International Journal of Minerals, Metallurgy and Materials
Accepted manuscript, https://doi.org/10.1007/s12613-024-3086-2

A novel smelting stage recognition method for converter steelmaking based on convolutional recurrent neural network

Zhangjie Dai[1)], *Ye Sun*[1)], *Wei Liu*[1),✉], *Shufeng Yang*[1,2),✉], *and Jingshe Li*[1)]

1) School of Metallurgical and Ecological Engineering, University of Science and Technology Beijing, Beijing 100083, China

2) State Key Laboratory of Advanced Metallurgy, University of Science and Technology Beijing, Beijing 100083, China

✉Corresponding Authors: Wei Liu E-mail: liuwei@ustb.edu.cn; Shufeng Yang, E-mail: yangshufeng@ustb.edu.cn

Abstract: The converter steelmaking process represents a pivotal aspect of steel metallurgical production, with the characteristics of the flame at the furnace mouth serving as an indirect indicator of the internal smelting stage. Effectively identifying and predicting the smelting stage poses a significant challenge within industrial production. Traditional image-based methodologies, which rely on a single static flame image as input, demonstrate low recognition accuracy and inadequately extract the dynamic changes in smelting stage. To address this issue, the present study introduces an innovative recognition model that preprocesses flame video sequences from the furnace mouth. Subsequently, it employs a convolutional recurrent neural network (CRNN) to extract spatiotemporal features and derive recognition outputs. Additionally, we adopt feature layer visualization techniques to verify the model's effectiveness and further enhance model performance by integrating the Bayesian optimization algorithm. The results indicate that the ResNet18 with convolutional block attention module (CBAM) in the convolutional layer demonstrates superior image feature extraction capabilities, achieving an accuracy of 90.70% and an area under the curve of 98.05%. The constructed Bayesian optimization-convolutional recurrent neural network (BO-CRNN) model exhibits a significant improvement in comprehensive performance, with an accuracy of 97.01% and an area under the curve of 99.85%. Furthermore, statistics on the model's average recognition time, computational complexity, and parameter quantity (Average recognition time: 5.49 ms, floating-point operations per second: 18260.21 M, parameters: 11.58 M) demonstrate superior performance. Through extensive repeated experiments on real-world datasets, the proposed convolutional recurrent neural network model is capable of rapidly and accurately identifying smelting stages, offering a novel approach for converter smelting endpoint control.

Keywords: Intelligent steelmaking; Flame state recognition; Deep learning; Convolutional recurrent neural networks

1. Introduction

The converter steelmaking process occupies a pivotal position in the metallurgical industry, with its intelligent manufacturing at the forefront of the sector and serving as a demonstration for

other [1-4]. Globally, advanced large and medium-sized steelmaking enterprises have extensively researched automatic control technologies for converters [5]. Endpoint control primarily focuses on the temperature and carbon content of molten steel, with its accuracy directly affecting product quality. Currently, traditional control models are categorized into static and dynamic types, with the former primarily based on energy and material balance calculations during the process to determine the addition of auxiliary materials and blowing conditions [6]. Nevertheless, owing to the inherent complexity and variability of the smelting process, these models exhibit significant deviations in effectively tracking and adjusting the process for optimization. Dynamic control models address the shortcomings of static models and are crucial for improving the accuracy of endpoint carbon and temperature control, as well as enhancing the quality of the steel [7-9]. The advancement of dynamic model control technology has largely been driven by the implementation of sub-lance detection and exhaust gas mass spectrometry, which have attained a relatively mature level of application [10-11]. However, challenges related to detection accuracy, equipment maintenance, and associated costs highlight the necessity for the development of a novel endpoint prediction model.

The optical information present in the furnace mouth flame provides a direct and real-time representation of the decarburization reaction progress occurring within the furnace [12]. In light of the rapid advancement of industrial big data, it is critically important to develop non-contact intelligent prediction models for smelting that leverage the characteristics of the furnace mouth flame and associated optical information. Current research mainly focuses on two directions: radiation spectral information methods [9,13-14] and flame image information methods [15-16]. Zhang et al. [13] collected spectral information of the converter flame using a USB2000 and a spectrometer, and simultaneously obtained continuous carbon content changes during the later stages of smelting using a flue gas analysis mass spectrometer, constructing a large sample dataset and establishing a prediction model. Zhao et al. [14] utilized the (Baseline estimation and sparse noise reduction) BEADS algorithm and genetic algorithm based on a dataset of converter mouth flame spectral data, and combined these with back propagation neural network (BPNN) to build a carbon content and temperature prediction model. The flame image method involves dividing the furnace mouth flame into multiple regions and using characteristic pixels appearing in these specific regions as inputs for pattern recognition, analyzing the characteristics of the blowing process, and introducing endpoint control [15]. Liu et al. [17] proposed an accurate and rapid multi-flame feature extraction method based on the generalized regression neural network (GRNN) and established a prediction model. Building on this, they also introduced a multi-scale color difference histogram feature weighted fusion method to describe the changes in the flame during the blowing process, which demonstrates good recognition rates and high computational speed, offering practical value in converter endpoint control [18].

Deep learning, recognized as a breakthrough technology within the realm of artificial intelligence, distinguishes itself from traditional machine learning approaches by obviating the need for data annotation prior to the execution of each learning task. This methodology is increasingly applied within the combustion industry [19]. However, the transformation of the flame during various phases of converter steelmaking constitutes a dynamic process characterized by fluctuations in flame oscillation and stroboscopic variations in brightness across the flame region at different temporal intervals, exerting significant temporal influences. Conventional convolutional networks (CNNs) that utilize static images of smelting flames as input are prone to

noise and other interferences, resulting in diminished accuracy. In contrast, video sequences encapsulate a broader spectrum of information and can more precisely reflect alterations in the smelting stage.

In the context of rapid advancements in deep learning, Niu et al. [20] employed CNN-LSTM to predict 3D temperatures from combustion flame dynamics, showing effective learning and forecasting. Lu et al. [21] applied convolutional recurrent neural network (CRNN) to compartment fire prediction, offering scientific guidance for the development of intelligent fire-fighting technologies. Similarly, Huang et al. [22] proposed a U-ConvLSTM model for the reconstruction of multi-dimensional combustion fields. Collectively, these studies underscore the considerable potential of the CNN-LSTM model in extracting spatiotemporal features of flames and predicting combustion states [23-24], thereby offering promising applications for predicting converter smelting stages. Currently, data-driven methods have been extensively studied and applied to explore the causal relationships between steelmaking components, processes, structures, and performances, aiding in the control of the steelmaking process [16,25]. However, there is a lack of research utilizing image data in this field. Therefore, against this backdrop, this paper proposes the use of a CRNN model and serialized mouth flame images to achieve the recognition of converter smelting stage.

To effectively leverage deep learning for the task of converter flame recognition, we initially gathered on-site smelting data and developed a corresponding dataset. Subsequently, we established the CRNN model and integrated it with visualization techniques and various evaluation metrics. Experimental results indicate that the proposed CRNN model exhibits superior performance in the process of flame recognition, providing a novel approach for endpoint control in converter smelting processes.

2. Data acquisition and description

Based on the characteristics of the chemical reactions occurring within the converter molten bath, the blowing process can be categorized into three distinct stages: early, middle, and late. The phenomenon of the mouth flame is fundamentally attributed to the combustion of CO, which is generated by the decarburization reaction occurring in the molten steel and ignited at the furnace mouth. Figure 1 illustrates the trend of changes in CO and CO_2 concentrations in the flue gas throughout the blowing process, with flame categories delineated based on variations in the carbon-oxygen reaction rate.

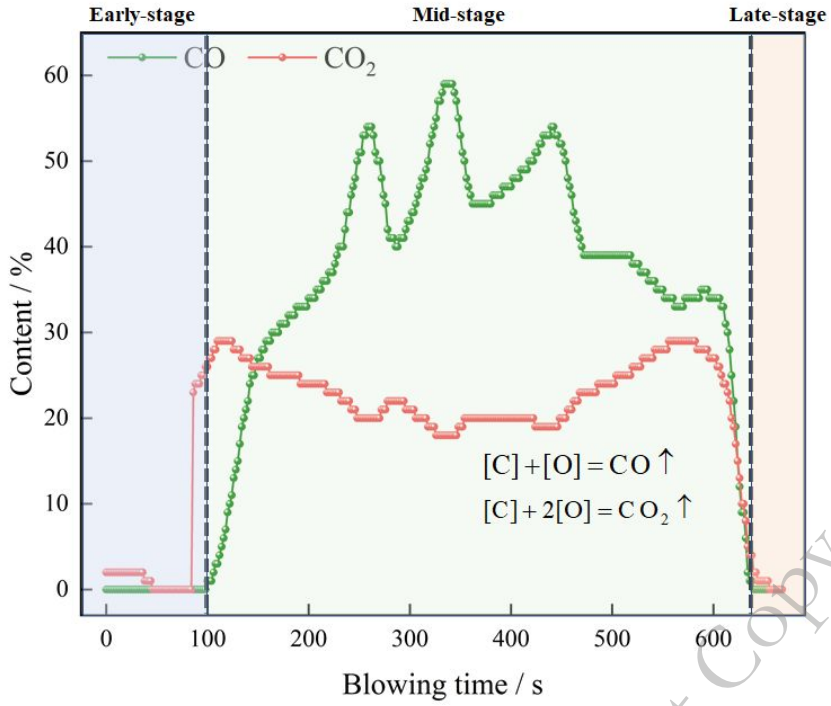

Fig. 1. The change curve of CO and CO_2 concentration with time in smelting process, and the basis of flame division in three stages.

The experimental data utilized in this study were collected from front-of-furnace cameras at a steelmaking plant, with all video recordings captured during the normal smelting process. To effectively differentiate the flame stages throughout the various blowing periods, dataset calibration was performed in conjunction with observed changes in flue gas patterns and insights from expert experience.

The flame morphologies corresponding to the three stages of smelting are depicted in Figure 2. In the early stage of smelting, the flame appears weak, exhibiting an overall conical shape characterized by a predominantly dark red hue, often approaching black, and the blowing process is frequently accompanied by black smoke. During the mid-stage of smelting, the flame color transitions from bright yellow to an even brighter yellow, with the flame body becoming softer and exhibiting more vigorous combustion. It displays a narrow base and an inverted trapezoidal top, yet its stability is poor, filling the entire mouth of the furnace. In the late stage of smelting, the flame shape at the furnace mouth stabilizes, with the area gradually contracting. The overall appearance is characterized by a soft yellow flame that exhibits a translucent and glowing visual effect.

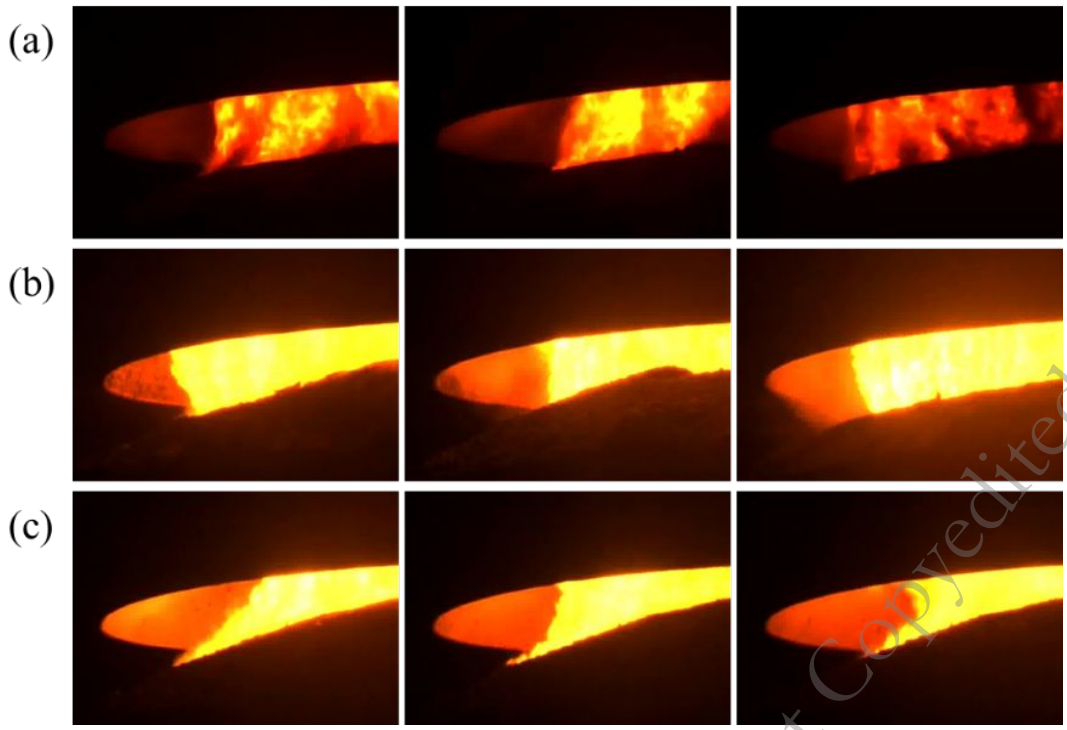

Fig. 2. Flame morphology during the main different stages of the converter smelting process: (a) early-stage, (b) mid-stage and (c) late-stage.

3. Methods

To achieve the recognition of flame states during the converter steelmaking process, the experiment employed the technical approach illustrated in Figure 3, which consists of two main parts. Firstly, Resnet is utilized to extract the spatial features of video frames from the smelting process. Secondly, the sequence of feature vectors extracted by Resnet is input into a long short-term memory (LSTM) model to extract temporal sequence features. Additionally, a convolutional block attention module (CBAM) attention mechanism is embedded in the image feature extraction part, and class activation mapping technology is employed for visualization, along with the tree-structured parzen estimator (TPE) optimization algorithm to search for hyperparameters. Ultimately, a three-classification result for the recognition of smelting stages is obtained. In this section, we will primarily introduce the specific methods used in the aforementioned process, as well as the CRNN model proposed in this paper.

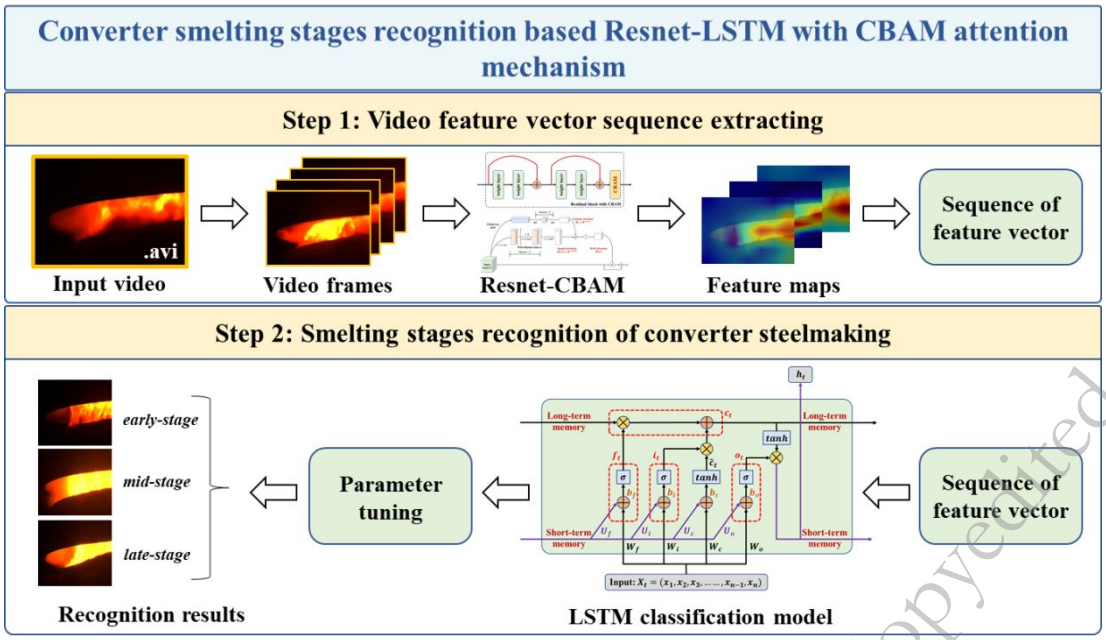

Fig. 3. Dilated CNN–LSTM with CBAM attention mechanism.

3.1 Extraction of spatial features by the Resnet block

CNNs possess the capability to extract image features from the local to the global level, enabling tasks such as image recognition. Deep network structures are capable of capturing richer and more complex features, typically exhibiting good adaptability to new tasks. However, due to issues such as vanishing gradients and network degradation, deeper networks are more difficult to train. To overcome the problem of network degradation, He et al. [26] proposed a deep residual network. Figure 4 shows the structure of the residual block used, which introduces an "identity mapping." This ensures that even when the network learns fewer features, its performance does not degrade, thereby maintaining relatively better performance.

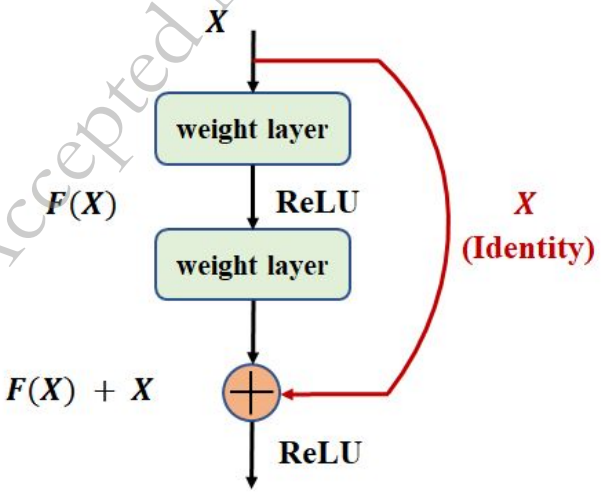

Fig. 4. Residual learning: a building block.

3.2 Extraction of time features by the LSTM block

RNNs are used to model problems with dynamic changes over time series, but they lack the ability to learn long-term dependencies. Therefore, in practical applications, the recurrent layer typically employs LSTM [27] structures or gated recurrent units (GRU) [28]. The unit structure of the LSTM model is shown in Figure 5, with the formulas presented in (1)-(6).

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{1}$$

$$f_t = \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right) \tag{2}$$

$$g_t = tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right) \tag{3}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

In the equations, h_t, c_t, and x_t represent the hidden state, cell state, and input at time t, respectively, while h_{t-1} is the hidden state of the layer at time $t-1$. i_t, f_t, g_t, and o_t correspond to the input gate, forget gate, cell gate, and output gate, respectively. W denotes the weight vectors, and b represents the bias vectors, with subscripts indicating the weights and biases for the corresponding units. σ represents the sigmoid function, and $\odot$ denotes element-wise tensor multiplication.

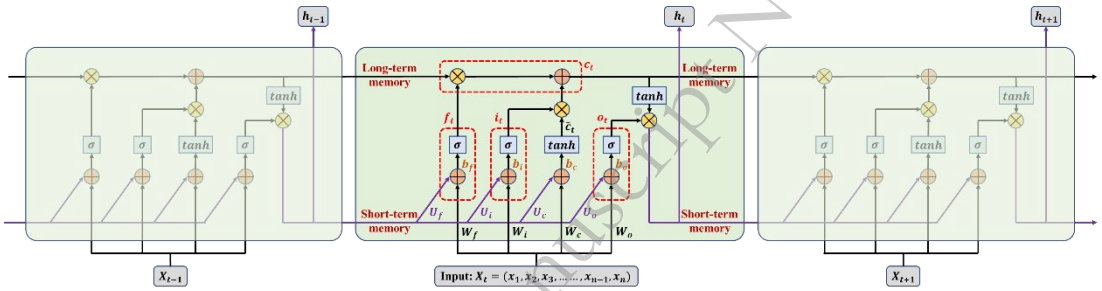

Fig. 5. LSTM network structure.

Specifically, the forget gate f_t regulates the information derived from the previous cell state, determining the extent to which this information should be retained or transmitted to the subsequent stage. The input gate i_t governs the incorporation of new information, integrating the outputs of both the forget gate and the input gate to update the current cell state. Finally, the output gate o_t merges the most recent cell state information with the input data to update the current hidden state, which functions as the output of the LSTM network. Throughout this entire process, the parameters of the LSTM are updated via backpropagation.

3.3 The CBAM attention module

Attention mechanisms selectively filter out a small amount of crucial information from a vast amount of data, focusing attention on this important information while disregarding the majority of less relevant content [29]. The CBAM is a lightweight and versatile attention module [30] that can be seamlessly integrated into any CNNs architecture and trained by end-to-end [31]. Figure 6 illustrates the structure of the CBAM attention mechanism, which dynamically adjusts channel and spatial feature weights in CNNs through channel and spatial attention modules, thereby enhancing the model's ability to perceive important features and improving performance in image

representation learning.

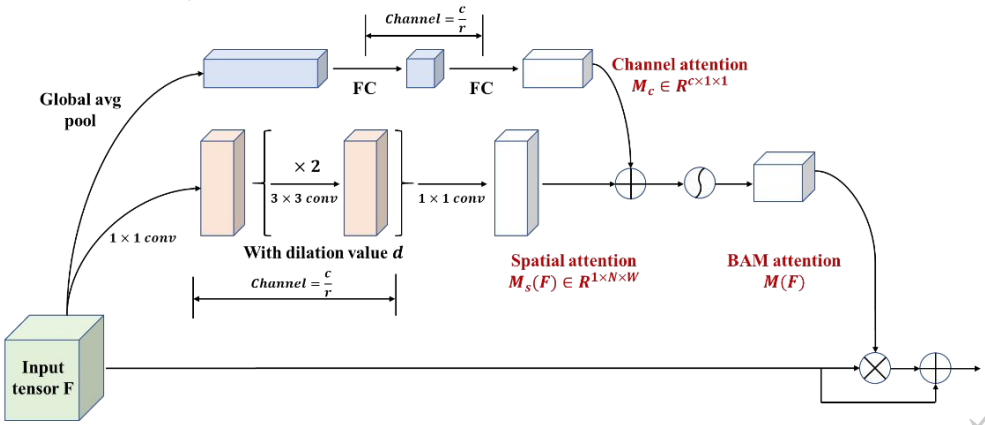

Fig. 6. Structure diagram of CBAM attention mechanism.

3.4 Class Activation Mapping

The class activation mapping (CAM) technique involves removing the fully connected layers after the final convolutional layer of CNNs and introducing a global average pooling (GAP) layer. By performing a series of Softmax and linear transformation operations on each feature map in the GAP layer, an activation map is ultimately generated. This process can be expressed by equation (7):

$$S^c = \sum_k \omega_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k \tag{7}$$

Where ω_k^c represents the weights between the GAP layer and the Softmax layer, c is the index of the target class, A is the feature map output by the last convolutional layer, k is the index of its channel dimension, and i and j are the indices for the width and height dimensions, respectively.

As an extension of CAM, Selvaraju et al. [32] proposed a gradient-weighted class activation mapping (Grad-CAM) technique, which rearranges the summation order to obtain the Grad-CAM map of an image. Generalizing the CAM algorithm, Grad-CAM can be applied to any CNN architecture without the need for retraining or modification of the network structure [33]. The Grad-CAM technique provides better visual explanations for deeply connected neural networks, and therefore, this paper will utilize Grad-CAM to offer deep interpretability insights for the converter smelting stage recognition task, providing additional information that aids in explaining the CNN's decisions.

3.5 The Proposed Network Architecture

The architecture of the stage recognition model proposed in this study is illustrated in Figure 7. The CRNN model is constructed from three essential modules: convolutional layers, recurrent layers, and output layer.

The convolutional layer processes a sequence of converter smelting images during each recognition cycle. To comprehensively extract feature information from the spatial dimensions of these images, the ResNet-CBAM architecture is employed to enhance feature representation. Following the convolutional operations, the resultant feature matrix is flattened and concatenated into vector form, which is then input into the LSTM network for the recurrent layer.

In the recurrent layer, a stacked architecture comprising two bidirectional LSTM structures is utilized to capture the temporal features present in the time series data. This stacked LSTM configuration significantly enhances the model's capacity compared to a conventional single-layer LSTM. Each layer within this framework is configured with an identical number of hidden nodes to ensure consistency in feature extraction.

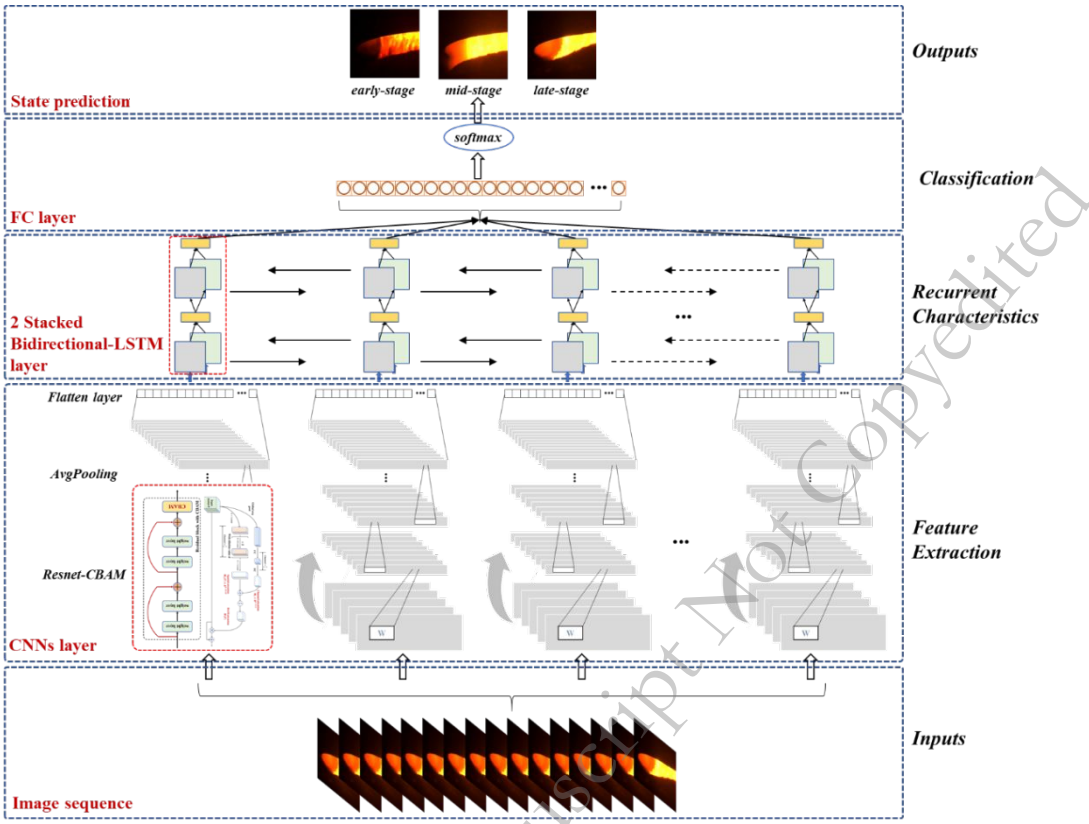

Fig. 7. Schematic diagram of the CRNN architecture proposed in this work. It primarily consists of three components: the convolutional layer, the recurrent layer, and the output layer.

The output layer is composed of a fully connected layer followed by a Softmax layer. The original flame sequence undergoes spatial feature extraction via convolutional layers and temporal feature extraction through recurrent layers, resulting in a feature vector v that integrates spatial image information and temporal relational characteristics. For the flame state classification task, a fully connected layer is first applied to v, performing a linear transformation to produce a new vector $v^{(f)}$. This transformed vector is subsequently passed through the Softmax layer, yielding the predicted probability p_{ic} for the image category during the recognition cycle, as expressed in the following equation:

$$p_{ic} = f\left(v^{(f)}\right) = e^{v_c^{(f)}} / \sum_{j=1}^{M} e^{v_j^{(f)}} \tag{8}$$

The Softmax layer maps the probabilities associated with each category to a range between 0 and 1, ensuring they sum to unity. The category with the highest probability is identified as the model's output, representing the predicted converter smelting stage at the given time.

4. Experimental analysis

The primary objective of this paper is to propose an automatic recognition model for the stage of converter steelmaking. The experiment encompasses the following content: (1) Collecting smelting video files through CCD cameras; (2) Extracting image information from video files; (3) Annotating all collected flame images according to the early, middle, and late stages of smelting; (4) Constructing and training a deep learning model; (5) Visualizing and comparing the evaluation of the converter smelting stage recognition model.

4.1 Dataset establishment

The establishment of the training dataset follows the processes and rules illustrated in Figure 8. For the video set collected on-site, the initial step involves manually selecting the regions of interest (RoI) based on the flame characteristics described in the preceding section, followed by segmenting the entire video into 10-second intervals. Subsequently, image frames are extracted from each small video segment, and the corresponding sequence lengths are determined. Finally, the extracted image frames are temporally ordered and labeled accordingly, thus completing the overall dataset construction. All image data utilized in this study are RGB three-channel with dimensions of 224×224.

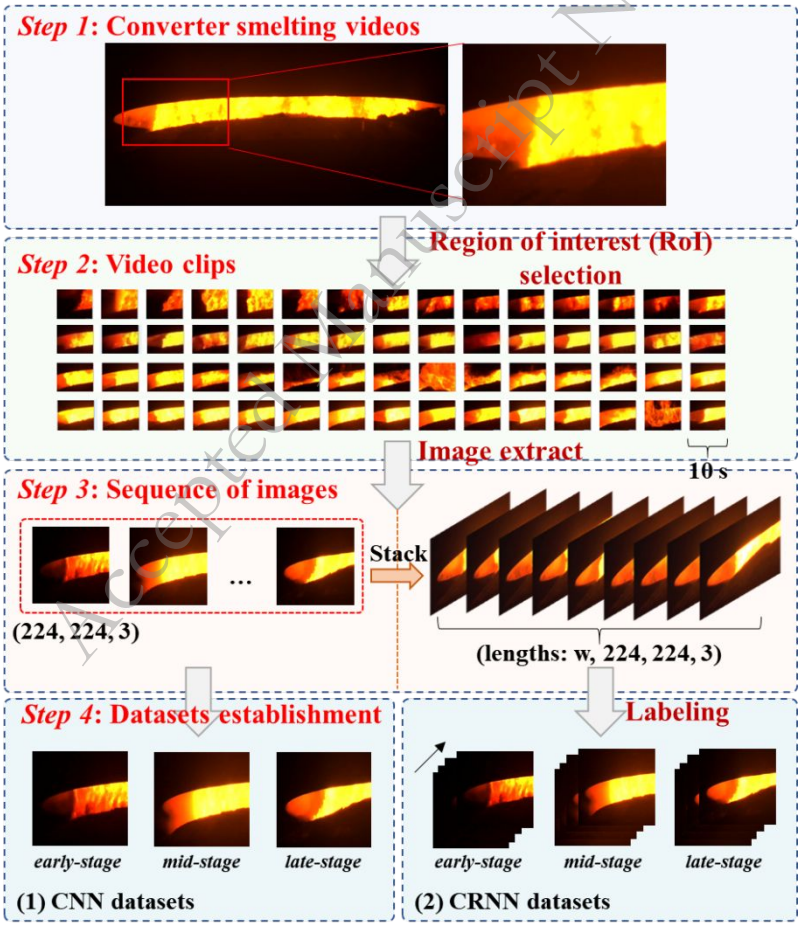

Fig. 8. Process and rules for establishing converter flame dataset.

In this work, 12,310 image data were extracted, comprising 4,970 samples from the

early-stage, 4,250 samples from the mid-stage, and 3,090 samples from the late-stage. The dataset utilized for time series prediction consists of 1,440 early-stage samples, 1,310 mid-stage samples, and 940 late-stage samples, amounting to a cumulative total of 3,690 samples. This dataset is partitioned such that 70% is allocated for training purposes, while the remaining 30% is designated for validation. Importantly, since only the training dataset was employed for model training, no duplicate data exists within the validation set.

4.2 Learning and evaluation

As mentioned above, the CRNN model proposed in this paper consists of three main elements: convolutional layers, recurrent layers, and an output layer. By flexibly combining convolutional layers and recurrent layers, this experiment mainly considers two different neural networks, the corresponding dataset establishment for which can be seen in Figure 8.

(1) CNN: This model consists only of convolutional layers and an output layer, without including recurrent layers. The purpose of using convolutional operations in the convolutional layer is to further learn features in the image spatial dimension, without considering the associated information in the continuous temporal dimension. The aim is to evaluate the impact of global information obtained from local image information after convolution on the accuracy of recognizing the smelting stage.

(2) CRNN: This model is the complete structure designed in this work, with its framework depicted in Figure 7. The purpose is to fully consider both the image spatial features and the time-related information, in order to further improve the accuracy of recognizing the state of the converter flame.

This experiment was conducted using Pytorch 2.0.0 under the Windows 11 operating system (CPU: Intel® i9-13980HX @ 2.20 GHz, GPU: NVIDIA GeForce RTX 4070 Laptop). During training, the loss calculation was performed using CrossEntropy, and the optimizer selected was SGD with a Learning rate of 0.001, a Momentum factor of 0.9, a Weight decay of 0.0005, and a maximum Epoch number of 100. The Batch size was set to 4. The same training set and hyperparameters were used for each training iteration, and the same validation set was used for model performance evaluation. Additionally, to enhance the model's ability to fit images with different angles, sizes, positions, and noise levels, data augmentation techniques such as image translation, flipping, and scaling were introduced to mitigate potential overfitting issues during the training process.

The evaluation of the classification model's accuracy generally employs four metrics: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The model's accuracy (ACC) represents the proportion of correctly classified results out of all classifications:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

In the context of multi-classification tasks, the use of average accuracy to evaluate model performance and generalization capabilities has certain limitations, as it does not reflect the recognition accuracy and error rates for each individual category. Therefore, this paper comprehensively considers the F1-score, confusion matrix, and the Receiver Operating Characteristic (ROC) curve to achieve a comprehensive assessment of the model's performance.

The F1-score is the harmonic mean of precision and recall, and its mathematical form is as follows:

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

Precision refers to the proportion of correctly classified events among all detected events:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

Recall indicates the proportion of events that are correctly classified out of all events:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

4.3 Results and discussion

4.3.1 Convolution layer comparison

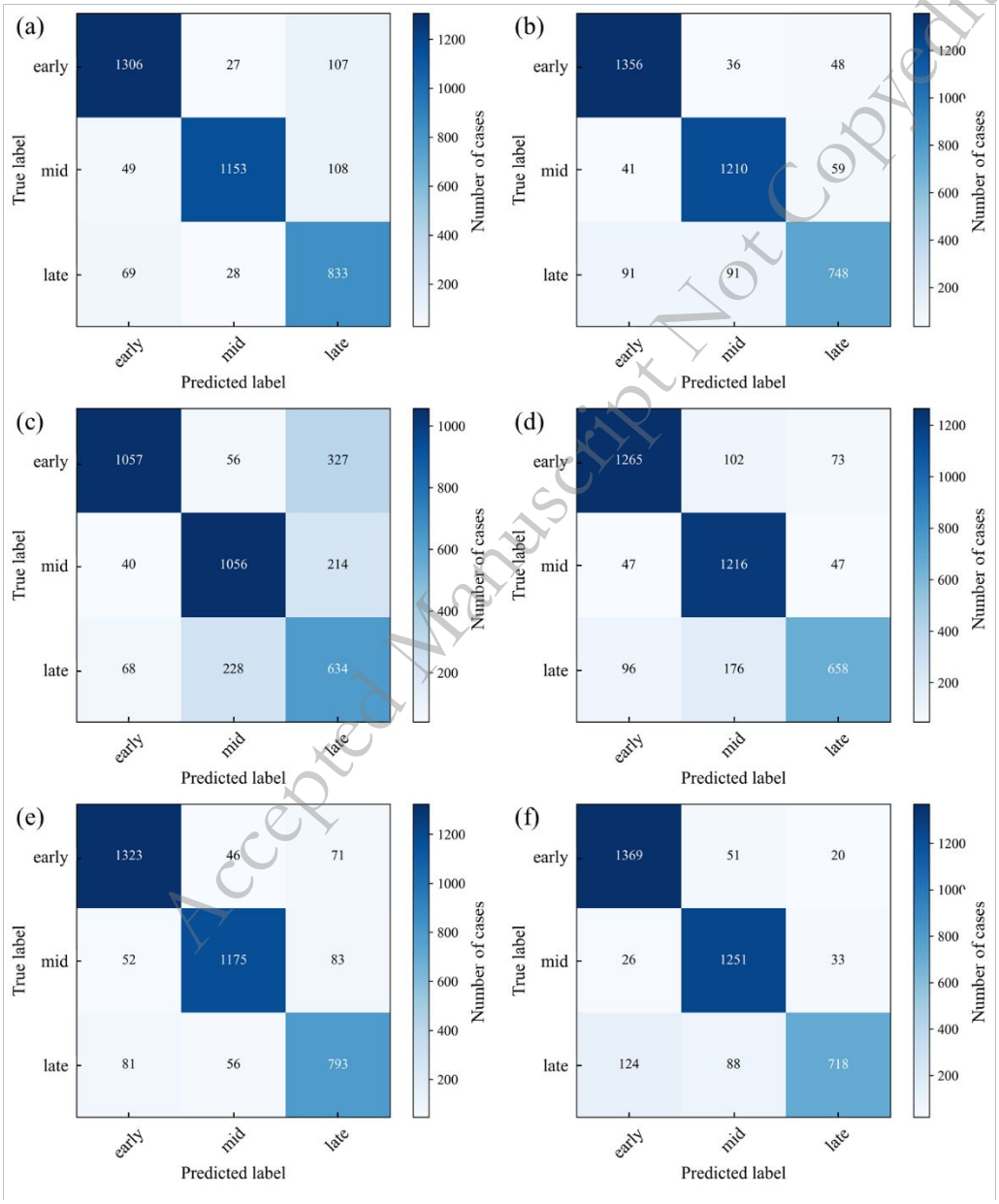

Fig. 9. Confusion matrix of different algorithms for classification test in each blowing period: (a) Resnet18, (b) Densenet121, (c) Transformer, (d) VGG16, (e) Resnet18-SE and (f) Resnet18-CBAM.

In this experiment, four CNN models were selected for training and validation on the furnace mouth flame dataset: Resnet18, Densenet121, Transformer, and VGG16. In addition, we considered the addition of SE-block and CBAM-block attention mechanisms within the Resnet18 network respectively. To ensure a fair comparison among all models, the experimental parameters were standardized.

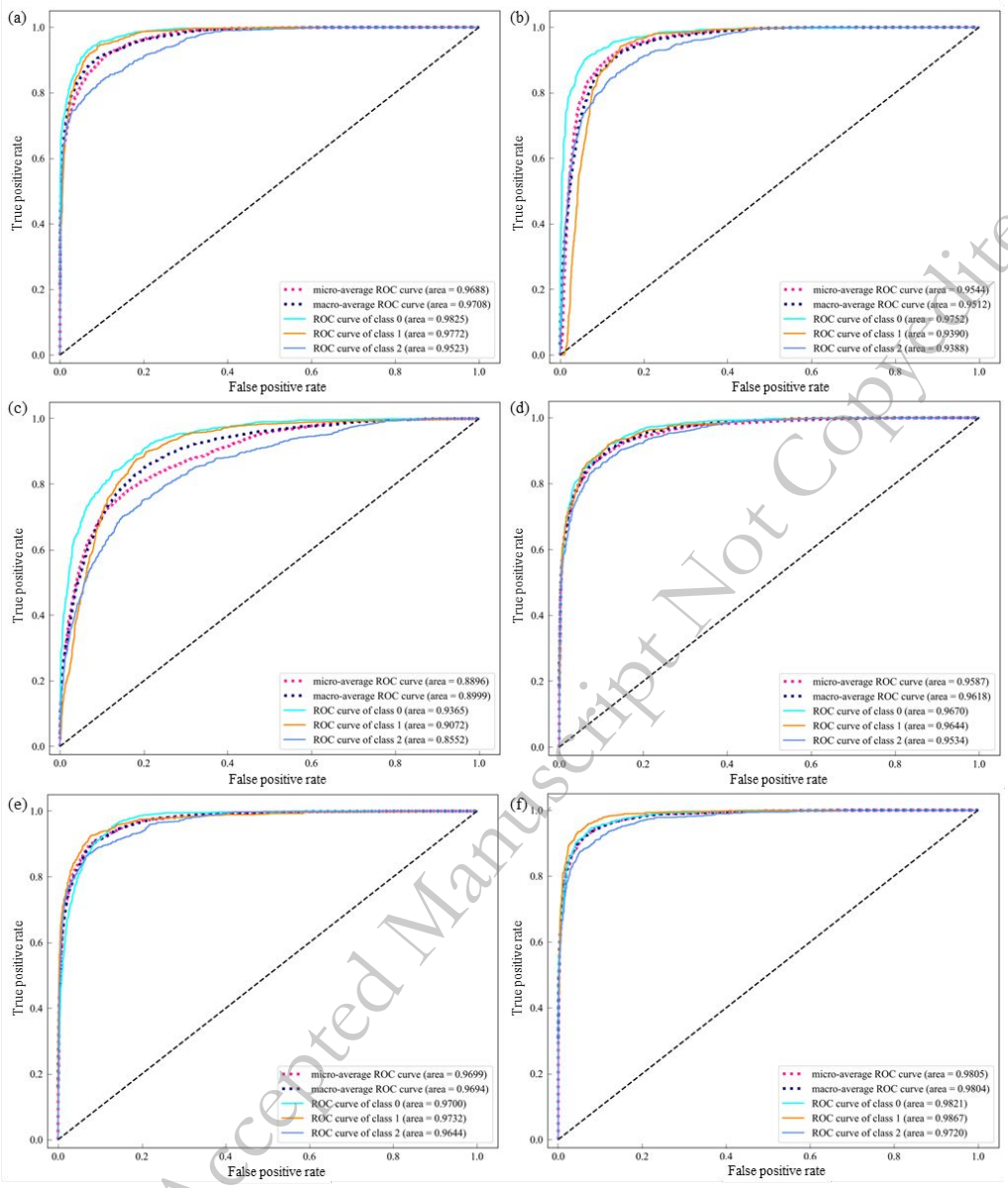

Fig. 10. Receiver operating characteristic (ROC) of different algorithms, with an area under the ROC curve (AUC): (a) Resnet18, (b) Densenet121, (c) Transformer and (d) VGG16, (e) Resnet18-SE and (f) Resnet18-CBAM.

Figure 9 presents the confusion matrices for the classification performance of different CNN models. By examining the confusion matrix, the classification effectiveness of each category can be intuitively analyzed. The larger and darker the values on the diagonal of the matrix, the better the performance of the classification model. Additionally, in Figure 10, we have plotted the ROC curve to further illustrate the recognition performance of each model. The classes are represented as class1-3 for the early, middle, and late stages of smelting, respectively, and are drawn with cyan, orange, and blue solid lines. The area under the ROC curve (AUC) closer to 1 indicates a

stronger recognition capability of the model. The results show that the AUC values of the models, from highest to lowest, are Resnet18-CBAM (98.05%), Resnet18-SE (96.99%), Resnet18 (96.88%), Densenet121 (95.44%), VGG16 (95.87%), and Transformer (88.96%). The experiment demonstrates that integrating spatial and channel attention mechanism modules significantly enhances the overall recognition performance.

We also discussed the impact of different depths of Resnet networks on recognition performance, with the accuracy and loss change curves of the training and validation processes depicted in Figure 11. As the network depth increases, the accuracy gradually decreases, while simultaneously, the model's computational complexity and parameter count significantly rise. Therefore, for the flame dataset used in this experiment, shallow network architectures exhibit better performance.

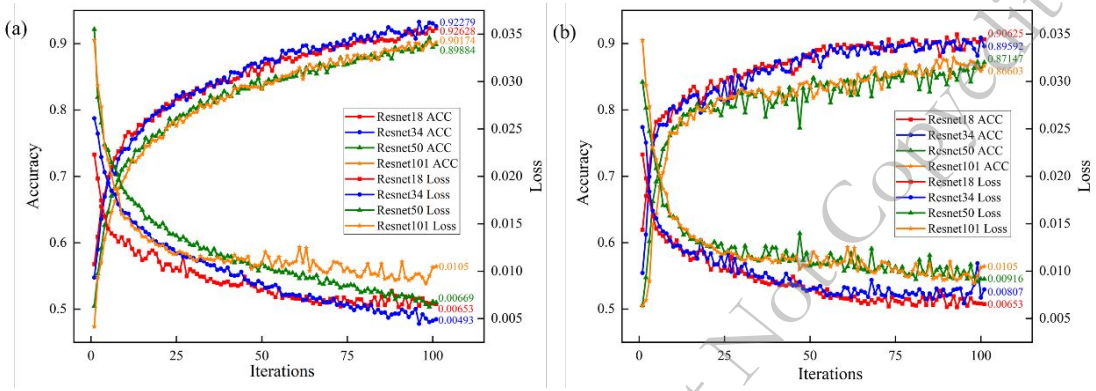

Fig. 11. Accuracy and loss curves of Resnet models with different depths: performance of (a) training sets and (b) validated sets.

4.3.2 Feature map visualization

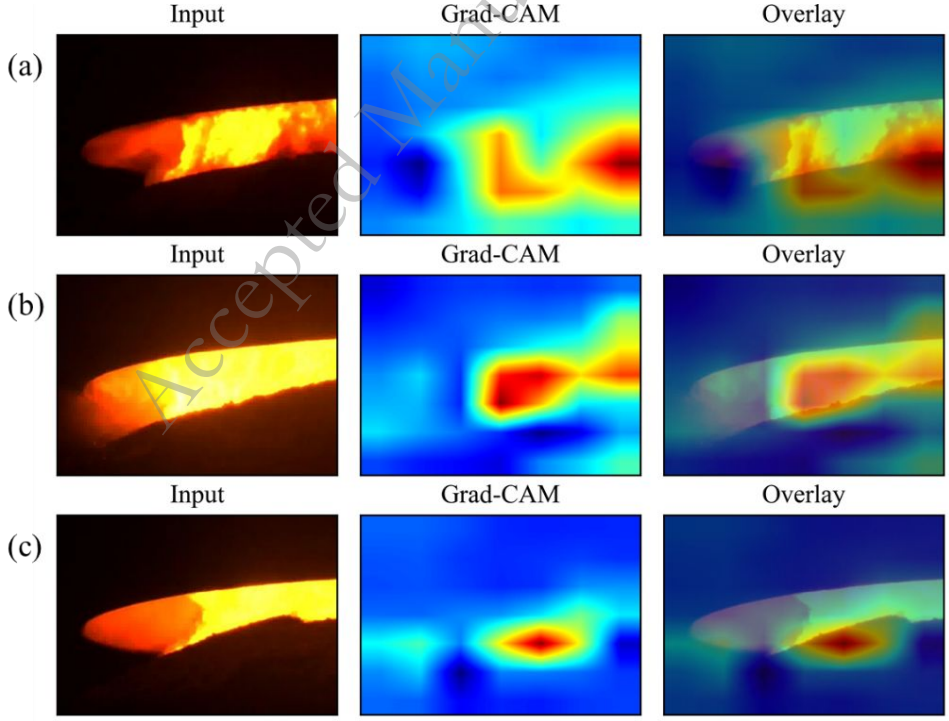

Fig. 12. Gradient-weighted class activation mapping (Grad-CAM) heatmaps of feature importance for predicting the flame state finding. (a) early-stage, (b) mid-stage and (c) late-stage.

We utilized the Grad-CAM algorithm to output the gradient heatmap of weights in the final convolutional layer and visualized the network model. Figure 12 displays the class activation mapping images of flame images from different stages extracted by Resnet, with (a)-(c) representing the visualizations and corresponding original images for the early, middle, and late stages, respectively. Regions with more intense red color in the visualizations indicate that those features play a more critical role in the category-specific direction. As illustrated in Figure 12(a), the model predominantly focuses on the relatively dark red regions of the flame, effectively capturing a comprehensive contour of the flame. Figure 12(b) demonstrates an increased emphasis on the brighter areas of the flame, indicating heightened sensitivity in feature extraction within regions of similar coloration. Additionally, Figure 12(c) reveals that the model concentrates on the corners of the flame to assess whether the area has contracted. The features extracted from the model have a certain correlation with the above flame features, which further proves the effectiveness of the method.

4.3.3 Recurrent layer and attention mechanism comparison

We considered network variant structures with different recurrent layers and compared them with Resnet18. The results indicate that the CRNN model exhibits higher accuracy in the task of converter flame recognition. The reason lies in the fact that Resnet18 inherently classifies single frames of images without incorporating dynamic information between adjacent frames, thus having fewer features. In contrast, the CRNN model not only extracts features from the image spatial dimension but also uses recurrent layers to extract temporal information from multiple frames of images. Therefore, it is more conducive to the recognition of converter smelting stage.

Table 1 summarizes the comprehensive performance of different models. The results indicate that Resnet18_BiLSTM exhibits the optimal recognition performance. GRU, as a mainstream variant of LSTM, was applied in this experiment. From the table, it is evident that GRU has fewer parameters and recognition time compared to LSTM, and when the convolutional layer uses Resnet18, it achieves almost the same accuracy as LSTM. However, as the depth of the convolutional network increases, the average accuracy of GRU is lower than that of LSTM. Additionally, it can be observed that both GRU and LSTM exhibit higher generalization compared to RNN and are capable of addressing long-term dependency problems. Furthermore, using bidirectional recurrent layers results in a higher recognition rate than unidirectional ones. In Figure 13, we have plotted the accuracy change curve of the CRNN models with different attention mechanisms on the validation set. Similarly, CBAM-block has a significant effect on improving accuracy.

Table 1. The recognition accuracy and performance of different models for classification test.

Algorithms		Accuracy (%)	Average recognition time (ms)	FLOPs (M)	Params (M)
Resnet18		90.62±0.54	2.55	1823.52	11.18
LSTM	Unidirectional	93.75±0.82	3.20	18251.03	12.75
	Bidirectional	94.56±0.27	3.88	18269.46	14.59
GRU	Unidirectional	94.29±0.30	3.27	18247.74	12.42
	Bidirectional	94.56±0.28	3.39	18261.57	13.81
RNN	Unidirectional	92.66±1.63	3.33	18241.13	11.77
	Bidirectional	92.93±1.36	3.58	18245.74	12.23

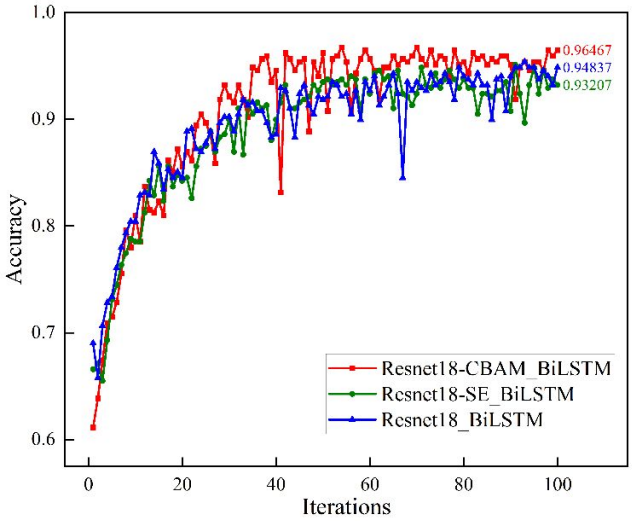

Fig. 13. The accuracy curve of CRNN models with different attention mechanisms in validated sets.

4.3.4 Optimization based on BO-CBAM-CRNN

For the hyperparameter optimization of feature extraction in the convolutional layer (latent_dim) and the recurrent layer (hidden_size), we employed the TPE optimization algorithm in this experiment to find the parameter combination corresponding to the optimal recognition accuracy. The ranges of the hyperparameters and the search results are provided in Table 2. Since the single training time of the CRNN for the converter flame video recognition problem is relatively long, we adopted the Early Stopping technique on this basis to reduce unnecessary iterations. The early stopping condition was set as ending the training round if the loss does not minimize within 15 iterations.

Table 2. Hyperparameter range domain and search results

Hyperparameter	Range domain	Results
latent_dim	[100, 512]	367
hidden_dim	[32, 256]	42

Table 3. The recognition accuracy, precision, recall, F1-score and AUC of different algorithms for classification test in each blowing period.

Algorithms	Evaluation metric					
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	micro-average AUC (%)	macro-average AUC (%)
CRNN	96.46	96.44	96.45	96.46	99.52	99.55
BO-CRNN	97.01 **(↑0.55)**	97.02 **(↑0.58)**	97.01 **(↑0.56)**	97.00 **(↑0.54)**	99.85 **(↑0.33)**	99.85 **(↑0.30)**

In Table 3, we compare the evaluation metrics of the BO-CRNN model optimized through Bayesian optimization with the benchmark model (Resnet18-CBAM_BiLSTM). The experiment shows a significant improvement in the performance of the optimized model, with accuracy, precision, recall, F1 score, micro-average AUC, and macro-average AUC increasing by 0.55%, 0.58%, 0.56%, 0.54%, 0.33%, and 0.30%, respectively. Additionally, in Table 3, we compare the comprehensive recognition performance of the optimized model and the benchmark model in

terms of the number of parameters and computational complexity. Similarly, the experiment indicates that the optimized hyperparameter combination has a reduced average recognition time. Finally, we present the confusion matrix and ROC curve of the BO-CRNN in Figure 14.

Table 4. Comparative analysis of comprehensive performance across different models.

Algorithms	Average recognition time (ms)	FLOPs (M)	Params (M)
CRNN	5.59	18290.75	14.68
BO-CRNN	5.49 **(↓0.10)**	18260.21 **(↓30.54)**	11.58 **(↓3.10)**

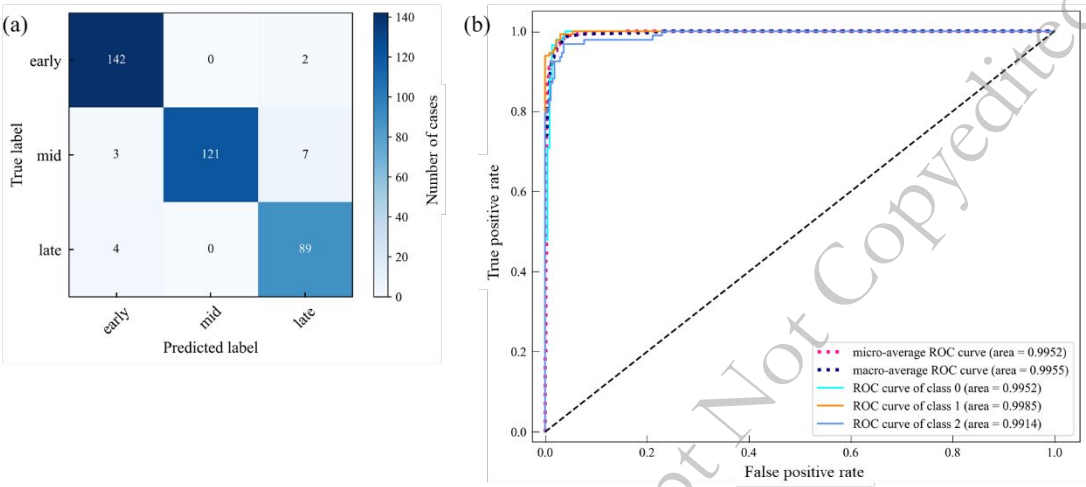

Fig. 14. Performance of BO-CRNN on validated on vaildated sets: (a) confusion matrix of algorithom predictions and (b) receiver operating characteristic (ROC) curve.

5. Conclusion

This paper proposes a converter smelting stage recognition model based on a CRNN. The model collects images of the mouth flame under various stages during the steelmaking process, utilizing the ResNet-CBAM module for spatial feature extraction. LSTM network is then employed to capture the flame's time-series characteristics, culminating in a Softmax classifier that predicts the current smelting stage. At the same time, hyperparameter optimization is achieved using the Bayesian optimization algorithm. Repeated experiments on a large dataset validate the model's generalization and robustness capabilities. The main achievements of this research are summarized as follows:

(a) A time-series modeling approach is used to consider the feature representation of the converter mouth flame in both the spatial and temporal domains, overcoming the limitations of traditional single-frame image recognition methods in terms of poor noise resistance.

(b) A comprehensive comparison of the impact of four convolutional networks (Resnet18, Densenet121, Transformer, and VGG16) and attention mechanisms on the model's recognition accuracy is conducted, with Resnet18-CBAM ultimately selected as the optimal convolutional layer algorithm.

(c) Through the Grad-CAM method, the convolutional feature layer is visually analyzed, further proving the rationality and effectiveness of the proposed model.

(d) By comparing different convolutional and recurrent layer structures and integrating Bayesian optimization, the proposed BO-CRNN model achieves the highest accuracy in recognizing converter smelting stage (97.01%), demonstrating strong robustness and generalization capabilities. It holds promise for practical industrial applications.

The CRNN model proposed in this paper offers a novel solution for the stage recognition in the converter steelmaking process. In future work, we plan to further incorporate converter endpoint control to predict the carbon content and temperature of the molten steel under different smelting stage.

Acknowledge

This work was supported by the National Natural Science Foundation of China (52374320).

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] K. Zhü, O. Cooper, S.-L. Yang, and Q.-X. Dong, An extension of the AHP dummy pivot modeling applied to the restructuring of the iron and steel industry in China, *IEEE Trans. Eng. Manag.*, 61(2014), No. 2, p. 370-380.

[2] R.H. Zhang and J. Yang, State of the art in applications of machine learning in steelmaking process modeling, *Int. J. Miner. Metall. Mater.*, 30(2023), No. 11, pp. 2055-2075.

[3] L. Zeng, Z. Zheng, X.Y. Lian, et al., Intelligent optimization method for the dynamic scheduling of hot metal ladles of one-ladle technology on ironmaking and steelmaking interface in steel plants, *Int. J. Miner. Metall. Mater.*, 30(2023), No. 9, pp. 1729-1739.

[4] Z.C. Xin, J.S. Zhang, K.X. Peng, et al., Explainable machine learning model for predicting molten steel temperature in the LF refining process, *Int. J. Miner. Metall. Mater.*, 31(2024), No. 12, pp. 2657-2669.

[5] Y. Han, C.J. Zhang, L. Wang and Y.C. Zhang, Industrial IoT for intelligent steelmaking with converter mouth flame spectrum information processed by deep learning, *IEEE Trans. Industr. Inform*., 16(2020), No. 4, p. 2640-2650.

[6] Z. Bai, G.-B. Huang, D. Wang, and M.B. Westover, Sparse extreme learning machine for classification, *IEEE Trans. Cybern*., 44(2014), No. 10, p. 1858-1870.

[7] B. Zhou, A. Khosla, A. Lapedriza, and A. Torralba, Learning deep features for discriminative localization, [in] *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2921-2929.

[8] X. Zhao, Y. Wu, G. Song, et al., A deep learning model integrating FCNNs and CRFs for brain tumor segmentation, *Med. Image Anal*., 43(2018), p. 98-111.

[9] J. Zhai and D. Song, Optimal instance subset selection from big data using genetic algorithm and open source framework, *J. Big Data*, 9(2022), No. 1, p. 87.

[10] T. Haque, R.T. Yazicigil, K.J.-L. Pan, and P.R. Kinget, Theory and design of a quadrature analog-to-information converter for energy-efficient wideband spectrum sensing, *IEEE Trans. Circuits*

Syst. I: Regul. Pap., 62(2014), No. 2, p. 527-535.
[11] H. Zhao and C. Zhang, An online-learning-based evolutionary many-objective algorithm, *Inf. Sci.*, 509(2020), p. 1-21.
[12] M.K. Ghalati, J. Zhang, G.M.A.M. El-Fallah, B. Nenchev, and H. Dong, Toward learning steelmaking—A review on machine learning for basic oxygen furnace process, *MGE Advances*, 1(2023), No. 1, p. e6.
[13] C.J. Zhang, Y.C. Zhang, and Y. Han, Industrial cyber-physical system driven intelligent prediction model for converter end carbon content in steelmaking plants, *J. Ind. Inf. Integration.*, 28(2022), p. 100356.
[14] B. Zhao, J. Zhao, W. Wu, et al., Research on prediction model of converter temperature and carbon content based on spectral feature extraction, *Sci. Rep*., 13(2023), No. 1, p. 14409.
[15] M. Zhou, Q. Zhao, and Y. Chen, Endpoint prediction of BOF by flame spectrum and furnace mouth image based on fuzzy support vector machine, *Optik*, 178(2019), pp. 575-581.
[16] K. Sun and Y. Zhu, A blowing endpoint judgment method for converter steelmaking based on improved DenseNet, [in] *2022 34th Chinese Control and Decision Conference (CCDC)*, Hefei, China, 2022, pp. 3839-3844.
[17] H. Liu, B. Wang, and X. Xiong, Basic oxygen furnace steelmaking end-point prediction based on computer vision and general regression neural network, *Optik*, 125(2014), No. 18, p. 5241-5248.
[18] H. Liu, Q. Wu, B. Wang, and X. Xiong, BOF steelmaking endpoint real-time recognition based on flame multi-scale color difference histogram features weighted fusion method, [in] *2016 35th Chinese Control Conference (CCC)*, Chengdu, China, 2016, pp. 3659-3663.
[19] U. Chadha, S.K. Selvaraj, and A. Raj, RETRACTED: AI-driven techniques for controlling the metal melting production: a review, processes, enabling technologies, solutions, and research challenges, *Mater. Res. Express*, 9(2022), No. 7, p. 072001.
[20] Z. Niu, H. Qi, and A. Sun, Efficient and robust CNN-LSTM prediction of flame temperature aided light field online tomography, *Sci. China Technol. Sci.*, 67(2024), pp. 271-284.
[21] X. Lu, M. He, Z. Wang, et al., Prediction of flashover time in a compartment fire by CNN-LSTM based deep neural network considering wearable data collection equipment, *J. Build. Eng.*, 97(2024), p. 110719.
[22] X. Huang, X. Hao, B. Pan, and P. Pei, Combustion field prediction and diagnosis via spatiotemporal discrete U-ConvLSTM model, *IEEE Photon. J.*, 16(2024), No. 2, pp. 1-10.
[23] A. Carreon, S. Barwey, and V. Raman, A generative adversarial network (GAN) approach to creating synthetic flame images from experimental data, *Energy AI*, 13(2023), p. 100238.
[24] T. Hai, M. Ashraf Ali, J. Zhou, et al., Feasibility and environmental assessments of a biomass gasification-based cycle next to optimization of its performance using artificial intelligence machine learning methods, *Fuel*, 334(2023), p. 126494.
[25] Y. Chen, J. Liu, and H. Xiong, Two-Stage recognition algorithm for untrimmed converter steelmaking flame video, [in] *Pattern Recognition and Computer Vision: 4th Chinese Conference (PRCV)*, Beijing, China, 2021, pp. 268-279.
[26] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, [in] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.
[27] X. Shu, L. Zhang, and Y. Sun, Host–parasite: Graph LSTM-in-LSTM for group activity recognition, *IEEE Trans. Neural Netw. Learn. Syst.*, 32(2020), No. 2, pp. 663-674.

[28] Y. Liu, A. Pei, F. Wang, et al., An attention-based category-aware GRU model for the next POI recommendation, *Int. J. Intell. Syst.*, 36(2021), No. 7, pp. 3174-3189.
[29] S. Song, S. Zhang, W. Dong, et al., Multi-source information fusion meta-learning network with convolutional block attention module for bearing fault diagnosis under limited dataset, *Struct. Health Monit.*, 23(2024), No. 2, pp. 818-835.
[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, CBAM: Convolutional block attention module, [in] *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 3-19.
[31] Y. Liang, Y. Lin, and Q. Lu, Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM, *Expert Syst. Appl.*, 206(2022), p. 117847.
[32] R.R. Selvaraju, M. Cogswell, A. Das, and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, [in] *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618-626.
[33] C. Van Zyl, X. Ye, and R. Naidoo, Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP, *Appl. Energy*, 353(2024), p. 122079.