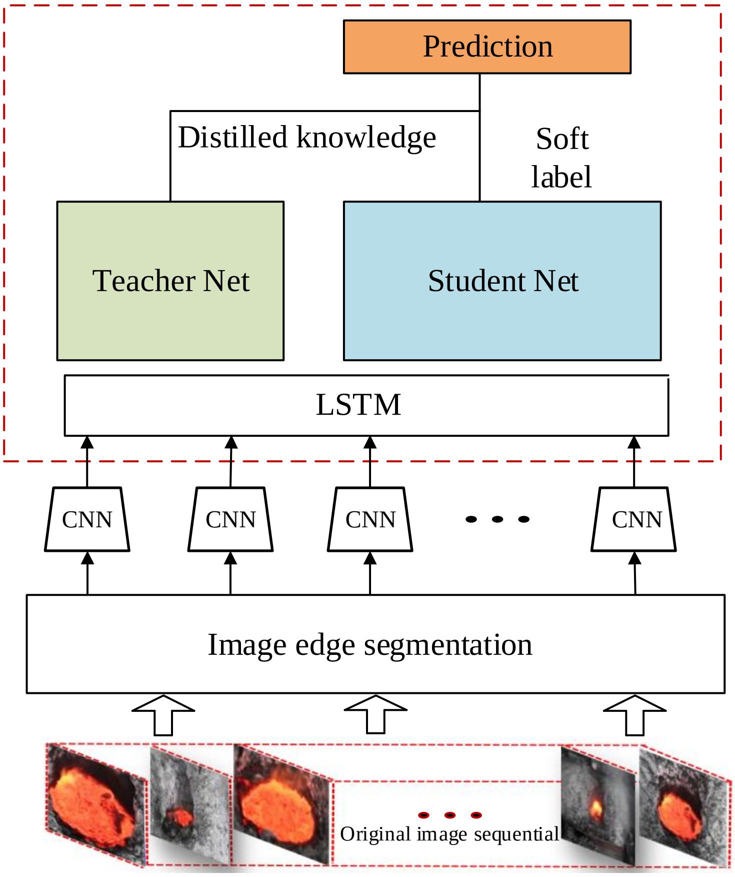
1.

Graphical Abstract

A self-supervised temporal temperature prediction method based on dilated contrastive learning

Yongxiang Lei,Xiaofang Chen,Yongfang Xie,Lihui Cen



Unlabeled sample from process industry

Highlights

A self-supervised temporal temperature prediction method based on dilated contrastive learning

Yongxiang Lei,Xiaofang Chen,Yongfang Xie,Lihui Cen

- A novel self-supervised architecture for temperature identification in the aluminum electrolysis process is proposed in this article. The model can achieve great accuracy with the constraint of the scarcity of labeled data.
- A new self-supervised loss is used in the proposed architecture. This loss can fully utilize the information contained in the unlabelled data, and cucumber the limit of the label data.
- A performance strategy for contrastive learning is used in the training of the model. The dual net of a student and teacher net is used to dilate the knowledge of the cosine loss.

A self-supervised temporal temperature prediction method based on dilated contrastive learning

Yongxiang Lei[a,c,1], Xiaofang Chen[a,b,*], Yongfang Xie[a,2] and Lihui Cen[a,b]

[a] School of Automation, Central South University, 410083, Changsha, China
[b] Research Center of Industry Intelligence, the Peng Cheng Laboratory, Shenzhen, China
[c] Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry, UK

ARTICLE INFO

Keywords:
Constrative self-supervised learning (CSSL)
Aluminum electrolysis industry
Long-short time memory (LSTM)
Temperature prediction
Dilated contrastive learning

ABSTRACT

Due to the scarcity of the labeled data, traditional supervised learning methods have a limited application scope, which caused the supervised-based model performance will greatly be decreased. In this paper, we propose a promising model based on self-supervised learning. To update the weight and the contrastive relation in the features, a new self-supervised loss, is introduced. First, the convolution neural network is used in the proposed network to extract the deep feature in the first processing. Second, the self-supervised long-short time memory (LSTM) sequential is constructed for further deal. At last, the teacher net and student net have coordinately fine-tuned the credibility of the temperature prediction. By the experimental comparison, our proposed CNN-SSDLSTM is competitive with other supervised and semi-supervised methods. The evaluation experiments achieve state-of-the-art performance in aluminum electrolysis temperature prediction applications.

1. Introduction

With the advent of artificial intelligence and information technologies, the productivity of the modern industrial system increases continuously [9, 10]. In aluminum electrolysis production, superheat degree is an important index for the indicator and monitoring of the whole production situation, lengthening the lifespan of the reduction cell, and improving the current efficiency [15, 16, 17, 38]. However, cell temperature detection suffers many limitations such as the hard environment, expensive device cost, and large delay. In the aluminum reduction reaction, some catalysts can be used to decrease the crystal temperature of the electrolysis so the current efficiency has also been improved. Recently, some data-driven methods have been proposed to detect the temperature situation [1]-[7]. The main strategy can be summarized as the following three categories: 1) the model-centric model, also the 'white' model, an apparent index is that the features and parameters can be visualized in the learning process. The model-based method needs to know the accurate dynamic mathematical model of the specific plant. Some works have also been reported about this field, such as in [8]. However, to sum up, those models' performance has largely relied on the real-time mathematical model and its generation ability is limited. Furthermore, in some industry scenarios, some mathematical model is hard to obtain, then the generation ability of these models is greatly limited. 2) the pair-wise data-centric intelligent model, also called the 'black' which uses the deep architecture to project the input and the output between the learning samples, some methods have been proposed in the literature, such as in [20]-[23].

Recently, deep learning and intelligent algorithms have played an important role in the industrial fields. With the abundant training data, the deep architecture can learn a good nonlinear projection of the data distribution, thus a remarkable performance is achieved. Some classification and regession tasks have been fully implemented in the related applications, such as [9]-[22]. However, because existing models are primarily based on the supervised learning method, their performance will suffer greatly when limited to labeled data. Although some semi-supervised learning methods have been proposed to solve this issue, the semi-supervised model relies on a few samples, if the few labeled samples have bad quality, the model has low robustness and performance. Furthermore, in the deep architecture, gradient descent and gradient explosion are substantial in training the model, resulting in over-fitting.

In the recent several years, self-supervised learning has been the hottest topic in the sub-domains of machine learning [18, 30, 36, 39]. In SSL, a learning machine captures the dependencies between input variables, some of which may be observed, denoted $x_1, ..., x_N$, and others not always observed, denoted $\{x_{N+1}, ..., x_{N+s}\}$. SSL pre-training has revolutionized natural language processing and is making rapid progress in speech and image recognition [13, 14, 28, 29]. SSL may enable machines to learn predictive models of the world through observation, and to learn representations of the perceptual world, thereby reducing the number of labeled samples or rewarded trials to learn a downstream task [30]. The core principle that resulted in the development of self-supervised learning is that under the constraint and limitation of the labeled data, how can we build some models autonomously to learn its feature and knowledge? The self-supervised learning is mainly classified into three categories: the basic self-supervised model, the contrastive model, and the learning model. The advantage of the proposed model individually learns its distortion, rotation, and angle iteratively. The large evolution of self-supervised learning will

[⋆] This document is the results of the research project funded by the National Science Foundation under grant No.62133016 and the China Scholarship of Council under grant No.202006370101.

[*] Corresponding author

leiyongxiang1205@gmail.com (Y. Lei); yfxie@csu.edu.cn (Y. Xie); lhcen@csu.edu.cn (L. Cen)

ORCID(s): 0000-0001-6375-8474 (Y. Lei)

[1] All the contributing authors belong to the School of Automation, Central South University,410083, Changsha, China.

be taken place in the world and so many experts and research apartments are working on this topic and many edge works have been done. For example, self-supervised learning extracts its original and naive manifold features rather than the labeled data intervention. Many methods have shown great accuracy in the computer vision domains, such as Jiasw [25], which is a picture distortion by using contrastive learning. Another self-supervised architecture is called the Moco, and some updated versions are called Moco-v2, also SimLR [8]. Some large-scale architectures with high performance have also been investigated in the front edge, such as transformer and attention mechanism. However, in the industrial aluminum electrolysis domain, self-supervised learning has shown an infant period, a few of the literature and works have been reported.

Some studies have been in the video domains based on self-supervised learning and semi-supervised learning. For example, Xu *et al.* proposes a self-supervised learning method for the space-time video data and applied it to style transfer [34]. For the time sequence applications, there are also some self-supervised learning methods that emerged, such as in [36]. In the literature, the unlabelled datasets are learned with self-supervised architectures and a competitive performance has been achieved. Only those data with highly confident labels are combined with original labeled data to train a new model [35]. Developed from [36], a CNN-LSTM model is given for the surgical phase recognition, their method can reduce the burden of manual annotation with the samples [19]. However, in the practical industrial process, it also needs to learn complex knowledge and improved the generation ability and accuracy. Chai *et al.* improves oblique random forests with dual-incremental learning capacity, it provides learning ability and can update effectively without laborious retraining from scratch [4]. Another customized soft sensor learning is given in [12], which utilizes a dual attention-based encoder-decoder to learn the sequential information between the different input variables and quality output variables, its efficacy is verified by a real cigarette production benchmark. Chang *et al.* also proposes a consistent-contrastive network with temporality awareness to obtain robust performance in industrial soft sensor application [5].

The current proposed state-of-the-art semi-supervised and self-supervised model on the following the self-training paradigm, where it first trains a Student Net model on the LSTM-layer representation and use it as a teacher to generate soft labels (also pseudo label) on 300M original unlabeled flame hole images in aluminum electrolysis [11, 32]. Some recent semi-supervised models that have been applied to aluminum electrolysis are shown in [31], [10]. Compared to semi-supervised learning, self-supervised learning remove the reliability of the labeled data [40]. The similarity of the soft label and permutation are jointly trained with the dual Teacher Net and Student Net [37]. Knowledge distillation as a student model based on labeled and soft labeled images [33]-[35]. We iterate this process by putting back the student as the teacher. During the pseudo labels generation, the teacher is not noised so that the soft labels are as accurate as possible. However, during the student's learning, we inject noise such as dropout, stochastic depth, and data augmentation via a random augmentation to the student to generalize better than the teacher network. The contrastive loss functions are developed to maximize the bound of the soft label and target output.

Based on the above analysis, we summarize the following challenges in the SD identification tasks.

(1) The aluminum electrolysis reduction lacks enough high-quality labels to train the network;
(2) The traditional artificial detection has large objectivity, the large delay, high cost, and inaccuracy exist;
(3) The dynamic relationship of the production variables is time sequential changing in the whole process;

Lei *et al.* presents a novel self-supervised learning method for aluminum temperature prediction, it implements an online framework by the serial LSTM. However, its architecture cannot extract the full features between the time-sequential images dataset. Developed from [24], we propose a novel self-contrastive method to detect the SD in aluminum electrolysis. First, the convolutional neural network is developed in the model to extract the image features as the initial feature representation of the flame hole video, further, an LSTM encoder architecture is directly linked to the top layer. In the top architecture, contrastive learning with a dual teacher net and student net are coordinated to improve the performance at a whole level. To sum up, the following contributions are given in this paper.

(1) A novel deep self-supervised architecture with contrastive learning is proposed with the unlabeled data, the approach for SD identification without requiring tedious manual supervision;
(2) A novel self-supervised loss is constructed, and the knowledge dilation networks are added in the top layers;
(3) Under the labeled data constraint, a robust SD identification method is proposed. The generations and robustness are also evaluated in the paper.

To verify the effectiveness, our proposed method is evaluated by an industrial electrolysis process application, some other comparative methods, which include the supervised methods and semi-supervised methods are fulfilled in the experiment. Compared to the existing competitive algorithms, the proposed method can achieve state-of-the-art performance.

The remainder of the paper is organized as follows. Sec 2 concludes some basic methods, Sec 3 gives the developing process of the proposed method. The experiment and Conclusion are given in Sec 4 and Sec 5, respectively.

2. Fundamentals

In this section, the basic modules are given, which include the aluminum production process, the CNN and LSTM unit, and the self-supervised learning method.

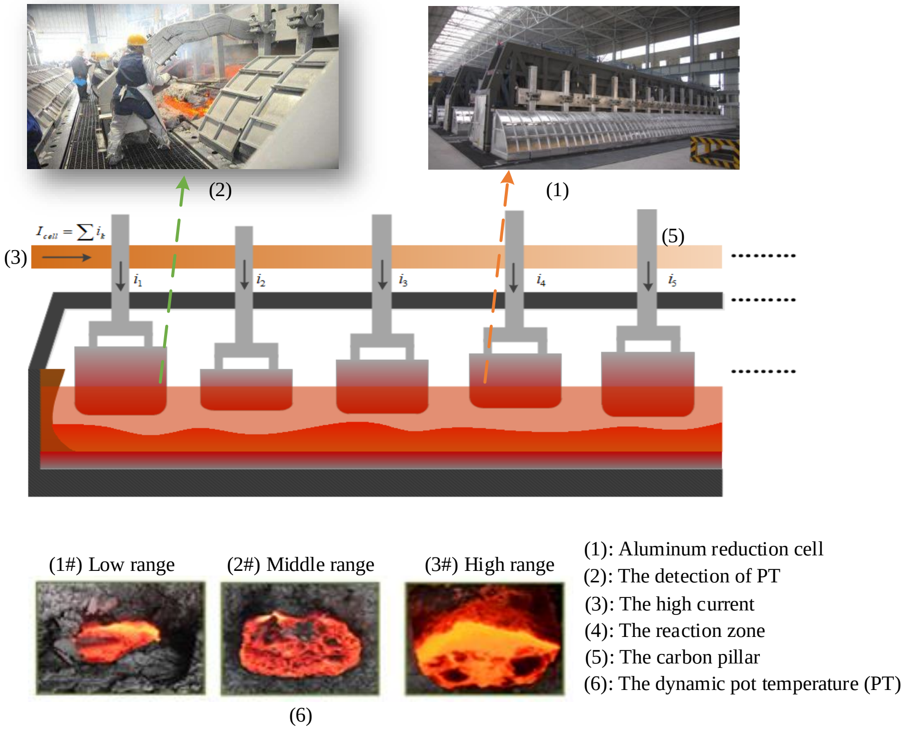


Figure 1: The flowchart of the aluminum production unit, covers a lot of quality variables and dynamic temperature range.

2.1. Aluminum production processes

The cell temperature is the crystal electrolyte of alumina powder. The prediction of temperature is significant because it can monitor the whole production, and further support the identification of superheating and control. Therefore, it maintains the stability of production. The temperature index is a critical index in the production of the aluminum electrolysis process. In a reduction cell, the alumina reacts with the carbon dioxide with a high current as the catalyst promotion. The main reaction can be described as follows:

$$2Al_2O_3 + 3C \xrightarrow[electrolyte]{(600-1200)^\circ} 4Al + 3CO_2 \tag{1}$$

Also, there are many other secondary reactions during the complex reactions. It is some icons and cons which is a neutral reaction:

$$3CO_2 + 2Al \rightarrow Al_2O_3 + 3CO \tag{2}$$

The definition of SD is the difference between the electrolysis temperature and the crystal temperature. Under the appropriate temperature, the reduction bath is based on a superior situation and can improve the current efficiency, extend the lifespan of the bath, and improve the purity of the aluminum. The bath temperature can lay the foundation for the further evaluation of superheating and guide the workers' operation. The existing method to detect the index of the temperature is the physical detection device, it cannot achieve the online and real-time temperature for the harsh production environment. Figure. 1 gives the whole intuitive production.

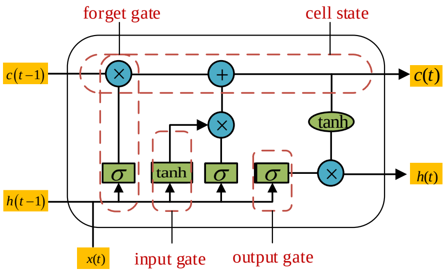


Figure 2: The basic unit of LSTM covers the forget gate, input gate, and output gate. σ represents the Sigmoid activation and the tangent activation kernel is given by *tanh*.

2.2. CNN and LSTM

In [22], the CNN is the basic unit for the deep feature extraction in the flame hole images. To make an accurate time sequence temperature prediction, the CNN has implicitly learned a representation of paws that can be further used as initialization for training a more accurate model.

Another unit that is used in the proposed model is the long-short time memory [27]. The LSTM copes with the long dependency by inducing the different gates so the useful information can be reserved for the next layer. As shown in Figure. 2, the main equation for information learning can be given as the following equation:

$$\hat{y} = \sigma\,(V \cdot h + b) \tag{3}$$

A lot of experiments and applications have shown LSTM's superior ability. Furthermore, many other extensive architectures such as deep LSTM, have been used in many domains. A widely accepted shortcoming of RNN is the long-term dependency and vanishing gradient problem. A special kind of RNN called LSTM is capable of learning long-term dependencies. LSTM is introduced by Hochreiter & Schmidhuber and is widely used in a large variety of problems. In the basic LSTM, three gates to protect and control the cell state and some repeating modules are added in an LSTM that contains four interacting layers. The LSTM network can implement temporal memory through the switch of the gates to prevent gradient varnishing. Denote the input vector $x(t)$ and the previous hidden state $h(t-1)$, and the external inputs are its previous cell state $c(t-1)$. Then, the forget gate is triggered as:

$$f(t) = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{4}$$

And, the input gate and new candidate vectors are computed as follows:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{5}$$

$$\tilde{C}_t = \tanh\left(W_{(C)} \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{6}$$

The new cell state in LSTM is updated by the following equation.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

The output gate vector can be calculated as follows:

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

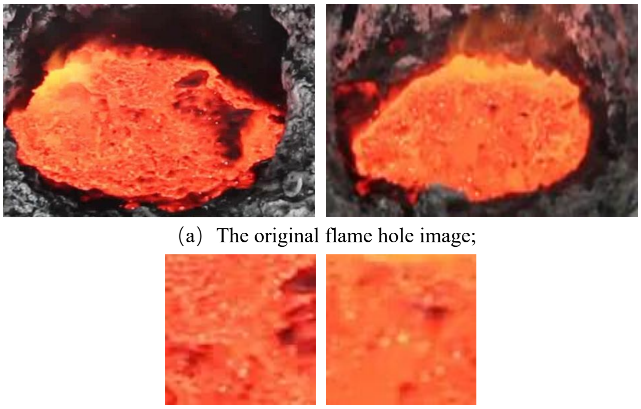
$$h_t = o_t * \tanh\left(C_t\right) \tag{9}$$

In the Eqs.(3)-(8), where σ is the nonlinear activation function, usually, the *sigmoid* function. *Tanh* represents the nonlinear tangent activation function. $*$ represents the pointwise multiplication operation. W_c, W_o, W_i are the related weights and b_i, b_c, b_o are the bias, respectively.

Compared to the traditional RNN, LSTM remains a cell state to accumulate the long-time feature from the time-sequential. So that the derivatives of the features from the front instant would be preserved. Further, to guarantee the temperature prediction performance, the information derivative over the time axis is very important, which promotes the proposed method in the sequel.

2.3. Self-supervised contrastive learning

With the difficulty and high cost of acquiring labeled data samples, there is an exponential increase in learning representation in unsupervised scenarios. Self-supervised learning is an active learning method that has shown great potential to learn the difference of the data sample for the constraint of labeled data. For instance, some rotations, angles, and distortions can be learned by the self-supervised model itself. In the proposed CNN-DSSLSTM model, a novel self-supervised loss with a knowledge dilation process is used for the iterative training process. The loss can be described as follows:



(a) The original flame hole image;

(b) The core deep temperature features after convolution.

Figure 3: The deep flame features of unsupervised images after the dilated convolution operation.

$$\mathcal{L}_\theta\left(\boldsymbol{x}_i, \left(\boldsymbol{n}_j, \boldsymbol{n}_k\right)\right) = L\left(\boldsymbol{z}_{i,j}, \boldsymbol{z}_{i,k}\right) \tag{10}$$

Mainstream SSL approaches, especially contrastive learning methods, often rely on a series of transformation functions $x_{i,k} = g(x_i, n_k)$ to produce multiple "views" $x_{i,k}$ of input instance x_i. The $n_k : k = 1, ..., K$ is the support of an "augmentation nuisances variable" n, representing a group of transformations such as flipping, scaling, cropping, or other augmentations acting on x_i via function g. Subscript k indexes all possible augmentations, i indicates the *ith* instance (each instance corresponds to a specific image input). Without labeling information, SSL relies on the invariance assumption so that deep feature extractor $z_{i,k} = f(x_i, k)$ is at least insensitive to such group of augmentation nuisances the network parameters. Existing work shows that, while the deep extractor $f(x_i, k)$ is trained to learn the shared semantics from distorted views x_i,k, such pretrained network parameters effectively benefit further downstream tasks [25]. SSL approaches based on such pairwise instance invariance assumptions can be generally summarized as penalizing some form of inconsistency between paired features z_i,j, z_i,k:

$$\begin{aligned}\mathcal{L}^* &= \mathbb{E}_{\mathbf{x}\in p_{\mathbf{z}}, n\in p^*}\left[\mathcal{L}_\theta(\mathbf{x}, \mathbf{n})\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\sum_{j,k} L\left(z_{i,j}, z_{i,k}\right) p^*\left(\mathbf{n}_j, \mathbf{n}_k\right)\end{aligned} \tag{11}$$

Inspired by [7], the unlabeled images directly be transformed to the target task. The fine-tuned network than as a teacher to impute labels to evaluate the soft labels' correctness. Specifically, the following distillation loss with soft labels are:

$$L_{distill} = -\sum_{x_i\in\pi(\cdot)}\left[\sum_y P^T\left(y\left|x_i;\tau\right.\right)\log P^S\left(y\left|x_i;\tau\right.\right)\right] \tag{12}$$

where $P\left(y\left|x_i\right.\right)=\exp\left(f\left(x_i\right)[y]/\tau\right)/\sum\limits_{y'}\exp\left(f\left(x_i\right)\left[y'\right]/\tau\right)$, and τ is the scalar parameters. The teacher network, which can produce $P^T\left(y\left|x_i\right.\right)$,this equation is fixed during the distillation process. The only network which is to be trained in the student net, is to produce $P^S\left(y\left|x_i\right.\right)$. Assume we have access to such oracle distribution p, and with equal sampling probability for x_i, the objective function is described as the above equation.

$$\mu[f](t,x)=\mu_0(x)\left(1+\gamma_\mu\rho_p[\rho(t,x)]\right) \tag{13}$$

The joint training process of the teacher network and student network can lead to a compact model which is used to improve the whole performance of the architecture.

3. The formalization of the proposed CNN-DSSLSTM method

This section introduces the formalization of the proposed CNN-DSSLSTM method.

3.1. Edge segmentation

To cultivate the burden of the computation, edge segmentation based on the threshold value is used in this architecture. Before the CNN learns the features, the original hole image is transmitted to an edge segmentation unit to extract a core unit. Since tedious annotations of the flame hole images are not available, so it is hard to train a multi-scale classifier to separate the categories, but instead employ self-supervised training. CNN is used to learn some complex feature representations of the original flame images. Assume that the original sequences are denoted as x_{st}, then a defined updated representation a_t is given as the output of the LSTM layer from the learned CNN-LSTM. Figure. 4 describes the self-supervised CNNDLSTM framework for learning the feature representation and its SD application. The building blocks for training are the CNN, the LSTM, and the final top-layer dual-dilated network. The output of the LSTM which links from the CNN in a consecutive frame by means of hidden states h_t and the nonlinear activation function can be rewritten as:

$$h_t=\sigma\left(W_hx_t+U_hh_{t-1}+b_h\right) \tag{14}$$

and output from the LSTM to the top layer representation can be given:

$$a_t=\sigma\left(W_ah_t+b_h\right) \tag{15}$$

In the proposed CNN-SSDLSTM framework, the KL divergence is used to compute the similarity in the training process. Since the label y_i is missing, the hidden features are constructed by the pair of (z_i,y_i). Then, the log-likelihood function can be written as the following constraint:

$$\begin{aligned}\max\quad & \mathrm{L}^s\left(\boldsymbol{x}_i,\boldsymbol{y}_i\right)= \\ & \mathrm{E}_{q(z)}\left[\ln p\left(\boldsymbol{x}_i,\boldsymbol{y}_i\mid\boldsymbol{z}\right)\right]-KL[q(\boldsymbol{z})\|p(\boldsymbol{z})]\end{aligned} \tag{16}$$

where $\mathrm{E}_q[\cdot]$ is the expectation of the distribution p, and $KL[q(z)\|p(z)]=\int q(z)\ln\frac{q(z)}{p(z)}dz$ is the Kullback-Leibler (KL) divergence to evaluate the dissimilarity between the two labels. Consider the following conditional distribution of x, y, and z,

$$p(z,y\mid x)=p(z\mid y,x)p(y\mid x) \tag{17}$$

$$p(\boldsymbol{x}\mid\mathbf{z},\boldsymbol{y})=p(\boldsymbol{x}\mid\boldsymbol{z}) \tag{18}$$

The static probability is derived as follows:

$$\begin{aligned}&\mathrm{L}^u\left(\boldsymbol{x}_i^u\right)= \\ &\mathrm{E}_{p(\boldsymbol{y}|\boldsymbol{x}_i^u)}\left[\mathrm{E}_{p(\boldsymbol{z}|\boldsymbol{x}_i^u,\boldsymbol{y})}\left[\ln p\left(\boldsymbol{x}_i^u\mid\boldsymbol{z}\right)\right]\right]+\mathrm{H}\left(p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)\right) \\ &-\mathrm{E}_{p(\boldsymbol{y}|\boldsymbol{x}_i^u)}\left[\boldsymbol{K}L\left[p\left(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{x}_i^u\right)\|p(\boldsymbol{z})\right]\right] \\ &+\mathrm{E}_{p(\boldsymbol{y}|\boldsymbol{x}_i^u)}\left[\mathrm{E}_{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}_i^u)}[\ln p(\boldsymbol{y}\mid\boldsymbol{z})]\right] \\ &=\mathrm{E}_{p(\boldsymbol{y}|\boldsymbol{x}_i^u)}\left[\mathrm{L}^s\left(\boldsymbol{x}_i^u,\boldsymbol{y}\right)\right]+\mathrm{H}\left(p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)\right)\end{aligned} \tag{19}$$

an interval term is introduced, the above equation then is expressed as follows:

$$\begin{aligned}&\mathrm{KL}\ \left[p\left(\boldsymbol{z},\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)\|p(\boldsymbol{z},\boldsymbol{y})\right] \\ &=\iint p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)p\left(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{x}_i^u\right)\ln\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}_i^u)p(\boldsymbol{y}|\boldsymbol{x}_i^u)}{p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})}d\boldsymbol{z}d\boldsymbol{y} \\ &=\iint p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)p\left(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{x}_i^u\right)\ln\frac{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}_i^u)}{p(\boldsymbol{z})}d\boldsymbol{z}d\boldsymbol{y} \\ &+\iint p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)p\left(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{x}_i^u\right)\ln p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)d\boldsymbol{z}d\boldsymbol{y} \\ &=\iint p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)p\left(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{x}_i^u\right)\ln p(\boldsymbol{y}\mid z)d\boldsymbol{z}d\boldsymbol{y} \\ &=\mathrm{E}_p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)\left[KL\left[p\left(\boldsymbol{z}\mid\boldsymbol{y},\boldsymbol{x}_i^u\right)\|p(\boldsymbol{z})\right]\right]-\mathrm{H}\left(p\left(\boldsymbol{y}\mid\boldsymbol{x}_i^u\right)\right) \\ &-\mathrm{E}_{p(\boldsymbol{y}|\boldsymbol{x}_i^u)}\left[\mathrm{E}_{p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x}_i^u)}[\ln p(\boldsymbol{y}\mid\boldsymbol{z})]\right]\end{aligned} \tag{20}$$

where z denotes the hidden variables, and E_p is the feature expectation of the output variables under the constraint of input x.

3.2. Contrastive knowledge dilation

In the aluminum electrolysis case study, the performance gap between supervised learning and the scarcity of labeled data should be narrowed. Therefore, to further improve the accuracy of the whole model, the knowledge dilation is applied to the proposed top-layer architecture to improve the accuracy of the proposed method. The dual teacher net and student network are joined to improve performance and reward the right action.

The output of LSTM can be further tackled with dual contrastive knowledge dilation networks to improve the performance of the proposed model. The specific mathematical equation is:

$$\widehat{y}_t=soft\max\left(W_Ta_t+W_Sa_t+b_h\right) \tag{21}$$

the matrix of W_T and W_s is the weight parameters that need to be fine-tuned with the distilled process, respectively.

The main flowchart of the proposed algorithm is given in the following steps:

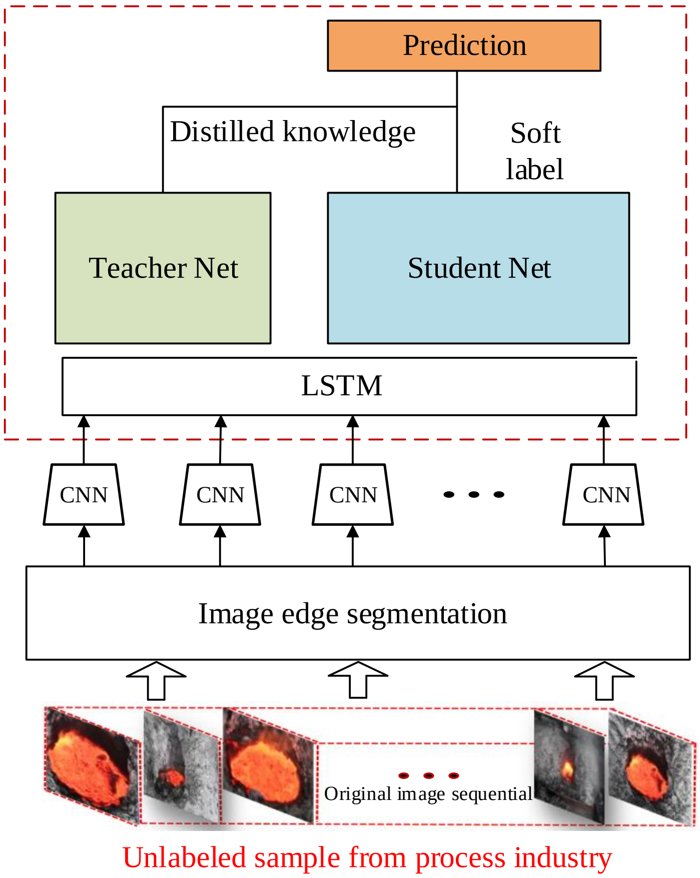


Figure 4: The proposed CNN-DSSLSTM framework for industrial electrolysis reduction. The original flame video was divided to piece of unlabeled images and then projected to CNN to extract the deep features. Self-supervised LSTM and top-layer dual distillation networks were coordinated to improve the accuracy and performance.

1) Collect sufficient continuous flame hole videos from the industrial plant scenarios;

2) Use the threshold edge segmentation to divide with the available data (∼1s for per image);

3) Split the data with the Training data, Testing data, and Validation data, initialize the network parameters such as learning rate α, the batch_size, the hidden layers of LSTM, and the epoch size;

4) Set up the CNN architecture, the Conv2d layer, and the max-pooling feed word network, and then train the CNN weight parameters with the training data.

5) Train the high-level LSTM network based on backpropagation through time (BPTT);

6) Use contrastive learning to train the joint Student Net and the Teacher Net to produce soft labels, the corresponding loss is given in equations (11)-(13);

7) Use the full-concat model to predict the cell temperature in the aluminum electrolysis reduction based on equation (3).

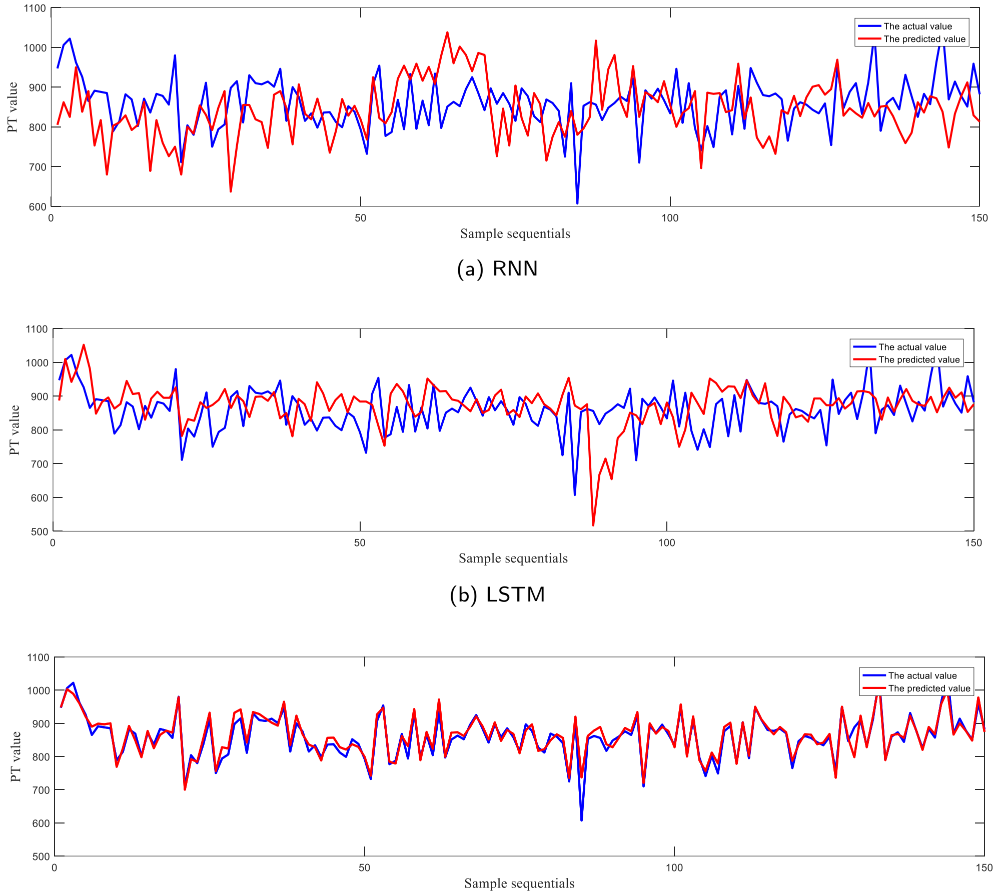
4. Experimental verification

In this section, some experiments are introduced to evaluate the performance of the proposed method. Firstly, a numeral case is used as the benchmark to verify the performance. Further, the industrial temperature experiment is also given. Some competitive algorithms are also compared in the experiments. To make a thorough comparison, the supervised learning method such as RNN, LSTM, and GRU are all tested in the simulation process. The $RMSE$ and MAE indexes are utilized to evaluate the model performance of the proposed methods. The mathematical expression of the $RMSE$ is:

$$RMSE = \sqrt{\sum_{i=1}^{N_u} \left(\hat{y}_t - y_t\right)^2 \Big/ N_u} \tag{22}$$

Another index in the comparison process is the mean absolute error (MAE). The specific expression can be written as the following equation:

$$MAE = \sum_{i=1}^{N_u} \left| \left(\hat{y}_t - y_t\right) \right| \Big/ N_u \tag{23}$$



(a) RNN

(b) LSTM

(c) SSDLSTM

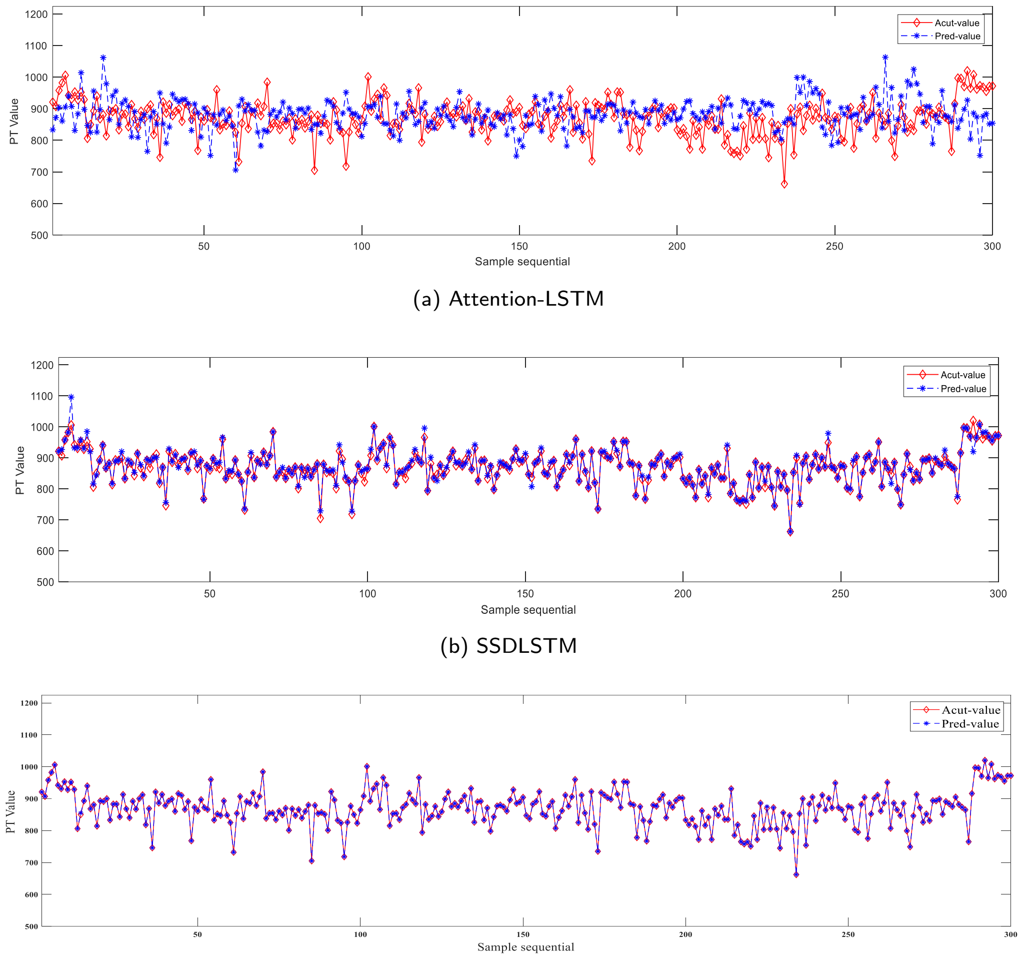
Figure 5: Testing error comparison with the different numbers of unlabeled samples for different algorithms: (a) RNN; (b) LSTM; (c) SSDLSTM.

The final evaluation indicator is mean absolute percentage error which is described as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y(t)-\hat{y}(t)}{y(t)}\right| \tag{24}$$

The experimental environment is conducted in a Ubuntu2020 Linux system with 64GB memory and a V100 GPU to accelerate the computation. Further, the python language and Pytorch framework are also used to realize the main algorithms. 56 videos of 3 categories with an average length of 6 minutes per video, which is in total $\tilde{5}$.5 hours recorded at 60 fps. In the segmentation process, the images have been divided into 1 second per image. To evaluate the performance, some temperature labels have been manually from the different videos and images. The features extracted from the image segmentation will be used as the input for the CNN, the results are given in Figure. 3. It consists of n Conv modules, where n is the number of frames of the input video. The main function of conv is to extract image features. This paper designs a three-layer convolution structure. It consists of two 3x3 convolutional layers, a maximum pooling layer, and a fully connected layer. The number of neurons in the fully connected layer is set to 256, and the input of the self-supervised LSTM model is 256 dimensions. As shown in [2], a similar initialization of LSTM with random weights and 512 hidden nodes is used. 12 sequences of flame hole videos are trained per batch and a random re-ordered version for each.

The original input is the flame video and position offset vector, and the dimension of flame hole video data is (n, 1080,1920,3), where n indicates that the video contains n frames of images, and (1080,1920,3) is the image format, indicating the width, height, and a number of channels. The dimension of the position offset vector is (n, 2), which means that each frame of the image contains two values. The design idea of the model is to extract the features of the flame hole video through the CNN module, and the position offset vector enters the self-supervised LSTM model after the convolution operation to extract the jitter state features and splicing the output sequence features with the features of the flame hole video after the maximum pooling. Therefore, the main reason why the position offset vector passes through the CNN layer first and then enters the self-supervised LSTM is that the position offset vector is not normalized, so the pooling layer is designed to normalize it. Finally, the predic-



(a) Attention-LSTM

(b) SSDLSTM

(c) Our CNN-SSDLSTM

Figure 6: Testing error comparison with the different numbers of unlabeled samples for different algorithms: (a) Attention-LSTM; (b) SSDLSTM; (c) Ours.

tion accuracy of tank temperature is further improved through the self-distillation model on the top floor. CNN consists of n modules, where n is the number of input frames. The main function of the con display is to extract image features. This proposed CNN-SSDLSTM designs a $v3-v$ three-layer video structure. For the fully connected layer, the input of the 256-dimensional self-supervised LSTM model is 256 dimensions.

Table 1
The competitive methods comparison for the temperature identification.

Methods	RMSE	MAE	MAPE	Train_time
LSTM	0.2345	0.2332	0.2621	21.67
RNN	0.3521	0.2546	0.2672	24.89
CNN-LSTM	0.1081	0.1665	0.1339	35.6
CNN-DLSTM	0.0473	0.1823	0.2645	37.99
Ours	**0.0335**	0.0221	**0.0045**	42.86

The dilated learning method is used for improving the accuracy of cell temperature prediction in the application of the plant-wide industry. From Table 2, after the successive rounds of training, the temperature accuracy has shown great potential. The highest pot temperature accuracy is up to 93%.

The self-supervised loss is constructed in the network training process. The rotation, augmentation policy is used for augmenting the amount of unlabeled data. Table. 1 gives a full comparison of the different algorithms. The train_time is evaluated with a second unit and RMSE, MAE, MAPE is both used to verify the effectiveness of the proposed methods. The commonly competitive methods are RNN, LSTM, CNN-LSTM, CNN-DLSTM and our methods, experiments describe that the integration of self-supervised can further lower the RMSE and improve the prediction accuracy at the same time. The accuracy of the final proposed model validation set is 0.823, and the accuracy of the position offset CNN-SSDLSTM model validation set is 0.845, respectively. In addition, CNN-LSTM and CNN-DLSTM's performances are superior to the basic RNN and LSTM units, which represents that CNN can learn better information from the flame hole videos. Our CNN-SSDLSTM achieves the

Table 2
Training times and variance for the different algorithms with different numbers of unlabeled samples.
A: low-temperature zone, B: middle-temperature zone, C: upper-temperature zone

Catergories	Te_acc			Tr_acc			Tr_time(s)		
	A	B	C	A	B	C	A	B	C
Supervised									
CNN	0.504	0.618	2.626	0.456	0.621	0.481	24.76	24.77	24.77
SimCLR [8]	0.362	1.023	0.128	0.523	0.722	0.813	4.73	5.33	5.22
BYOL[14]	0.233	0.732	0.625	0.616	0.714	**0.881**	6.88	7.31	7.37
Semi-supervised									
SWPPLSR [6]	0.342	0.534	0.636	0.324	0.566	0.698	3.71	3.45	4.23
CNN-LapsELM [21]	0.645	0.423	0.245	0.601	0.711	0.748	3.42	3.44	3.28
Self-supervised									
MoCo [26]	0.674	0.828	0.867	0.086	0.785	0.761	6.21	6.20	6.52
ARTMAP [3]	0.562	0.623	0.648	0.456	0.725	0.737	7.93	8.40	7.22
SSDLSTM [24]	0.693	0.832	0.845	0.656	0.895	0.841	2.59	2.34	2.53
Ours	0.863	**0.934**	**0.905**	0.820	**0.905**	0.891	31.59	25.34	33.53

lowest RMSE and MAE, 0.0335 and 0.0221, respectively, which is the result of CNN, self-supervised loss, and constructive design. While it also costs a longer training time, due to the independent CNN training process.

In the comparison experiment, after two rounds of iterative, an observed rapid convergence and accuracy gain has been shown. The contrastive joint training of the top dual network learns a superior representation of the unlabeled flame images. Compared to the traditional RNN and LSTM unit, the prediction performance has increased by a large margin, up to 0.92.

5. Novelty analysis

In the above experiments, the proposed CNN-SSDLSTM provides state-of-the-art performance compared with the other competitive methods. The advantages of the proposed method can be specified as the following items:

Economic profit saving. The proposed method is a fully data-driven method that utilizes the data to train the algorithms rather than the physical devices with a lot of profit savings.

Online real-time prediction. The method in the paper proposes an online real-time prediction for SD in the aluminum reduction cell. It can be also extended to use this method for other similar industrial plant-wide applications.

Break the constraint of labeled samples. This method uses self-supervised contrastive learning to narrow the gap between supervised learning and the scarcity of labeled samples. Considering that the practical industrial plants are mainly the categories of unlabeled data, the generation of the proposed method is greatly improved.

6. Conclusion

In this paper, an improved self-supervised LSTM method is proposed for aluminum temperature prediction. The deep architecture of the proposed algorithm leverages the loss function with contrastive learning. Further, a novel unit with a self-supervised loss function is used in the proposed method. The knowledge dilation with dual teacher net and student net is also jointly trained for the proposed architecture. The experimental results demonstrate that the performance is superior to the other existing comparisons. The proposed self-supervising function is vividly utilized in the CNN-SSDLSTM models and got a competitive experimental and application verification. Some extended exploration of this method can be fault soft sensor domains, which have the great potential to deal with fault prediction under the limitation of sufficient labels.

A. My Appendix

This work is supported by the National Natural Foundation of China under grant No.62133016 and the China Scholarship Council under grant No.202006370101. The authors would like to thank the anonymous reviewers for their valuable comments to improve the quality of this paper.

CRediT authorship contribution statement

Yongxiang Lei: Idea, Conceptualization of this study, Methodology, Software, Validation, Original draft writing, Review&Editing. **Xiaofang Chen:** Data curation, Experiment platform and environment, Supervision, Funding acquisition. **Yongfang Xie:** Data curation, Visualization, Supervision. **Lihui Cen:** Resource, Supervision.

References

[1] Aleotti, F., Tosi, F., Zhang, L., Poggi, M., Mattoccia, S., 2020. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation, in: European Conference on Computer Vision, Springer. pp. 614–632.

[2] Brattoli, B., Buchler, U., Wahl, A.S., Schwab, M.E., Ommer, B., 2017. Lstm self-supervision for detailed behavior analysis, in: Pro-

ceedings of the IEEE conference on computer vision and pattern recognition, pp. 6466–6475.

[3] Carpenter, G.A., Grossberg, S., Reynolds, J.H., 1991. Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural networks 4, 565–588.

[4] Chai, Z., Zhao, C., 2020. Multiclass oblique random forests with dual-incremental learning capacity. IEEE transactions on neural networks and learning systems 31, 5192–5203.

[5] Chang, S., Zhao, C., Li, K., 2021. Consistent-contrastive network with temporality-awareness for robust-to-anomaly industrial soft sensor. IEEE Transactions on Instrumentation and Measurement 71, 1–12.

[6] Chen, J., Gui, W., Dai, J., Jiang, Z., Chen, N., Li, X., 2021. A hybrid model combining mechanism with semi-supervised learning and its application for temperature prediction in roller hearth kiln. Journal of Process Control 98, 18–29.

[7] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E., 2020a. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems 33, 22243–22255.

[8] Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

[9] Deng, Z., Chen, X., Xie, S., Xie, Y., Sun, Y., 2021. Distributed process monitoring based on joint mutual information and projective dictionary pair learning. Journal of Process Control 106, 130–141.

[10] Deng, Z., Chen, X., Xie, S., Xie, Y., Zhang, H., 2022. Semi-supervised discriminative projective dictionary pair learning and its application for industrial process monitoring. IEEE Transactions on Industrial Informatics .

[11] Feng, L., Zhao, C., Huang, B., 2021. Adversarial smoothing tri-regression for robust semi-supervised industrial soft sensor. Journal of Process Control 108, 86–97.

[12] Feng, L., Zhao, C., Sun, Y., 2020. Dual attention-based encoder–decoder: A customized sequence-to-sequence learning for soft sensor development. IEEE Transactions on Neural Networks and Learning Systems 32, 3306–3317.

[13] Georgescu, M.I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M., 2021. Anomaly detection in video via self-supervised and multi-task learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12742–12752.

[14] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

[15] Huang, K., Wu, Y., Long, C., Ji, H., Sun, B., Chen, X., Yang, C., 2021a. Adaptive process monitoring via online dictionary learning and its industrial application. ISA transactions 114, 399–412.

[16] Huang, Z., Yang, C., Chen, X., Huang, K., Xie, Y., 2020. Adaptive over-sampling method for classification with application to imbalanced datasets in aluminum electrolysis. Neural computing and applications 32, 7183–7199.

[17] Huang, Z., Yang, C., Chen, X., Zhou, X., Chen, G., Huang, T., Gui, W., 2021b. Functional deep echo state network improved by a bi-level optimization approach for multivariate time series classification. Applied Soft Computing 106, 107314.

[18] Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[19] Kim, T., Kim, H.Y., 2019. Forecasting stock prices with a feature fusion lstm-cnn model using different representations of the same data. PloS one 14, e0212320.

[20] Kolesnikov, A., Zhai, X., Beyer, L., 2019. Revisiting self-supervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1920–1929.

[21] Lei, Y., Chen, X., Min, M., Xie, Y., 2020a. A semi-supervised laplacian extreme learning machine and feature fusion with cnn for industrial superheat identification. Neurocomputing 381, 186–195.

[22] Lei, Y., Chen, X., Xie, Y., 2020b. An improved cell situation identification approach with convolutional neural network and wavelet extreme learning machine. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering , 0959651820935667.

[23] Lei, Y., Karimi, H.R., Cen, L., Chen, X., Xie, Y., 2021. Processes soft modeling based on stacked autoencoders and wavelet extreme learning machine for aluminum plant-wide application. Control Engineering Practice 108, 104706.

[24] Lei, Y., Karimi, H.R., Chen, X., 2022. A novel self-supervised deep lstm network for industrial temperature prediction in aluminum processes application. Neurocomputing 502, 177–185.

[25] Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717.

[26] Purushwalkam, S., Gupta, A., 2020. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. Advances in Neural Information Processing Systems 33, 3407–3418.

[27] Ren, J.C., Liu, D., Wan, Y., 2021. Modeling and application of czochralski silicon single crystal growth process using hybrid model of data-driven and mechanism-based methodologies. Journal of Process Control 104, 74–85.

[28] Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G., 2018. Time-contrastive networks: Self-supervised learning from video, in: 2018 IEEE international conference on robotics and automation (ICRA), IEEE. pp. 1134–1141.

[29] Shi, T., Li, L., Wang, P., Reddy, C.K., 2020. A simple and effective self-supervised contrastive learning framework for aspect detection. arXiv preprint arXiv:2009.09107 .

[30] Song, J., Zhang, H., Li, X., Gao, L., Wang, M., Hong, R., 2018. Self-supervised video hashing with hierarchical binary auto-encoder. IEEE Transactions on Image Processing 27, 3210–3221.

[31] Wang, J., Xie, Y., Xie, S., Chen, X., 2022. Optimization of aluminum fluoride addition in aluminum electrolysis process based on pruned sparse fuzzy neural network. ISA transactions .

[32] Wang, Y., Yang, H., Yuan, X., Shardt, Y.A., Yang, C., Gui, W., 2020. Deep learning for fault-relevant feature extraction and fault classification with stacked supervised auto-encoder. Journal of Process Control 92, 79–89.

[33] Xu, G., Liu, Z., Li, X., Loy, C.C., 2020. Knowledge distillation meets self-supervision, in: European Conference on Computer Vision, Springer. pp. 588–604.

[34] Xu, K., Wen, L., Li, G., Qi, H., Bo, L., Huang, Q., 2021. Learning self-supervised space-time cnn for fast video style transfer. IEEE Transactions on Image Processing 30, 2501–2512.

[35] Yang, C., An, Z., Cai, L., Xu, Y., 2021. Hierarchical self-supervised augmented knowledge distillation. arXiv preprint arXiv:2107.13715 .

[36] Yengera, G., Mutter, D., Marescaux, J., Padoy, N., 2018. Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of cnn-lstm networks. arXiv preprint arXiv:1805.08569 .

[37] Yue, J., Fang, L., Rahmani, H., Ghamisi, P., 2021. Self-supervised learning with adaptive distillation for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 60, 1–13.

[38] Yue, W., Gui, W., Chen, X., Zeng, Z., Xie, Y., 2020. Evaluation strategy and mass balance for making decision about the amount of aluminum fluoride addition based on superheat degree. Journal of Industrial & Management Optimization 16, 601.

[39] Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L., 2019. S4l: Self-supervised semi-supervised learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1476–1485.

[40] Zhou, L., Chen, J., Song, Z., Ge, Z., 2015. Semi-supervised plvr models for process monitoring with unequal sample sizes of process variables and quality variables. Journal of Process Control 26, 1–16.