applied sciences

MDPI

Article

Multivariate Time Series Data Prediction Based on ATT-LSTM Network

Jie Ju and Fang-Ai Liu *

School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China; ju111jie@163.com
* Correspondence: lfa@sdnu.edu.cn

Abstract: Deep learning models have been widely used in prediction problems in various scenarios and have shown excellent prediction effects. As a deep learning model, the long short-term memory neural network (LSTM) is potent in predicting time series data. However, with the advancement of technology, data collection has become more accessible, and multivariate time series data have emerged. Multivariate time series data are often characterized by a large amount of data, tight timeline, and many related sequences. Especially in real data sets, the change rules of many sequences will be affected by the changes of other sequences. The interacting factors data, mutation information, and other issues seriously impact the prediction accuracy of deep learning models when predicting this type of data. On the other hand, we can also extract the mutual influence information between different sequences and simultaneously use the extracted information as part of the model input to make the prediction results more accurate. Therefore, we propose an ATT-LSTM model. The network applies the attention mechanism (attention) to the LSTM to filter the mutual influence information in the data when predicting the multivariate time series data, which makes up for the poor ability of the network to process data. Weaknesses have greatly improved the accuracy of the network in predicting multivariate time series data. To evaluate the model's accuracy, we compare the ATT-LSTM model with the other six models on two real multivariate time series data sets based on two evaluation indicators: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The experimental results show that the model has an excellent performance improvement compared with the other six models, proving the model's effectiveness in predicting multivariate time series data.

Keywords: multivariate time series data prediction; attention mechanism; LSTM

Citation: Ju, J.; Liu, F.-A. Multivariate Time Series Data Prediction Based on ATT-LSTM Network. *Appl. Sci.* **2021**, *11*, 9373. https://doi.org/10.3390/app11209373

Academic Editor: Krzysztof Koszela

Received: 3 September 2021
Accepted: 5 October 2021
Published: 9 October 2021

1. Introduction

With the advancement of today's social science and technology, the structure of data has become more complex. The amount of data has been ever-increasing, which marks our entry into the era of big data [1]. As data has become complex and massive, its hidden information, value, and laws have also increased. Therefore, only by fully mining and analyzing a large number of details in life can the knowledge and value contained in it be extracted and further fed back to people's daily production and life. The analysis of big data is usually divided into two parts. On the one hand, a summary conclusion is obtained through the study of big historical data in the past few months or years. The other kind of analysis is performed through the study of big historical data digging and predicting the state of work in the next few months. With the continuous improvement of social productivity and the increasing demand of people for future cognition, predicting the future state through the analysis of big historical data has become the top priority in the considerable current data analysis work [2].

The development of deep learning, especially neural networks, has shown strong performance and brought new vitality to extensive data analysis and prediction, and has achieved excellent results. Nowadays, deep learning, especially with neural networks,

is widely used in extensive data analysis and forecasting in various industries, such as weather forecasting [3], traffic flow forecasting [4], financial forecasting [5], medical forecasting [6], etc. LSTM [7] networks have shown sound prediction effects and stood out among many neural networks. The network solves the problem of gradient disappearance and gradient explosion in the long sequence training process of the recurrent neural network (RNN) [8]. The network has certain advantages in sequence modeling and has a long-term memory function. At present, LSTM is widely used in big data prediction, machine translation, image processing, encoding/decoding, and other fields.

With the advancement of technology, data collection has become more accessible, and multivariate time series data [9] came into being. This type of data often has the characteristics of massive data, close time points, and many related variables. The change of a sequence often depends on the changes of several or even dozens of sequences. Multivariate time series data often have multiple time-related variables, and each variable depends not only on its past value but also on other variables. These kinds of mutual influence rules between sequences allows us to grasp more information, capture the law of sequence changes, and improve prediction accuracy. The generation of multiple time series data has brought significant challenges to extensive data analysis and forecasting. A single neural network has not achieved a good prediction effect in the prediction of multiple time series data, and it also exposed some problems in the neural network. The generation of the attention mechanism [10] makes up for the shortcomings of a single neural network's insufficient processing capability for multiple time series data. The attention mechanism can assign necessary weights to different representations, ignoring noise and redundancy in the input. The attention mechanism can also directly evaluate which inputs are preferred for the next task by checking the weights. In short, the attention mechanism can remove redundant information and noise in the multivariate time series data, and filter the multivariate time-related variables, assign different weights, and then filter the input of the following neural network. The main contributions of our model are as follows:

(1) We propose a long short-term memory neural network (ATT-LSTM) based on the attention mechanism for multivariate time series data prediction.
(2) Use an attention mechanism to process multiple time series data. The attention mechanism can reduce the effect of irrelevant information on the results and enhance the influence of related information by assigning different weights and improve prediction results' accuracy.
(3) We compare the proposed ATT-LSTM model with the other six models on two real multivariate time series data sets based on two evaluation indicators: MAE and RMSE. The results prove the effectiveness of the model in predicting multivariate time series data.

The main structure is as follows: the second part introduces the related work of the LSTM and attention mechanism. The third part explains the ATT-LSTM model. The fourth part is the part of experimenting on two data sets. The fifth part is the summary part, which summarizes the results of our work and the existing shortcomings.

2. Related Work

Deep learning [11] is an algorithm tool in the current era of big data. It has become a research hotspot in recent years, and it has also achieved breakthrough development. Deep learning, especially neural networks, has achieved excellent results in search technology, data mining, machine translation, natural language processing, multimedia learning, and other fields. As deep learning develops, especially artificial neural networks, predictive models based on artificial neural networks have begun to appear in extensive data analysis. Among them, the prediction model based on LSTM [12] has shown particularly excellent prediction performance. Chen K [13] et al. used LSTM to model and predict stock returns, showing the predictive performance of LSTM. Altché F [14] et al. used LSTM to achieve the first step of consistent trajectory prediction and successfully and accurately predict the future longitudinal and lateral traces of vehicles on highways. Li Y F [15] et al. applied

LSTM to tourism flow forecasting and achieved good forecasting results. As the amount of data increases and the complexity of the data continues to grow, the LSTM networks have gradually exposed its deficiencies in the complex data processing. To solve this problem, researchers further improved the LSTM or combined it with other methods to further enhance its prediction performance. Bai Y [16] et al. proposed an integrated long short-term memory neural network (E-LSTM) based on the LSTM for PM2.5 concentration prediction. To predict the concentration of air pollutants, Qi Y [17] et al. also proposed a hybrid model based on a deep learning method, which combines graph convolutional neural networks and LSTM to predict PM2.5 concentration. Xie G. [18] et al. proposed a trajectory prediction method based on a sequential model, combining a convolutional neural network (CNN) and LSTM to accurately monitor the surrounding environment. Zhou J. [19] et al. proposed a water quality prediction method based on an improved grey relational analysis (IGRA) algorithm and LSTM. The emergence of time series data has once again brought considerable challenges to big data prediction. To cope with this challenge, researchers combined the attention mechanism with neural networks to analyze and predict multivariate time series data. Qin Y. [20] et al. combined the attention mechanism with the recurrent neural network and proposed a two-stage recurrent neural network (DA-RNN) based on attention. In the first stage, the attention mechanism was introduced to extract the sequence adaptively. In the second stage, the time attention mechanism was used to select the hidden state of the encoder. The model effectively made predictions on two data sets and achieved good prediction results. Zheng C. [21] et al. proposed a graph multi-attention network (GMAN) based on time and space factors to predict traffic conditions at different locations on the road network map ahead of time. To solve the problem of the insufficient ability of LSTM to process multi-feature data, Li Y. [22] et al. proposed an LSTM training method based on evolutionary attention, combined with a competitive random search for multivariate time series prediction. To carry out a long-term forecast of multivariate time series, Liu Y. [23] et al. proposed DSTP-based RNN (DSTP-RNN and DSTP-RNN-II) prediction models. Through studying the results of many researchers in the field of multivariate time series data analysis and prediction, it is found that, although the analysis and forecast of multivariate time series data have achieved remarkable results, it still faces several enormous challenges:

(1) The diversity of multiple time series data. Various sequences often determine changes in one sequence, and at the same time, changes in one sequence often affect various sequences. This makes the data analysis process particularly complicated.
(2) The time series of multiple time series data. Multivariate time series data are data that changes with time and often have a changing law within a certain period, which will significantly affect the overall analysis of the data by the model.
(3) The instability of multiple time series data. Multivariate time series data are often realistic data sets, and most of the data contain strange information, such as missing values and data mutations. How to deal with this strange information is also an excellent challenge for multivariate time series data prediction.
(4) How to correctly grasp the mutual influence information between sequences. The interaction information between the sequences can make the data analysis and prediction process more complicated on the one hand. It can also improve the accuracy of the prediction results on the other hand. The key question is how to remove irrelevant influence information and grasp the interaction information between sequences.

3. ATT-LSTM Model

3.1. Long Short-Term Memory Neural Network

Long short-term memory neural network (LSTM) is a special kind of recurrent neural network (RNN). LSTM was first proposed by Hochreiter and Schmidhuber [7], which effectively solves the RNN network time delay and gradient disappearance. LSTM [24] is widely used in text generation, machine translation, speech recognition, generated image description, and video tagging. LSTM is more potent than ordinary RNN because it can

selectively record or forget the input information. On the one hand, it has a powerful memory block, which mainly contains three gates (memory gate, forgetting gate, and output gate), on the other hand, it also has a memory unit, which can control the transfer of information to the next moment [25]. The network structure of the LSTM is shown in Figure 1.

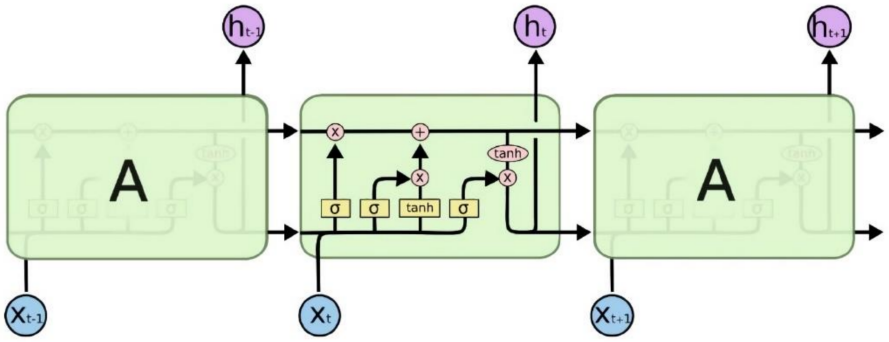

Figure 1. LSTM network structure diagram.

The unit structure of LSTM is mainly composed of input gate, forget gate, output gate, and unit state [26]. The unit structure diagram of the LSTM is shown in Figure 2.

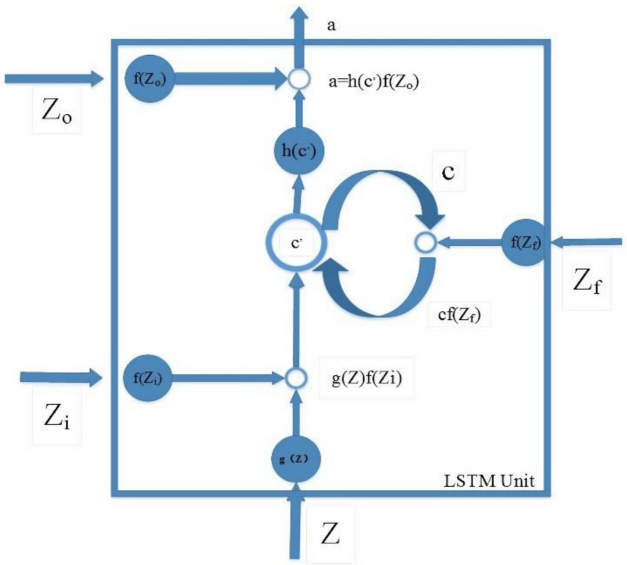

Figure 2. LSTM unit structure diagram.

Z represents the input, and Z_i represents the control signal of the input gate. Z_f represents the control signal of the forget gate, Z_o represents the control signal of the output gate, and the $f(x)$ function is usually the Sigmod function:

$$f(x) = \frac{1}{1+e^{-x}}. \tag{1}$$

This function is used to indicate the degree of opening of the door, and the value range is within [0, 1]. Both $g(x)$ and $h(x)$ are activation functions. First, Z passes the activation function to get $g(Z)$, Z passes the Sigmod function to get $f(Z_i)$ and then multiplies it to get $g(Z)f(Z_i)$. Z_f gets $f(Z_f)$ through the Sigmod function and then multiplies it with the value a of the memory unit at the last moment to get $cf(Z_f)$. Finally, update the value of the memory unit to:

$$c' = g(Z)f(Z_i) + cf\left(Z_f\right). \tag{2}$$

c' passes the activation function to get $h(c')$, Z_o passes the Sigmod function to get $f(Z_o)$, and multiplies it to get the output:

$$\mathrm{a} = h(c')f(Z_o). \tag{3}$$

The input x_t at time t and the output h_{t-1} of the hidden layer neuron at time $t-1$ are jointly used as the input part of the hidden layer at time t, multiplied by different weight vectors. The activation function is used to obtain the control signals of the three gates Z_f, Z_i, Z_o, and the input value Z. The formula is as follows:

$$Z_f = \omega_f \cdot [h_{t-1}, x_t] + b_f \tag{4}$$

$$Z_i = \omega_i \cdot [h_{t-1}, x_t] + b_i \tag{5}$$

$$Z_o = \omega_o \cdot [h_{t-1}, x_t] + b_o \tag{6}$$

$$Z = \omega_x \cdot [h_{t-1}, x_t] + b_x. \tag{7}$$

Among them, b_f, b_i, b_o, and b_x are the biases of different connection weights. After the LSTM unit is operated, the value c of the memory unit is updated Equation (1) and the Formula (8) is obtained. The output of hidden layer neurons is Formula (9):

$$\begin{aligned} c' = \; & g(\omega_x \times [h_{t-1}, x_t] + b_x) f(\omega_i \times [h_{t-1}, x_t] + b_t) \\ & + c f(\omega_f \times [h_{t-1}, x_t] + b_f) \end{aligned} \tag{8}$$

$$\begin{aligned} h_t &= h(c')f(Z_o) \\ &= h(g(\omega_x \cdot [h_{t-1}, x_t] + b_x) f(\omega_i \cdot [h_{t-1}, x_t] + b_i) \\ &\quad + c f(\omega_f \cdot [h_{t-1}, x_t] + b_f)) f(\omega_o \cdot [h_{t-1}, x_t] + b_o) \end{aligned} \tag{9}$$

It can be seen from the above equations that the input of LSTM not only includes the output h_{t-1} of the hidden layer neuron at the last moment, but also includes the value of the memory unit in the LSTM unit. The LSTM network can effectively avoid the occurrence of gradient disappearance, can memorize long-term historical information, and fit long-term time series data more effectively [27].

3.2. Attention Mechanism

The attention mechanism has been widely used in various fields of deep learning in recent years. It is a resource allocation scheme that is the primary means to solve information overload [28]. The attention mechanism is similar to the artificial neural network and originated from the human behavior mechanism. The attention mechanism is derived from human vision and draws on the human visual attention mechanism. The core idea is to select more critical information to the current task goal from a lot of information [29]. The attention mechanism also has a variety of classification methods, which can be divided into spatial attention and temporal attention according to space and time; it can be divided into soft attention and complex attention according to the nature of work, as shown in Figure 3.

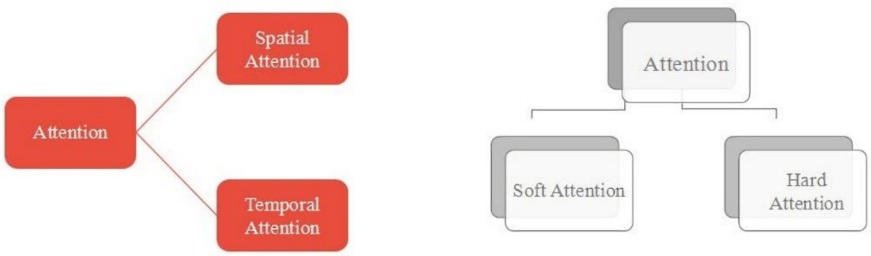

Figure 3. Attention mechanism classification.

The essence of the attention mechanism has two aspects: (1) Determine which part of the data the model needs to focus on from the input data and assign different weights; (2) Assign limited information processing resources according to the importance weight vector [30,31].It is shown in Figure 4.

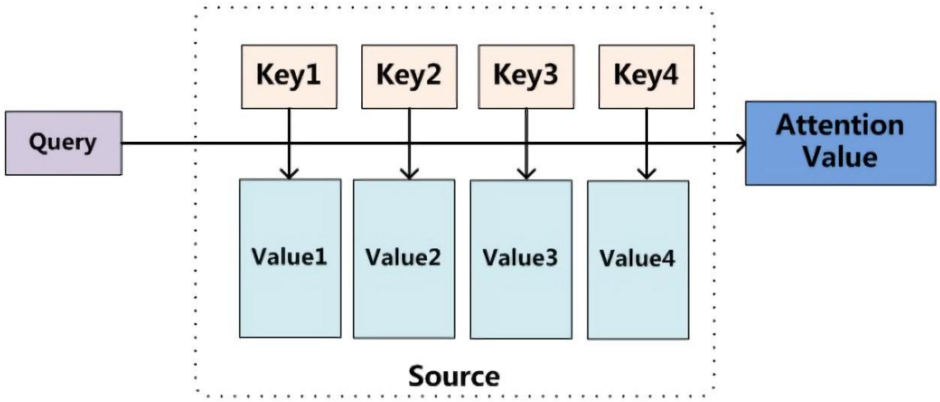

Figure 4. Attention mechanism.

The attention mechanism imagines the constituent elements in the data as a series of <Key, Value> data pairs. The attention mechanism is to perform a weighted summation of the value of the elements in the data, where key and query are used to calculate the weight coefficients of the corresponding value, as shown in the following formula:

$$Attention(Query, Source) = \sum_{i=1}^{L_x} similarity(Query, Key_i) * Value_i. \quad (10)$$

Among them, L_x represents the length of the data source. Attention mechanisms are widely used in natural language processing, classification, recommendation systems, and extensive data analysis. The analysis and prediction of multivariate time series data have also been further developed because of the emergence of the attention mechanism.

3.3. ATT-LSTM Model

In this section, we will describe in detail the ATT-LSTM model structure. First, this section will explain the data processing and experimental procedures of the entire ATT-LSTM model. Secondly, this section will give a detailed introduction to the composition of the ATT-LSTM model and the composition of each module. We aim to analyze and predict multivariate time series data. To solve the large scale and high complexity characteristics of current multivariate time series data, we apply the attention mechanism to the LSTM and proposes the ATT-LSTM network model. The ATT-LSTM model has five processes in the experiment, and the specific process is shown in Figure 5.

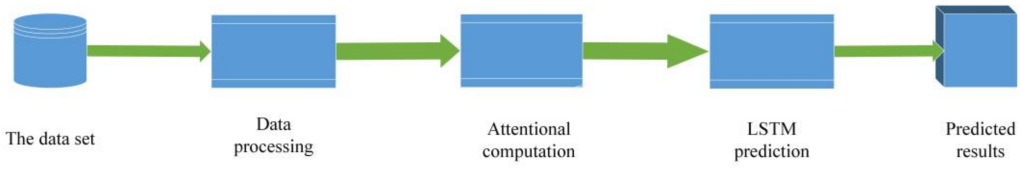

Figure 5. Model flow chart.

As shown in Figure 5, the ATT-LSTM model is divided into the following five steps in data processing and data analysis and prediction:

(1) Data set selection. We select two real multivariate time series data sets for follow-up experiments. We will elaborate on the details of each data set in the experimental section.

(2) Data processing. Since we choose an actual multivariate time series data set, the processed data set can be input to the model. The data set processing mainly includes two parts: ① missing value processing; ② mutation information processing.
(3) Attention mechanism processing. Multivariate time series data has several or even dozens of related sequences, and changes in each sequence will affect the target sequence. According to historical lows, the attention mechanism calculates the attention value of each sequence, which is the distribution weight of each sequence. Finally, analyze and predict based on the attention value of each sequence.
(4) LSTM network prediction. LSTM analyzes and predicts data based on the input data and the attention value processed by the attention mechanism.
(5) Model output. According to the data set's attributes, the analysis and prediction results of the ATT-LSTM model are output intuitively. The model is compared with other models according to the two evaluation indicators of RMSE and MAE.

As shown in Figure 6, the ATT-LSTM model is mainly composed of the input, ATT-LSTM module, and output. Among them, the input data are multivariate time series data. The red box represents the ATT-LSTM module, which is also the main structure of the ATT-LSTM model. The processed time series data are first input to the ATT module for attention value calculation and the state of the LSTM module at the last moment is also input to the ATT module for attention calculation. The ATT module calculates the attention value and inputs the data to the LSTM module. The LSTM network analyzes and predicts based on the data processed and calculated by the ATT module and outputs the prediction result.

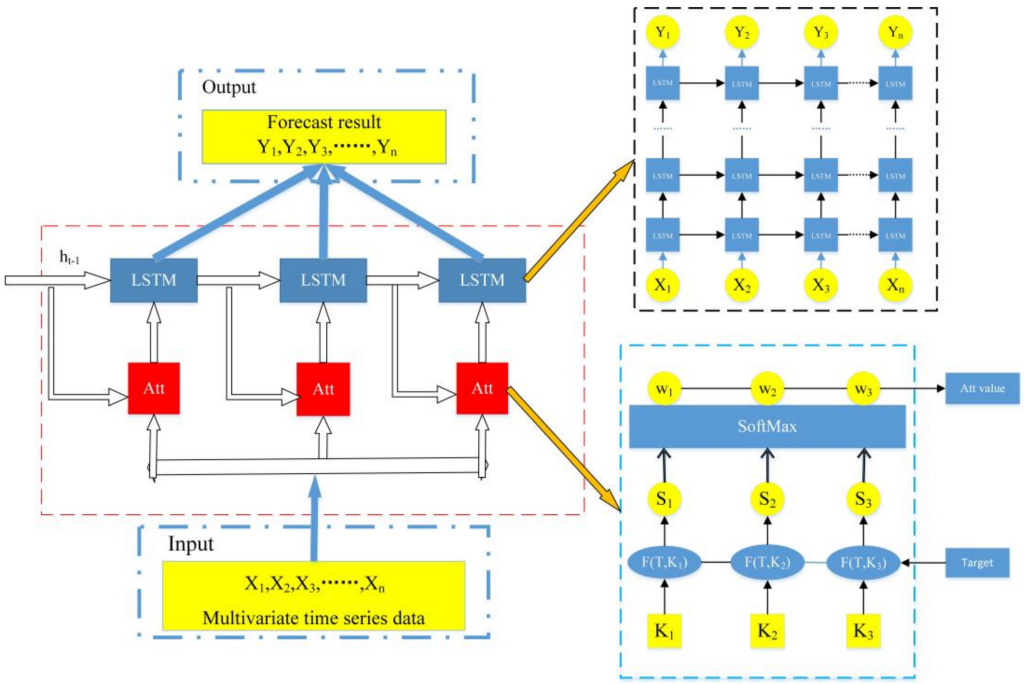

Figure 6. ATT-LSTM model structure diagram.

The blue box on the right represents the detailed structure of the ATT module. K_1, K_2, $K_3 \ldots K_n$ represent non-target sequences of multivariate time series data. The target on the right represents the target sequence. In the first stage, the function is used to calculate the correlation between the non-target sequence K and the target list sequence T:

$$Similarity(Target, K_n) = Target \cdot K_n. \tag{11}$$

Among them, the commonly used functions are dot product, Cosine of two vectors and neural network, and Formula (11) uses dot product function for calculation.

In the second phase, the result of stage one is converted through the Softmax function. The converted weight value is between 0–1, and the sum is 1. Doing so can make the sequences related to the Target sequence more prominent. The weight value of each element is calculated as follows:

$$w_i = Softmax(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_i}}. \quad (12)$$

where w_i is the attention value calculated by the ATT module. After processing by the ATT module, the non-target sequences are assigned their respective Attention values and then input into the LSTM module for analysis and prediction. The black box on the right side of Figure 6 represents the LSTM structure. The whole represents an LSTM module with an uncertain number of layers. Among them, the number of LSTM layers is determined according to the specific effects of the experiment. Each layer of the LSTM consists of several LSTM units connected to each other. Among them, the unit state of each LSTM unit can either be input to the adjacent LSTM unit of the same layer or input to the LSTM unit of the next layer.

4. Experiment

This section mainly introduces the experimental part in detail. It is divided into three parts. The first part is a detailed introduction to the data set and model evaluation indicators. The second part will introduce the comparison model and experimental parameter settings of the comparison experiment. The third part describes the experimental results, which visually display the prediction results of the ATT-LSTM model and the comparison test results with the comparison model.

4.1. Dataset and Evaluation Indicators

4.1.1. Dataset

The data sets are two real multivariate time series data sets, the Nasdaq 100 stock data set and the Beijing PM2.5 data set. The following is a detailed description of the two data sets, and the relevant information of the two data sets is shown in Table 1:

Table 1. Dataset Details.

Dataset	NASDAQ 100	PM2.5 of Beijing
Target series	NDX	PM2.5
Related series	104	8
Time	2016/07/26–2017/04/28	2020/01/01–2015/12/31
Time Intervals	1 min	1 h
The amount of data	74,501	52,584
Train/Validation/Test	56875/7626/10000	39438/3146/10000

(1) Nasdaq 100 stock data set: This data set consists of the stock prices of 104 companies under the Nasdaq 100 and the index value of the Nasdaq 100. The data collection frequency was 1 min. The data set contains 191 days of closing data from 26 July 2016 to 28 April 2017, and includes 390 data points per day. We take the Nasdaq 100 index as the target sequence and the stock prices of other 104 companies as the correlation sequence.

(2) Beijing PM2.5 concentration dataset: This dataset is a PM2.5 concentration dataset sampled by the US Embassy in Beijing. The frequency of data collection was 1 h. The data set includes the PM2.5 concentration from 1 January 2020 to 31 December 2015 and other related factors (temperature, humidity, wind direction, wind force, rainfall, etc.). We take PM2.5 concentration as the target sequence and other related factors as the correlation sequence.

4.1.2. Model Evaluation Indicators

We used Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as the evaluation indicators of the model. At the same time, MAE and RMSE were also used as evaluation indicators for comparison experiments with other models. The calculation formulas of these two evaluation indicators are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - x_i| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2} \tag{14}$$

where N represents the length of the test set, y_i represents the model's predicted value, and x_i represents the actual value.

4.2. Comparison Model and Parameter Configuration

To evaluate the effect of the ATT-LSTM model, we compare the model with the following models:

VAR: The vector autoregressive model (VAR) adopts the simultaneous form of multiple equations and is not based on economic theory. It is an unstructured multiple equation model. This model is often used to predict relevant time series data.

LSTM: Long Short-Term Memory Neural Network (LSTM) [32] is a variant of Recurrent Neural Network (RNN), which solves the long-term dependence problem of RNN. It is a classic time series data prediction network.

GRU: Gated Recurrent Unit (GRU) [33] is a very effective variant of the LSTM network. GRU has a more straightforward structure than LSTM.

LSTM+Zoneout: Adding Zoneout technology [34] to a single-layer LSTM network can significantly improve the learning ability and improve the LSTM network's predictive ability.

LSTM-RNN: This model was proposed by Abdel-Nasser M. [35] and others, which further reduces the prediction error and improves the accuracy of the prediction.

Attention-RNN: This model was proposed by Wang F. [36] and others. It applies the attention mechanism to the Recurrent Neural Network (RNN) so that RNN can focus on different parts of input and output.

Since the data sets are two realistic multivariate time series data sets and contain economic and financial series and time-temperature series, the stationarity of the data needs to be tested in subsequent experiments. Only a stationary series can establish a VAR model. We will perform an ADF test on the two data sets respectively to determine whether there is a unit root in the data set, and whether the data set is a stationary series. If the data set is a non-stationary sequence, the difference operation is performed until the differenced sequence reaches a plateau and a VAR model is established.

For the LSTM and the ATT-LSTM model, many parameters (time step, number of hidden units, batch size, etc.) need to be set and debugged during model training to achieve the best prediction effect. The final parameter settings of this experiment are shown in Table 2.

Table 2. Parameter settings.

Dataset	Time Step	Units	Batch Size	Learning Rate
NASDAQ 100	10	128	256	0.001
PM2.5 of Beijing	10	64	128	0.001

4.3. Experimental Results and Analysis

In this section, we conduct a practical evaluation of our proposed model. First, compare the ATT-LSTM model with the other six models (VAR, LSTM, GRU LSTM+Zoneout, LSTM-RNN, Attention-RNN) on two data sets. Then use the ATT-LSTM model to predict and visualize the prediction results to further prove the model's effectiveness.

4.3.1. Model Comparison

We compare the ATT-LSTM model with the other six models (VAR, LSTM, GRU, LSTM+Zoneout, LSTM-RNN, Attention-RNN) to prove the effectiveness of the ATT-LSTM model. Based on two evaluation indicators, MAE and RMSE, we show the best results of the model. The results are shown in Table 3. The comparison result is shown in Figure 7.

Table 3. Comparison result table.

Model	Dataset			
	NASDAQ 100		PM2.5 of Beijing	
	MAE	RMSE	MAE	RMSE
VAR	0.7012	0.9134	0.6538	0.7986
LSTM	0.6974	0.8965	0.6328	0.7245
GRU	0.9567	0.9981	0.6728	0.7457
LSTM+Zoneout	0.3665	0.4203	0.5769	0.6867
LSTM-RNN	0.5439	0.6268	0.6364	0.7842
Attention-RNN	0.4056	0.4938	0.4241	0.5178
ATT-LSTM	0.1948	0.2633	0.2134	0.2956

According to the comparison results in Table 3 and Figure 7, we can draw the following conclusions:

(1) Three baseline models (VAR, LSTM, GRU). On the Nasdaq 100 dataset, the LSTM performed best, followed by the VAR model. The gated recurrent unit (GRU), as a simpler variant of the LSTM, has not achieved better prediction results on this data set. The LSTM is still the best performer on the Beijing PM2.5 dataset, but the difference between the three models is insignificant.

(2) Three variant models (LSTM+Zoneout, LSTM-RNN, Attention-RNN). The LSTM+Zoneout model performed best on the Nasdaq 100 dataset, followed by the Attention-RNN model. The combination of the attention mechanism and RNN has greatly improved the ability of the RNN to process multivariate data. However, due to the characteristics of the RNN, the processing of multivariate time series data has not reached the desired effect.

(3) The performance of the three baseline models (VAR, LSTM, GRU) on the Nasdaq 100 dataset is worse than that on the Beijing PM2.5 dataset because the three baseline models do not analyze the correlation between multivariate sequences. The correlation sequence of the NASDAQ 100 dataset is far greater than that of the Beijing PM2.5 dataset, which causes some interference to the three baseline models, making the effect of the three baseline models on the NASDAQ 100 data set low the impact of the Beijing PM2.5 dataset. On the contrary, the three variant models (LSTM+Zoneout, LSTM-RNN, Attention-RNN) analyze related sequences. The overwhelming number of related sequences in the Nasdaq 100 dataset provides the three models with more robust learning capabilities, making the three models perform better on the Nasdaq 100 dataset than the PM2.5 dataset.

(4) The ATT-LSTM model has achieved good results on both datasets. On the Nasdaq 100 dataset, compared to the best-performing LSTM baseline model, the ATT-LSTM model has about 70% improvement, compared to the best-performing LSTM+Zoneout change model, ATT-LSTM model has about 40% improvement. On the Beijing PM2.5 dataset, the ATT-LSTM model has also achieved performance improvements compared to other models, but the improvement is lower than in the Nasdaq 100 dataset.

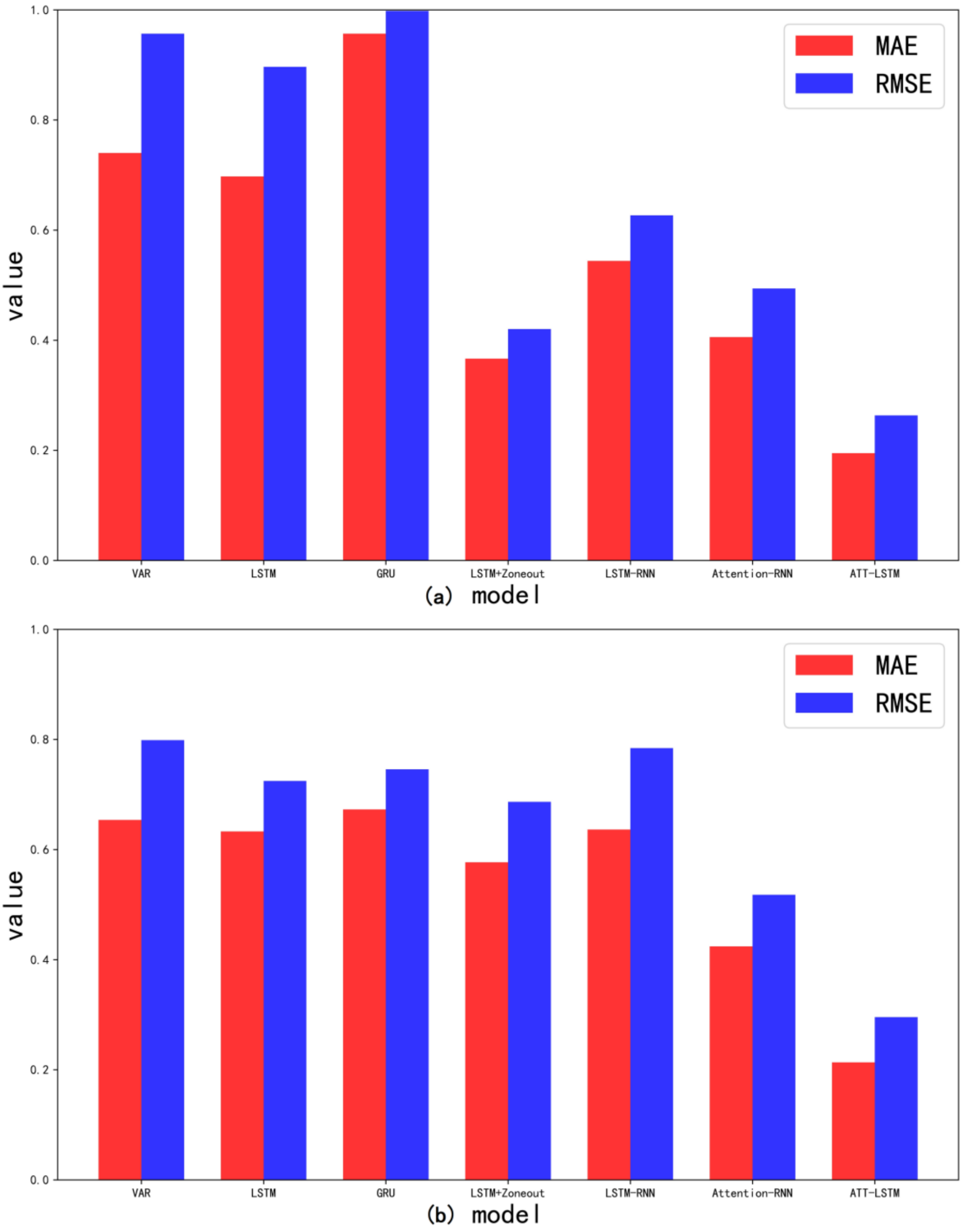

Figure 7. (**a**) NASDAQ100 dataset model comparison chart. (**b**) PM2.5 dataset model comparison chart.

4.3.2. Model Prediction

The experiment was divided into three parts. Firstly, we trained the ATT-LSTM model on the training dataset. Then, we further trained and adjusted the parameters on the validation set, made a preliminary evaluation of the model, and finally tested on the test set. Evaluate the model's predictive ability according to the two evaluation indicators of MAE and RMSE. After training and testing, we got the final ATT-LSTM model. We took a part of the continuous time nodes in the test set of the two data sets, used the trained ATT-LSTM

model to predict this part of the data, and output the prediction result. Then, we compared the actual data with the predicted results. The two data sets' prediction comparison results are shown in Figure 8a,b.

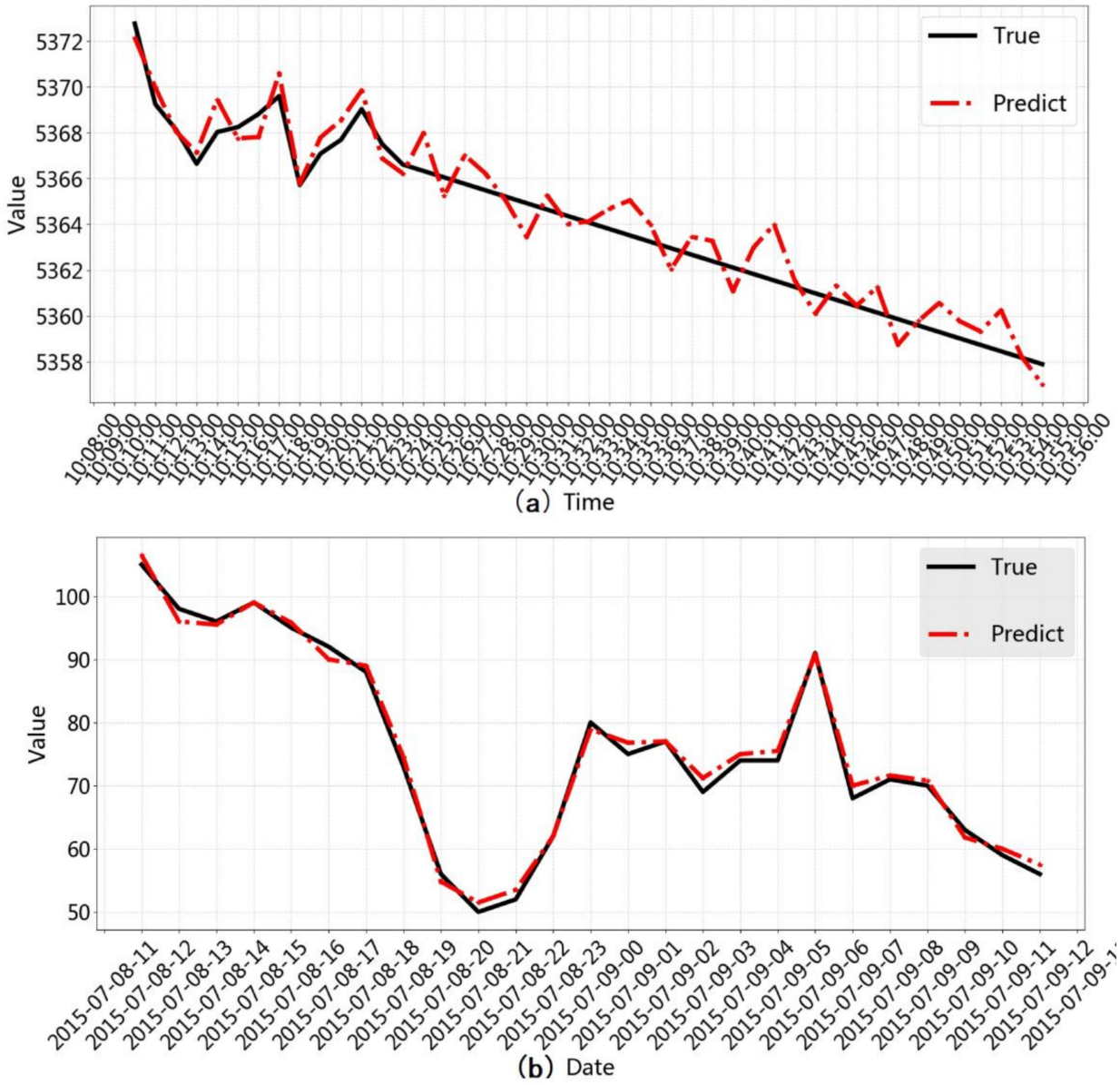

Figure 8. (**a**) Nasdaq100 dataset prediction results. (**b**) Beijing PM2.5 dataset prediction results.

Figure 8 clearly and directly shows the prediction effect of the ATT-LSTM model. It can be seen from Figure 8 that the overall model still has an error between plus and minus 0.5–2. Figure 8a is the prediction result of some continuous data on the Nasdaq 100 dataset and the comparison of actual data. This part of the data has two small fluctuations in the early period, and the latter is stable data, and the overall data value has not changed much. The prediction result of the model on the data conforms to the trend of data fluctuation. In the later stage of stable data, the prediction result of the model has an error of about plus or minus one. Figure 8b shows the prediction results of some continuous data on the Beijing PM2.5 dataset and comparing actual data. The data of this part fluctuate significantly as a whole, with both significant mutations and multiple small mutations. The prediction results of the model on this part of the data are in line with the trend of data fluctuations as a whole, indicating that the model can successfully capture the mutation information in the data set and learn.

5. Conclusions

We propose an ATT-LSTM model based on attention mechanism and LSTM. This model applies the attention mechanism to the LSTM, enabling the LSTM to screen multiple sequences, remove irrelevant redundant information, and to capture the interaction information between sequences. It greatly enhances the ability of LSTM to analyze and predict multiple time data. We conducted experiments on two real multivariate time series datasets of Nasdaq 100 and Beijing PM2.5. At the same time, we compare the ATT-LSTM model with the other six models based on two evaluation indicators: MAE and RMSE. The experimental results show that the ATT-LSTM model has improved performance compared with the other six models. Finally, we select part of the continuous data of the two data sets, use the trained model to predict the output, and directly compare it with the actual data. We further demonstrate the effectiveness of the ATT-LSTM model.

Author Contributions: J.J. writing—original draft, writing—review and editing. F.-A.L. conceptualization, methodology, software, data curation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 61772321. This research was funded by Shandong Natural Science Foundation, grant number ZR202011020044.

Institutional Review Board Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Trnka, A. Big data analysis. *Eur. J. Sci. Theol.* **2014**, *10*, 143–148.
2. Labrinidis, A.; Jagadish, H.V. Challenges and opportunities with big data. *Proc. VLDB Endow.* **2012**, *5*, 2032–2033. [CrossRef]
3. Karevan, Z.; Suykens, J.A.K. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Netw.* **2020**, *125*, 1–9. [CrossRef]
4. Zhao, Z.; Chen, W.; Wu, X.; Chen, P.C.Y.; Liu, J. LSTM network: A deep learning approach for short-term traffic forecast. *IET Intell. Transp. Syst.* **2017**, *11*, 68–75. [CrossRef]
5. Pang, X.; Zhou, Y.; Wang, P.; Lin, W.; Chang, V. An innovative neural network approach for stock market prediction. *J. Supercomput.* **2020**, *76*, 2098–2118. [CrossRef]
6. Kırbaş, I.; Sözen, A.; Tuncer, A.D.; Kazancıoğlu, F. Şinasi Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos Solitons Fractals* **2020**, *138*, 110015. [CrossRef]
7. Hochreiter, S.; Schmidhuber, J. Long short-time memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
8. Medsker, L.R.; Jain, L.C. Recurrent neural networks. *Des. Appl.* **2001**, *5*, 64–67.
9. Singhal, A.; Seborg, D.E. Clustering multivariate time-series data. *J. Chemometr. J. Chemometr. Soc.* **2005**, *19*, 427–438. [CrossRef]
10. Newman, J.; Baars, B.J. A neural attentional model for access to consciousness: A global workspace perspective. *Concepts Neurosci.* **1993**, *4*, 255–290.
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
12. Greff, K.; Srivastava, R.K.; Koutnik, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef]
13. Chen, K.; Zhou, Y.; Dai, F. A LSTM-based method for stock returns prediction: A case study of China stock market. In Proceedings of the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 2823–2824.
14. Altché, F.; de La Fortelle, A. An LSTM network for highway trajectory prediction. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 353–359.
15. Li, Y.F.; Cao, H. Prediction for tourism flow based on LSTM neural network. *Proc. Comput. Sci.* **2018**, *129*, 277–283. [CrossRef]
16. Bai, Y.; Zeng, B.; Li, C.; Zhang, J. An ensemble long short-term memory neural network for hourly PM2.5 concentration forecasting. *Chemosphere* **2019**, *222*, 286–294. [CrossRef] [PubMed]
17. Qi, Y.; Li, Q.; Karimian, H.; Liu, D. A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Sci. Total Environ.* **2019**, *664*, 1–10. [CrossRef] [PubMed]
18. Xie, G.; Shangguan, A.; Fei, R.; Ji, W.; Ma, W.; Hei, X. Motion trajectory prediction based on a CNN-LSTM sequential model. *Sci. China Inf. Sci.* **2020**, *63*, 1–21. [CrossRef]
19. Zhou, J.; Wang, Y.; Xiao, F.; Wang, Y.; Sun, L. Water quality prediction method based on IGRA and LSTM. *Water* **2018**, *10*, 1148. [CrossRef]
20. Qin, Y.; Song, D.; Chen, H. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv* **2017**, arXiv:1704.02971.

21. Zheng, C.; Fan, X.; Wang, C.; Qi, J. GMAN: A graph multi-attention network for traffic prediction. In Proceedings of the Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 1234–1241.
22. Li, Y.; Zhu, Z.; Kong, D.; Han, H.; Zhao, Y. EA-LSTM: Evolutionary attention-based LSTM for time series prediction. *Knowl.-Based Syst.* **2019**, *181*, 104785. [CrossRef]
23. Liu, Y.; Gong, C.; Yang, L.; Chen, Y. DSTP-RNN: A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. *Expert Syst. Appl.* **2020**, *143*, 113082. [CrossRef]
24. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
25. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Trans. Smart Grid* **2019**, *10*, 841–851. [CrossRef]
26. Karim, F.; Majumdar, S.; Darabi, H.; Chen, S. LSTM fully convolutional networks for time series classification. *IEEE Access* **2018**, *6*, 1662–1669. [CrossRef]
27. Wu, Y.; Yuan, M.; Dong, S.; Lin, L.; Liu, Y. Remaining useful life estimation of engineered systems using vanilla LSTM neural networks. *Neurocomputing* **2018**, *275*, 167–179. [CrossRef]
28. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10705–10714.
29. Chorowski, J.; Bahdanau, D.; Serdyuk, D. Attention-based models for speech recognition. *arXiv* **2015**, arXiv:1506.07503.
30. Qiu, J.; Wang, B.; Zhou, C. Forecasting stock prices with long-short term memory neural network based on attention mechanism. *PLoS ONE* **2020**, *15*, e0227222. [CrossRef] [PubMed]
31. Li, Z.; Li, Y. A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 143.
32. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef] [PubMed]
33. Dey, R.; Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600.
34. Krueger, D.; Maharaj, T.; Kramár, J. Zoneout: Regularizing rnns by randomly preserving hidden activations. *arXiv* **2016**, arXiv:1606.01305.
35. Abdel-Nasser, M.; Mahmoud, K. Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Comput. Appl.* **2019**, *31*, 2727–2740. [CrossRef]
36. Wang, F.; Tax, D.M.J. Survey on the attention based RNN model and its applications in computer vision. *arXiv* **2016**, arXiv:1601.06823.