Attention based Long Short-Term Memory Network for Coastal Visibility Forecast

Rui Min[1], Ming Wu[1*], Mengqiu Xu[1], Xun Zhu[1]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
mr1098546972@163.com, wuming@bupt.edu.cn, xumengqiu@bupt.edu.cn, zhuxun@bupt.edu.cn

Abstract: Visibility prediction in coastal areas has always been an important issue affecting the safety of residents and the efficiency of urban transportation. The visibility prediction methods currently used by meteorological centers are mainly based on the statistical forecast with relatively low prediction accuracy and high computational complexity. These methods cannot work well with large amounts of data. However, with the rapid development of deep learning technology, the use of deep learning has become a primary trend.

In this paper, we propose our visibility prediction model based on (Long Short-Term Memory) LSTM network and self-attention mechanism. The model takes Medium-range Forecasts Data from European Centre for Medium-range Weather Forecasting (ECMWF) which we use EC data to refer it for simplicity and observatory visibility data as input to predict and uses the LSTM network as the backbone to extract time series information. We also use self-attention mechanism to process the input data before the data is input to the model to let the model better focus on the valuable information for prediction. Compared with the predicted visibility in EC data, our proposed method improved the 3-hour prediction accuracy by 20%, 1.5 times, and 8 times for high-range, medium-range, and low-range visibility, respectively. We also find the data imbalance will greatly affect the prediction accuracy for low-visibility data and use the weighted-loss and mix-up data augmentation strategy model in our model training. We improved the accuracy of low-visibility data by 1.2 times while the prediction results of high-visibility and medium-visibility data remained almost the same. In addition, we conduct several experiments to verify the effectiveness of our model design and the rationality of data augmentation.

Keywords: Coastal visibility prediction; Deep learning; Long short-term memory; Self-attention; Data imbalance

1 Introduction

In the field of meteorology, coastal visibility is a very important meteorological variable that has a significant impact on people's daily life. For example, low visibility affects road traffic, daily travel, and aircraft takeoff. People need to forecast visibility for the coming period to better make advance decisions, therefore, an accurate and robust coastal visibility prediction method is urgently needed. In fact, observatory visibility prediction is always considered a time series prediction problem. A time series is a set of observations obtained by sampling a process in the time dimension [1], and the study of time series data has been going on for a long time. According to [2], the essence of time series prediction is using the patterns of variation exhibited by the past time data to predict the data over a future period. Time series prediction plays a significant role in both academic and industrial fields. Examples are the prediction of stock price changes [3], the numerical prediction of wind speed [4][5], and the prediction of weather [6][7].

Research on visibility forecasting has been going on for a long period and many traditional time series forecasting approaches such as AR models, ARIMA models, and time series decomposition models [8][9] were developed to solve this problem [10]. These models are powerful, with many experiments and previous works showing their capacity. [11] used ARIMA models for visibility prediction, [12] used the numerical model and output numerical weather prediction (NWP), [13] introduced a regression algorithm for visibility prediction. However, most of these methods use some statistical-based time series analysis for visibility prediction that are difficult to deal with complex data. It is worth mentioning that nowadays the visibility prediction task has become more complicated since the current meteorological data always consists of numerous meteorological factors and the size of the data volume is even in the order of petabytes [14]. Therefore, a single statistical model is not sufficient to perform an accurate and fast visibility prediction.

Recently, deep learning techniques are getting popular in many fields. [15][16][17] achieved good performance on many tasks of computer vision (CV), and [18][19][20] accomplished accurate language translation in the natural language processing (NLP). Similarly, deep learning techniques are also used in the time series analysis and solve many temporal tasks. [21] used LSTM networks for visibility prediction and employed re-sampling to improve the model's prediction accuracy for low-visibility data. [22] proposed a powerful multi-step time series prediction model by combining LSTM networks with attention mechanism [19]. [23] proposed ARRNN by exploiting the correlation between meteorological factors and surpassed the performance of ordinary LSTM. These innovative methods improved the effectiveness of the vanilla RNN model and inspired us to make use of other techniques to customize our model for specific tasks.

In this paper, we propose a coastal observatory visibility prediction framework based on LSTM and self-attention

978-1-6654-7735-2/22/$31.00 ©2022 IEEE

mechanism. Our framework is shown in Fig.1. Compared with the prediction accuracy of meteorological centers, our framework can predict the observatory visibility with much higher accuracy and is more flexible to handle heterogeneous data. Our main contributions are as follows.

1. We design a complete visibility prediction framework which introduces observed visibility data to assist in improving prediction accuracy.

2. We introduce the self-attention mechanism to give different weights for the meteorological data which reflect different importance of the input data for prediction results. The self-attention mechanism allows the model to pay attention to the essential part of the input data which improves the overall prediction effect. We further use weighted loss and mix up data augmentation to solve the data imbalance problem.

3. Extensive experiments show that our model improvs the accuracy of the visibility prediction largely compared to the baseline prediction result, which achieves TS scores 0.281, 0.627, and 0.789 on low-visibility data, medium-range data, and high-range visibility respectively.

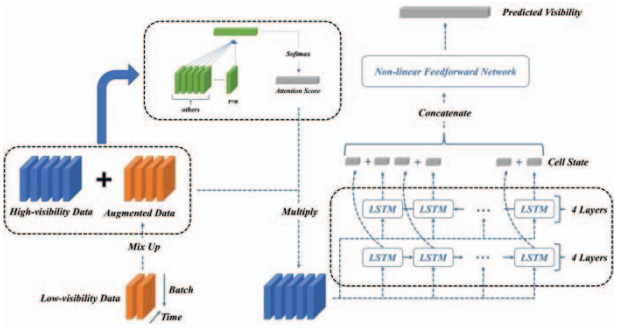


Figure 1 Procedure of visibility prediction

2 Data

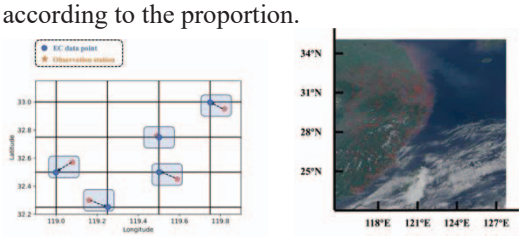
The data set is obtained from two sources, one is EC data, and another is from coastal observation stations. The EC data we used is from the European Centre for Medium-Range Weather Forecasts (ECMWF), which is a common data set used for weather forecasting with quality control (QC) and quality assurance (QA). EC data contains the meteorological prediction data in the future at 0:00 and 12:00 GMT each day. The prediction interval in the early stage is 3 hours and gradually increases to 6 hours with the increase of time. Another data source we used is the real-time observations obtained from the coastal observatory observations. Different from the EC data, this visibility data is obtained from real observations. This data records hour-by-hour observations at each station, providing us with the trends of coastal visibility over time in the local area.

Unlike previous prediction methods that only use EC data, we combine EC model forecast data and visibility data obtained from real observation stations according to their corresponding temporal and spatial locations. This is because we find that the visibility sequence itself has strong correlational information which will improve the visibility prediction. Specifically, in the time dimension, we take the forecast data of the first 12 hours of EC data at each forecasting time a day and form a time series with a 3-hour interval. We take the visibility data of coastal stations at the corresponding time and combine it with the time series of meteorological variables. In the spatial dimension, since the EC data are grid data and visibility data are related to the specific location of the coastal observatory sites, which are often not at specific grid points, we use the nearest neighbor algorithm to associate the nearest EC data point with the observatory sites, as shown in Fig.2(a).

We select data from coastal observation stations for our visibility prediction task in the spatial range of 23°N to 33°N, 116°E to 126°E, and the time range of January 1, 2020, to August 31, 2021. Fig.2(b) shows the real site location. In the process of data processing, to handle the missing data, we select the data of other time points closest to this time point for filling. This has two advantages, compared to directly deleting the time point, we do not remove any possible important information, and we also make the time interval of the time data consistent. In contrast to some strategies that use mean replacement, the proximity padding can maintain as much of the original time series relationship as possible without destroying the original time series structure. In addition to data cleaning, we also perform data normalization to speed up the convergence of the model. Specifically, we turn the feature data into data with a mean of 0 and a variance of 1 that conforms to the standard normal distribution, with the following normalization formula:

$$x\_norm = (x - E(x))/Var(x) \quad (1)$$

where E(x) and Var(x) are the mean value and standard deviation of data respectively. And in terms of data set split, we split our training set, validation set, and test set according to the ratio of 8:1:1. To ensure that our validation set, and test set can well reflect the effect of our model, we divide the overall data set into multiple sub data sets. On each sub-dataset, we divided the three data sets according to the proportion.



(a) **(b)**

Figure 2 (a) Associate the nearest EC data point spatially with the observatory sites **(b)** The location of the coastal observatory sites ranging from 23°N to 33°N, 116°E to 126°E

3 Methods

3.1 Long Short-Term Memory Network

Long Short-Term Memory (LSTM) network is a variant of vanilla RNN network which solves the problems of gradient explosion and gradient diffusion by adopting the

design of cell state. Since the network was proposed in 1995 [24], LSTM has derived many variants [25][26] and becomes the backbone model of many time series tasks [27]. Fig.3 shows the modern LSTM structure with three important gate structures, namely forget gate, input gate and output gate. The formula of LSTM is as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] \quad (2)$$
$$i_t = \sigma(W_i[h_{t-1}, x_t]) \quad (3)$$
$$o_t = \sigma(W_o[h_{t-1}, x_t]) \quad (4)$$
$$\tilde{C}_t = tanh(W_C[h_{t-1}, x_t]) \quad (5)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (6)$$
$$h_t = o_t \tanh(C_t) \quad (7)$$

Where f is the forget gate, i is the input gate and o is the output gate. The forget gate is used to control the flow of information from the past node's cell state and select what information to leave behind. The input gate selects the information to update the cell state of the current node. The output gate is used to calculate the hidden state of the current node and control the information flowing into the next LSTM node.

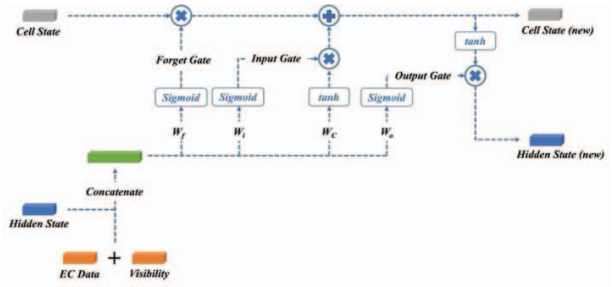


Figure 3 The architecture of LSTM network

3.2 Self-attention Mechanism

Self-attention mechanism was first proposed in [28]. It improves the traditional attention mechanism [19] which models time series dependencies by calculating the attention score among the data. Fig.4 shows the process of calculating the attention score of the data at a specific time point. The calculation formula is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (8)$$

Q, K and V are query, key, and value respectively, which are transformed by the input parameter matrix. This attention scores reflect the relative importance of different parts of the input data and forces the model to focus on important data parts and ignore some irrelevant information.

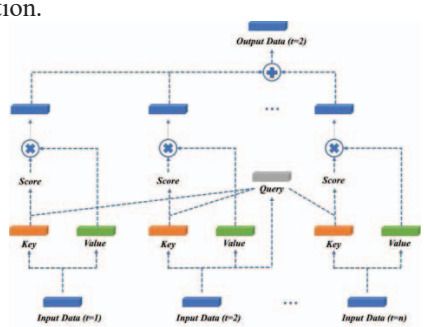


Figure 4 The details of self-attention mechanism

3.3 Attention-based LSTM Network

We propose a new time series prediction framework for the visibility prediction of coastal observation stations and utilize LSTM as the backbone for visibility prediction to extract temporal information in the visibility sequence. Compared with other temporal prediction model, LSTM is simple but effective which performs well in time series prediction tasks without consuming much computational resources. While other models [28] show more potential, we simply introduce LSTM to illustrate the effectiveness of deep-learning techniques and leave more complex structures to the future work. To make final prediction, we use a two-layer nonlinear feedforward network and utilize the cell state output from the last time node as the input to make a one-step visibility prediction. We also introduce the self-attention mechanism to preprocess the input features before the input data gets through LSTM. Different from the original self-attention calculation method, we only calculate the attention score of the data closest to the prediction point with data at other time points. The reason why we adopt this design is that we believe that the correlation between the visibility to be predicted and the input meteorological data in the past is different, and this kind of correlation decreases gradually as the time interval between past and current time points increases. According to this prior assumption, we use the self-attention mechanism to dynamically give weights to input data at different time points. We observe that larger weights are given to the input data having a high correlation with the predicted value, otherwise, the weights are small. This enables our LSTM network to focus on more useful information for the prediction process.

3.4 Evaluation Metrics

In the process of model training and testing, we use the loss function MSE loss to measure the difference between our predicted visibility and the real visibility. The formula of MSE is:

$$L = \frac{1}{2n}\sum_{i=1}^{n}(y - f(x_i))^2 \quad (9)$$

Where y is the real visibility from the observation station, x is the input data and f is our model. Through this loss function, we can see the difference between our predicted results and the ground truth. In addition to MSE loss, a more commonly used visibility evaluation standard in the meteorological field is the TS score, which is a commonly used evaluation standard to measure the classification quality of classification models. The calculation formula for TS score is as follows:

$$\text{TS score} = \frac{TP}{(TP+FP+FN)} \quad (10)$$

TP is the true positive samples, FP is the false positive samples, FN is the false negative samples, and the value of TS score is between 0-1. The larger the TS score is, the more accurate the classification result is. Here we divide

the visibility into three categories: less than 1km, between 1-10km and more than 10km. We get a corresponding TS score for each class and measure the quality of our model according to this value.

3.4 Data Imbalance

We classify the data with visibility less than 1km, visibility between 1-10km, and visibility greater than 10km. According to the statistical results, we found that the data with visibility less than 1km only accounted for less than 2.5% of the whole training data, and the other two types of visibility data accounted for 36.26% and 61.33% respectively. The model learns a misguided prior assumption from imbalanced data, which leads our model to predict high-visibility results in most of the training process and weakens the ability of the model to map the data with the real visibility label. As a result, our model has a good prediction effect for high-visibility weather, but a poor prediction effect for low-visibility weather. Many methods have proposed solutions to the data imbalance phenomenon, such as OHEM [30] and focal loss [31]. Inspired by the modification on loss function, we adopt weighted loss from the perspective of the loss function which could be formulated as:

$$L(x) = \sum_{i=1}^{n} \alpha_i * MSE(f(x_i), y) \quad (11)$$

where x_i represents one piece of input series and y is the ground truth data, and f is our temporal model. α is a hyperparameter predefined by the user. We set high α to low-visibility samples to make the model pay more attention to these samples and low α for the high-visibility (higher than 1km) ones. In addition, we adopt the commonly used multiple data augmentation strategy mix up [29] by taking convex combination of two samples from training set to form a new data sample. According to this principle, we fuse our low-visibility samples to obtain an expanded data set. The data augmentation method is as follows:

$$\tilde{x} = \lambda * x_i + ((1 - \lambda)) * x_j \quad (12)$$

where λ is a random value ranging from 0 to 1. Although this data augmentation method is linear, we find that this method can greatly improve the prediction accuracy of our model in low-visibility data.

4 Experiment Results

In our experiments, we use the visibility prediction results from the meteorological center as our baseline result and compare our prediction results with them. Our experimental results are shown in table 1. We choose Bi-LSTM as our backbone with 4 layers and the size of the cell state is set to 50. For the input data, we select EC data as the model input which contains 66 meteorological features at each time point and set the batch size to 64. We use 48-hour historical data with a 3-hour interval between adjacent times, and the total time length is 16. In the optimization part, we use random gradient descent (SGD) as our optimizer, and the learning rate is set to 0.001.

We also adopt the early stop method in the training process. The early stop method is a simple and effective strategy to prevent the model from overfitting. When the error on the validation set does not decrease after predefined epochs, the model training will be stopped. In our experiments, the patient epoch size is set to 20. After around 70 epochs, the model converges. The results demonstrate that the use of the LSTM network can indeed improve the prediction accuracy, especially in medium-range and high-range visibility data but have little effect on the prediction of low-visibility data.

Table I TS score within different visibility ranges

Results of Experiment				
	Range	TS score	FP rate	FN rate
Baseline	< 1km	0.031	0.215	0.754
	1-10km	0.246	0.391	0.363
	> 10km	0.656	0.175	0.169
LSTM	< 1km	0.037	0.895	0.068
	1-10km	0.548	0.276	0.176
	> 10km	0.741	0.085	0.174
LSTM + Visibility	< 1km	0.123	0.86	0.017
	1-10km	0.636	0.197	0.168
	> 10km	0.791	0.081	0.128
LSTM + Self-Attention	< 1km	0.142	0.841	0.016
	1-10km	0.629	0.208	0.162
	> 10km	0.788	0.077	0.134

We then introduce the visibility sequence combined with the original meteorological input to our model without changing other settings. The experimental results show that the prediction effect has been greatly improved in all visibility ranges, increasing the prediction accuracy by 20%, 1.5 times, and almost 3 times for high-range, medium-range, and low-range visibility compared to the baseline result. One reason for this is that the visibility sequence has an internal relevance and visibility from the current time point has a strong correlation with the time points before. Therefore, we can effectively improve the accuracy of the model by introducing visibility data as input. We then introduce the self-attention mechanism further to process the input data. We think the introduction of the self-attention mechanism makes the model pay more attention to the part of the input that is helpful for prediction. The experimental results show that self-attention is more effective in the prediction of low-visibility data which increases its TS score from 0.123 to 0.142. We think when the visibility is low, the weather conditions change a lot, and the introduction of self-attention can learn some key information, filtering out redundant information and thus improving the effectiveness of the model.

However, we find that the overall prediction effect of low-visibility data is still very poor, which is due to the imbalance of data caused by too few low-visibility samples. We introduce the mix-up mechanism and

weighted loss to solve this problem and reconduct the experiments. The experimental results show that by introducing mix-up and weighted loss, we can effectively solve the problem of uneven data distribution. We increase the low-visibility data to different multiples using this strategy to test the performance of our model and the result is shown in Fig.5. The overall accuracy improvement can be up to 8 times as we increase the dataset size, and we also notice that as the dataset size becomes larger, the TS score for low-visibility data rises first and then gradually becomes flat. We make a tradeoff between prediction accuracy and training time, and we choose to augment the low-visibility samples by 3 times. This enables our prediction framework to provide accurate visibility prediction data in various visibility ranges and improves the practical availability of our algorithm. The results are shown in Fig.6 which increase the prediction accuracy by 20%, 1.5 times, and 8 times for three types of visibility data compared with the baseline.

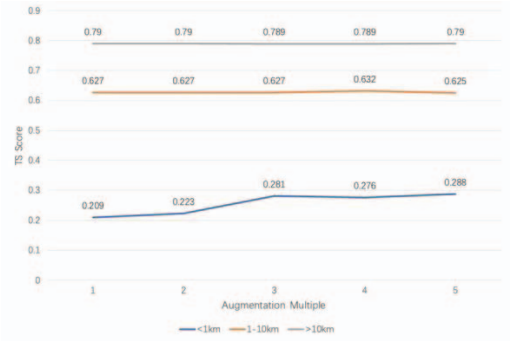


Figure 5 Relationship between augmentation multiple and model performance on low-visibility data

5 Conclusion

In this paper, we propose a time series prediction framework based on the LSTM network combined with self-attention for visibility prediction of coastal observation stations. Our experimental results show that the LSTM network is effective in visibility prediction and the introduction of self-attention mechanism can make the model selectively pay attention to the input data useful for prediction. Further, we find that there is a serious data imbalance phenomenon in visibility data. In the case of less low-visibility, we introduce weighted loss and mix up to expand the low-visibility data samples, and after using the augmented data set for training, we can increase the TS score of the low-visibility prediction by 1.2 times. The overall experimental results show that the proposed visibility prediction framework can improve the overall prediction accuracy and can also be used in scenarios with unbalanced actual data, which can be used as a tool for meteorologists. While our model is focused on the prediction of a single time point, we believe a multi-step visibility prediction model is more required in real application scenarios and we leave the study of a more general and multi-step prediction framework in our future work.

6 Discussion

Our paper aims to provide an accurate and robust visibility prediction framework to the public. For scientific field, our method promotes the research on visibility prediction and can be further adopted in other time-series related field. In industrial aspect, with precise visibility prediction, people can make critical decisions in advance to avoid potential risks and unnecessary spending. Our future work will still focus on using deep-learning techniques to solve problems in real scenarios to drive the development of other fields.

Acknowledgements

Gratefully acknowledge the funding provided by the National Key Research and Development Program of China (2019YFC1510102) and Shanghai Typhoon Research Foundation (TFJJ202109). Also thanks to the reviewers for their efforts.

References

[1] Liu, Z., Zhu, Z., Gao, J. and Xu, C., 2021. Forecast methods for time series data: a survey. IEEE Access, 9, pp.91896-91912.

[2] Cryer, J.D., 1986. Time series analysis (Vol. 286). Boston: Duxbury Press.

[3] Ariyo, A.A., Adewumi, A.O. and Ayo, C.K., 2014, March. Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (pp. 106-112). IEEE.

[4] Mohandes, M.A., Rehman, S. and Halawani, T.O., 1998. A neural networks approach for wind speed prediction. Renewable Energy, 13(3), pp.345-354.

[5] Mohandes, M.A., Halawani, T.O., Rehman, S. and Hussain, A.A., 2004. Support vector machines for wind speed prediction. Renewable energy, 29(6), pp.939-947.

[6] Hewage, P., Trovati, M., Pereira, E. and Behera, A., 2021. Deep learning-based effective fine-grained weather forecasting model. Pattern Analysis and Applications, 24(1), pp.343-366.

[7] Sharma, U. and Sharma, C., 2022, January. Deep Learning Based Prediction Of Weather Using Hybrid_stacked Bi-Long Short Term Memory. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 422-427). IEEE.

[8] Appiah, T., 2015. Regression and Time Series Analysis of Loan Default At Minescho Cooperative Credit Union. Tarkwa, 4(08), pp.188-195.

[9] Dagum, E.B. and Bianconcini, S., 2016. Seasonal adjustment methods and real time trend-cycle estimation. Berlin/Heidelberg, Germany: Springer International Publishing.

[10] Mahmoud, A. and Mohammed, A., 2021. A survey on deep learning for time-series forecasting. In Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges (pp. 365-392). Springer, Cham.

[11] Salman, A.G. and Kanigoro, B., 2021. Visibility forecasting using autoregressive integrated moving average (ARIMA) models. Procedia Computer Science, 179, pp.252-259.

[12] Singh, A., George, J.P. and Iyengar, G.R., 2018. Prediction of fog/visibility over India using NWP Model. Journal of Earth System Science, 127(2), pp.1-13.

[13] Bari, D., 2018, October. Visibility prediction based on kilometric nwp model outputs using machine-learning regression. In 2018 IEEE 14th international conference on e-Science (e-Science) (pp. 278-278). IEEE.

[14] Agapiou, A., 2017. Remote sensing heritage in a petabyte-scale: satellite data and heritage Earth Engine© applications. International Journal of Digital Earth, 10(1), pp.85-102.

[15] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[16] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[17] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[18] Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

[19] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[20] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[21] Deng, T., Cheng, A., Han, W. and Lin, H.X., 2019, February. Visibility Forecast for Airport Operations by LSTM Neural Network. In ICAART (2) (pp. 466-473).

[22] Meng, Y., Qi, F., Zuo, H., Chen, B., Yuan, X. and Xiao, Y., 2020, July. Multi-step LSTM prediction model for visibility prediction. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[23] Jonnalagadda, J. and Hashemi, M., 2020, August. Forecasting atmospheric visibility using auto regressive recurrent neural network. In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 209-215). IEEE.

[24] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[25] Huang, Z., Xu, W. and Yu, K., 2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

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

[27] Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. and Schmidhuber, J., 2016. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), pp.2222-2232.

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[29] Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

[30] Shrivastava, A., Gupta, A. and Girshick, R., 2016. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 761-769).

[31] Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).