symmetry

MDPI

Article

Atmospheric Temperature Prediction Based on a BiLSTM-Attention Model

Xueli Hao, Ying Liu, Lili Pei *, Wei Li and Yaohui Du

School of Information Engineering, Chang'an University, Xi'an 710064, China
* Correspondence: peilili@chd.edu.cn

Abstract: To address the problem that traditional models are not effective in predicting atmospheric temperature, this paper proposes an atmospheric temperature prediction model based on symmetric BiLSTM (bidirectional long short-term memory)-Attention model. Firstly, the meteorological data from five major stations in Beijing were integrated, cleaned, and normalized to build an atmospheric temperature prediction dataset containing multiple feature dimensions; then, a BiLSTM memory network was used to construct with forward and backward information in the time dimension. And the limitations of the traditional LSTM method in long-term time series analysis were solved by introducing the attention mechanism to achieve the prediction analysis of atmospheric temperature. Finally, by comparing the prediction results with those of BiLSTM, LSTM-Attention, and LSTM, it is revealed that the proposed model has the best prediction effect, with a MAE value of 0.013, which is 0.72%, 0.41%, and 1.24% lower than those of BiLSTM, LSTM-Attention, and LSTM, respectively; the R^2 value reaches 0.9618, which is 2.73%, 1.23%, and 4.98% higher than BiLSTM, LSTM-Attention, and LSTM, respectively. The results show that the symmetrical BiLSTM-Attention atmospheric temperature prediction model can effectively improve the prediction accuracy of temperature data, and the model can also be used to predict other time series data.

Keywords: bidirectional long short-term memory network; attention mechanism; machine learning; temperature prediction

Citation: Hao, X.; Liu, Y.; Pei, L.; Li, W.; Du, Y. Atmospheric Temperature Prediction Based on a BiLSTM-Attention Model. *Symmetry* **2022**, *14*, 2470. https://doi.org/10.3390/sym14112470

Academic Editor: Mikhail Sheremet

Received: 6 October 2022
Accepted: 10 November 2022
Published: 21 November 2022

1. Introduction

The study of temperature prediction affects many fields, and the ability to accurately predict atmospheric temperatures is important for urban flood and drought prevention, resource use, and agricultural development; as such, this has become a topic that needs to be addressed and further researched [1,2]. In response to this, early researchers have developed models to predict atmospheric temperatures based on historical meteorological data combined with statistical knowledge [3], which are more interpretable than people's subjective experiences. However, the accuracy of such prediction methods is not high, and atmospheric temperature data are generally subject to a variety of different influencing factors, and the data exhibit strong randomness and uncertainty [4]. The use of neural network models trained on a large amount of historical data will allow for more accurate learning of data fluctuations and better extraction of data features, thus largely improving the accuracy of predictions [5].

For historical data with time series characteristics, people have started to use time series prediction models to carry out temperature prediction, such as LSTM (long short-term memory) [6–8], which are based on the memory function of neural networks, and can achieve good temperature data prediction for a large amount of historical data in a time series. However, symmetry-based BiLSTM networks can better avoid the shortcomings of unidirectional LSTM networks. The attention mechanism is widely used to improve the problem of unfocused and time-consuming feature extraction by allocating computational resources to neural networks. Therefore, in the field of time series prediction, the attention mechanism is also of very good use [9].

In order to further improve the prediction accuracy of temperature data, this paper combines a BiLSTM network with an attention mechanism and proposes a BiLSTM (bidirectional long short-term memory)-Attention model, with the following main innovations.

(1) Firstly, by extracting time series information in both directions through a bidirectional long short-term memory network with symmetry, the problem of forward and backward time dependence of series data is well solved, and more accurate prediction results can be obtained through the training of the neural network.
(2) Secondly, by adding an attention mechanism to the network and using it to reasonably allocate the attention resources in the model, the impact of key sequence information on the prediction results during temperature prediction is highlighted.

The prediction analysis is carried out for several districts in Beijing, and finally compared with several models, such as the BiLSTM, to verify the effectiveness of the BiLSTM-Attention model, which will be a good guide for meteorologists to conduct more complete temperature prediction research.

The article is organized as follows. Section 2 of the article describes the related work. Section 3 introduces the dataset used in this paper. Section 4 introduces the methods used in this paper, mainly including BiLSTM, the attention mechanism, and BiLSTM-Attention model. Section 5 presents the analysis and discussion of experimental results, and Section 6 is the summary and outlook.

The overall framework is shown in Figure 1. This paper selects the temperature datasets of five major stations in Beijing (China), constructs effective features based on the characteristics of temperature data, and divides the training set and test set proportionally; then, the training set is input into the constructed temperature prediction model for training, and then the test set is input into the already trained model for testing, after which the data need to be renormalized to finally obtain the prediction results of the temperature data. The model is also analyzed and compared with BiLSTM, the LSTM-Attention model, and LSTM model, which finally proves the certain superiority of the BiLSTM-Attention model.

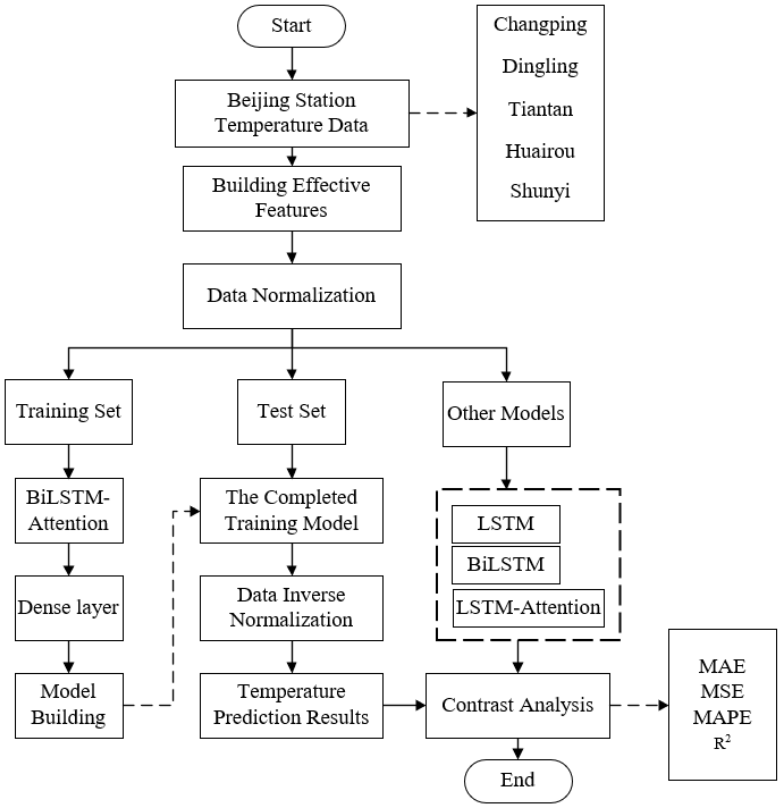

Figure 1. Thesis flow chart.

2. Literature Review

Several scholars have already made many contributions to the field of temperature prediction [10,11]. More traditional machine learning methods have been used in the past [12,13]. Zhou et al. proposed a grey-Markov temperature prediction model based on seasonal indices based on the interannual cyclical and seasonal variation patterns of historical temperature data, and obtained more accurate prediction results by predicting the average temperature of Guangzhou city [14]. Raviprased et al., formulated a compound-specific prediction model that could better predict the critical temperature of superconductors to solve the problem of predicting the effectiveness of the decision tree approach in this area, which was finally demonstrated after comparison with other multiple models [15]. Hou et al. developed a BP (back propagation) neural network model for climate change prediction by integrating factors such as atmospheric CO2 emissions, heat dissipation of the Earth, and changes in ocean surface temperature over the years, and produced a more accurate prediction of temperature changes in the future years [16]. Cai et al. established an SVM (support vector machine) model to predict the indoor temperature of buildings, and compared it with a BP neural network, and found that the SVM prediction accuracy effect was better than that of the BP neural network, which proved the applicability of SVM prediction method in the field of prediction [17]. From the above studies related to temperature prediction, it can be seen that prediction is mostly based on traditional machine learning models [18–20], which can better achieve the prediction of temperature data, but the prediction accuracy of the model is affected by many factors, such as the quality of the data, the extraction of data features, the configuration of model parameters, etc., and its prediction accuracy needs to be further improved.

Among existing temperature prediction studies, time series prediction models represented by LSTM networks have been widely used in temperature prediction, and can reliably predict long-term time series data [21]. Qiu et al., used LSTM models to predict daily river temperatures and, through experimental analysis of data from the Three Gorges reservoir system, captured the daily average variation of the thermal system more accurately, demonstrating that the LSTM outperformed other methods in predicting the daily average water temperature of rivers [22]. MASOOMA et al. used an LSTM model based on a spatial attention mechanism to accurately capture the space and time of multiple meteorological features to predict temperature, and discovered that spatial feature attention captured the interaction of input features on target features, and the study maintained a better prediction accuracy [23]. Song et al., proposed a temporal prediction model, based on LSTM and Kalman filtering, for predicting observations in atmospheric quality datasets, and found that the LSTM–Kalman model had better prediction results when compared with the LSTM model [24]. Liu et al. analyzed the time dependence of ocean temperature variability at multiple depths, and proposed a new method for ocean temperature time series prediction, namely the time-dependent ocean-temperature-prediction-based long short-term memory network (TD-LSTM), which confirmed that the TD-LSTM outperformed other methods and performed well in different regions and depths [25].

The above studies on temperature prediction are summarized in Table 1. All of the above methods provide good solutions for temperature prediction.

Table 1. Comparison of temperature prediction methods.

Literature	Method	Overall Evaluation of the Method
[17]	Grey-Markov	Combining longitudinal and cross-sectional analysis for non-stationary data prediction, but the method is traditional and the accuracy is not high.
[18]	Decision Trees	Multiple conventional and non-conventional models are used for superconductor critical temperature prediction, demonstrating the benefits of decision trees, however, the prediction accuracy needs to be improved.

Table 1. *Cont.*

Literature	Method	Overall Evaluation of the Method
[19]	BP	A wide range of atmospheric factors affecting temperature change are considered to predict climate conditions over the next 25 years, but the method is single and the accuracy needs to be improved.
[20]	SVM	A support vector machine SVM model for indoor temperature prediction is shown to outperform a back propagation neural network BPNN model, but the model is traditional, the contrast is single.
[22]	LSTM	Neural networks predict daily water temperatures and quantify trends, resulting in a significant improvement over traditional models, however, the model is single and the feature extraction is not sufficient.
[23]	LSTM-Attention	Accurately capture the spatial and temporal relationships of multiple meteorological features, but the feature extraction is not sufficient.
[24]	LSTM-Kalman	Kalman filtering added to data series processing, however, the data features are not sufficiently extracted and the prediction accuracy needs to be verified.
[25]	TD-LSTM	A time-varying parameter matrix based on the fusion of historical observations is proposed, but more models need to be compared and the accuracy can be further improved.

3. Dataset Construction and Data Quality Improvement

3.1. Data Sources and Their Visualization

This paper focuses on the prediction of Beijing temperature data, which were taken from the multi-site meteorological dataset of Beijing in the machine learning database UCI (University of California-Irvine). A total of five major sites were selected for the actual study; the five sites are Changping, Dingling, Tiantan, Huairou, and Shunyi. Each site contains atmospheric monitoring data from March 2013 to February 2017, and each site has about 35,064 pieces of data; the data therefore contain a total of 35,064 × 5 samples. The dataset is recorded for each characteristic value for 24 h within each day. The sample data are shown in Table 2.

Table 2. Tiantan Station air monitoring dataset (sample).

No	Year	Month	Day	Hour	PM2.5	PM10	SO_2	NO_2	CO	O_3	PRES	DEWP	RAIN	WSPM	TEMP	Station
1	2013	3	1	0	6	6	4	8	300	81	1024.5	−21.4	0	5.7	−0.5	Tiantan
2	2013	3	1	1	6	29	5	9	300	80	1025.1	−22.1	0	3.9	−0.7	Tiantan
3	2013	3	1	2	6	6	4	12	300	75	1025.3	−24.6	0	5.3	−1.2	Tiantan
4	2013	3	1	3	6	6	4	12	300	74	1026.2	−25.5	0	4.9	−1.4	Tiantan
5	2013	3	1	4	5	5	7	15	400	70	1027.1	−24.5	0	3.2	−1.9	Tiantan

PM2.5, PM10, SO_2, NO_2, CO, and O_3 are air pollutant indicators in Table 1. Several other features in Table 1 are meteorological terminology, and the meaning of these terminologies is explained in Table 3.

Table 3. The meaning of the eigenvalues.

Features	Eigenvalue Meaning
PRES	Atmospheric pressure intensity
DEWP	Dew point temperature
RAIN	Rainfall
WSPM	Wind speed
TEMP	Atmospheric temperature
Station	Five stations

Based on the year and temperature characteristics provided in the dataset, a graph of the average temperature trend for 2013–2016 can be obtained, as shown in Figure 2. In the year 2013, the highest average temperature value (15.96 °C) was observed, and the average

temperature showed a decreasing trend in 2014–2015, while the average temperature increased in 2016.

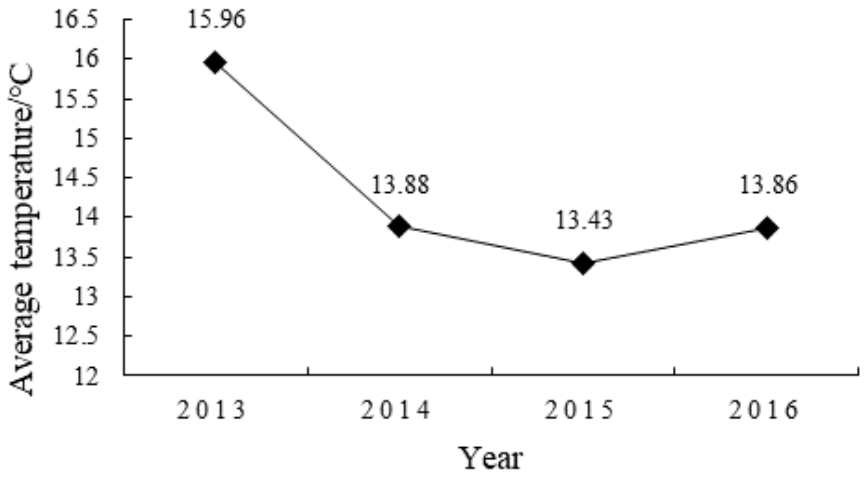

Figure 2. Trend of temperature change from 2013 to 2016.

3.2. Data Cleaning

3.2.1. Removing Invalid Attributes

The dataset of Tiantan City is used as the main sample for the experiment, which has a total of 35,064 records and 17 attributes. The attributes that are not relevant to this data prediction, such as "No", and "station", can be deleted directly when conducting data pre-processing.

3.2.2. Fill Missing Values

The missing values in the original dataset are indicated by "NA", and the missing values of the valid features are counted. We found that there are a certain number of missing values for each feature; for example, 597 missing values for PM10, 1118 missing values for SO_2, 744 missing values for NO_2, etc. The results of missing value statistics are shown in Figure 3. According to the missing value statistics, the missing values are filled by mean interpolation.

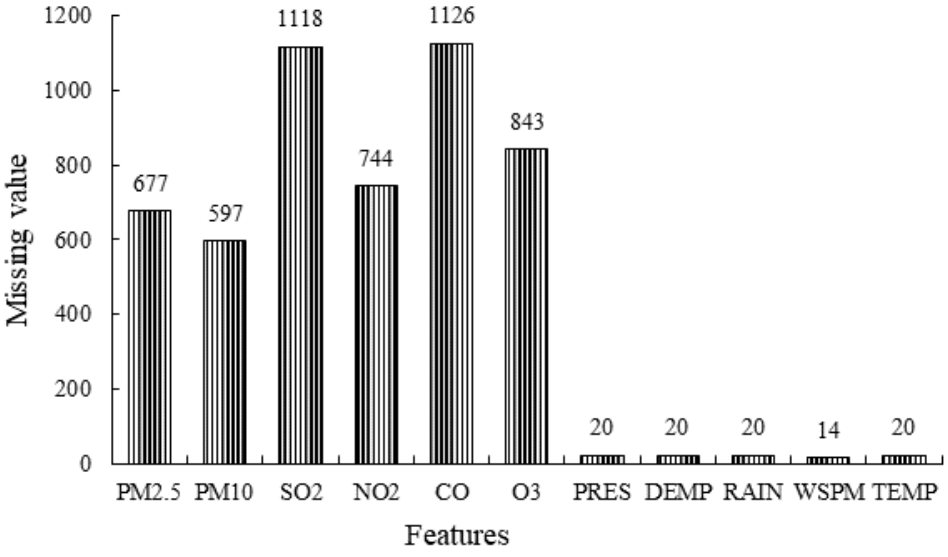

Figure 3. Missing value statistics chart.

3.3. Feature Importance Analysis

In general, feature importance measures the weight and value of a feature in the construction of a model; the higher the score of a feature used as an input, the more important it is relatively, and conversely, the less important it is. The importance scores for the above features of PM2.5, PM10, SO_2, etc., are ranked, as shown in Figure 4. It can be seen that DEWP (dew point temperature) has the highest importance score, PRES (atmospheric pressure) has the next highest importance score, and CO (carbon monoxide) and RAIN (rainfall) have relatively low importance scores.

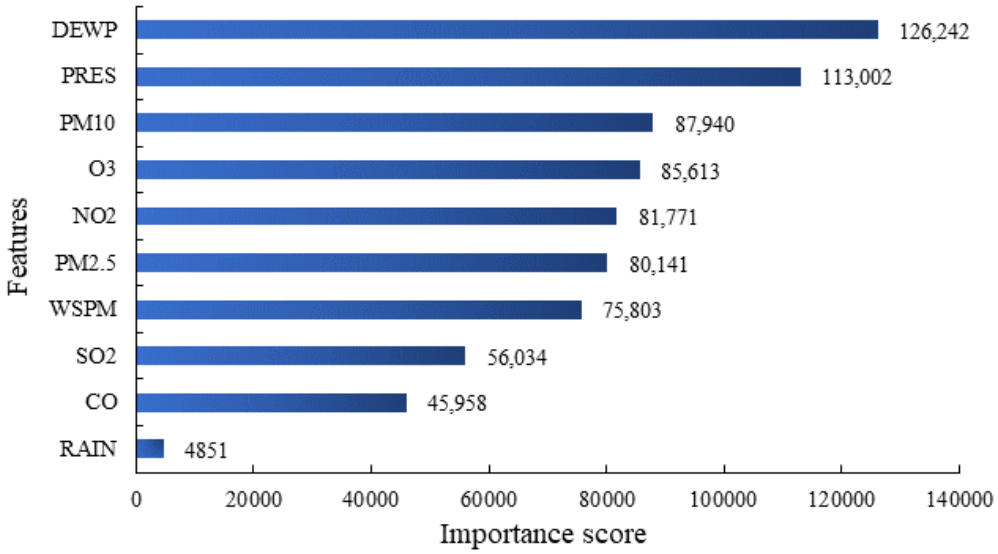

Figure 4. Feature importance ranking.

3.4. Data Normalization Process

The data prediction accuracy is affected by the data dimensionality, and to eliminate the effect of dimensionality on the experimental results, the data need to be normalized. The normalization operation can transform all the data with magnitudes into dimensionless data, i.e., all lie within [0, 1] [26], and can also improve the training accuracy and speed of the model. In this paper, the minimum–maximum normalization conversion was performed for PM2.5, PM10, CO, O_3, TEMP, DEWP, and RAIN attributes, and the conversion method is shown in Equation (1).

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

In the above equation, x_{max} is the maximum data in the column, and x_{min} is the minimum data in the column. x^* is the value of the entire column normalized by the highest value.

The feature box visualization type plots before and after normalization are shown in Figures 5 and 6, respectively. It can be seen from the figures that the original eigenvalue data fluctuate greatly, and the normalized eigenvalue distributions are all between [0, 1], at which point, the data need to be restored using the inverse normalization operation.

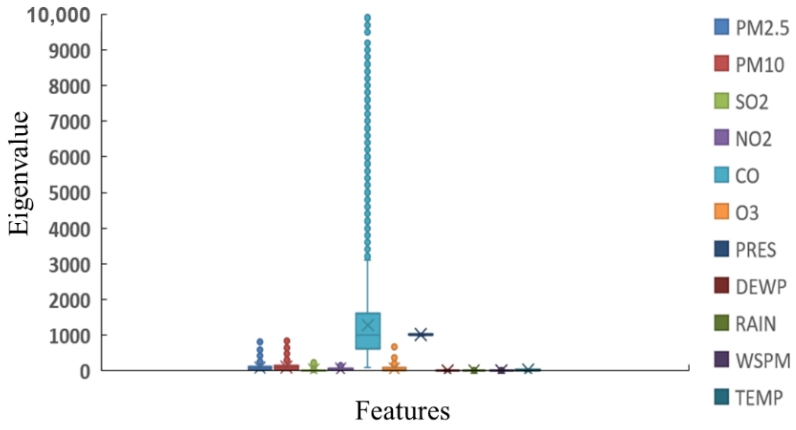

Figure 5. Eigenvalue distribution before normalization.

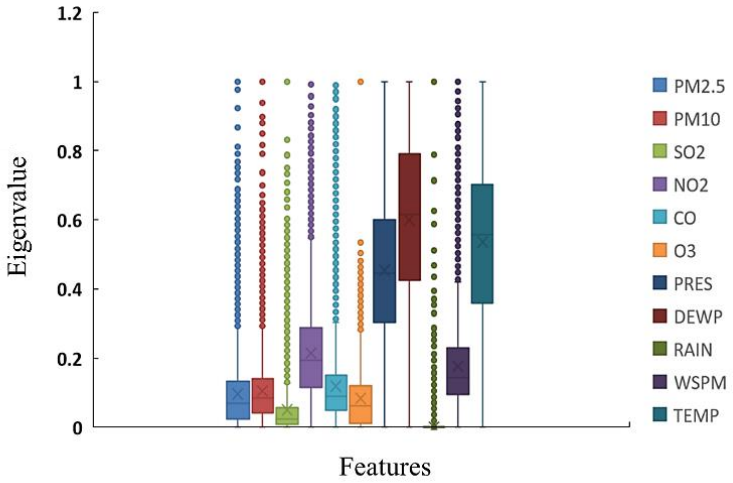

Figure 6. Eigenvalue distribution after normalization.

4. Temperature Data Series Prediction Method

4.1. BiLSTM Network

The LSTM network is a deep learning network model evolved from a recurrent neural network [27] that can improve some shortcomings in the recurrent neural network model, such as gradient disappearance [28]. The LSTM network contains a total of four structures: memory unit, forgetting gate f_t, input gate i_t, and output gate o_t Its structure is shown in Figure 7.

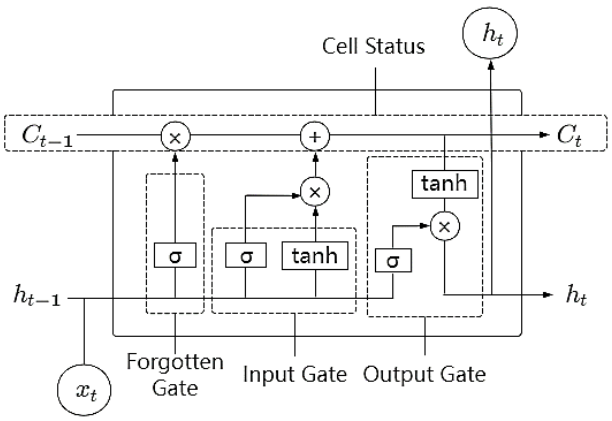

Figure 7. Basic structure of LSTM.

As seen in Figure 6, the LSTM network has added cell states with three gate components compared to the RNN (recurrent neural network) [29]. The forgetting gate f_t in the LSTM structure is responsible for determining what percentage of information is left in that network, and is calculated as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2}$$

In the above equation, x_t is the input sequence; h_{t-1} is the state memory of the previous moment; $\sigma(\cdot)$ is the sigmoid activation function; W_f is the weight matrix of the forgetting gate; b_f is the bias of the forgetting gate; and f_t is the state of the forgetting gate.

The input gate i_t is responsible for selectively memorizing the new information in the cell state, and is calculated as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

$$C_t = f_t * C_t + i_t * \widetilde{C}_t \tag{5}$$

In the above equation, $\widetilde{C}_t$ is the cell state candidate; C_t is the new cell state; $\tanh(\cdot)$ is the hyperbolic tangent function; W_i is the weight matrix of the input gate; W_c is the weight matrix of the cell state; b_i is the bias of the input gate; b_c is the bias of the cell state; and i_t is the state of the input gate.

The output gate o_t is responsible for determining the current state of the output information, and is calculated as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

In the above equation, W_o is the weight matrix of the output gate; b_o is the bias of the output gate; and o_t is the state of the output gate.

In this paper, we focus on the BiLSTM model, which is a bidirectional long short-term memory model combining forward and backward information, i.e., it can process information in both directions. Both forward and backward directions have hidden layers, and these hidden layers can extract the forward and backward key information together in a given time [30]; thus, we can obtain more adequate temperature data features and help to improve the prediction accuracy of the model. The BiLSTM network structure is shown in Figure 8.

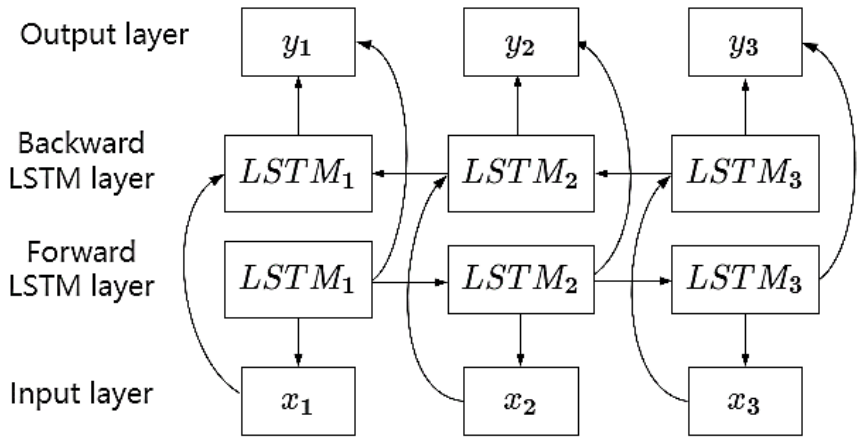

Figure 8. BiLSTM network structure.

4.2. Attentional Mechanisms

The attention mechanism is mainly designed to quickly extract more valid information from the large volume of information, reduce the influence of invalid information on the training effect of the model, and achieve the purpose of improving prediction accuracy [31].

There are generally hard and soft attention mechanisms for machine learning. The hard attention mechanism is a random selection of the information in the input sequence. Since the selection probability is difficult to quantify, which increases the difficulty of model training, in this paper, we choose to use a soft attention mechanism. Combined with the sequence data information, the input information is calculated as a weighted average, and then input into the network for training, which can effectively improve the attention of the model to the input information, and achieve a reasonable allocation of resources, which is suitable for predicting temperature sequence data in this paper [32]. The flow structure of the soft attention mechanism is shown in Figure 9.

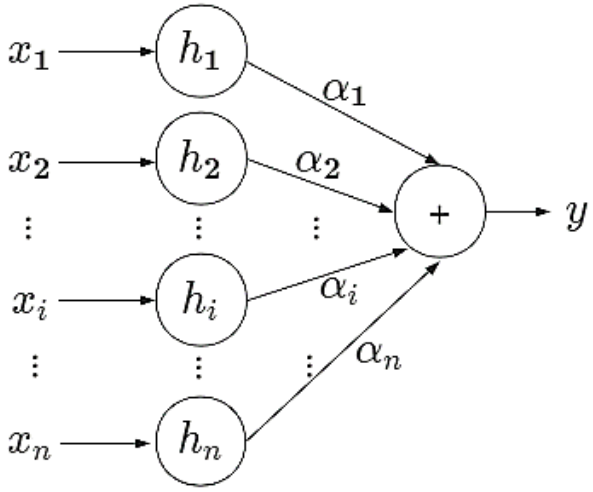

Figure 9. Structure of attentional mechanism.

In Figure 9, x_i is the input of the BiLSTM layer embedded with the attention mechanism; h_i is the output of the BiLSTM layer; α_i is the different weights of the different channels of the BiLSTM obtained after making calculations based on the attention mechanism; y is the final output of the neural network model.

The main formulas of the attentional mechanism are as follows:

$$e_t = utanh(w \cdot h_t + b) \tag{8}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t=1}^{i} \alpha_t \cdot h_t} \tag{9}$$

$$s_t = \sum_{t=1}^{i} \alpha_t \cdot h_t \tag{10}$$

where e_t is the attention distribution value at moment t; u and w are weight coefficients; b is bias; α_t is the different weights of different channel information in BiLSTM; s_t is the output h_t and weight matrix of BiLSTM layer after weighting.

The attention mechanism is mainly manifested in the operation of the weight coefficients of different channels, which can be updated and optimized to adjust the allocation ratio of the model to the channel information [33], producing the best training effect of the model in the current computing environment.

4.3. BiLSTM-Attention Model

Combined with the temperature dataset used in this paper, the BiLSTM-Attention model is proposed, which can fully utilize the advantages of the bidirectional memory network and the attention mechanism. The BiLSTM network structure can process the input of the network based on both forward and backward directions simultaneously, and obtain the information of the previous moment and the next moment at a particular time. Moreover, BiLSTM has a unique bidirectional network structure, so it can extremely enhance the information memory of the model at the beginning and end phases of the input information during the training process [34]. Based on the bidirectional memory network structure, the means of the attention mechanism are embedded, which makes the model channels reasonably assign weights and strengthen the attention of the key information; therefore, using the improved model, the prediction effect of temperature data can be improved. The network structure of the BiLSTM-Attention model is shown in Figure 10.

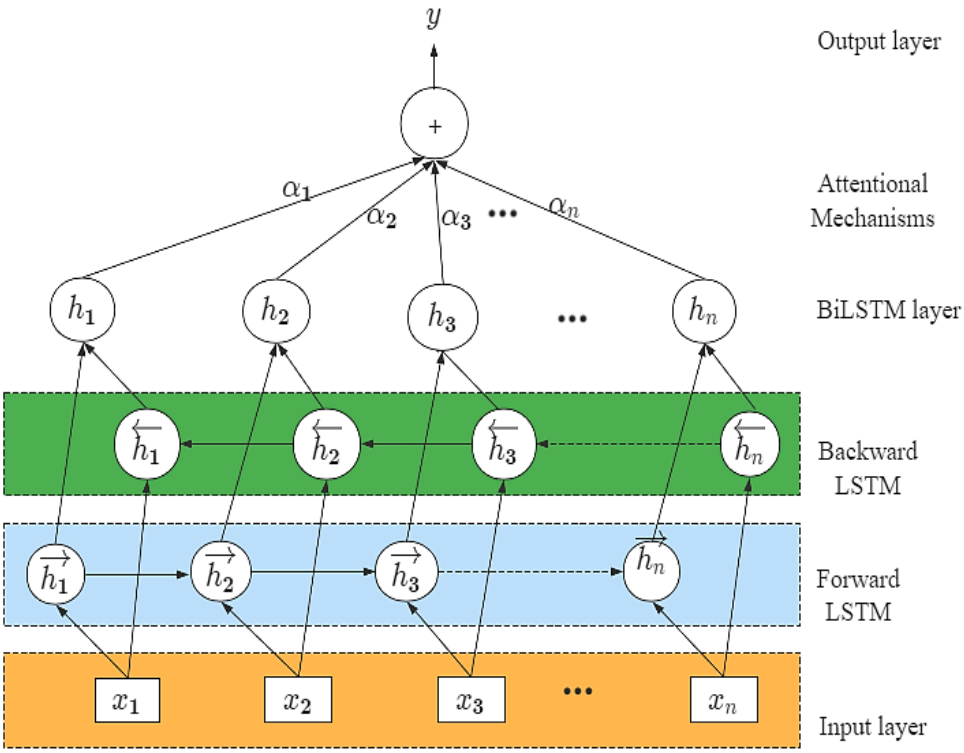

Figure 10. BiLSTM-Attention network structure diagram.

The BiLSTM-Attention model is divided into four parts: the feature vector input layer, the BiLSTM layer, the attention layer, and the output layer. As can be seen from Figure 10, the first layer is the input layer, x_i (i = 1, 2, . . . , n) is the input of the input layer; the second layer is the bidirectional LSTM layer, which is further divided into forward LSTM and backward LSTM layers; the third layer is the attention mechanism layer, α_i (i = 1, 2, . . . , n) values are the different weights of different channels of information; the fourth layer is the output layer, in which y which is the final output of the network.

(1) Input layer: This refers to the input feature vectors. The input layer of this paper focuses on pre-processing the atmospheric temperature datasets into the form of feature vectors that can be directly accepted and processed by the BiLSTM layer.

(2) BiLSTM layer: The BiLSTM layer consists of forward and backward LSTM layers, which have the role of capturing the information features before and after. The forward LSTM layer computation vector is denoted as $\overrightarrow{h_i}$ (i =1,2, . . . ,n), and the

backward LSTM layer computation vector is denoted as $\overleftarrow{h_i}$ (i =1,2, . . . ,n), which yields the output h of the BiLSTM layer at moment t, as follows.

$$h_t = \alpha\overrightarrow{h_i} + \beta\overleftarrow{h_i} \tag{11}$$

$$h = \sigma(h_t) \tag{12}$$

In the above equation, α and β are constants and the sum of α and β is 1.

(3) Attention layer: In temperature data prediction, the neural network is trained to focus on certain key features through the attention mechanism, the core of which is the weight coefficient. The first step is to learn the importance of each feature, and then assign the corresponding weight to each feature according to the importance. Equation (9) enables the transition from the input initial state to the new attention state, after which the final output state vector s_t is obtained through Equation (10). Finally, s_t is integrated with the dense layer as an output value input into the final output layer.

(4) Output layer: The input into the output layer is the output of the attention mechanism layer in the implicit layer, which in this paper is mainly the set of predicted y-vectors of atmospheric temperature.

By continuously optimizing and updating the weights and biases, the cost function in the model structure gradually becomes smaller, and the network model becomes better during the training period.

Based on the measured and predicted temperature values, the prediction results of the model are evaluated based on commonly used prediction metrics such as MAE (mean absolute error), MSE (mean squared error), MAPE (mean absolute percentage error), and R^2 (linear correlation coefficient) [35]. The formulas, meanings, and evaluation criteria of each metric are shown in Table 4, where n is the total number of measured values, y_i is the temperature predicted measurements, $\widetilde{y_i}$ is the predicted value of the temperature prediction and $\overline{y_i}$ is the average value of y_i.

Table 4. Model evaluation metrics.

Model Evaluation Metrics	Equation	Evaluation Criteria
MAE	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert(y_i - \widetilde{y_i})\right\rvert$	Average absolute error, the more the value tends to 0, the better the model.
MSE	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{y_i})^2$	Mean square error, the more the value tends to 0, the better the model.
MAPE	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \widetilde{y_i}}{y_i}\right\rvert$	Average absolute percentage error, the more the value tends to 0, the better the model.
R^2	$R^2 = 1 - \frac{\sum_i^n (y_i - \widetilde{y_i})^2}{\sum_i^n (y_i - \overline{y_i})^2}$	Linear correlation coefficient, the more the value tends to 1, the better the model.

5. Temperature Data Prediction Based on BiLSTM-Attention Model

5.1. Experimental Materials

5.1.1. Experimental Datasets

The experimental process used the dataset once it had undergone the processing stage in Section 3.1, including real-time monitoring data from five major stations, with 35,064 sample records for each dataset, including several features, such as PM2.5, PM10, NO, PRES, etc. For this dataset, the model is trained by randomly dividing it into an 80% training set and a 20% test set.

5.1.2. Experimental Environment

The configurations used for the experiments, such as software and hardware, are shown in Table 5.

Table 5. Experimental environment.

Configuration Items	Configuration Conditions
Software environment	Pycharm, Jupyter Notebook
Hardware environment	Win10, GPU, 8 GB of memory, GTX960M graphics card
Language	Python
Frames	Keras
Evaluation indicators	MAE, MSE, MAPE, R^2

5.2. Input and Output Variables

The model mainly solves the learning-based problem of mapping features between input and output variables. For the atmospheric temperature prediction experiment, the input and output variables are determined to conform to the requirements of the model. In this experiment, the model input is a time series variable consisting of a time step and an independent variable, and the output is a predicted value formed by a one-dimensional array.

(1) Model input: The model input is the input data X. $M_{(t)}$ represents the indicator data at moment t, and consists of a set of multiple factors selected to satisfy the variation in the influencing temperature. The change in continuous values $M'_{(t)}$ in $M_{(t)}$ at moment t in the future is fitted by the current indicator value $M_{(t-1)}$ at moment $t-1$. The data from moment $t-1$ and moment t form the temporal input variable X. The formula is calculated as follows. In this experiment, the input variables of the model are: PM2.5, PM10, SO_2, NO_2, CO, O_3, PRES, DEWP, RAIN, and WSPM.

$$M'_{(t)} = \sigma LSTM\left(M_{(t-1)}, M_{(t)}\right) \tag{13}$$

$$X = concat\left(M_{(t-1)}, M'_{(t)}\right) \tag{14}$$

(2) Model output: The model output is the prediction result Y', a one-dimensional array of predictions. In this experiment, the output variable of the model is temperature.

$$Y = [y_1, y_2, \ldots y_i \ldots y_n](i = 1, 2, \ldots, n) \tag{15}$$

$$Y' = softmax\ (\text{BiLSTM} - \text{Attention}(\text{X}, \text{Y})) \tag{16}$$

5.3. Optimizer Selection

For the training of large-scale data, optimizers are generally needed to speed up the model learning rate and make the model converge faster. In machine learning, optimizers are generally used mainly for solving the gradient descent problem; the principle of gradient descent is shown in Equation (17).

$$\theta^{n+1} = \theta^n - \eta \cdot \nabla\theta^{J(\theta)} \tag{17}$$

In the above equation, η is the learning rate, θ^n is the parameter before the update; θ^{n+1} is the parameter after the update; and $\nabla\theta^{J(\theta)}$ is the derivative of the current parameter.

For the SGD (stochastic gradient descent) optimizer, the parameters of the gradient descent can be updated once with a single piece of data, but the amount of data used to update the parameters of this optimizer is extremely small, and the amplitude of the gradient update is extremely large. As for the ADAM optimizer (adaptive moment estimation), it can store the exponential decay average of the squared v_t of the historical gradient as well as keep the exponential decay average of the past gradient m_t, which enables adaptive learning for each parameter.

$$g_t = \nabla\theta^{J(\theta_{t-1})} \tag{18}$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{19}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) {g_t}^2 \tag{20}$$

β_1 is the exponential decay rate, which controls the weight assignment (momentum vs. current gradient) and usually takes a value close to 1, with a default of 0.9.

β_2 is the exponential decay rate, which controls the influence of the previous gradient squared.

In this paper, two optimizers, SGD and ADAM, are used for comparative analysis, and MSE is used as the model evaluation index. As shown in Figure 11a, when the SGD optimizer is used, the error of both the training and test sets gradually decreases and tends to 0. However, the error of the test set is higher than that of the training set within 0–400 iterations; thus, the model has the problem of overfitting. In Figure 11b, when the ADAM optimizer is used, the errors of the training and test sets are also gradually reduced to zero, and the error line fits better; consequently, the ADAM optimizer is finally chosen for the optimization of the prediction model.

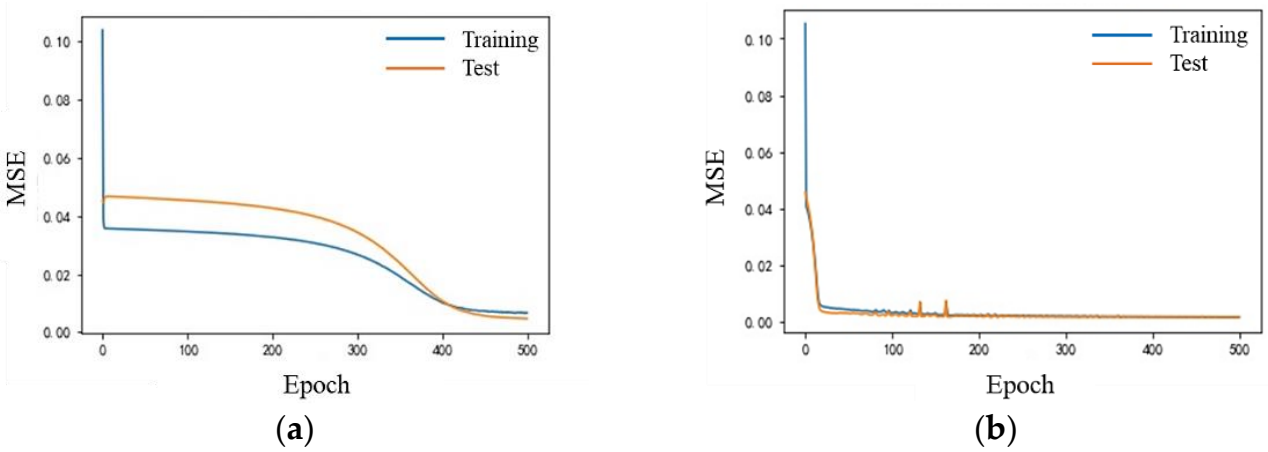

Figure 11. Loss curves for different optimizers: (**a**) SGD optimizer; (**b**) ADAM optimizer.

5.4. Model Parameters Configuration

In this paper, the training is based on the Sklearn framework for the Python platform. To avoid the problem of overfitting the model, the Dropout layer is added for improvement. By setting the value of the dropout parameter to 0.01, which means that each layer randomly discards the neuron weights of the network built in each layer with a probability of 0.01, the purpose of improving the generalization ability of the built model can be achieved.

In addition, during the training process of the model, the recurrent neural network may generally execute for a longer time, and the training duration is generally governed by the size of the input sample set, the size of the set epoch, batch size, and other parameters, hence the model hyperparameters should be set reasonably to achieve efficient training.

Here, the process of setting the dropout parameter is used as an example to illustrate the process of setting the parameter, and the other parameters are also adopted in a similar way. When the dropout is set to 0.01, 0.1, 0.5, and 0.9 respectively, the other parameters are kept constant and the model is trained. It can be seen that the model achieves the highest R^2 value on both the training set and the test set when dropout = 0.01, and the accuracy error on both the training set and the test set is the smallest, indicating that the model fits best when dropout = 0.01. Therefore, the dropout parameter is set to 0.01. The change in prediction accuracy with dropout is shown in Figure 12.

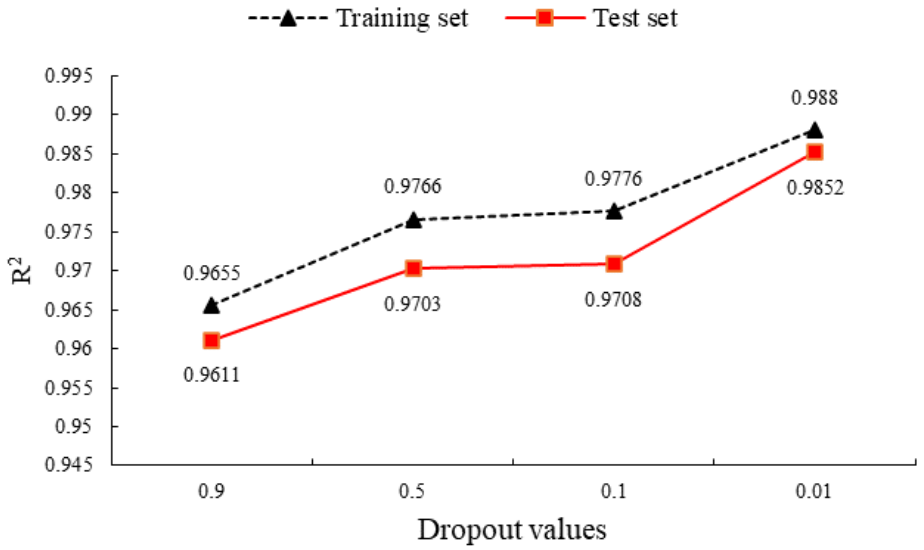

Figure 12. Dropout values selection.

During the experiments, the model was trained using the control variable method, and the input parameters of the model, i.e., dropout, learning rate, epoch, batch size, and window, were configured separately, where dropout was chosen from 0–1 with four common values of 0, 0.01, 0.5, and 0.9 for the experiments. The learning rate was experimented with in decreasing order of 0.1, 0.01, and 0.001; epoch was experimented with in increasing order of 100; batch size was experimented with in increasing order of 128; and window was experimented with in increasing order of 3. As shown in Table 6, the meanings of the main parameters of the BiLSTM-Attention model and their values are provided.

Table 6. BiLSTM-Attention model parameter configuration.

Model Parameters	Meaning	Takes Values
window	Frames the time series according to the specified unit length, i.e., the size of the sliding window.	5
lstm_units	Dimensions of the hidden layers within the LSTM cell.	16
learning_rate	Learning rate, which is negatively correlated with model training time, the higher the learning rate the shorter the training time.	0.001
dropout	The activation values of neurons stop working with a certain probability value, making the model more generalizable.	0.01
epoch	Number of training rounds.	500
input_size	Number of features of the input variable x.	11
output_size	Number of output variables y.	1
batch_size	Number of samples selected for one training session.	256
optimizer	Faster model learning and faster model convergence.	ADAM

5.5. Model Training Process

For the experiments, the training and test sets were divided into a ratio of 8:2, with a sample size of 27,245 × 5 × 11 for the training set and 7713 × 5 × 11 for the test set. The experimental procedure is divided into the following steps.

(1) Input a training set sample of size 27,245 × 5 × 11, with a step size of 5 and a dimension of 11.
(2) Randomly initialize model parameters, including dropout, learning rate, epoch, etc.

(3) The training data are learned by BiLSTM and the feature vector of (none, 5, 11) is output, connecting the weights of the temporal states by the attention mechanism layer, and finally, the prediction result of the atmospheric temperature value is output by the softmax function.

(4) The output predictions are compared with the true labels to calculate the cross-entropy loss, and the weight parameters in the network are updated by the ADAM optimizer to calculate the error loss gradient reverse optimization according to the set learning rate size.

(5) Repeat the model training process (1)–(4) over and over again according to the number of training steps.

The flow chart for model training is shown in Figure 13.

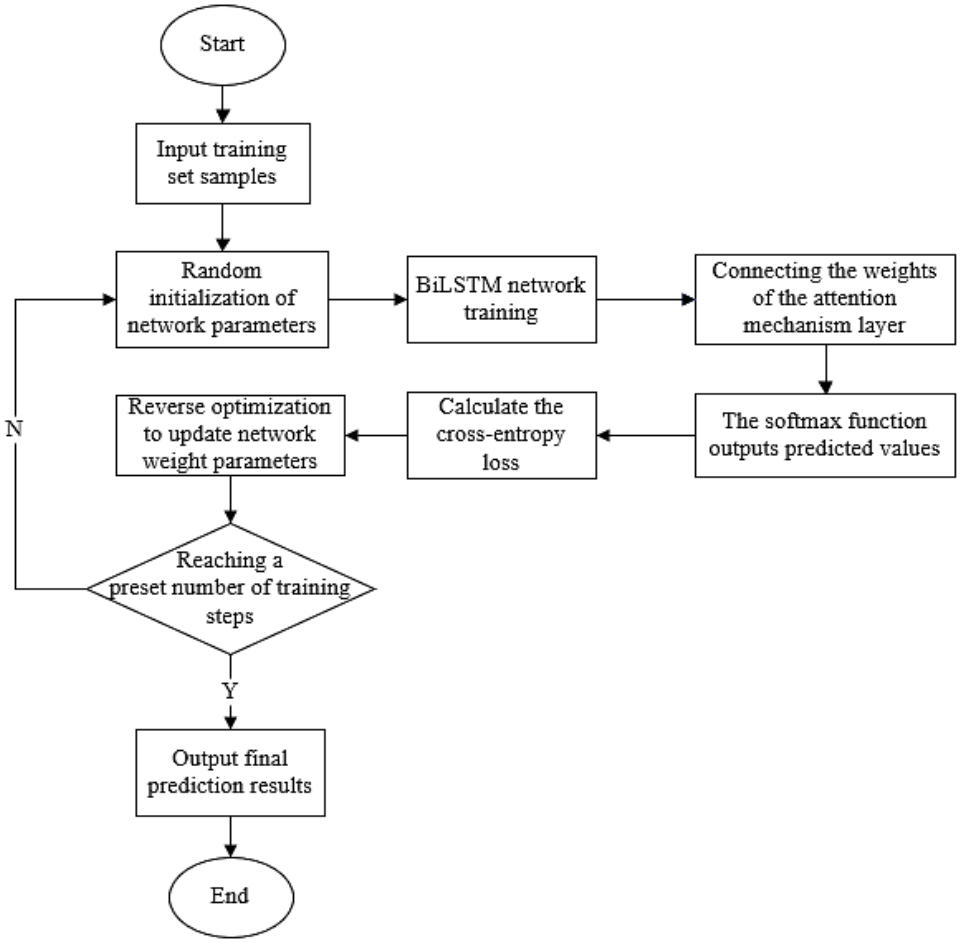

Figure 13. Model training flow chart.

5.6. Predictive Effectiveness Evaluation

5.6.1. Loss Curve

The loss curve plots of the BiLSTM-Attention model training are shown in Figure 14a. When the number of iterations is 500 rounds, the model has the best fitting effect on the loss values of the training and test sets, and the model has the best stability, i.e., the losses of both the training and test sets have converged and the difference between them is small, and the fitting effect is the best. The curve plots of BiLSTM, LSTM-Attention, and LSTM loss profiles are shown in Figure 14b–d, respectively, and it can be seen that the fitting effects of the comparison models are noisy and fluctuating, and the loss values of the test set are sometimes good and at other times bad, and the fitting effect is poor.

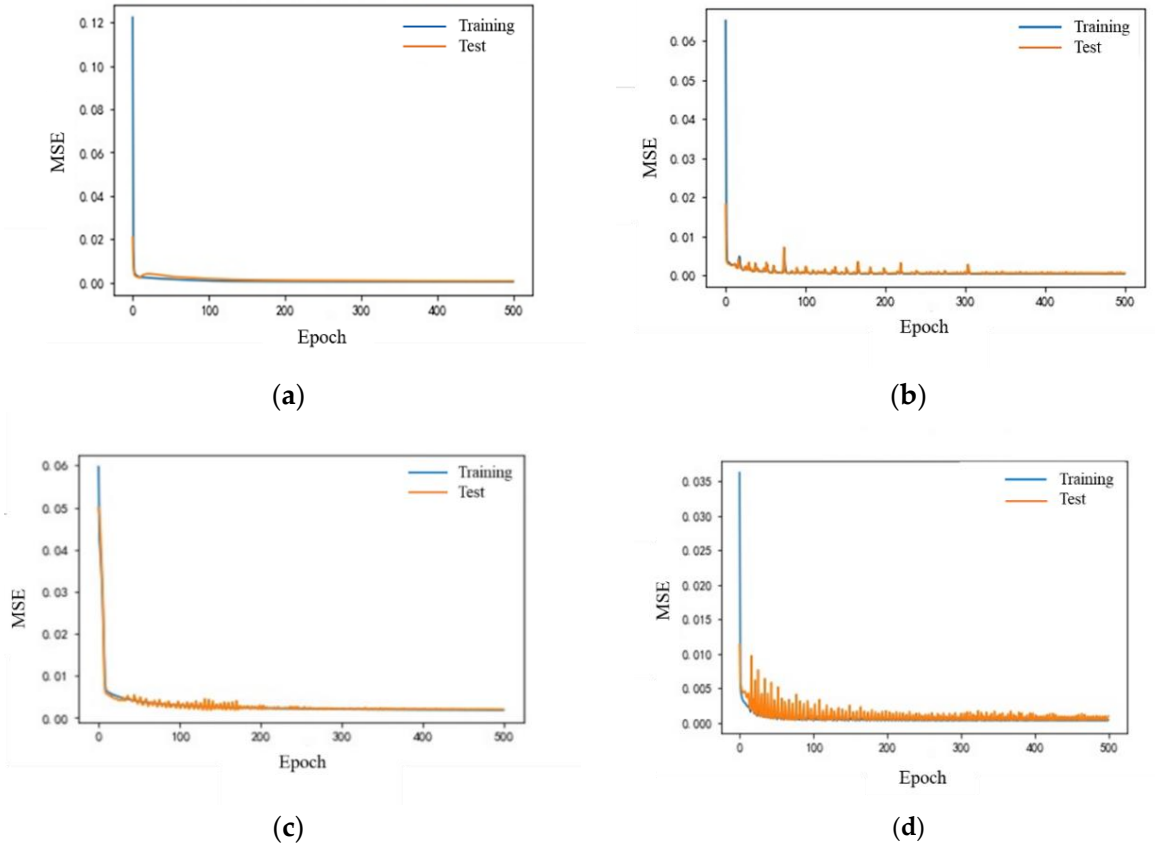

Figure 14. Loss curves for different models: (**a**) BiLSTM-Attention; (**b**) LSTM-Attention; (**c**) BiLSTM; (**d**) LSTM.

5.6.2. Visualization of Prediction Results

After the actual prediction using the BiLSTM-Attention model, it is revealed that the improved model shows very good prediction results, both in the test set and in the training set. A line graph of the model's prediction results on the test set is shown in Figure 15, where it can be seen that the errors between the true and predicted values are very small, and the curves are very close, i.e., the predicted values can reflect the magnitude of the true values well.

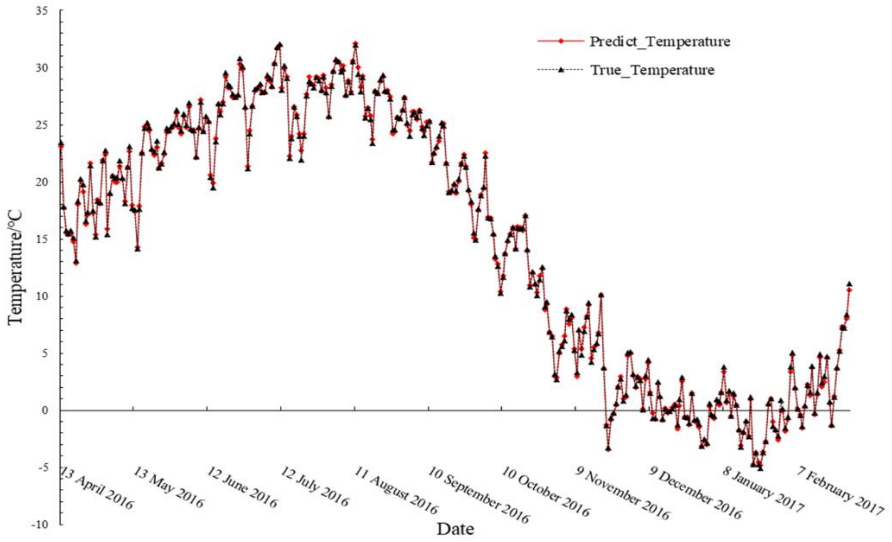

Figure 15. BiLSTM-Attention prediction results.

In order to better express the goodness of the model prediction results, the errors between the real temperature values and the predicted temperature values can be displayed using box plots, as shown in Figure 16, from which it can be seen that the errors between the predicted and real values are basically all between 0.1 and 0.3, and very few errors are slightly larger, but they are also between 0 and 1. Therefore, it is shown that the overall prediction errors are extremely small, and the model has a very good prediction effect. The model is therefore suitable for the field of atmospheric temperature prediction.

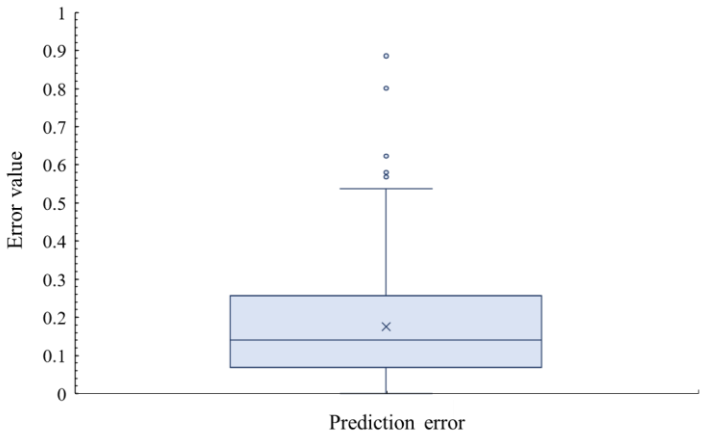

Figure 16. Prediction error box plot.

5.6.3. Comparative Testing of Models

To quantify the effectiveness of the BiLSTM-Attention model for temperature prediction, the prediction evaluation indexes of BiLSTM, LSTM-Attention, and LSTM were compared with those of Tiantan City as an example, and the results are shown in Table 7.

Table 7. Analysis of evaluation indices of each model.

Contrast Model	Training Set				Test Set				Training Time/s
	R^2	MAE/°C	MSE/°C	MAPE/°C	R^2	MAE/°C	MSE/°C	MAPE/°C	
BiLSTM-Attention	0.9922	0.0117	0.0003	3.5886	0.9618	0.0130	0.0004	4.2370	204
LSTM-Attention [23]	0.9884	0.0161	0.0004	3.6735	0.9495	0.0171	0.0006	5.0966	210
BiLSTM [34]	0.9875	0.0159	0.0005	4.4780	0.9345	0.0202	0.0007	7.0559	207
LSTM [22]	0.9809	0.0214	0.0007	4.9403	0.9120	0.0254	0.0010	7.4216	207

It can be seen that the BiLSTM-Attention model outperforms the other models in both the test set and the training set, and the prediction accuracy of the model reaches 0.9618, while the mean square error is only 0.0004, the average absolute error is only 0.0130, and the average absolute percentage error is only 4.2370. In terms of time, the average training time of each model is 207 s, while the execution time of BiLSTM-Attention model is relatively short at 204 s, which shows that the training time of this model does not increase due to the addition of the attention mechanism.

As shown in Figures 17 and 18, the BiLSTM-Attention model, both for the training and test sets, shows the highest R^2 value and its error value is the smallest. Comparing the prediction results of BiLSTM-Attention with BiLSTM, LSTM-Attention, and LSTM models, its MAE values are reduced by 0.72%, 0.41%, and 1.24% compared with BiLSTM, LSTM-Attention, and LSTM, respectively; furthermore, its MSE values are reduced by 0.03% compared with BiLSTM, LSTM- Attention, and LSTM by 0.03%, 0.02%, and 0.06%, respectively, and the R^2 values are improved by 2.73%, 1.23%, and 4.98% compared to BiLSTM, LSTM-Attention, and LSTM, respectively. This shows that the stability of the proposed improved model is better and the model relevance is stronger. This experimental result also effectively verifies that the BiLSTM-Attention model has some superiority in the field of temperature prediction.

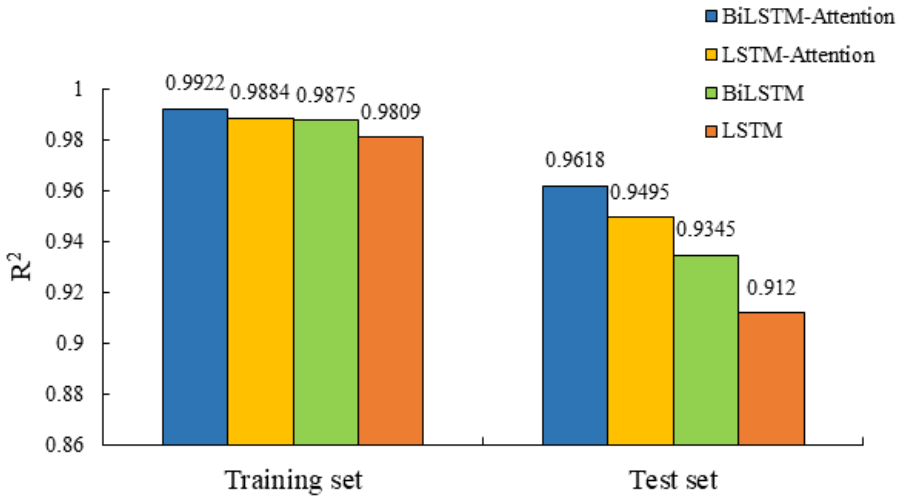

Figure 17. Comparison of R^2 metrics by model.

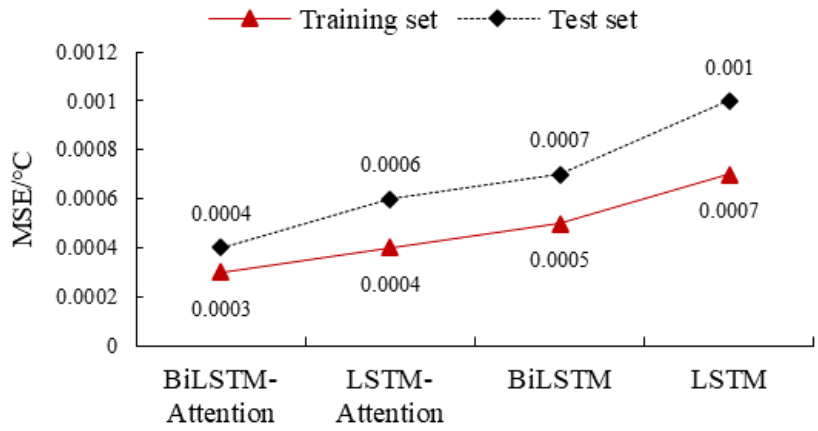

Figure 18. Comparison of MSE metrics by model.

Experiments were conducted separately for multi-site temperature data, and the prediction results based on the model were obtained for each site, as shown in Table 8. The R^2 value on the test set was maintained at 0.9638, MAE at 0.0133, MSE at 0.0004, and MAPE at 4.4130, indicating that the BiLSTM-Attention model has good model generalization ability and portability on different regional monitoring datasets.

Table 8. Evaluation index analysis of each station.

Stations	Training Set				Test Set			
	R^2	MAE/°C	MSE/°C	MAPE/°C	R^2	MAE/°C	MSE/°C	MAPE/°C
Changping	0.9886	0.0145	0.0004	3.3236	0.9632	0.0132	0.0004	4.5364
Dingling	0.9886	0.0144	0.0004	3.4486	0.9600	0.0143	0.0005	4.8303
Tiantan	0.9922	0.0117	0.0003	3.5886	0.9618	0.0130	0.0004	4.2370
Huairou	0.9905	0.0127	0.0004	2.8664	0.9649	0.0138	0.0004	4.2658
Shunyi	0.9904	0.0137	0.0004	3.4345	0.9691	0.0122	0.0004	4.1956
Mean	**0.9900**	**0.0134**	**0.0004**	**3.3323**	**0.9638**	**0.0133**	**0.0004**	**4.4130**

5.7. Discussion

From Table 7, it can be seen that the BiLSTM [34] tends to have better prediction accuracy than the LSTM [22], which also indicates that the bidirectional network structure can more fully consider the complete information of the sequence data in the forward and backward directions, thus improving the prediction accuracy of the model. At the same

time, it can be seen that the improved network based on the attention mechanism [23] has better accuracy, which also indicates that the attention mechanism can assign different attention weights to different stages of temperature change during the training process of the model, so that the model can focus on the key sequence information as much as possible, thus achieving the purpose of enhancing the improvement of temperature prediction accuracy.

The execution time of the model can also be used as a criterion to evaluate the goodness of the model. Controlling the same sample set and the same hardware environment, different comparison models were input for experiments and the training time of the model was recorded, as shown in Table 7. It can be seen that the execution times of the four major models are relatively close to each other, and the average execution time is 207 s. However, the execution time of the BiLSTM-Attention model is relatively shorter, indicating a faster execution speed.

6. Conclusions

In this paper, an improved symmetric BiLSTM network is proposed for the prediction of atmospheric temperature data, and Beijing temperature data are used as an example for validation and analysis, with the following main findings.

(1) The proposed BiLSTM-Attention model enables the model to efficiently extract feature data in a specific time step through a bidirectional LSTM network structure while retaining complete information between the past and the future. It then continuously and dynamically adjusts the weight values of different channels based on the attention mechanism, which in turn enables efficient allocation of computational resources, and can effectively improve the model's temperature prediction accuracy.
(2) The model is used for temperature data prediction and compared with BiLSTM, LSTM-Attention, and LSTM models, and it was found that the proposed BiLSTM-Attention model has the highest prediction accuracy and the lowest error, reduced by 0.72%, 0.41%, and 1.24%, respectively; the R^2 value reached 0.9618, which improved by 2.73%, 1.23%, and 4.98% compared with BiLSTM, LSTM-Attention, and LSTM, respectively. Thus, it is shown that the BiLSTM-Attention model has good theoretical value and practical application significance, and can provide better solutions in the field of temperature prediction.

The following directions exist for future research that deserve further study.

(1) Consider using more efficient and fast hyperparameter optimization methods to optimize the model parameters to obtain a more suitable parameter configuration for the temperature prediction domain.
(2) In future research, more regional temperature datasets should be collected for a more complete prediction analysis to make the model more reliable and adaptable.
(3) More detailed comparison experiments can be attempted in future studies to prove the superiority of the model.

Author Contributions: Conceptualization, Y.L. and X.H.; methodology, Y.L. and X.H.; validation, Y.D. and Y.L.; writing—original draft preparation, Y.L. and L.P.; writing—review and editing, X.H., L.P. and Y.L.; project administration, X.H., L.P. and Y.L.; funding acquisition, X.H. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by National Natural Science Foundation of China (51908059), Fundamental Research Funds for the Central Universities, CHD (300102240206), Key R&D Projects in Shaanxi Province (2022JBGS3-08) and Chang'an University Ph.D. Candidates' Innovative Capacity Development Grant Program (300203211241).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, D.; Cao, Z.; Chen, B.; Ni, Y. Multivariate Time Series Local Support Vector Regression Forecast Methods for Daily Temperature. *J. Syst. Simul.* **2016**, *28*, 654–660.
2. Zhu, J.; Zhao, X.; Wu, S.; Hui, W.; Xing, C.; Center, H.C. Temperature Forecast Model Based on Support Vector Machine Method. *Nat. Sci. J. Hainan Univ.* **2016**, *34*, 40–44.
3. Cong, L.; Cai, J. The Application of Auto-Regressive and Moving Average Model in Harbin Temperature Fore-cast. *Math. Pract. Theory* **2012**, *42*, 190–195.
4. Zhang, W.Y.; Xie, J.F.; Wan, G.C.; Tong, M.S. Single-step and Multi-step Time Series Prediction for Urban Temperature Based on LSTM Model of TensorFlow. In Proceedings of the 2021 Photonics & Electromagnetics Research Symposium (PIERS), Hangzhou, China, 21–25 November 2021; pp. 1531–1535.
5. Ma, D.; Ma, S.; Chen, Q.; Yang, C. Temperature Prediction Algorithm based on Spatio-temporal Prediction. In Proceedings of the 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China, 4–6 May 2022; pp. 151–157.
6. Choi, H.-M.; Kim, M.-K.; Yang, H. Abnormally High Water Temperature Prediction Using LSTM Deep Learning Model. *J. Intell. Fuzzy Syst.* **2021**, *40*, 8013–8020. [CrossRef]
7. Kun, X.; Shan, T.; Yi, T.; Chao, C. Attention-based long short-term memory network temperature prediction model. In Proceedings of the 2021 7th International Conference on Condition Monitoring of Machinery in Non-Stationary Operations (CMMNO), Guangzhou, China, 11–13 June 2021; pp. 278–281.
8. Park, I.; Kim, H.S.; Lee, J.; Song, C.H. Temperature prediction using the missing data refinement model based on a long short-term memory neural network. *Atmosphere* **2019**, *10*, 718. [CrossRef]
9. Wang, X.; Li, Z.; Zhang, J.; Liu, H.; Qiu, C.; Cai, X. An LSTM-attention wind power prediction method considering multiple factors. In Proceedings of the 8th Renewable Power Generation Conference (RPG 2019), Shanghai, China, 24–25 October 2019; IET: London, UK, 2019; pp. 1–7.
10. Cao, Z. Daily Temperature and Drought Index Support Vector Regression Prediction Methods. Master's Thesis, Nanjing University of Information Engineering, Nanjing, China, 2015.
11. Jiang, W.; Wang, Y.; Hao, X.; Li, F. Application of Decision Tree in Temperature Predicition. *Comput. Appl. Softw.* **2012**, *29*, 141–144.
12. Dan, Y.; Dong, R.; Cao, Z.; Li, X.; Niu, C.; Li, S.; Hu, J. Computational prediction of critical temperatures of superconductors based on convolutional gradient boosting decision trees. *IEEE Access* **2020**, *8*, 57868–57878. [CrossRef]
13. Tao, H.; Junjie, L.; Yu, S.; Yongjian, C.; Zhenyu, L. Predictive analysis of indoor temperature and humidity based on BP neural network single-step prediction method. In Proceedings of the 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 27–29 September 2020; pp. 402–407.
14. Zhou, H.; Li, H.; Sun, X.; Yan, W. A Temperature Forecasting Model of Grey-Markov Based on Seasonal Index. *Math. Pract. Theory* **2016**, *46*, 167–173.
15. Raviprasad, S.; Angadi, N.A.; Kothari, M. Tree Based Models for Critical Temperature Prediction of Superconductors. In Proceedings of the 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 27–29 May 2022; pp. 1–5.
16. Hou, H. Global Climate Change Prediction Model Based on BP Neural Network. *Sci. Technol. Innov.* **2021**, *9*, 10–11.
17. Qi, C.; Wenbiao, W.; Siyuan, W. Application of indoor temperature prediction based on SVM and BPNN. In Proceedings of the 27th Chinese Control and Decision Conference (2015 CCDC), Qingdao, China, 23–25 May 2015; pp. 2883–2887.
18. Xu, B.; Dan, H.C.; Li, L. Temperature prediction model of asphalt pavement in cold regions based on an improved BP neural network. *Appl. Therm. Eng.* **2017**, *120*, 568–580. [CrossRef]
19. Zhang, K.; Guliani, A.; Ogrenci-Memik, S.; Memik, G.; Yoshii, K.; Sankaran, R.; Beckman, P. Machine learning-based temperature prediction for runtime thermal management across system components. *IEEE Trans. Parallel Distrib. Syst.* **2017**, *29*, 405–419. [CrossRef]
20. Yang, L.; Shang, L.; Yang, L. Improved Prediction of Markov Chain Algorithm for Indoor Temperature. In Proceedings of the 2016 International Symposium on Computer, Consumer and Control (IS3C), Xi'an, China, 4–6 July 2016; pp. 809–812.
21. Kim, M.; Yang, H.; Kim, J. Sea surface temperature and high water temperature occurrence prediction using a long short-term memory model. *Remote Sens.* **2020**, *12*, 3654. [CrossRef]
22. Qiu, R.; Wang, Y.; Rhoads, B.; Wang, D.; Qiu, W.; Tao, Y.; Wu, J. River water temperature forecasting using a deep learning method. *J. Hydrol.* **2021**, *595*, 126016. [CrossRef]
23. Suleman, M.A.R.; Shridevi, S. Short-Term Weather Forecasting Using Spatial Feature Attention Based LSTM Model. *IEEE Access* **2022**, *10*, 82456–82468. [CrossRef]
24. Song, X.; Huang, J.; Song, D. Air quality prediction based on LSTM-Kalman model. In Proceedings of the 2019 IEEE 8th Joint International In-formation Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 24–26 May 2019; IEEE: New York, NY, USA, 2019; pp. 695–699.
25. Liu, J.; Zhang, T.; Han, G.; Gou, Y. TD-LSTM: Temporal dependence-based LSTM networks for marine temperature prediction. *Sensors* **2018**, *18*, 3797. [CrossRef] [PubMed]
26. Yu, T.; Pei, L.; Li, W.; Sun, Z.Y.; Huyan, J. Prediction of Pavement Surface Condition Index Based on Random Forest Algorithm. *J. Highw. Transp. Res. Dev.* **2021**, *38*, 16–23. [CrossRef]

27. Aziz, N.; Abdullah, M.H.A.; Zaidi, A.N. Predictive Analytics for Crude Oil Price Using RNN-LSTM Neural Network. In Proceedings of the 2020 International Conference on Computational Intelligence (ICCI), Bandar Seri Iskandar, Malaysia, 8–9 October 2020; IEEE: New York, NY, USA, 2020; pp. 173–178.
28. Chandriah, K.K.; Naraganahalli, R.V. RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting. *Multimed. Tools Appl.* **2021**, *80*, 26145–26159. [CrossRef]
29. Chen, D.; Zhang, J.; Jiang, S. Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and LSTM neural networks. *IEEE Access* **2020**, *8*, 91181–91187. [CrossRef]
30. Zhao, G.; Jiang, P.; Lin, T. Remaining Life Prediction of Rolling Bearing Based on CNN-BiLSTM Model with Attention Mechanism. *Mech. Electr. Eng. Mag.* **2021**, *38*, 1253–1260.
31. Cheng, Q.; Li, H.; Wu, Q.; Meng, F.; Xu, L.; Ngan, K.N. Learn to Pay Attention Via Switchable Attention for Image Recognition. In Proceedings of the 2020 IEEE Confer-ence on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 6–8 August 2020; IEEE: New York, NY, USA, 2020; pp. 291–296.
32. Mahato, N.K.; Dong, J.; Song, C.; Chen, Z.; Wang, N.; Ma, H.; Gong, G. Electric Power System Transient Stability Assessment Based on Bi-LSTM Attention Mechanism. In Proceedings of the 2021 6th Asia Conference on Power and Electrical Engineering (ACPEE), Chongqing, China, 8–11 April 2021; IEEE: New York, NY, USA, 2021; pp. 777–782.
33. Liu, J.; Liu, J.; Luo, X. Research Progress in Attention Mechanism in Deep Learning. *Chin. J. Eng.* **2021**, *43*, 1499–1511.
34. Zhai, M. Research on the Prediction Effect of LSTM/BILSTM-ARMA Model Based on Signal Decomposition for Influenza in Shanxi Province. Shanxi Medical University: Taiyuan, China, 2021.
35. Pei, L.; Sun, Z.; Hu, Y.; Li, W.; Gao, Y.; Hao, X. Neural Network Model for Road Aggregate Size Calculation Based on Multiple Features. *J. South China Univ. Technol.* **2020**, *48*, 77–86.