled target samples. We can distinguish the feature from the source or target domain based on the entropy value,

$$h(f) = \begin{cases} 1, & \text{if } H(C_1(f)) \gg \gamma \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where γ is a threshold of entropy to determine domain label. In our method, classifier C_1 outputs the probability of the furnace conditions. Eq. (2) can be rewritten as

$$\begin{aligned} d_{H\Delta H}(s,t) &= 2 \sup_{h\in H} \left| \underset{x\sim s}{E} [h(f_s) = 1] - \underset{x\sim t}{E} [h(f_t) = 1] \right| \\ &= 2 \sup_{C_1\in C} \left| \underset{x\sim s}{E} [H(C_1(f_s)) \geq \gamma] - \underset{x\sim t}{E} [H(C_1(f_t)) \geq \gamma] \right| \\ &\leq 2 \sup_{C_1\in C} \underset{x\sim t}{E} [H(C_1(f_t)) \geq \gamma] \end{aligned} \tag{15}$$

Since we have a lot of labeled source domain data, and train the whole network to minimize the classification loss, the entropy on the source sample is very small. Therefore, we assume that $\underset{x\sim t}{E} [H(C_1(f_t)) \geq \gamma] \geq \underset{x\sim s}{E} [H(C_1(f_s)) \geq \gamma]$. This inequality states that the domain divergence is determined by the proportion of target samples with entropy bigger than γ. An upper bound on the domain divergence can be obtained by training C to achieve the maximum entropy on the unlabeled target features. Furthermore, our goal is to find the discriminative features to achieve the low-density separation. then the objective can be rewritten as

$$\min_{f_t} \max_{C_1\in C} \underset{x\sim t}{E} [H(C_1(f_t)) \geq \gamma] \tag{16}$$

The working process of C2 is basically the same as that of C1, and will not perform repeated analysis. To summarize, the process of training the classifier to increase entropy can be viewed as measuring domain divergence, while the process of training the feature classifier to decrease entropy can be viewed as minimizing domain divergence.

In addition, under the condition that the feature is fixed, ε in Eq. (1) is ignored because it is also fixed. However, for deep transfer learning-based methods, features are generally determined by the fully connected layers of the neural network, which means that when the parameters of the neural network are changed, it is possible for ε to change to a large value, so we need to reconsider this term.

Due to the lack of labels in the target domain, ε cannot be directly calculated, so we use pseudo labels to evaluate it approximately. We consider the pseudo-labeled target dataset X_{tl} given false labels at a ratio ρ, the Eq. (1) can be re rewritten as

$$\forall h \in H, R_t(h) \leq R_s(h) + \frac{1}{2} d_{H\Delta H}(s,t) + \varepsilon \leq R_s(h) + \frac{1}{2} d_{H\Delta H}(s,t) + \varepsilon' + \rho$$

where ε' denotes the error of ideal joint hypothesis h^* on the X_{tl} and X_s. Different from traditional self-training methods, we use co-training to further reduce ρ.

On the other side, if we consider h and h' as C1 and C2 respectively, $\underset{x\sim s}{E} [h(x) \neq h'(x)]$ should be very low since training is based on the same labeled samples. For the same reason, $\underset{x\sim t}{E} [h(x) \neq h'(x)]$ is expected to be low even though we use the pseudo-labeled target dataset X_{tl}. Therefore, our method considers both the domain divergence $d_{H\Delta H}(s,t)$and the error ε of an ideal joint hypothesis in Eq. (1) to get more discriminative features in the target domain, which means that our model has better generalization performance in the target domain.

4. Experiment Results and Comparisons

The blast furnace production process operates under high temperature and pressure, which is a complex process with various physical and chemical reactions occurring all the time. Once a fault occurs, the economic loss and safety hazard will be very serious. Therefore, to ensure the safe and reliable operation of BF, timely fault diagnosis is indispensable.

4.1. Dataset

In this paper, experiments are conducted using actual BF production data. The fault samples include the two types of common BF fault, *i.e.*, channeling and hanging.

(1) Channeling: Channeling is the phenomenon of excessive development of airflow in a local area within the blast furnace. Before the channeling occurs, there is a gradual imbalance in air pressure and air volume, the air pressure decreases, the permeability index increases relatively, and the air volume has a tendency to increase automatically and fluctuate beyond normal levels.
(2) Hanging: When the charge stops falling and lasts for more than two batches of charge, it is known as hanging. Before hanging occurs, air pressure continues to rise, the air volume automatically drops, the top pressure and permeability index decreases, *etc.* are all signs of hanging occurring in the blast furnace.

We used the actual BF production data from October and November 2017 from a steel plant as the experimental datasets. The observed variables of BF data at each moment are shown in Table 2, and the observation frequency of BF data is every 10 seconds. We take a matrix consisting of observed variables at 35 consecutive moments, *i.e.*, a 35 × 35 matrix as a blast furnace sample. We denote the production data sampled from October and November as dataset Oct and dataset Nov, respectively.

The data in dataset Oct and Nov are from three furnace conditions: normal, hanging, and channeling. The total number of samples in Oct is 2435, and the corresponding samples belonging to the above three types of furnace conditions are: 1391, 412, and 632, respectively. The total number of samples in Nov is 2329, and the corresponding sample numbers of the three furnace conditions are: 1288, 923, and 118. As mentioned in Section 2, under the premise of large fluctuation of BF data, we consider the datasets Oct and Nov sampled from different time periods as drawn from different data distributions.

Table 2
The observed variables of BF data

No.	Variable	No.	Variable
1	Oxygen enrichment rate/%	19	Hot blast pressure/MPa
2	Blast furnace permeability index	20	Actual blast velocity/m·s^{-1}
3	CO volume/%	21	Hot blast temperature (1)/°C
4	H_2 volume/%	22	Hot blast temperature (2)/°C
5	CO_2 volume/%	23	Top temperature (Southeast)/°C
6	Blast velocity at tuyere of blast furnace/m·s^{-1}	24	Top temperature (Northeast)/°C
7	Enriching oxygen flow/m^3·h^{-1}	25	Top temperature (Northwest)/°C
8	Cold blast flow/m^3·h^{-1}	26	Top temperature (Southwest)/°C
9	Blast momentum/kJ	27	Coal injection set value/t·h^{-1}
10	BF bosh gas volume/m^3	28	Actual coal injection rate/t·h^{-1}
11	BF bosh gas index	29	Actual coal injection in last hour/t
12	Theoretical combustion temperature/°C	30	Enriching oxygen pressure/MPa
13	BF top pressure (Southeast)/°C	31	Total pressure drop/kPa
14	BF top pressure (Northeast)/°C	32	Drag coefficient
15	BF top pressure (Northwest)/°C	33	Cold blast temperature (1)/°C
16	BF top pressure (Southwest)/°C	34	Cold blast temperature (2)/°C
17	Cold blast pressure (1)/MPa	35	Blast humidity/g·m^{-1}
18	Cold blast pressure (2)/MPa		

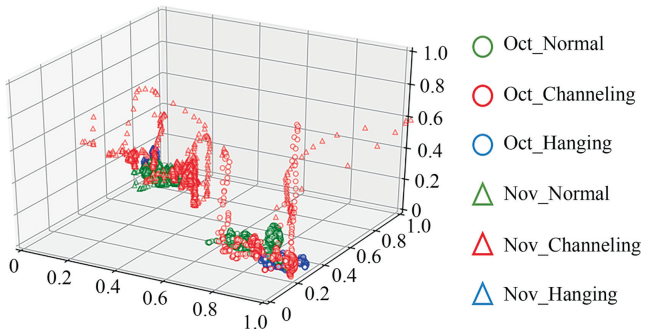

Fig. 2. t-SNE visualization of initial datasets.

As shown in Fig. 2, we use the t-distributed stochastic neighbor embedding (t-SNE) [27] technique to visualize the difference between the initial data distribution of the dataset Oct and Nov.

4.2. Results of Experiments

As shown in Table 3, we evaluated the performance of methods in the transfer fault diagnosis experiments. The datasets before and after the arrows represent the source and target domains, respectively. We divide the unlabeled target sample into two halves, with half of the unlabeled target samples and the labeled source samples forming the training set for the model to update its parameters, and the remaining target samples serving as the test set to evaluate the model performance.

The detailed parameters of the MMEC are set as follows. In the feature extractor, the parameters of the CNN are listed in Table 1. To take advantage of the powerful feature extraction capability of deep neural networks, we removed the last linear layer to build a feature extractor F and set up two K-way linear classification C1 and C2 with random initialized weight matrices. At each iteration, the input consists of a batch of labeled BF samples and a batch of unlabeled BF samples to be monitored. According to Eqs. (11) and (12), we update parameters θ_F and θ_C for adversarial learning. We set a gradient reversal layer to achieve the inversion of the gradient of the entropy loss on the unlabeled target samples between the classifiers and the feature extractor during backpropagation. The momentum of stochastic gradient descent (SGD) is set to 0.9. The learning rate of SGD is calculated by the formula, $0.02/(1 + 10 \times p)^{0.75}$ [17,18]. In all experiments, we set the trade-off parameter μ to 0.6. This is determined by the verification performance of the experiment, and the result is showed in Fig. 3. It can be seen that the performance of the model is sensitivity to μ.

The parameters of the other methods were optimally searched with the grid search method. After ten repetitions of each experiment, the accuracy of the proposed MMEC method exceed 92% in all fault diagnosis experiments, which verifies the effectiveness of the proposed method.

4.3. Analysis of experimental results

In the transfer fault diagnostic experiments, six approaches are compared with the proposed MMEC to demonstrate that it is effective in diagnosing BF faults. The six methods used for comparison are CNN trained with source samples, DDC, Deep Coral, DANN, BDA, and DAAN.

Average classification accuracies and corresponding standard deviations of ten repetitions of each experiment are shown in

Table 3
Classificationn resulsts of methods

Method	Testing accuracy ± STD/%		Average accuracy
	Oct → Nov	Nov → Oct	
CNN	0.422 ± 0.037	0.472 ± 0.024	0.447
DDC [12]	0.517 ± 0.064	0.583 ± 0.088	0.550
Deep Coral [14]	0.603 ± 0.036	0.619 ± 0.044	0.611
DANN [15]	0.732 ± 0.032	0.745 ± 0.014	0.739
BDA [13]	0.806 ± 0.013	0.728 ± 0.025	0.767
DAAN [16]	0.874 ± 0.035	0.883 ± 0.028	0.879
Self-training + co-training	0.529 ± 0.016	0.635 ± 0.023	0.582
Minmax entropy + softmax	0.788 ± 0.047	0.842 ± 0.068	0.815
MMEC(ours)	0.931 ± 0.079	0.926 ± 0.027	0.926

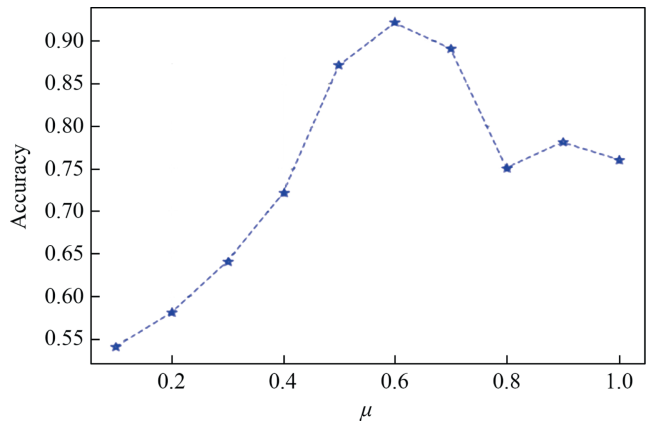

Fig. 3. Fault diagnosis accuracy with different value of μ.

Table 3. The experimental results show that the accuracy of the proposed MMEC is significantly higher than other methods in all BF fault diagnosis experiments. Furthermore, we are able to make the following observations by comparing the results.

For the transfer fault diagnosis task of unlabeled data in the target domain, the methods based on transfer learning significantly outperform the classic method without transfer learning such as CNN. This shows that transfer learning has obvious advantages for dealing with unsupervised problems with different data distributions.

Compared with DDC, Deep Coral and BDA, MMEC is different in that it adopts an adversarial learning approach for domain adaptation instead of using the MMD, *etc.* distance formula to measure domain differences.

Compared with DANN and DAAN, DANN adapts the marginal distribution using a domain discriminator, and DAAN has extra domain discriminators for each type of working condition that also adjust the conditional distribution. The difference of MMEC is that it adopts minmax entropy for adversarial learning instead of the domain classifier.

To illustrate the effectiveness of the proposed method, we conducted ablation experiments to compare methods including: 1) Retaining co-training and replacing only the minmax entropy-based training method with the self-training method. 2) Retaining the minmax entropy-based training method and replacing co-training with a single softmax classifier. From the experimental results, it can be seen that co-training can achieve knowledge transfer compared with the traditional CNN, showing a certain improvement effect. But there is still a large gap compared with the MMEC model, which indicates that the minmax entropy method can effectively achieve knowledge transfer. And after replacing the co-training in MMEC with softmax, the fault diagnosis accuracy of the model is reduced.

This strongly supports that co-training can effectively improve the weak classifier performance in the target domain, which ensure that the feature vectors are clustered towards the correct category centers in the target domain with high confidence to realize the improvement of diagnosis performance.

In all experimental results, MMEC achieved better results which indicates that our adversarial learning strategy is an effective technique to improve performance in blast furnace fault diagnosis.

To illustrate that the features extracted by MMEC are discriminative in the target domain. First, we analyze the eigenvalues of the feature covariance matrix on the target domain obtained in the output of the Fc1. The eigenvectors and the eigenvalues represent the components of features, and the corresponding amount of information they contain, *i.e.*, contributions, respectively. In general, the more discriminative the features are, the fewer feature components are required. Therefore, in this case, the first few eigenvalues that contain a lot of information should be large, and the eigenvalues that contain less information will show a rapid decay. In our method, the features are concisely presented with fewer components, as seen in Fig. 4(a). Second, the entropy on the unlabeled target samples is shown in Fig. 4(b). Although Deep Coral reduces entropy more quickly than our method, it performs poorly in the diagnosis. This suggests that the Deep Coral method incorrectly raises the prediction's confidence while our method simultaneously achieves higher accuracy.

Furthermore, in experiment Oct → Nov, we plot the features extracted by various methods using t-SNE in Fig. 5. Fig. 5(a)–(f) visualize the features in the target and source domains, where

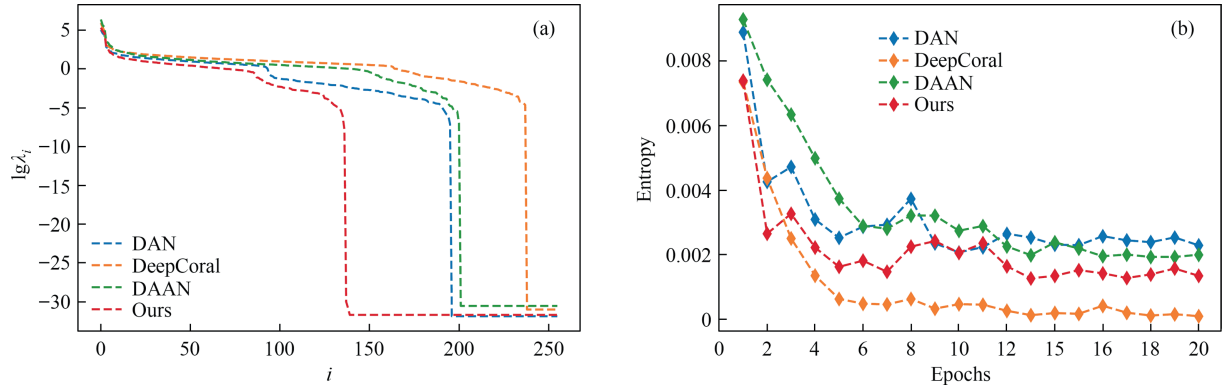

Fig. 4. (a) Eigenvalues of the feature covariance matrix on the target domain (in our approach, eigenvalues decrease rapidly, demonstrating that features are more discriminative than other approaches. (b) Our method achieves lower entropy values, although the model has a process of maximizing entropy).

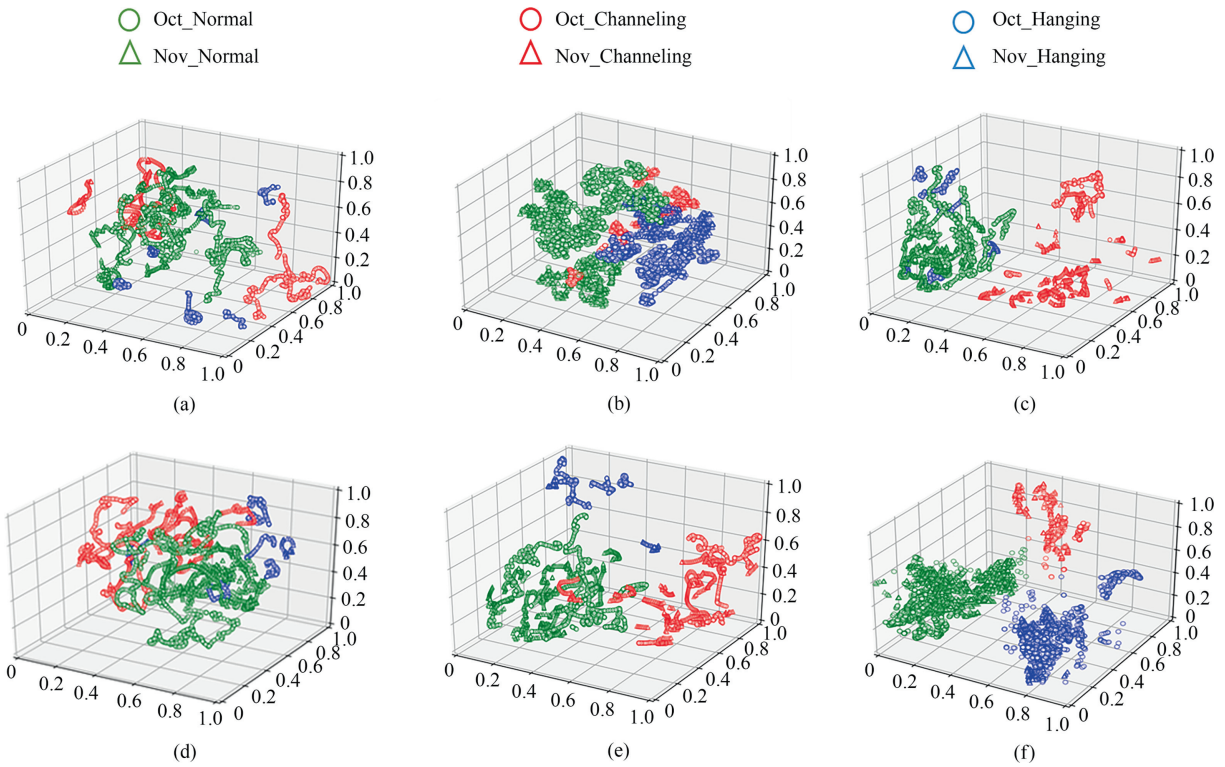

Fig. 5. t-SNE visualization of learned features: (a) CNN, (b) Deep Coral, (c) DANN, (d) BDA, (e) DAAN, (f) the proposed method.

the colors represent furnace conditions and the shapes represent domains. Using our method, the data belonging to same furnace condition from different domains are well aligned, and the clustering effect is significantly better than other methods. In summary, the experimental results prove the superiority of our method to minimize domain divergence and obtain discriminative features, enabling low-density separation.

5. Conclusions

This paper introduces minmax entropy into the unsupervised transfer learning and proposes a novel transfer learning-based fault diagnosis method for BF called MMEC. The structure of MMEC includes a dual-view feature extractor, followed by two classifiers that compute the features' cosine similarity to representative vectors of each class. Knowledge transfer is achieved by alternately increasing and decreasing the entropy of unlabeled target samples with the classifiers and the feature extractor, respectively. Transfer BF fault diagnosis experiments illustrated the effectiveness of our method.

MMEC can achieve high accuracy fault diagnosis of BF data under the prerequisite that the categories of furnace conditions in the source and target domain are the same. However, it will be misclassified if a brand-new kind of furnace condition that isn't covered by the source domain shows up in the target domain. How to successfully recognize the unknown faults is the focus of the research direction of transfer learning-based fault diagnosis methods.

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.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61933015), in part by the Central University Basic Research Fund of China under Grant K20200002 (for NGICS Platform, Zhejiang University).

References

[1] P. Zhou, D.W. Guo, H. Wang, T.Y. Chai, Data-driven robust M-LS-SVR-based NARX modeling for estimation and control of molten iron quality indices in blast furnace ironmaking, *IEEE Trans. Neural Netw. Learn. Syst.* 29 (9) (2018) 4007–4021.
[2] L.M. Zhang, C.C. Hua, J.P. Li, X.P. Guan, Operation status prediction based on top gas system analysis for blast furnace, *IEEE Trans. Control Syst. Technol.* 25 (1) (2017) 262–269.
[3] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, H. Ohno, Combined multivariate statistical process control, *IFAC Proc.* 37 (1) (2004) 281–286.
[4] J. Min, C.Y. Lee, S.W. Choi, P.A. Vanrolleghem, I.-B. Lee, Nonlinear process monitoring using kernel principal component analysis, *Chem. Eng. Sci.* 59 (1) (2004) 223–234.
[5] L.M. Liu, A.N. Wang, M. Sha, Feng-yun Zhao, Multi-class classification methods of cost-conscious LS-SVM for fault diagnosis of blast furnace, *J. Iron Steel Res. Int.* 18 (10) (2011) 17–33.
[6] L.M. Liu, A.N. Wang, M. Sha, X.Y. Sun, Y.L. Li, Optional SVM for fault diagnosis of blast furnace with imbalanced data, *ISIJ Int.* 51 (9) (2011) 1474–1479.
[7] Z.Q. Ge, Z.H. Song, F.R. Gao, Review of recent research on data-based process monitoring, *Ind. Eng. Chem. Res.* 52 (10) (2013) 3543–3562.
[8] P. Zhou, C.Y. Wang, M.J. Li, H. Wang, Y.J. Wu, T.Y. Chai, Modeling error PDF optimization based wavelet neural network modeling of dynamic system and its application in blast furnace ironmaking, *Neurocomputing* 285 (2018) 167–175.
[9] H. Zhou, C.J. Yang, Y.X. Sun, A collaborative optimization strategy for energy reduction in ironmaking digital twin, *IEEE Access* 8177570–177579.
[10] R.Q. An, C.J. Yang, Y.J. Pan, Graph-based method for fault detection in the ironmaking process, *IEEE Access* 840171–40179.
[11] L. Shao, F. Zhu, X.L. Li, Transfer learning for visual categorization: A survey, *IEEE Trans. Neural Netw. Learn. Syst.* 26 (5) (2015) 1019–1034.
[12] M. Long, J.M. Wang, Learning transferable features with deep adaptation networks, In: International Conference on Machine Learning, Lile, France, 97–105 (2015).

[13] J.D. Wang, Y.Q. Chen, S.J. Hao, W.J. Feng, Z.Q. Shen, Balanced distribution adaptation for transfer learning, In: 2017 IEEE International Conference on Data Mining, New Orleans, LA, USA. IEEE, 1129–1134.
[14] B.C. Sun, K. Saenko, Deep CORAL: Correlation alignment for deep domain adaptation, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 433–450 (2016).
[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, *Domain-Adversarial Training of Neural Networks* vol. 17 (2017) 189–209.
[16] C.H. Yu, J.D. Wang, Y.Q. Chen, M.Y. Huang, Transfer learning with dynamic adversarial adaptation network, In: 2019 IEEE International Conference on Data Mining. Beijing, China. IEEE, 778–786.
[17] W.P. Wang, Z.R. Wang, Z.F. Zhou, H.X. Deng, W.L. Zhao, C.Y. Wang, Y.Z. Guo, Anomaly detection of industrial control systems based on transfer learning, *Tsinghua Sci. Technol.* 26 (6) (2021) 821–832.
[18] L. Wen, L. Gao, X. Li, A new deep transfer learning based on sparse auto-encoder for fault diagnosis, *IEEE Transactions on Systems, Man, and Cybernetics. Systems,* 49(1) (2019) 136–144.
[19] L. Guo, Y.G. Lei, S.B. Xing, T. Yan, N.P. Li, Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data, *IEEE Trans. Ind. Electron.* 66 (9) (2019) 7316–7325.
[20] W. Zhang, Z.W. Wang, X. Li, Blockchain-based decentralized federated transfer learning methodology for collaborative machinery fault diagnosis, *Reliab. Eng. Syst. Saf.* 229 (2023) 108885.
[21] W.W. Qian, S.M. Li, P.X. Yi, K.C. Zhang, A novel transfer learning method for robust fault diagnosis of rotating machines under variable working conditions, *Measurement* 138 (2019) 514–525.
[22] D.L. Gao, X.Z. Zhu, C.J. Yang, X.K. Huang, W.H. Wang, Deep weighted joint distribution adaption network for fault diagnosis of blast furnace ironmaking process, *Comput. Chem. Eng.* 162 (2022) 107797.
[23] W. Zhang, X. Li, H. Ma, Z. Luo, X. Li, Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning, *IEEE Trans. Ind. Inform.* 17 (12) (2021) 7957–7967.
[24] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, *Mach. Learn.* 79 (1–2) (2010) 151–175.
[25] K. Saito, D. Kim, S. Sclaroff, T. Darrell, K. Saenko, Semi-supervised domain adaptation *via* minimax entropy, In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), IEEE, 8049–8057 (2019).
[26] T. Jafar, M. van Someren, H. Afsarmanesh, Ensemble based co-training, In: 23rd Benelux Conference on Artificial Intelligence, 223–231 (2011).
[27] L. van der Maaten, Accelerating t-SNE using tree-based algorithms, *J. Mach. Learn. Res.* 15 (2014) 3221–3245.