

Research on Short- and Medium-Term Airport Visibility Prediction Based on the TimeGAN-GA-LSTM Model

Pengcheng Shen*

College of Sciences, China Jiliang University, Hangzhou 310018, China

* Corresponding author: Pengcheng Shen (Email: pcshen@cjlju.edu.cn)

Abstract: Accurate airport visibility prediction is crucial for flight safety, as reduced visibility may lead to flight delays, cancellations, or even accidents, which impact the efficiency of air transportation. To improve prediction accuracy, this paper proposes a medium-to-short-term visibility prediction model based on TimeGAN-GA-LSTM. Using nearly 10 years of daily meteorological and pollutant data from Xiaoshan International Airport, feature selection is conducted through Spearman correlation analysis. The training dataset is enhanced using TimeGAN, and the LSTM hyperparameters are optimized using a Genetic Algorithm (GA). The results show that under the condition of a 1-day sliding window, the TimeGAN-GA-LSTM model achieves the best prediction accuracy for airport visibility. For a 580-day prediction dataset, the RMSE is 2.01, MAE is 1.49, and MAPE is 16%, which are reductions of 1.01, 0.72, and 15.2%, respectively, compared to the LSTM model, and the R^2 increases to 0.83. This study provides an effective method for intelligent airport visibility prediction, contributing to flight schedule optimization and safety enhancement

Keywords: Visibility Prediction, TimeGAN, GA-LSTM, Feature Selection.

1. Introduction

Airports, as essential transportation hubs, play a pivotal role in modern air transport systems. Their operational efficiency and flight punctuality directly impact aviation safety and passenger satisfaction. In recent years, the development of intelligent airports has become a key objective in the aviation industry. Intelligent airports aim to enhance air transport efficiency and safety, particularly by improving flight operations under complex and changing weather conditions. Among various meteorological factors, visibility stands out as a critical indicator that significantly affects flight scheduling, takeoff, landing, and overall airport management. Low visibility conditions can lead to flight delays, cancellations, and even accidents, emphasizing the need for accurate and reliable visibility prediction.

With the rapid advancement of deep learning techniques, visibility prediction has emerged as a prominent research focus. Traditional statistical methods often struggle with complex time series data, prompting researchers to explore deep learning-based models for improved prediction accuracy. Among these models, Long Short-Term Memory (LSTM) networks have gained considerable attention due to their effectiveness in capturing temporal dependencies. Wu et al. [1] proposed an LSTM model to predict airport visibility over the next three hours based on historical meteorological data, demonstrating satisfactory predictive performance.

However, LSTM models often encounter challenges such as data sparsity and overfitting, especially when dealing with complex and variable time series data from airports. To address these limitations, Genetic Algorithm-optimized LSTM (GA-LSTM) models have been proposed, integrating GA to fine-tune model hyperparameters. For instance, Kara et al. [2] applied a GA-LSTM model for soil moisture prediction, achieving significant improvements in accuracy, while Chen et al. [3] utilized the approach for smart grid energy consumption forecasting, showing enhanced predictive performance. Despite these advancements, the application of

GA-LSTM models to airport visibility prediction remains limited, primarily due to difficulties in capturing intricate time series features[4].

To overcome these challenges and enhance the accuracy and generalization of visibility prediction, this paper proposes a novel prediction model based on TimeGAN and GA-LSTM. TimeGAN is employed to augment the time series data, effectively addressing data sparsity and enriching the training dataset. Simultaneously, GA is utilized to optimize the hyperparameters of the LSTM model, resulting in an efficient and robust model tailored for medium-to-short-term airport visibility prediction. This integrated approach aims to provide more reliable visibility forecasts, thereby supporting flight safety and operational decision-making at intelligent airports.

2. Feature Selection Based on Spearman Correlation Coefficient

This study focuses on Hangzhou Xiaoshan International Airport, utilizing daily visibility data and meteorological influencing factors sourced from the U.S. National Oceanic and Atmospheric Administration (NOAA) database (<http://www.cdc.noaa.gov>), alongside daily air pollution data obtained from the Air Quality Online Monitoring Platform (<https://www.aqistudy.cn/>). These datasets provide a comprehensive basis for analyzing the factors affecting airport visibility and enable the development of a robust prediction model.

In the process of feature selection, choosing an appropriate correlation analysis method is crucial. Since the Kolmogorov-Smirnov test indicates that some data do not follow a normal distribution, the Spearman correlation coefficient method is used. This method is not limited by data distribution and can effectively measure the nonlinear relationships between variables, making it suitable for the dataset in this study.

The Spearman correlation coefficient is a non-parametric statistic primarily used to measure the monotonic relationship between two variables [5]. Its calculation involves converting

the original data into rank data, sorting the variables based on their values, and using the ranked values as substitutes for the original values. The correlation is determined by calculating the ranking differences between two variables, with the formula:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (1)$$

where n represents the number of paired observations, and d_i is the difference between the ranks of the i -th pair of data points in the two variables. The coefficient ρ_s ranges from -1 to 1.

By calculating the Spearman correlation coefficient, a correlation matrix for all indicators is obtained, and the correlation heatmap is plotted, as shown in Figure 1.

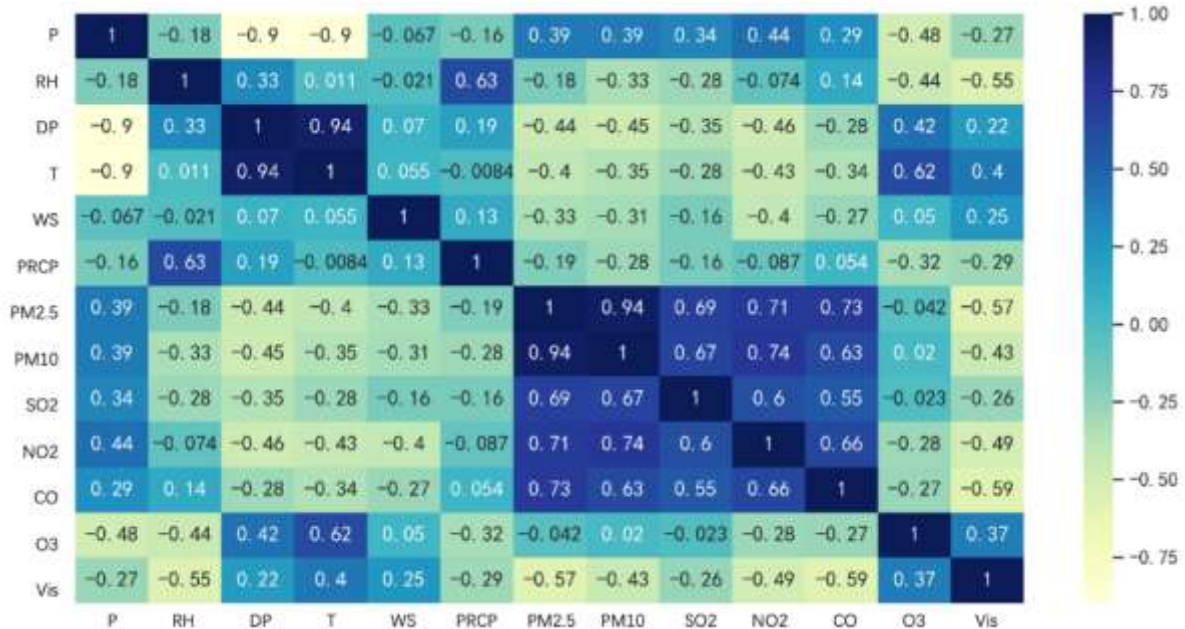


Figure 1. Heatmap of Spearman Correlation Coefficients Between Various Factors

3. Construction of the TimeGAN-GA-LSTM Model

3.1. TimeGAN-GA-LSTM Model

Based on the principles mentioned earlier, this study constructs an airport visibility time series prediction model based on TimeGAN-GA-LSTM. The specific steps are as follows:

3.1.1. Model Input and Output

Determining the input and output of the model is a crucial step to ensure that the prediction model can accurately solve the problem and generate practical value. The input features selected for the model are those with significant impacts on visibility: temperature, PM2.5, PM10, CO, and NO2 (with units of %, °C, µg/m3, µg/m3, mg/m3, and µg/m3, respectively). The output of the model is the visibility of Xiaoshan International Airport for the next 1 and 2 days, measured in kilometers.

3.1.2. Training and Testing Set Division

The prediction model in this study is trained based on multi-feature, single-output time series data. The model is trained using the training set and evaluated using the testing

set. The division of sample points should align with the time series characteristics. The training and testing sets are divided in a 7:3 ratio, as shown in Table 1.

Table 1. Dataset Sample Division

Sample	Time Range	Sample Size
Training Set	2013/12/2 to 2022/9/5	3200
Testing Set	2022/9/6 to 2024/4/7	580

3.1.3. TimeGAN Data Augmentation

In this study, TimeGAN is used for data augmentation to avoid overfitting and improve prediction accuracy. TimeGAN can generate synthetic data with similar temporal characteristics to the original data, thus expanding the dataset and enhancing the model's robustness [6-9]. Based on the 3200 data points in the training set, a 5-fold data augmentation is performed. To verify the authenticity and effectiveness of the generated data, principal component analysis (PCA) is applied for dimensionality reduction, followed by quality evaluation. Figure 2 presents the visual comparison of the dimensionality-reduced original and generated data. The results show that the generated data closely overlap with the original data, indicating that the generated data has successfully learned the temporal features of the original data.

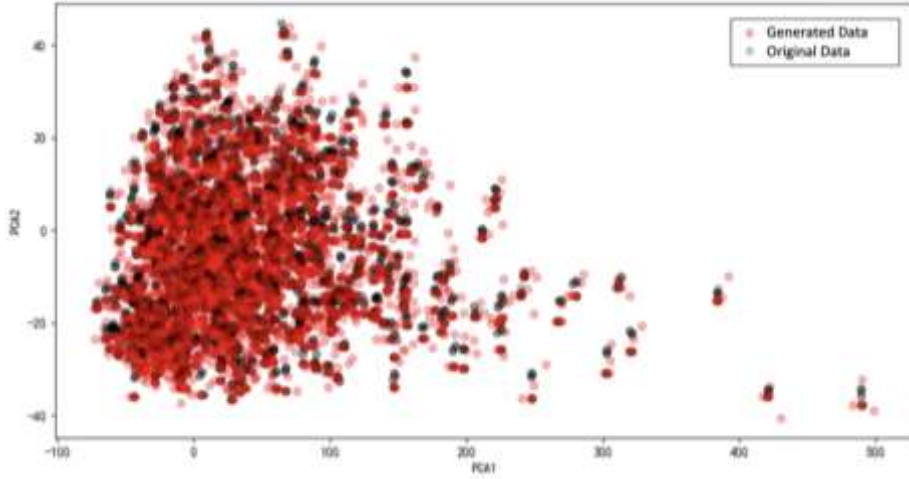


Figure 2. Visual Comparison of Dimensionality-Reduced Original and Generated Data

As seen in Figure 2, the generated data closely overlap with the original data, indicating that the synthetic data has learned the temporal characteristics of the original data. This suggests that the data augmented by TimeGAN from the visibility-influencing factors is highly valuable.

3.1.4. Data Normalization

The dimensional differences among various features in the data may impact the training performance of the model, so normalization is required. In this study, the Min-Max normalization method is used, and the formula is as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

where X represents the original data, X_{max} and X_{min} are the maximum and minimum values of the original data, respectively.

3.1.5. Hyperparameter Optimization and Initialization of GA Parameters

The prediction accuracy of the LSTM model depends on the settings of the hyperparameters in the neural network. The selection of hyperparameters directly affects the learning efficiency, training speed, and prediction performance of the entire model. In this study, the Gelu function is chosen as the activation function, and the MSE is selected as the loss function. GA is used to find the optimal values for the number of neurons in the hidden layers SSS, the number of iterations TTT, and the learning rate RRR. The constraints for the hyperparameters are based on the literature [10].

$$\begin{cases} 10 \leq S \leq 150 \\ 10 \leq T \leq 350 \\ 0.001 \leq R \leq 0.01 \end{cases} \quad (3)$$

GA is a global optimization probabilistic search algorithm with powerful global optimization capabilities. In this study, based on literature references, the number of hidden layers is set to 2. Therefore, the number of neurons in the two hidden layers S1 and S2, the number of iterations T, and the learning rate R are treated as optimization parameters. The fitness function $fit(x)$ is selected as the coefficient of determination R2.

When using GA for hyperparameter optimization, the initialization parameters are listed in Table 2:

Table 2. Initialization Parameters of GA

Parameter	Population Size	Termination Generation	Crossover Probability	Mutation Probability
Value	20	50	0.5	0.1

3.1.6. Model Construction and Solution

After obtaining the GA-optimized hyperparameters, these parameters are assigned to the LSTM model, and the model is evaluated using the test dataset. Finally, the predicted values are transformed into actual visibility predictions through inverse normalization. Based on the aforementioned modeling process, Figure 3 illustrates the training process of the TimeGAN-GA-LSTM model, presenting the complete workflow from data input to prediction output.

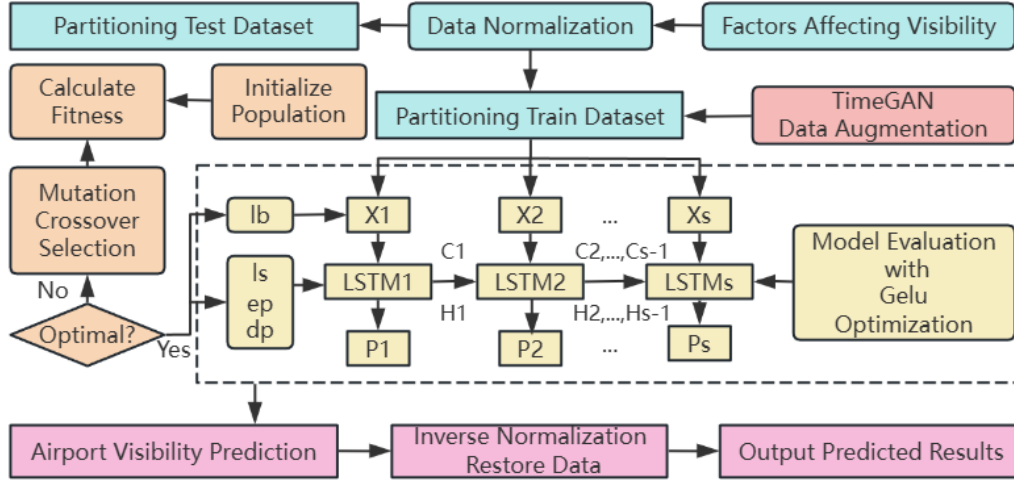


Figure 3. Training Process of the TimeGAN-GA-LSTM Model

3.2. Analysis of experimental results

To further evaluate the predictive performance of the model, four evaluation metrics are employed: coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) [7].

R^2 measures the model's goodness of fit to the observed data, calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (|\hat{y}_i - y_i|)^2}{\sum_{i=1}^n (|\bar{y}_i - y_i|)^2} \quad (4)$$

MAE reflects the average distance between the predicted values and the actual values, expressed as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (5)$$

MAPE calculates the percentage of prediction error relative

to the actual value, formulated as:

$$MAPE = \frac{100\%}{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (6)$$

RMSE measures the average distance between predicted and actual values, defined as:

$$RMSE = \sqrt{\frac{1}{n} (\hat{y}_i - y_i)^2} \quad (7)$$

These evaluation metrics comprehensively analyze the predictive performance of the model and provide a basis for further optimization.

4. Results Analysis

4.1. Analysis of Key Feature Indicators

This study analyzes three key features that significantly impact visibility: humidity, PM2.5, and CO, to verify the rationality of feature selection.

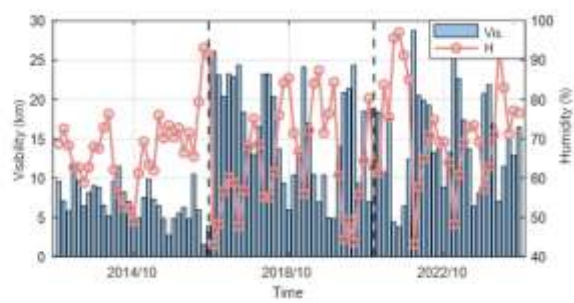
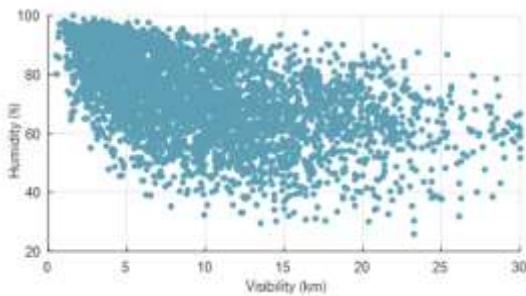


Figure 4. Scatter Plot and Trend of Visibility vs. Relative Humidity

Figure 4 shows a significant negative correlation between visibility and relative humidity. As humidity increases, visibility decreases, and vice versa. From 2014 to 2022, the

construction of an intelligent meteorological system in Hangzhou steadily improved visibility.

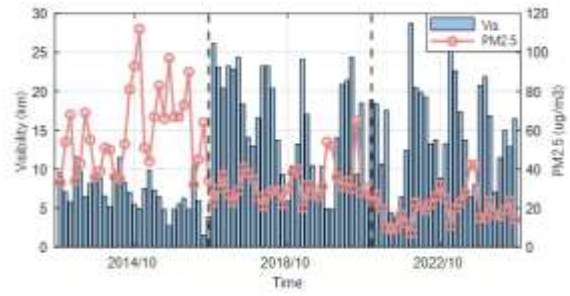
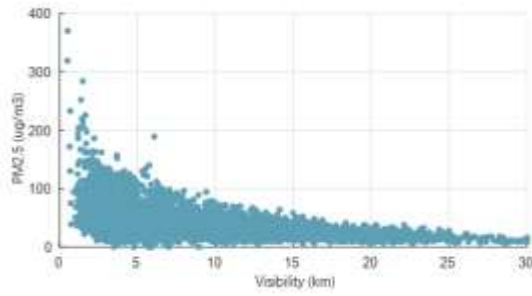


Figure 5. Scatter Plot and Trend of Visibility vs. PM2.5

Figure 5 shows a negative correlation between visibility and PM2.5 concentration. From 2014 to 2022, the annual decrease in PM2.5 concentration in Hangzhou resulted in

improved visibility, reflecting the effectiveness of atmospheric pollution control measures.

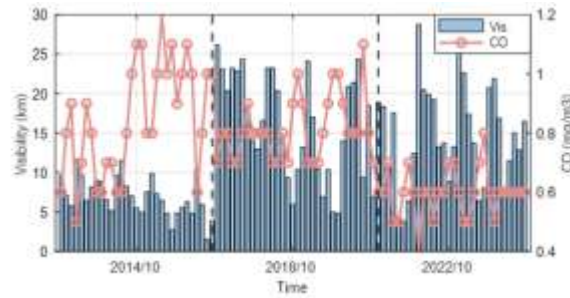
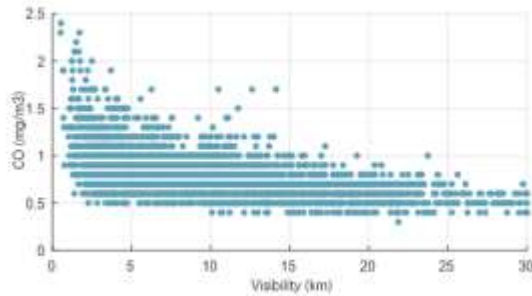


Figure 6. Scatter Plot and Trend of Visibility vs. CO

Figure 6 shows that visibility and CO concentration also exhibit a negative correlation. Hangzhou has effectively reduced CO emissions and improved atmospheric transparency through wetland protection and exhaust control.

fitness of the TimeGAN-GA-LSTM model with a prediction step size of 1 day is 0.868. The GA algorithm converges after 28 iterations. The optimal hyperparameters of the LSTM model are shown in Table 3. The iterative graph of the GA algorithm is shown in Figure 7.

4.2. Analysis of Hyperparameter Optimization Results

When the GA terminates at the 50th generation, the optimal

Table 3. Optimal Hyperparameters of the LSTM Model

Model	Input Step Size	Prediction Step Size	Activation Function	K	S1	S2	T	R
LSTM	15	1 or 2	Gelu	2	104	118	245	0.0062

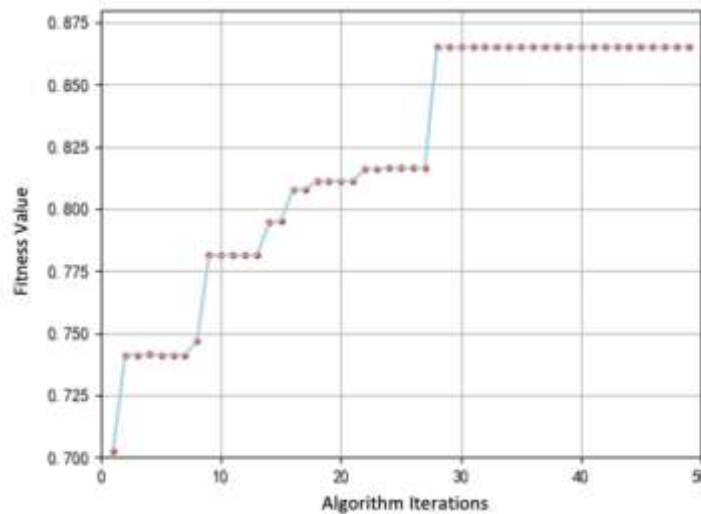


Figure 7. Iteration Graph of GA Algorithm

4.3. Analysis of Visibility Prediction Results

This study compares the predictive performance of four

models: LSTM, GA-LSTM, TimeGAN-LSTM, and TimeGAN-GA-LSTM, for prediction periods of 1 day and 2 days.

Table 4. Performance of Four Models with a 1-Day Prediction Period

Model	1 Day					
	Training Set			Testing Set		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE
LSTM	2.17	29.2%	2.96	2.21	31.3%	3.02
GA-LSTM	2.06	23.7%	2.44	1.97	24.1%	2.57
TimeGAN-LSTM	1.92	19.4%	2.49	1.83	18.3%	2.41
TimeGAN-GA-LSTM	1.34	17.3%	2.08	1.49	16.1%	2.01

Table 4 shows that with a 1-day prediction period, the TimeGAN-GA-LSTM model achieves the best performance on both the training and testing sets, with MAE of 1.34 and 1.49, MAPE of 17.3% and 16.1%, and RMSE of 2.08 and

2.01, respectively, significantly outperforming other models. In contrast, the LSTM model has higher errors, while GA-LSTM and TimeGAN-LSTM models show improvement but still lag behind TimeGAN-GA-LSTM.

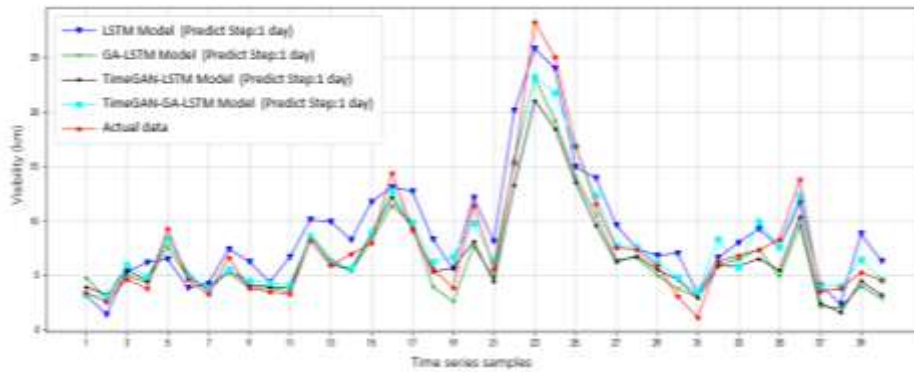
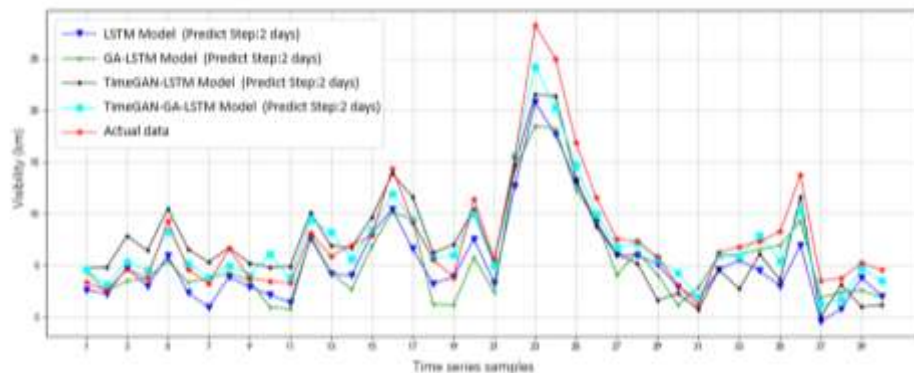
Table 5. Performance of Four Models with a 2-Day Prediction Period

Model	2 Days					
	Training Set			Testing Set		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE
LSTM	3.02	30.8%	3.54	3.11	32.7%	3.76
GA-LSTM	2.51	28.2%	2.77	2.59	30.7%	2.82
TimeGAN-LSTM	2.16	24.5%	2.81	2.21	26.7%	2.96
TimeGAN-GA-LSTM	1.72	22.4%	2.43	1.80	21.1%	2.51

Table 5 shows that when the prediction period is 2 days, the prediction error of all four models increases, but the TimeGAN-GA-LSTM model still performs the best, with MAE of 1.72 and 1.80, MAPE of 22.4% and 21.1%, and RMSE of 2.43 and 2.51. The LSTM model still has the highest errors, while GA-LSTM and TimeGAN-LSTM models show some improvement, but the performance still

falls short of TimeGAN-GA-LSTM.

Figures 8 and 9 respectively compare the predictive performance of the four models with 1-day and 2-day prediction steps. Overall, the TimeGAN-GA-LSTM model demonstrates higher predictive accuracy and stability under different prediction steps, proving its significant advantage in capturing visibility time-series features.

**Figure 8.** Comparison of Prediction Performance of Different Models (1-Day Prediction Step)**Figure 9.** Comparison of Prediction Performance of Different Models (2-Day Prediction Step)

5. Conclusion

This study proposes a novel TimeGAN-GA-LSTM hybrid

model to improve medium-to-short-term visibility prediction at airports. By integrating TimeGAN for time series data augmentation and genetic algorithms for optimizing LSTM

hyperparameters, the model achieves superior prediction accuracy and robustness for 1-day and 2-day forecasting horizons compared to conventional models such as standalone LSTM and GA-LSTM. The results demonstrate the model's strong potential in practical airport operation settings, contributing to enhanced flight safety, more accurate scheduling, and better resource allocation.

Furthermore, analysis of the visibility data from Hangzhou Xiaoshan International Airport highlights the positive impact of meteorological protection and air pollution control policies on atmospheric visibility improvement in recent years. This underscores the importance of combining advanced predictive modeling with environmental management efforts to achieve sustainable development goals in the aviation sector.

For future work, the study will focus on incorporating additional meteorological and environmental variables to capture more complex factors influencing visibility. Also, exploring more advanced deep learning architectures, such as attention mechanisms and transformer models, along with hybrid optimization methods, is expected to further boost the model's long-term forecasting capabilities and generalization performance. This framework lays a solid foundation for the development of intelligent airport weather prediction systems, which can significantly improve operational efficiency and flight safety on a broader scale.

References

- [1] Zixuan W, Qingchi Y, Zhihong Y, et al. Visibility prediction of plateau airport based on lstm[C]//2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2021: 1886-1891.
- [2] Kara A, Pekel E, Ozcetin E, et al. Genetic algorithm optimized a deep learning method with attention mechanism for soil moisture prediction[J]. *Neural Computing and Applications*, 2024, 36(4): 1761-1772.
- [3] Chen Y, Dong Z, Wang Y, et al. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history[J]. *Energy Conversion and Management*, 2021, 227: 113559.
- [4] Wang K, Hua Y, Huang L, et al. A novel GA-LSTM-based prediction method of ship energy usage based on the characteristics analysis of operational data[J]. *Energy*, 2023, 282: 128910.
- [5] Croux C, Dehon C. Influence functions of the Spearman and Kendall correlation measures[J]. *Statistical methods & applications*, 2010, 19: 497-515.
- [6] Zhang Y, Zhou Z, Liu J, et al. Data augmentation for improving heating load prediction of heating substation based on TimeGAN[J]. *Energy*, 2022, 260: 124919.
- [7] Fochesato M, Khayatian F, Lima D F, et al. On the use of conditional TimeGAN to enhance the robustness of a reinforcement learning agent in the building domain[C]//Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. 2022: 208-217.
- [8] Han W, Ying W, Hu W, et al. Time series anomaly detection based on TimeGAN and LSTM neural network[C]//2023 Eleventh International Conference on Advanced Cloud and Big Data (CBD). IEEE, 2023: 122-127.
- [9] Liu Y, Liang Z, Li X. Enhancing short-term power load forecasting for industrial and commercial buildings: A hybrid approach using TimeGAN, CNN, and LSTM[J]. *IEEE Open Journal of the Industrial Electronics Society*, 2023, 4: 451-462.
- [10] Kassa Y, Zhang J H, Zheng D H, et al. A GA-BP hybrid algorithm based ANN model for wind power prediction[C]//2016 IEEE Smart Energy Grid Engineering (SEGE). IEEE, 2016: 158-163.