Prediction of Mechanical Properties of Cold Rolled Steel Using Fusion of Multi-Head Attention

Qiwen Zhang, and Rongping Guo

***Abstract*—Cold rolling of steel is a complex nonlinear process, and there exist intricate spatiotemporal correlations among its process parameters, chemical composition, and mechanical properties. However, traditional long short-term memory networks cannot distinguish the importance of different input features and often only consider the temporal information during feature extraction, thus neglecting the spatial associations of the feature data. To better extract features from cold-rolled steel data and enhance the model's prediction accuracy, in this paper a fusion of improved multi-head attention modules is proposed. These modules enable the adaptive allocation of weights to different influencing factors, allowing the accurate distinction of the importance of different features. Additionally, a feed-forward neural network is introduced into the memory to further explore and extract the spatiotemporal features of the data. Experimental results on an industrial dataset demonstrate that the proposed model outperforms five other advanced models in terms of prediction performance.**

***Index Terms*—spatiotemporal correlation, mechanical properties, long short-term memory, multi-head attention, feed-forward neural network**

I. Introduction

THE carbon steel cold rolling process is a highly nonlinear, unknown, and complex time-varying system that is difficult to model mathematically. Establishing the exact relationship between process parameters, chemical composition, and mechanical properties is a challenging task. Therefore, the prior prediction of the mechanical properties of cold-rolled steel is of great practical and research significance for reducing production costs and improving product quality.

Currently, mechanical properties are primarily predicted using metallurgical mechanism models, machine learning, and deep learning methods. While metallurgical mechanism models have a solid theoretical foundation, they have complex structures and rely on numerous dependencies, requiring detailed analysis and tedious calculations [1]. In recent years, machine learning models have been widely applied in the field of predicting mechanical properties [2]. Zhao et al. [3] addressed the issue of unstable prediction accuracy for multiple steel grades through the adoption of random forest and factor analysis to achieve feature dimension reduction and decoupled processing. Cheng et al. [4] focused on tensile strength and constructed a mechanical property prediction model based on principal component analysis of the feature data and gradient-boosted decision trees. Shi et al. [5] utilized Bayesian optimization to optimize the hyperparameters of a mechanical property prediction model based on Bo-XGBoost, and the generated model exhibited excellent generalization capability.

Although machine learning-based methods have led to certain achievements, the process parameters of the evolving techniques adopted by the steel industry exhibit multivariate and strongly nonlinear characteristics. Moreover, the relationships between input parameters have become more complex. Therefore, machine learning-based methods often fail to meet the accuracy requirements when establishing mechanical property prediction models.

Long short-term memory (LSTM) [6] has demonstrated excellent performance and effectiveness in numerous prediction fields due to their ability to handle complex nonlinear problems and large volumes of data [7]-[8]. Marani et al. [9] designed an LSTM model with two layers and eight hidden units for predicting tool flank wear during machining. Sagheer et al. [10] proposed a deep LSTM network and optimized its structure using a genetic algorithm. They then applied their model on nonlinear prediction problems in the petroleum industry. Wang et al. [11] proposed an LSTM model based on a quadruple loss function, combining the squared error loss function with distance metric learning between sample features. This model was used to predict soil temperature on different dates. Li et al. [12] introduced evolutionary attention into LSTM through parameter sharing and trained the model using a competitive random search strategy. This approach enabled multivariate time series prediction. Ding et al. [13] combined an LSTM with a dynamic attention mechanism in both temporal and spatial dimensions to model and perform interpretable analysis on flood prediction. Wei et al. [14] proposed a traffic flow prediction method combining an autoencoder with an LSTM. The autoencoder was used to extract features from traffic flow data and capture their internal relationships, while the LSTM utilized the extracted features and historical data to predict complex linear traffic flow patterns. The results demonstrated that this method exhibited good stability and higher prediction accuracy. Xu et al. [15] utilized an LSTM, a

Manuscript received July 14, 2023; revised November 8, 2023.

This work was supported in part by the National Natural Science Foundation of China 62162040, 62063021.

Qiwen Zhang is an associate professor of School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China (e-mail: 823869941@qq.com).

Rongping Guo is a postgraduate student of School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China (corresponding author to provide phone: +86-17694834825; e-mail: 212085400220@lut.edu.cn).

gated recurrent unit (GRU), and Gaussian Process Regression models to predict the mechanical properties of steel. They discussed the prediction accuracy and learning efficiency of different models and proposed an online relearning method for transfer learning models.

As demonstrated from the above analysis of the different application examples, it is evident that existing LSTM models have achieved good results in diverse areas such as humidity prediction and flood forecasting. However, for the prediction of the mechanical properties of cold-rolled steel, LSTM faces challenges in distinguishing the importance of different chemical compositions and process parameters on mechanical performance effectively. Additionally, LSTM is unable to handle spatial and temporal features among the input data simultaneously. To address this issue, in this paper a Multi-Head Attention Feed Forward LSTM (MHA-FLSTM) is proposed that combines multiple attention heads, enabling high-precision prediction of cold-rolled steel mechanical properties.

II. Theoretical Foundation

A. Long Short-Term Memory Network

The LSTM was proposed by Schmidhuber in 1997 to address the vanishing gradient problem in Recurrent Neural Networks (RNNs). It replaced the memory unit in RNNs with a storage unit utilizing a gating mechanism composed of input, forget, and output gates. The structure of the LSTM model is illustrated in Fig. 1 Each LSTM unit primarily consists of three stages. The forget stage is the first stage, where using the first σ unit in Fig. 1, the forget factor f_t of the forget gate is calculated at time *t*, thus determining which values from the previous state layer c_{t-1} need to be forgotten. The second stage is the selective retention stage, which uses the second σ unit in Fig. 1, to calculate the value i_t of the input gate, which determines which values need to be updated. The *tanh* unit calculates the candidate state vector $\tilde{c}_t$ at time *t*. Then, the old state c_{t-1} and the current candidate state $\tilde{s}_t$ are selectively or retained, respectively. The two are then added to obtain the updated value of the state layer c_t at time *t*.

The final stage is the output stage, which uses the third σ unit in Fig. 1 to calculate the value o_t of the output gate at time *t*, thus determining which values need to be output from the cell state. c_t is compressed using the *tanh* function. The result of this process is the hidden layer value h_t at the current time step. The equations for the above calculations as below:

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

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$\tilde{c}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \tag{4}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t \otimes \tanh\left(c_t\right) \tag{6}$$

where W_f, W_i, W_c and W_o represent the weight parameters for the forget, input, and output gates, and the candidate state, respectively, while b_f, b_i, b_c and b_o represent the corresponding bias terms.

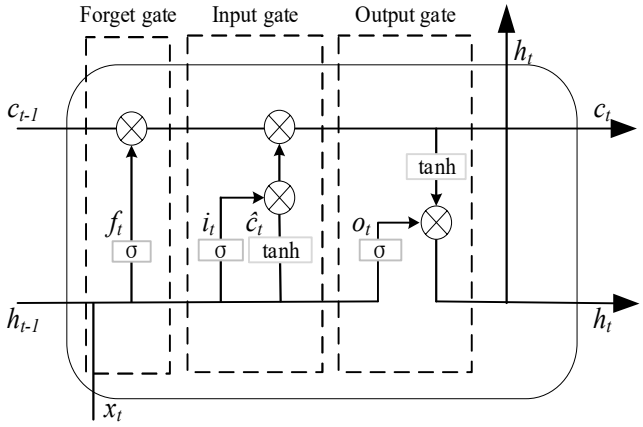


Fig. 1. Basic LSTM structure.

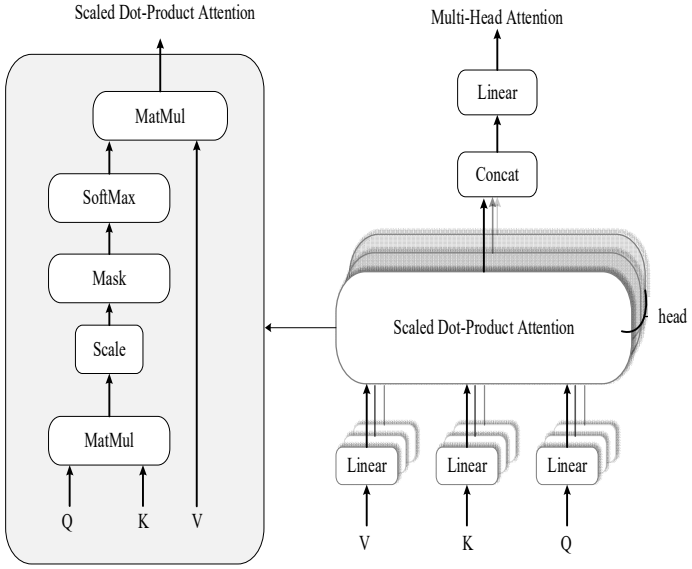


Fig. 2. The basic MHA mechanism.

B. Multi-head Attention

The multi-head attention (MHA) mechanism was proposed by Vaswani et al. [16] in 2017, and is essentially a combination of multiple self-attention mechanisms. The MHA mechanism allows the model to utilize different attention mechanisms on different attention heads, and thus feature interactions in different feature subspaces. Its structure is schematically shown in Fig. 2, where *Q*, *K*, and *V* represent the query, key, and value, respectively. After applying linear transformations to *Q*, *K*, and *V*, the transformed results are split into multiple heads, on which independent attention mechanisms operate. The calculation results of the multiple sub-attention mechanisms are then concatenated and transformed through linear transformations to obtain the output of the MHA. The functional expression is as follows:

$$Attention\left(Q, K, V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

$$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (8)$$

III. Enhanced Multi-Head Attention

The mechanical properties of cold-rolled steel depend on its microstructure, which, in turn, is intricately and non-linearly affected by the composition and processing of the steel. There exists a complex mapping relationship among chemical compositions, process parameters, and mechanical properties, wherein distinct process parameters and chemical compositions exhibit varying degrees of influence on mechanical performance. In order to predict the mechanical properties of cold-rolled steel effectively, it is crucial to represent the importance of different chemical compositions and process parameters accurately. However, LSTM networks cannot distinguish the importance of multiple input features accurately, as they treat all input features in the same manner and fail to capture the correlations and variations in influence among them. As a result, the accuracy and reliability of conventional LSTM models are limited.

To address the aforementioned issues, inspired by GAT [17], a feature-enhanced MHA mechanism was designed, where the input and output share the same dimensions, while the output contains richer information. In the task of predicting the mechanical properties of cold-rolled steel, there exist both positive and negative correlations among input features. Therefore, the LeakyReLU activation function is applied to the output features of the attention mechanism before softmax, allowing both positive and negative correlations to be reflected during the weight update process. This interaction helps capture the complex relationships among input features from a global perspective, resulting in output features with more comprehensive information. The attention mechanism also enables the adaptive allocation of weights to different influencing factors, selectively focusing on key features within the input and assisting the model in creating more accurate representations of the complex relationships embedded in the input data.

Suppose there are n inputs, $X_T = (X_T^1, X_T^2, ..., X_T^n)$, each of length T. The inputs are transformed linearly to obtain the query, key, and value matrices Q, K, and V, respectively.

$$Q = W_q \bullet X_T^i$$
$$K = W_k \bullet X_T^i \quad (9)$$
$$V = W_v \bullet X_T^i$$

where $W_q \in \mathbb{R}^{d_k}$, $W_k \in \mathbb{R}^{d_k}$ and $W_v \in \mathbb{R}^{d_v}$ represent the weights of the linear mapping. Each head performs a scaled dot-product operation on Q and K, followed by applying the LeakyReLU activation function to the output features of the attention mechanism before softmax. This yields the weighted output of matrix V. Fig. 3 illustrates the computation process of the enhanced MHA. The attention calculation equation is updated from Equation (7) to the following:

$$Attention(Q,K,V) = \frac{\exp\left(Leaky\,\mathrm{Re}\,LU\left(QK^T\right)\right)}{\exp\left(Leaky\,\mathrm{Re}\,LU\left(\sqrt{d_k}\right)\right)}V \quad (10)$$

Finally, each head concatenates the output vectors weighted by self-attention, forming the final output of the MHA, which can be represented as follows:

$$y_{mha} = Concat(head_1,...,head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (11)$$

where for the weight matrices, we have $W_i^Q \in \mathbb{R}^{d_n \times d_k}$, $W_i^V \in \mathbb{R}^{d_n \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_n}$.

To further demonstrate the ability of the MHA mechanism to differentiate the importance of different input features, visual analysis was conducted on both the original sample data and the MHA output. Fig. 4 is a visualization of the two forms. Ten random sample data points were selected for analysis; the x-axis represents the fourteen different input features, and the y-axis represents the ten sample data points. The color variations reflect the differences between input the features. From the figure, it can be seen that while there are some differences between different input features in the original sample data, the differences are relatively subtle. However, after passing through the MHA mechanism, the transformed data capture the correlations among input features more clearly, resulting in richer feature representations and more pronounced differences.

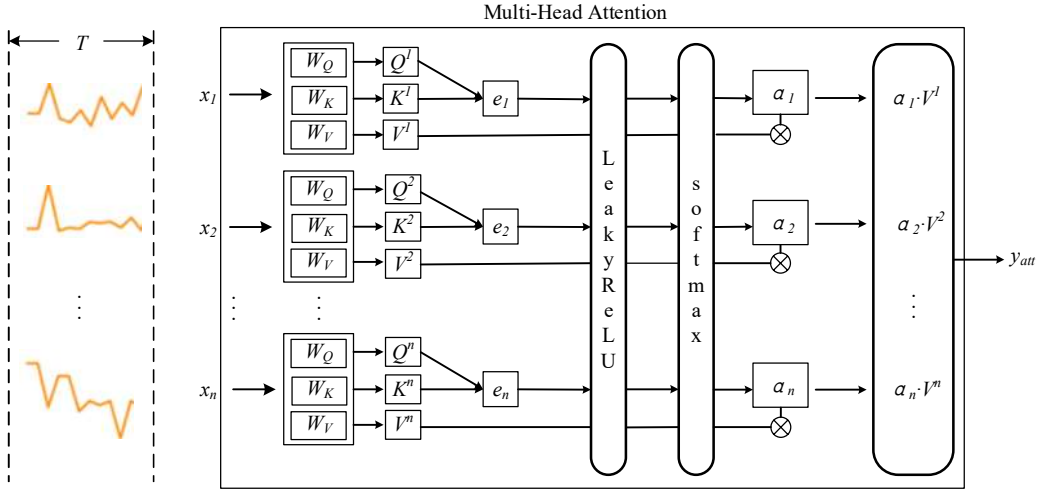
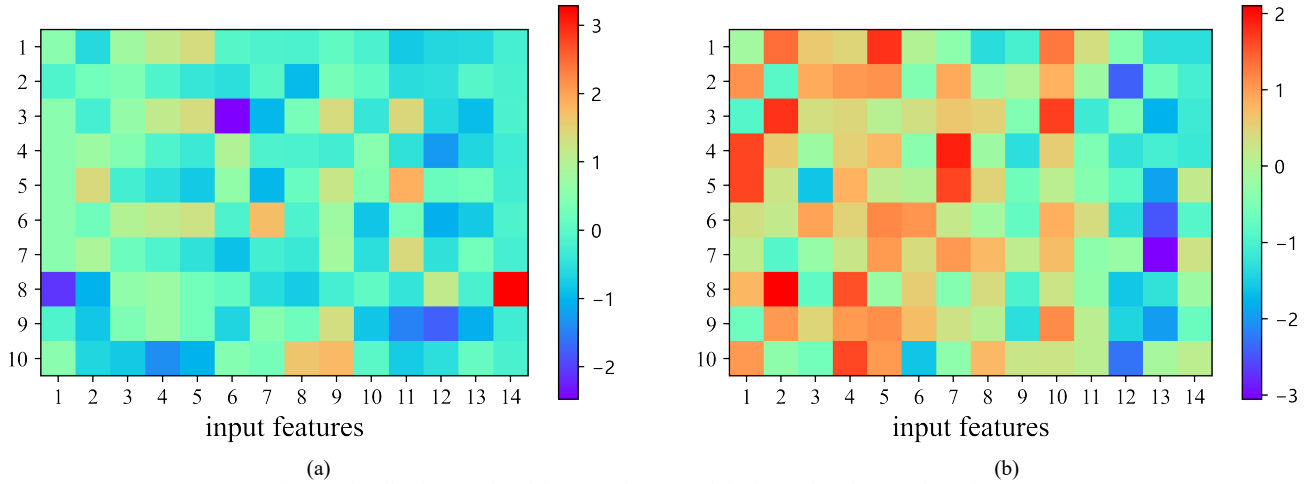


Fig. 3. The calculation process of the improved MHA.



(a) (b)

Fig. 4. Visualization results of data samples: (a) Original samples, (b) Samples using MHA.

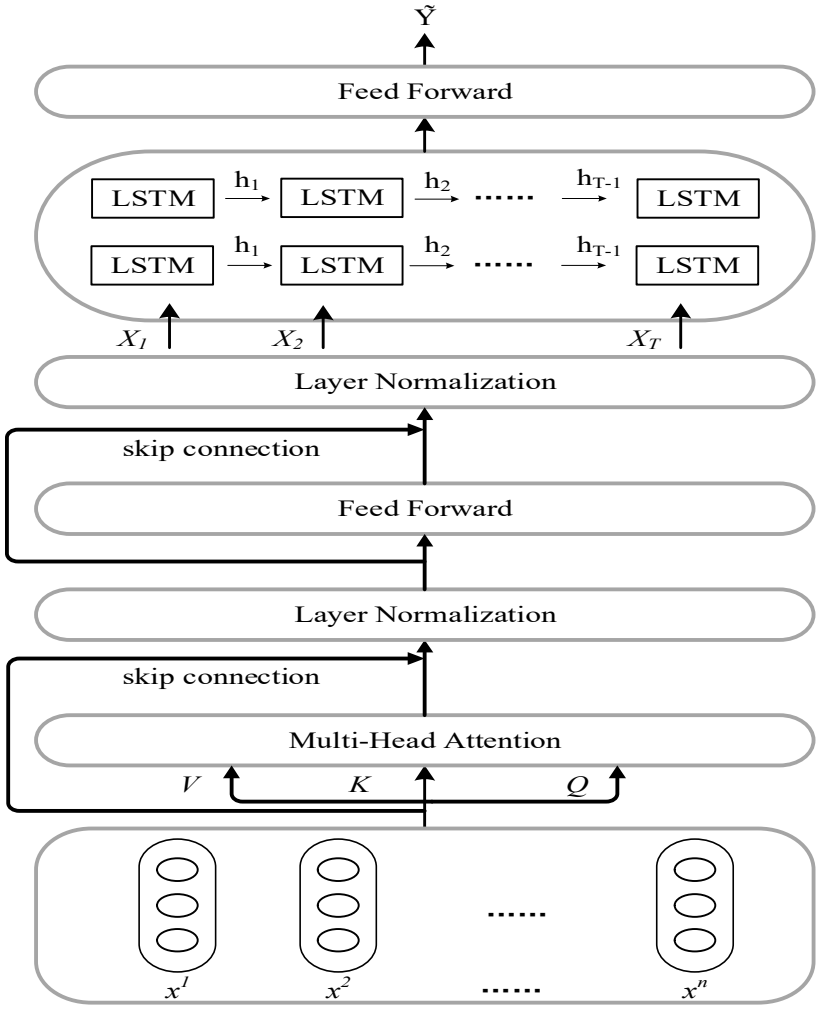


Fig. 5. MHA-FLSTM model structure.

IV. MHA-FLSTM MODEL

Due to the complex spatiotemporal correlations among different process parameters and chemical compositions, LSTM models often focus only on capturing temporal information during the feature extraction stage while neglecting spatial information. This results in a relatively weak spatial feature extraction capability. To address this, a feed-forward neural network (FFNN) is introduced before and after LSTM. The FFNN implicitly extracts spatial features through non-linear transformations and mapping operations between layers. Additionally, to prevent difficulties in gradient calculation caused by the increased network depth and to retain more global information, skip connections were introduced between the first FFNN, the MHA, and the LSTM. This multi-level feature representation helps the network capture complex patterns and relationships in the input data more effectively, thereby improving the predictive performance of the network.

First, an output vector $y_{mha} = (y_1, y_2, ..., y_n)$ with the same dimension as the input is obtained using the MHA mechanism. Then, the original input X_T and the output vector y_{mha} pass through a skip connection and are input to a batch normalization layer, which helps the network better adapt to different input distributions, improves its generalization capability, and adjusts the mean and variance of the input data to be consistent, thereby accelerating model convergence. Subsequently, the input is fed into the input layer of the FFNN and connected to the hidden layer neurons through weight parameters. The output of the FFNN is obtained through the application of a ReLU activation function, thus yielding the final result. The equation is as follows:

$$F(X) = \mathrm{Re}\,LU\left(W_f\left(LayerNorm(X_T + y_{mha})\right) + b_f\right) \quad (12)$$

Finally, a stacked LSTM with two layers is employed to extract the complete features from the input data and obtain the final prediction of the mechanical properties. The basic structure of the LSTM is described in Section II.A. The specific steps are as follows: First, the integrated new input features are passed to the LSTM, which then calculates the hidden states. Next, the hidden states of the LSTM are fed into a two-layer feed-forward layer. The ReLU non-linear activation function is applied in the first layer to further enhance the model's non-linear expressive ability. The final prediction result can be expressed as follows:

$$\tilde{Y}_k = LSTM\left(\vec{H}\right), H = [H_1, H_2, ..., H_T], k \in \{1, ..., T\} \quad (13)$$

Using this approach, the improved MHA mechanism from Section III and the FFNN-based LSTM model were combined to formulate the final prediction model, namely MHA-FLSTM. Benefiting from the improved MHA mechanism, which allows for feature interactions from different perspectives, MHA-FLSTM addresses the issue of the LSTM's inability to encompass the importance of different features. Through the incorporation of feed-forward layers before and after the LSTM, MHA-FLSTM extracts spatio-temporal features from the data effectively. Additionally, the adoption of the skip connections helps preserve more global information. The MHA-FLSTM achieves high-precision prediction of the mechanical properties of cold-rolled steel. The overall structure of the model is depicted in Fig. 5.

V. Experimental Results and Analysis

A. Cold-rolled steel dataset

The dataset used in this article was collected from a commercial cold-rolled steel plate production line at JISCO Carbon Steel Sheet Factory. The dataset included the mechanical properties of cold-rolled steel under different process parameters and chemical compositions. After handling missing and abnormal data, a total of 13,485 samples were selected. The training data accounted for 80% of the total and consisted of 10,788 data samples, while the testing data accounted for 20% and consisted of 2,697 data samples. Each data sample included 5 process parameters, 9 chemical compositions, and 3 major mechanical performance indicators: Yield Strength (YS), Tensile Strength (TS), and Elongation (EL). The names and descriptions of the chemical compositions and process parameters are shown in Table I.

TABLE I
Input Feature Description of Cold-rolled Steel Plates

Variable	Description	Variable	Description
C	mass fraction of carbon element	Als	mass fraction of Als element
Si	mass fraction of silicon element	Cu	mass fraction of copper element
Mn	mass fraction of manganese element	Al	mass fraction of aluminum element
P	mass fraction of phosphorus element	Ni	mass fraction of nickel element
S	mass fraction of sulfur element	FT1	rolling temperature
FT2	final rolling temperature	FT3	cold-rolled steel thickness
FT4	cold-rolled steel thickness	FT5	leveling machine elongation

B. MHA-FLSTM Hyperparameter Settings and Evaluation Metrics

The MHA-FLSTM model has four key parameters: the time step of the sliding window (T), the number of hidden neurons in the LSTM (H), the number of layers in the network (n), and the number of heads in the MHA mechanism (h). The values of the above key parameters and other hyperparameters were determined through experimentation as shown in Table II. In the MHA-FLSTM model, the MHA mechanism is crucial for improving model performance and the number of heads is a key hyperparameter for accelerating convergence speed and enhancing model accuracy. To investigate the impact of the number of heads in the MHA mechanism on model accuracy and training time and ensure a fair comparison between different configurations, the other parameters of the network were fixed as the number of heads varied. Since the input features need to be evenly distributed among each head, the number of heads must be divisible by the number of input features. In this study, 14 features were selected as input features, so the numbers of heads tested were 1, 2, 7, and 14.

TABLE II
Hyperparameter Settings for MHA-FLSTM

Parameter	Values	Parameter	Values
T	2	h	7
H	64	learning rate	0.01
n	2	Epoch	800

Table III presents the prediction results for different numbers of heads. The experimental results demonstrate that the multi-head structure of the model outperforms the single-head structure in both evaluation metrics. However, as the number of heads in the MHA mechanism increases, the training time also increases. When the number of heads reaches 7, the model achieves its optimal performance in predicting all three mechanical performance indicators. However, when the number of heads increases to 14, the model's performance starts to decline, which indicates that the model's performance does not always improve from an increased number of heads. The main reason for this may be that the attention mechanism focuses more on the input features, leading the model to learn redundant information, which results in a decrease in prediction performance.

To assess the effectiveness of various prediction methods, four different evaluation metrics were adopted in this study: the root mean square error (RMSE), the mean squared error

(MSE), the mean absolute error (MAE), and R2. Let y_t, $\hat{y}$, and $\bar{y}$ be the target, predicted, and mean target values. Then, the definitions of these metrics are as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_t^{\,i} - \hat{y}_t^{\,i})^2} \tag{14}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_t^{\,i} - \hat{y}_t^{\,i}\right)^2 \tag{15}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_t^{\,i} - \hat{y}_t^{\,i}\right| \tag{16}$$

$$R^2 = 1 - \frac{\sum_i\left(\hat{y}_t^{\,i} - y_t^{\,i}\right)^2}{\sum_i\left(y_t^{\,i} - \bar{y}\right)^2} \tag{17}$$

Among these metrics, smaller RMSE, MSE, or MAE values indicate that the predicted values are closer to the true ones, indicating better model performance. An R^2 value of closer to 1 indicates better model performance.

C. Results and Comparison

To demonstrate the effectiveness of the MHA-FLSTM, the model was compared with five other well-performing models, including a plain LSTM, a GRU [18], which is an improvement over LSTM, the Attention-based Bidirectional LSTM Network (ATT-BiLSTM)[19], a 1D-CNN and the Convolutional LSTM Network (CNN-LSTM)[20]. Table IV lists the key parameters of these models. To ensure fairness, all the LSTM-related comparative models were set with the same window size.

Table V shows the experimental results of the MHA-FLSTM model and the other five compared models on the test dataset. For YS, MHA-FLSTM achieved a decrease in MSE of 51.04%, 42.16%, 41.97%, 18.69%, and 26.09% compared to LSTM, ATT-BiLSTM, GRU, CNN, and CNN-LSTM, respectively. For TS, MHA-FLSTM achieved a decrease in MSE of 55.19%, 42.40%, 62.34%, 48.47%, and 45.15% compared to the respective models mentioned above. For the prediction results of EL, MHA-FLSTM achieved a decrease in MSE of 55.93%, 42.86%, 55.93%, 53.28%, and 54.12% compared to the respective models.

Furthermore, the MHA-FLSTM obtained the best results in the remaining three evaluation metrics compared to the other models of the comparison. This is because the MHA-FLSTM not only utilizes the MHA mechanism to provide reliable input features through the adaptive allocation of weights, but also enhances the model's spatial feature extraction and non-linear expression capabilities through the inclusion of skip connections and FFNN.

To visually demonstrate the effectiveness of the proposed model, Fig. 6 illustrates the model's fitting results for the three mechanical performance indicators against their corresponding true values. It can be observed that the predicted values of the proposed model closely align with the true ones, indicating a strong fitting capability and accurate capturing the underlying patterns in the data.

TABLE III
MODEL PREDICTIONS WITH DIFFERENT NUMBERS OF HEADS

Heads	YS			TS			EL		
	MSE	MAE	Time	MSE	MAE	Time	MSE	MAE	Time
1	31.887	4.071	**6.913**	18.233	3.116	**6.869**	0.907	0.696	**6.930**
2	20.727	3.120	7.0336	15.751	2.743	6.890	1.133	0.789	6.984
7	**16.775**	**4.096**	7.173	**11.980**	**2.288**	7.004	**0.684**	**0.587**	7.153
14	20.362	3.034	7.942	16.303	2.820	7.783	0.853	0.651	7.765

TABLE IV
PARAMETER SETTINGS OF COMPARISON MODELS

Models	Parameters	Models	Parameters
LSTM	T = 2, H = 64, n = 2 learning rate = 0.01 epoch = 800	CNN	conv1 = 32, kernel size =1, pool1 = 2 conv2 = 64, kernel size =1, pool2 = 2 conv3 = 128, kernel size =1, pool3 = 2 learning rate = 0.01, epoch = 800
ATT-BiLSTM	T = 2, H = 64, n = 2 learning rate = 0.01 epoch = 800	CNN-LSTM	conv1 = 32, kernel size = 1, pool1 = 2 conv2 = 64, kernel size = 1, pool2 = 2 conv3 = 128, kernel size = 1, pool3 = 2
GRU	T = 2, H = 64, n = 2 learning rate = 0.01 epoch = 800		T = 2, H = 64, n = 2 learning rate = 0.01 epoch = 800

TABLE V
EXPERIMENTAL RESULTS OF DIFFERENT MODELS COMPARED

Target value	Evaluation metric	Model					
		LSTM	ATT-BiLSTM	GRU	CNN	CNN-LSTM	MHA-FLSTM
YS	MSE	34.261	29.003	28.907	20.632	22.696	**16.775**
	RMSE	5.853	5.386	5.377	4.542	4.764	**4.096**
	R^2	0.785	0.795	0.805	0.857	0.831	**0.910**
	MAE	4.284	3.855	3.828	2.902	2.920	**2.583**
TS	MSE	26.736	20.799	31.808	23.248	21.841	**11.980**
	RMSE	5.171	4.561	5.640	4.822	4.673	**3.461**
	R^2	0.804	0.847	0.804	0.829	0.839	**0.935**
	MAE	3.833	3.300	3.833	3.699	3.554	**2.288**
EL	MSE	1.552	1.197	1.552	1.464	1.491	**0.684**
	RMSE	1.246	1.094	1.264	1.210	1.221	**0.827**
	R^2	0.879	0.906	0.879	0.886	0.884	**0.947**
	MAE	0.948	0.813	0.944	0.939	0.930	**0.587**

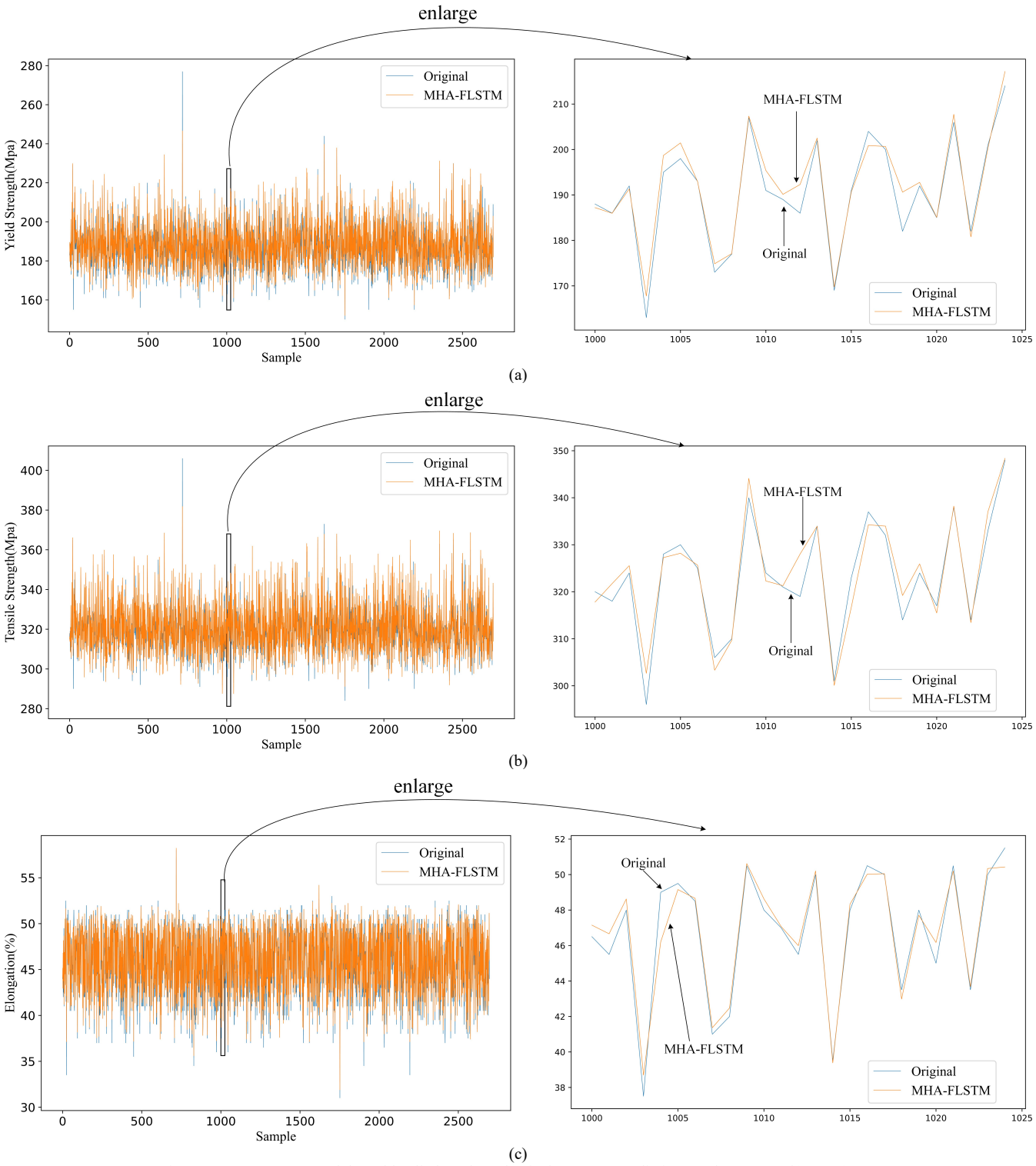


Fig. 6. MHA-FLSTM model's cold-rolled steel mechanical property prediction results: (a) YS (b) TS (c) EL.

D.Ablation Experiment

To demonstrate the importance of each module in the proposed model, the prediction results of MHA-FLSTM were compared with three other models: the MHA-based LSTM (MHA-LSTM), the feed-forward neural network-based LSTM (FLSTM), and plain LSTM. Table VI presents the prediction results of these four models.

The experimental results indicate that the improved MHA mechanism and FFNN contribute to further enhancing the predictive performance of the network. By allocating greater weights to key features of the input through the attention mechanism and enhancing the spatial feature extraction and non-linear expression capabilities of the LSTM model through the FFNN, the proposed model exhibits superior predictive performance compared to the other models. This comparison highlights the significance of each module and demonstrates their collective contribution to its enhanced predictive capabilities.

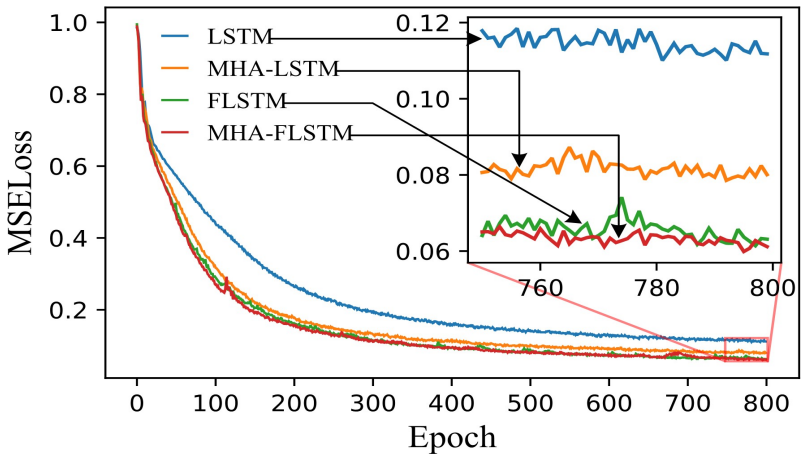
The training loss of the model was also analyzed, as is a crucial indicator for measuring the convergence of the model. Fig. 7 depicts the changes in training loss during the training process for the LSTM, MHA-LSTM, FLSTM, and MHA-FLSTM for the three mechanical performance

indicators, under the same number of iterations.

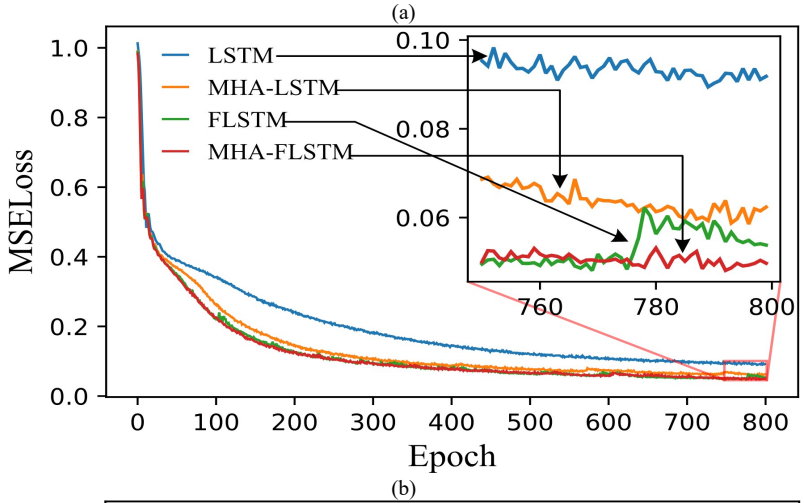
For the three indicators shown in Fig. 7, MHA-FLSTM and MHA-LSTM exhibit faster convergence and higher accuracy compared to LSTM. This indicates that the MHA mechanism is capable of precisely distinguishing key features, thereby enabling the LSTM to achieve higher learning efficiency during training. Furthermore, compared to MHA-LSTM, both FLSTM, and MHA-FLSTM demonstrate faster convergence and higher accuracy. The primary reason for this is that FLSTM and MHA-FLSTM introduce feed-forward layers, which allow for more comprehensive feature extraction. Additionally, the skip connections establish direct connections between layers, enabling lower-level features to be directly transmitted to deeper network layers. This facilitates the network's training and further enhances its non-linear expression capabilities and gradient propagation effects.

TABLE VI
ABLATION EXPERIMENT RESULTS

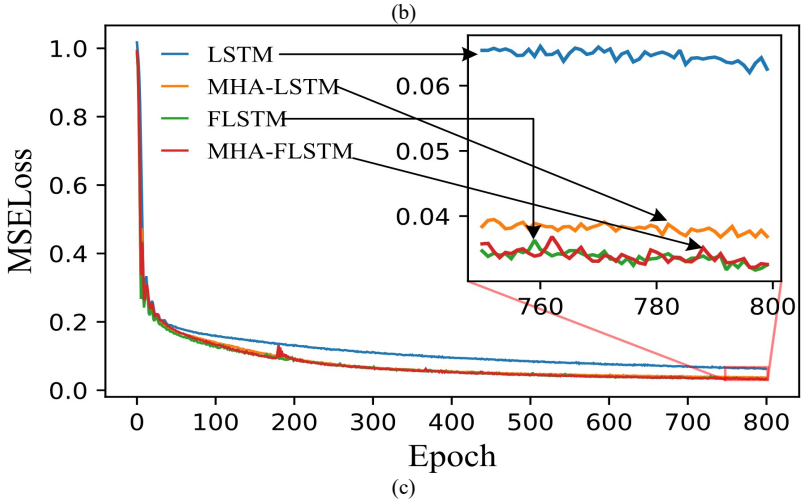
Target value	Evaluation metric	Models			
		LSTM	FLSTM	MHA-LSTM	MHA-FLSTM
YS	MSE	34.261	19.128	19.771	**16.775**
	RMSE	5.853	4.374	4.447	**4.096**
	R^2	0.785	0.868	0.863	**0.910**
	MAE	4.284	2.977	3.108	**2.583**
TS	MSE	26.736	16.231	18.683	**11.980**
	RMSE	5.171	4.029	4.322	**3.461**
	R^2	0.804	0.881	0.873	**0.935**
	MAE	3.833	2.911	3.150	**2.288**
EL	MSE	1.552	1.385	1.048	**0.684**
	RMSE	1.246	1.177	1.024	**0.827**
	R^2	0.879	0.892	0.919	**0.947**
	MAE	0.948	0.847	0.743	**0.587**



(a)



(b)



(c)

Fig. 7. Changes in losses during training: (a) YS, (b) TS, (c) EL.

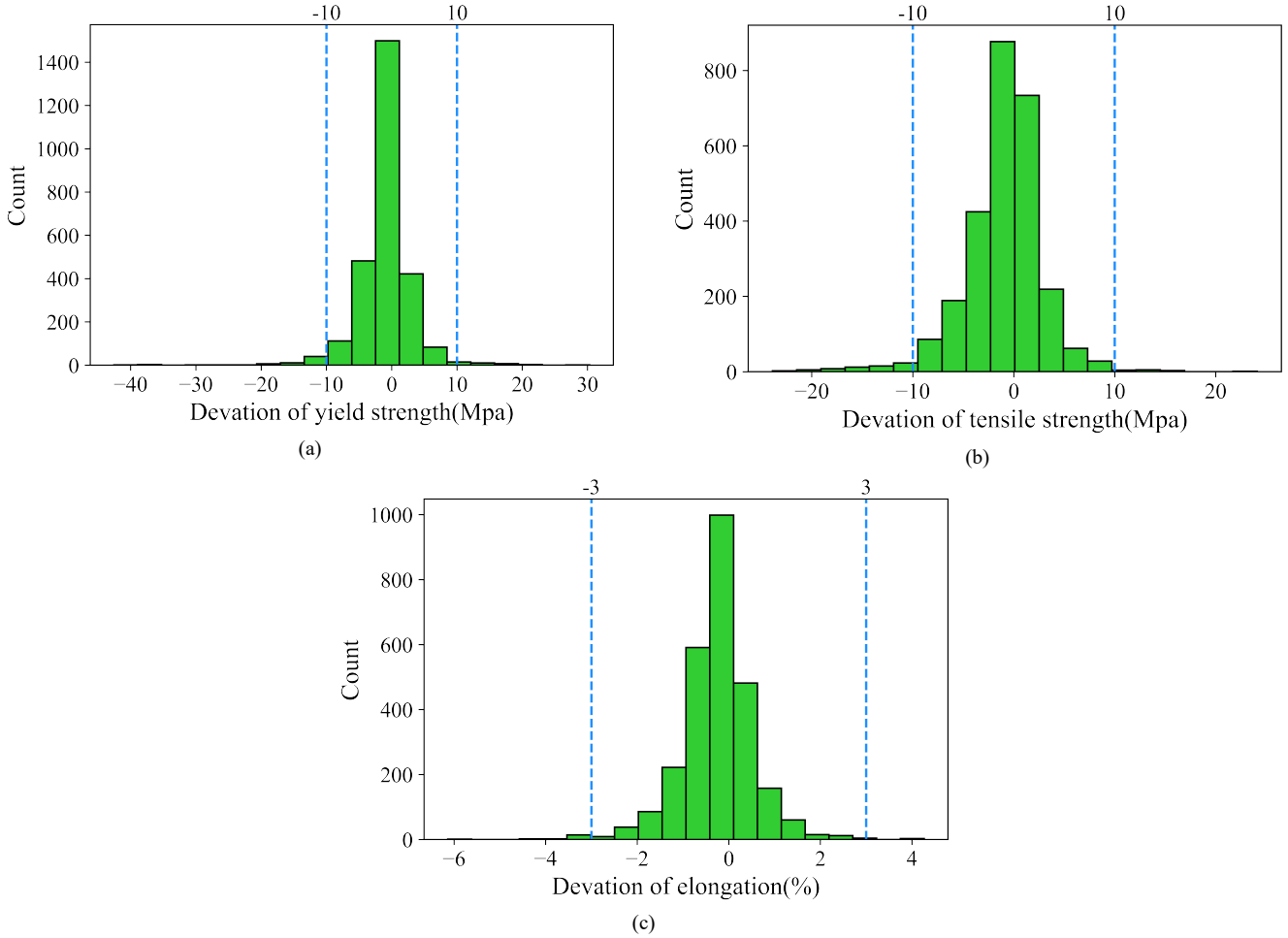


Fig. 8. MHA-FLSTM error histogram: (a) YS, (b) TS, (c) EL.

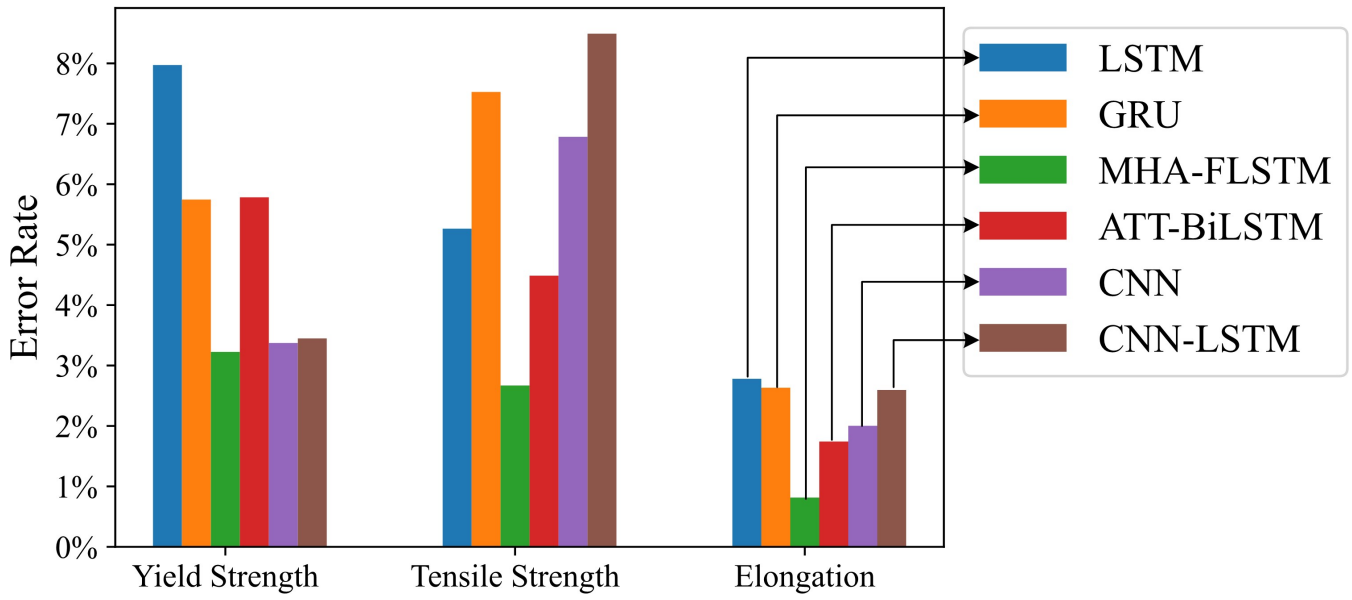


Fig. 9. Prediction error rates of different models.

E. Model Accuracy Analysis

During actual production processes, the three mechanical performance indicators, YS, TS, and EL, are allowed deviations of ±10 MPa, ±10 MPa, and ±3%, respectively. An analysis was conducted on the prediction results of the 2,697 test data samples, and the corresponding MAE for the three target variables were found to be 2.630 MPa, 2.684 MPa, and 0.587%, respectively. It is evident that the predicted values and deviations of the MHA-FLSTM model are mostly within the allowed range. This indicates that the proposed model holds promising prospects for practical applications, as it demonstrates good accuracy in predicting the mechanical performance indicators within the specified tolerances.

Fig. 8 shows the error distribution histograms between the predicted and true values of the three mechanical

performance indicators (YS, TS, and EL) for the MHA-FLSTM model. It can be observed from the graphs that the errors of the three indicators are uniformly distributed around zero, and approximate a normal distribution. Specifically, 96.7% of the YS errors fall within the range of ±10 MPa, 97.3% of the TS errors lie within ±10 MPa, and 99.2% of the EL errors lie within ±3%. This further confirms the excellent predictive performance of the proposed MHA-FLSTM model.

To demonstrate the superiority of the proposed model over other methods in practical applications, the error rates of exceeding the allowed tolerance range for the proposed model and five other comparative models were also analyzed. Fig. 9 illustrates the different models' error rates exceeding the allowed tolerance range across the three mechanical performance indicators.

VI. Conclusions

In this article, we present an MHA-FLSTM model designed to predict the mechanical properties of cold-rolled steel. Initially, the data are input into an improved MHA mechanism to accurately differentiate the importance of different features. Then, skip connections are introduced between the initial input and the attention outputs, and the outcomes of these skip connections are channeled into a batch normalization layer. Afterward, the results are fed into an FFNN to capture the spatial features in the data effectively, with the continued application of the same skip connections and batch normalization operations. Subsequently, the results are input into an LSTM to extract the temporal features, and the final predictions of the mechanical properties of cold-rolled steel are derived through a feedforward layer. Experiments conducted on an industrial dataset have substantiated the efficacy and superiority of the MHA-FLSTM model, thereby providing empirical evidence of its aptitude in addressing the intricate engineering challenge posed by the prediction of the mechanical properties of cold-rolled steel.

References

[1] W. G. Li, W. Yang, Y. T. Zhao, and H. F. Hu, "Mechanical property prediction model of hot-rolled strip via big data and metallurgical mechanism analysis." *Journal of Iron and Steel Research* 30.4 (2018): 301-308.

[2] Z. H. Deng, H. Q. Yin, et al, "Machine-learning-assisted prediction of the mechanical properties of Cu-Al alloy." *International Journal of Minerals, Metallurgy and Materials* 27 (2020): 362-373.

[3] Y. B. Zhao, Y. Song, F. F. Li, and X. L. Yan, "Prediction of mechanical properties of cold rolled strip based on improved extreme random tree." *Journal of Iron and Steel Research International* 30.2 (2023): 293-304.

[4] T. Q. Cheng, and G. C. Chen, "Prediction of mechanical properties of hot-rolled strip steel based on PCA-GBDT method." *Journal of Physics: Conference Series*. Vol. 1774. No. 1. IOP Publishing, 2021.

[5] Z. X. Shi, L. X. Du et al, "Prediction Model of Yield Strength of V–N Steel Hot-rolled Plate Based on Machine Learning Algorithm." *JOM* 75.5 (2023): 1750-1762.

[6] S. Hochreiter, and J. Schmidhuber, "Long short-term memory." *Neural Computation* 9.8 (1997): 1735-1780.

[7] J. Schmidhuber, "Deep learning in neural networks: An overview." *Neural Networks* 61 (2015): 85-117.

[8] A. Shrestha, and A. Mahmood, "Review of deep learning algorithms and architectures." *IEEE Access* 7 (2019): 53040-53065.

[9] M. Marani, M. Zeinali, V. Songmene, and C. K. Mechefske, "Tool wear prediction in high-speed turning of a steel alloy using long short-term memory modelling." *Measurement* 177 (2021): 109329.

[10] A. Sagheer, and M. Kotb, "Time series forecasting of petroleum production using deep LSTM recurrent networks." *Neurocomputing* 323 (2019): 203-213.

[11] X. Z. Wang, W. H. Li, Q. L. Li, and X. N. Li, "Modeling Soil Temperature for Different Days Using Novel Quadruplet Loss-Guided LSTM." *Computational Intelligence and Neuroscience* 2022 (2022).

[12] Y. R. Li, Z. F. Zhu, D. Q. Kong, H. Han, and Y. Zhao, "EA-LSTM: Evolutionary attention-based LSTM for time series prediction." *Knowledge-Based Systems* 181 (2019): 104785.

[13] Y. K. Ding, Y.L. Zhu, J. F, P. C. Zhang, and Z. R. Cheng, "Interpretable spatio-temporal attention LSTM model for flood forecasting." *Neurocomputing* 403 (2020): 348-359.

[14] W.Y. Wei, H.H. Wu, and H.D. Ma, "An autoencoder and LSTM-based traffic flow prediction method." *Sensors* 19.13 (2019): 2946.

[15] X. Gang, J.S. He, Z.M. Lü, M. Li, and J. W. Xu, "Prediction of mechanical properties for deep drawing steel by deep learning." *International Journal of Minerals, Metallurgy and Materials* 30.1 (2023): 156-165.

[16] A. Vaswani, et al, "Attention is all you need." *Advances in Neural Information Processing Systems* 30 (2017).

[17] P. Velickovic, et al, "Graph attention networks." *Stat* 1050.20 (2017): 10-48550.

[18] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[19] Q. Zhang, R. Wang, Y. Qi, and F. Wen, "A watershed water quality prediction model based on attention mechanism and Bi-LSTM." *Environmental Science and Pollution Research* 29.50 (2022): 75664-75680.

[20] T. Y. Kim, and S. B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks." *Energy* 182 (2019): 72-81.

Qiwen Zhang was born on February 23, 1975, master degree, master tutor. He graduated from Gansu University of Technology, majoring in computer and applied technology, with a bachelor's degree in engineering in 1999. He received the M.S. degree in computer and applied technology from Lanzhou University of Technology in 2005. He presided over or participated in the completion of a number of projects such as the National Science and Technology Support Program, the National Natural Science Foundation of China, and the Natural Science Foundation of Gansu Province, and more than 20 horizontal projects, and won the Gansu Province University Science and Technology Progress Award (first prize). Published more than 40 academic papers. Main research directions: intelligent information processing, knowledge discovery, computational intelligence.

Rongping Guo was born in 1998. He is currently a master student in computer technology at Lanzhou University of Technology. His main research areas are artificial intelligence and pattern recognition.