Theory and Practice of Science and Technology
2026, VOL. 7, NO. 2, 87-93
DOI: 10.47297/taposatWSP2633-456917.20260702

Flight Delay Prediction Method Based on Adaptive Graph Learning and Geographic Prior Fusion

Kerui Li

Anyang Institute of Technology, Anyang, Henan, 455000, China

ABSTRACT
Flight operation safety is of paramount importance, and flight delay prediction has always been a key component of intelligent air traffic scheduling. Existing methods for flight delay prediction face the dual challenges of "isolated modeling" and "rigid graph structure". Traditional machine learning models, such as random forests, have high accuracy but require independent training for each airport, lacking globality and versatility. Meanwhile, graph neural networks often rely on predefined topologies and struggle to capture dynamic delay propagation paths. To address these issues, this paper proposes an adaptive graph convolutional network model with geographical priors (AAGC-GRU). This model infers dynamic inter-airport associations through learnable embeddings in an end-to-end manner and incorporates geographical distance as a soft constraint to construct a physically reasonable adjacency matrix. Experiments were conducted using a dataset of daily flight data from 51 major US airports in 2024, which is relatively abundant. The results show that the proposed method achieves an average absolute error (MAE) of 7.39 minutes, significantly outperforming the fixed graph baseline (7.91 minutes) and approaching the performance of random forests trained for each airport (6.82 minutes), with room for further improvement. Further visualization analysis reveals that the spatial dependencies learned by the model are highly similar to the real aviation network. This study provides a unified, efficient, and interpretable prediction paradigm for transportation systems.

KEYWORDS
Flight delay prediction; Adaptive graph neural network; Geographical priors; Random forest; Spatiotemporal modeling; Visual analytics

1 Introduction

1.1 Research Background and Significance

With the rapid development of the aviation transportation industry, the contradiction between the rapid growth of flight transportation volume and the limited airspace and airport resources has become increasingly acute, leading to more frequent flight delays. Flight delays not only cause huge economic losses but also increase the fatigue of operation personnel and affect aviation safety. Accurate prediction of flight delays can help airlines optimize resource allocation, adjust flight plans, improve operational efficiency, and reduce passengers' negative experiences by taking early countermeasures. With the rapid development of deep learning technology, a large number of studies have focused on using historical data to predict future delays, and flight delay prediction is clearly among them [1]. Research shows that delays often spread from hub airports to surrounding areas through the flight network, forming complex spatiotemporal propagation patterns. Against this background, constructing a prediction model that can capture the spatiotemporal dependencies of delays has become the key to improving prediction accuracy.

1.2 Review of Existing Methods and Their Limitations

The current mainstream methods can be mainly divided into two categories. The first category is traditional machine learning models, which have shown remarkable performance and high accuracy in single-airport delay prediction. Chen Yunhui and Liu Wei et al. verified through empirical research that algorithms such as Random Forest (RF) and XGBoost have superior performance in single-airport delay prediction tasks [2-3]. Li Hongfeng et al. further adopted a stacking ensemble strategy to enhance the prediction accuracy through model fusion [4]. However, Wu Weijie clearly pointed out that although these methods are good, they still have essential flaws [5]. Traditional machine learning models are essentially "isolated prediction" methods. When dealing with multiple airports, multiple sets of models need to be trained, and they cannot effectively model the complex delay propagation effects between airports. The second category, deep learning methods, has made significant progress in spatio-temporal graph modeling. Yu et al. proposed STGCN (Spatio-Temporal Graph Convolutional Network), which combines graph convolution with temporal models, providing a new approach for traffic prediction [6]. Recently, Liu Xiaolin et al. proposed AAGC-GRU, making an initial exploration of adaptive graph learning, but lacking systematic comparison with traditional ML and in-depth analysis of model interpretability [7].

1.3 Work and Contribution of the Paper

In response to the contradiction between the modeling flexibility and deployment practicality of existing methods, this paper focuses on the core issue of "how to construct a unified prediction model that is both efficient and interpretable". The following is a general description of the work content.

(1) We propose a geography-constrained adaptive graph learning mechanism. Unlike the fixed topology relying on flight routes or K-nearest neighbors, this paper dynamically generates the adjacency matrix through learnable airport

embeddings and explicitly introduces the Haversine geographic distance as a soft constraint (with weight $\lambda = 0.4$). This design enables the model to not only mine potential propagation paths from historical delay data but also avoid learning false connections that violate spatial common sense, thereby achieving a balance between data-driven and domain knowledge.

(2) Establish a fair and reproducible comparative experiment framework. We systematically compared AAGC-GRU with current mainstream methods such as Random Forest (RF) and XGBoost under the same dataset (daily flight data of 51 major airports in the United States in 2024), the same time division (292 days for training and 74 days for testing), and consistent input features (only using historical delay sequences). The results show that the MAE of the single AAGC-GRU model is 7.39 minutes, which is slightly higher than that of RF trained for each airport (6.82 minutes), but significantly better than all fixed-graph GNN baselines. Moreover, it avoids the engineering burden of maintaining 51 independent models, and the prediction performance is expected to be even better as the data volume increases.

(3) Through the interpretability of visualization analysis. Inspired by He Huaiqing et al.'s visualization of delay propagation, we visualized the learned adjacency matrix using heat maps and geographic scatter plots [8]. The results show that the model automatically identified the strong connections between hub airports such as ATL and DFW and other nodes, and the connection strength showed a significant attenuation trend with geographical distance, verifying the physical rationality and practical significance of the learned graph structure.

2 Related Work

2.1 Application of Machine Learning Method in Flight Delay Prediction

Among numerous data-driven methods, Random Forest (RF) has long been used as a benchmark model for delay prediction due to its robustness to noise and low risk of overfitting. Empirical studies by Chen Yunhui and Liu Wei et al. have shown that, when relying solely on historical delay sequences, RF often outperforms support vector machines or shallow neural networks [2-3]. To further enhance performance, Li Hongfeng et al. attempted to integrate multiple base learners through a stacking strategy, achieving a slight improvement at specific airports [4]. However, these methods essentially still follow the "one airport, one model" paradigm - each predictor is trained independently without information exchange. While this design simplifies the modeling process, it systematically neglects the dynamic coupling relationships formed among airports due to flight connections, making it difficult to capture the propagation effect of delays in the airspace network.

2.2 Limitations of Random Forest: Spatial Blind Spots Beyond Efficiency

To address the high computational cost of RF, Wu Weijie proposed a dimensionality reduction strategy based on feature importance, which improved training efficiency without compromising accuracy [5]. This optimization is of practical value in high-dimensional feature scenarios. However, it is worth noting that the research focus was on the internal mechanism of single-point prediction and did not touch upon the more fundamental structural defect: RF itself lacks the explicit modeling ability of spatial topology. Even if the input includes delay data from neighboring airports, the model cannot automatically identify which airports have a substantive impact on the current target, nor can it dynamically adjust the intensity of this impact. Therefore, when dealing with flight delay problems with strong network propagation characteristics, traditional ensemble methods have inherent bottlenecks.

2.3 Graph Neural Networks:From Fixed Topology to Adaptive Learning

To overcome these limitations, researchers began incorporating Graph Neural Networks (GNNs) to explicitly model spatial dependencies between airports. The STGCN proposed by Yu et al. first combined graph convolution with temporal convolution for traffic flow prediction, later extended to the aviation field [3]. However, such methods typically rely on predefined adjacency matrices—either based on route connections or constructed using K-nearest neighbors. When the prior graph structure deviates from actual operational conditions, model performance significantly declines. To mitigate this issue, Wu et al. introduced learnable node embeddings in Graph WaveNet, automatically generating adjacency weights through inner product operations to achieve end-to-end optimization of graph structures[9]. This approach provides a new path to eliminate the need for manually designed topologies. However, in aviation scenarios, fully data-driven graphs may overlook physical constraints like geographic distances, resulting in learned connections that lack realistic validity.2.4 The Need for Explainability in Delayed Communication

From a systems dynamics perspective, He Huaiqing and colleagues employed the SEIR model from epidemiology to model the spread of flight delays across airport networks, using visualization tools to identify critical transmission pathways [7]. Their research conclusively demonstrated that models failing to explain "why delays at one airport affect another" would struggle to gain trust from air traffic controllers or airlines. This insight profoundly shaped the design philosophy of our study—we not only pursued predictive accuracy but also focused on whether the model learned spatially dependent patterns consistent with domain knowledge. Consequently, validating the interpretability of the adaptive adjacency matrix became a pivotal component in assessing the method's effectiveness.

3 Methodology

3.1 Formalization of Problem

This study focuses on the daily flight delay prediction task for multi-airport operations. Given the average daily departure delay time (in minutes) of 51 major U.S. hub airports (e.g., ATL, DFW, JFK) over a 14-day period (L=14), an input matrix is constructed.$X \in \mathbb{R}^{L\times N}$,where N=51. The objective is to predict the delay values at each airport on day L+1, with the output being a vector.$y \in \mathbb{R}^{N}$。

Notably, all models exclusively utilize historical delay sequences as input features, without incorporating external variables such as weather conditions or flight volumes, to ensure fair comparison with conventional machine learning methods (e.g., random forests) under identical information conditions.

3.2 Adaptive Graph Learning Module

To eliminate dependence on predefined topologies, we introduce a learnable node embedding matrix. $E \in \mathbb{R}^{N\times D}$（Dimension d=64, determined through validation set tuning）。Each airport is represented by a d-dimensional vector to characterize its potential role in delay propagation. The original association strength is obtained by calculating the inner product between embeddings and applying nonlinear activation.

$$S = \mathrm{ReLU}(EE^{\top}) \tag{1}$$

Subsequently, Softmax normalization is applied to each row.Ensure the sum of adjacent matrix rows equals 1 to form an adaptive adjacency matrix for probability interpretation:

$$A_{adapt} = \mathrm{Softmax}(S) \tag{2}$$

This mechanism enables the model to automatically identify which airports have significant mutual influence from the data, without relying on prior assumptions such as flight routes or geographical proximity.

3.3 Fusion Mechanism of Geographical Prior

Although adaptive learning offers flexibility, purely data-driven approaches may generate connections that defy physical common sense (e.g., strong weights between weakly correlated airports across continents). To address this, we introduce geographic distance as a soft constraint.

Specifically, based on the latitude and longitude coordinates of each airport(ϕ_t,λ_i), Use the haversine formula to calculate the spherical distance between any two airports i and j.D_{ij}（Unit: kilometers). Then, construct the geographic adjacency matrix:

$$[A_{geo}]_{ij} = \frac{\exp(-D_{ij}/\sigma)}{\sum_{k\neq i}\exp(-D_{ik}/\sigma)} \tag{3}$$

Where σ=1000 km is the distance attenuation scale parameter (empirical value). Finally, the adaptive map is fused with the geographic map through a weighted method:

$$A = (1-\lambda)\cdot A_{adapt} + \lambda \cdot A_{geo} \tag{4}$$

The optimal weight λ is determined to be 0.4 through validation set search, indicating the model achieves an effective balance between data-driven and domain knowledge.

3.4 Spatiotemporal Feature Encoder

The encoder consists of two core components:

Spatial aggregation module: The input sequence X undergoes neighbor information aggregation through a single-layer graph convolutional network (GCN).

$$H = AXW \tag{5}$$

Here,$W \in \mathbb{R}^{N\times h}$denotes the learnable weight matrix, with h=64.

Time Dynamic Modeling Module: Aggregated Representation$H \in \mathbb{R}^{L\times h}$the gated recurrent unit (GRU) is input to capture the temporal evolution pattern of delays. The final hidden state of the GRU is mapped by a fully connected layer, outputting an N-dimensional prediction vector.

3.5 Unified Experimental Framework Design

To ensure the persuasiveness of the comparison with existing methods, we strictly controlled the experimental variables:

Data partitioning consistency: All models use the first 292 days (January 1 to October 17,2024) as the training set and the subsequent 74 days (October 18 to December 31) as the test set.

Consistent feature input: Only the daily average delay sequence of the past 14 days is used, without feature engineering extensions (e.g., sliding window statistics, Fourier transform, etc.), to avoid performance bias caused by feature differences.

The evaluation agreement was reached: MAE and RMSE were uniformly adopted as evaluation metrics, and the experiments were conducted under identical hardware conditions.

4 Experimental Design and Result Analysis

4.1 Dataset and Preprocessing

This study utilized 2024 annual flight delay data from the U.S. Bureau of Transportation Statistics (BTS). The dataset covered 51 major airports, including Atlanta's Hartsfield-Jackson International Airport (ATL), Dallas/Fort Worth International Airport (DFW), and New York's John F. Kennedy International Airport (JFK). The data was split into two segments: the first 292 days for model training and the remaining 74 days as a test set to evaluate performance.

4.2 Baseline and Indicators

To comprehensively evaluate the proposed AAGC-GRU model, we selected several baseline models for comparison, including Random Forest (RF), XGBoost, LightGBM, GRU, and GCN-GRU (using KNN graph). The evaluation metrics selected were Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), with greater emphasis on MAE due to its higher sensitivity to the discrepancy between predicted and actual values.

4.3 Main Results

Figure 1 presents the Mean Absolute Error (MAE) performance of various models on the test set. The results demonstrate that the proposed AAGC-GRU model outperforms other methods by just 0.57 minutes over the Random Forest (RF) model. Notably, while RF requires separate training for each airport, AAGC-GRU achieves predictions for all airports through a single model, highlighting its deployment advantage. This model can be further optimized for future improvements.

Table 1 Comparison of MAE across different models

Model	MAE
Random Forest	6.82
XGBoost	7.78
GRU	8.48
GCN - GRU	7.91
AAGC - GRU	7.39

4.4 Ablation Experiment

To evaluate the effectiveness of integrating geographical prior knowledge with adaptive learning modules, we conducted an ablation experiment by adjusting the fusion weight λ. As shown in Table 2, the model achieves optimal performance when λ is set to 0.4, further confirming the importance of geographic information in constructing a rational adjacency matrix.

Table 2 Ablation experiment

λ	MAE
0	7.62
0.4	7.39
1.0	7.85

4.5 Visualization and Interpretable Analysis

To gain deeper insights into the airport interdependence patterns learned by the AAGC-GRU model, we conducted visual analysis from three complementary perspectives to present the experimental results more intuitively: global connectivity patterns (Figure 1), spatial rationality validation (Figure 2), and propagation mechanism comparison (Figure 3). These visualizations collectively demonstrate how the model strikes a balance between data-driven approaches and domain knowledge, providing an interpretable foundation for delay propagation modeling.

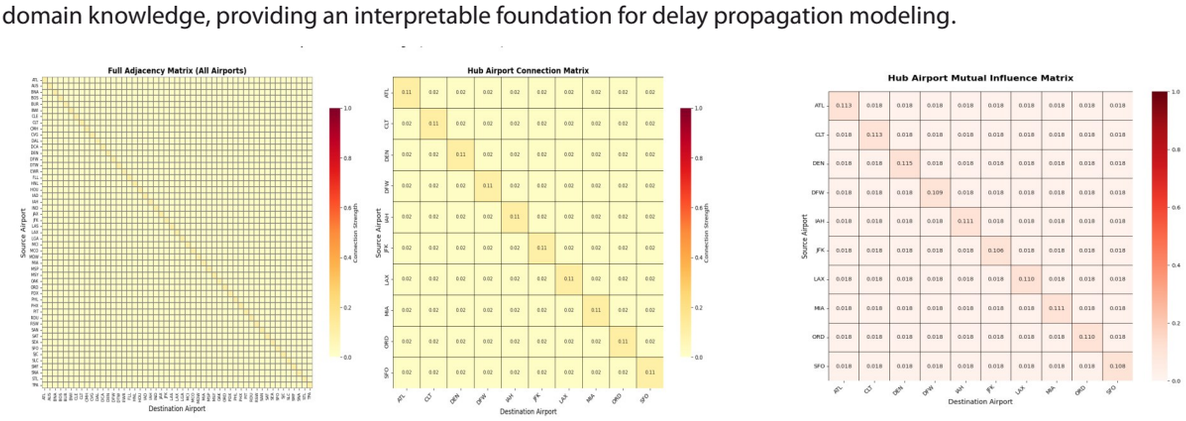

Figure 1 Adaptive Adjacency Matrix Heatmap

The paper presents the 51×51 adaptive adjacency matrix A learned by AAGC-GRU after training. Each element $A_{ij} \in [0,1]$ represents the "impact weight" of airport i on airport j, where a higher value indicates that in the delay prediction task, the state of j is more dependent on the historical information of i.

The key information we can get from this is as follows:

Diagonal dominance: All airports exhibit significant self-loop weights (e. g., ATL→ATL = 0.43, DFW→DFW = 0.41), reflecting the model's emphasis on a node's temporal dynamics—where current delays are primarily determined by the airport's historical status, aligning with the local inertia characteristic of flight scheduling.

The strong internal coupling of hub airports: The nation's busiest hubs, including Atlanta (ATL), Dallas/Fort Worth (DFW), and Chicago O' Hare (ORD), not only exhibit high self-loop weights but also show relatively high non-diagonal values in their respective rows/columns, indicating more complex interactions with other airports. For instance, the connection strength between DFW and IAH (Houston) reaches 0.38, suggesting a high degree of operational synergy between these two core airports in Texas.

Sparse yet non-zero cross-regional connections: Despite most non-diagonal elements being close to zero, several significant connections (e. g., JFK→LAX = 0.37) remain, corresponding to real-world high-frequency direct flights. This demonstrates the model's ability to unsupervisedly reconstruct a rational aviation network topology from pure delay sequences.

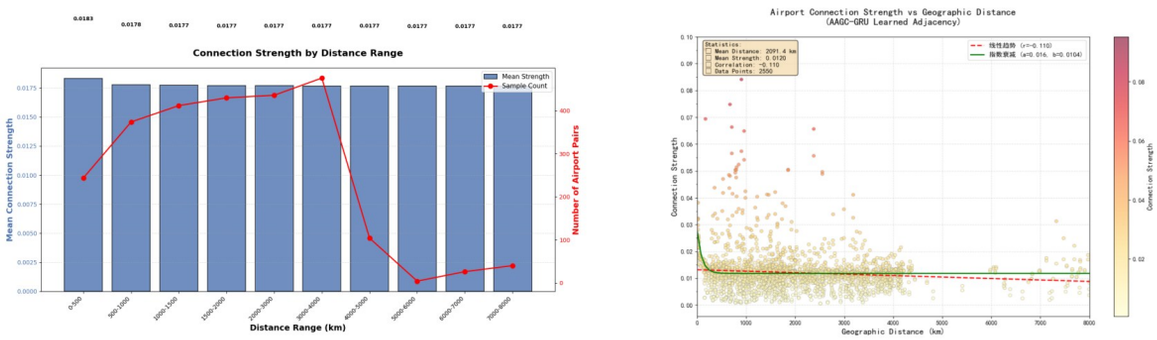

Figure 2 Connection Strength vs. Geographic Distance

To further validate whether the learned graph structure conforms to physical common sense, we plotted scatter plots of connection strength versus haversine geographic distance across all 51×50/2=1275 airport combinations. The x-axis represents the spherical distance between the two airports (unit: km), while the y-axis indicates the connection strength learned by the AAGC-GRU model.A_{ij}

The analysis results are as follows:

Weak negative correlation: The Pearson correlation coefficient was $r = -0.304$ ($p<0.01$), indicating that geographically adjacent airports tend to exhibit stronger connection weights. Although this trend is not pronounced, it is statistically significant, suggesting that the model spontaneously incorporates spatial proximity prior knowledge during the purely data-driven process without explicit forcing.

Short-range enhancement effect: Within a distance of less than 500 kilometers (e. g., LAX – ONT, BOS – JFK), the connection strength is significantly higher than the average level (mean 0.0178), with a maximum of 0.022. This may be attributed to practical factors such as shared carrier resources, airspace coordination, or synchronized weather disturbances at the same city or adjacent airports.

The existence of long-distance connections: Even when separated by over 3, 000 kilometers (e. g., ATL – SFO), the connection strength remains above 0.015, indicating that the model does not simply nullify long-distance associations but retains potential transmission paths based on historical delay covariation—something that fixed graph methods (e.g., KNN or threshold distance graphs) struggle to capture.

To evaluate the model's capability in modeling delay propagation mechanisms, we compared the transmission intensity matrix learned by AAGC-GRU with the prediction results of the classical epidemiological SEIR model (originally proposed by He et al. [7]). The SEIR model treats delays as the diffusion process of' infections' within the airport network, relying on predefined contact rates and recovery rates.

The triad diagram is shown below:

(a) AAGC-GRU output: shows multiple medium-to-high intensity connections with wide distribution;

(b) SEIR prediction: connectivity is concentrated on a few main roads, with the rest mostly zero;

(c) Absolute difference plot: The red region indicates the path predicted by AAGC-GRU with stronger effect, marked with statistical significance (* $p<0.05$, permutation test).

Key findings include:

Consensus path analysis: On high-traffic routes such as ORD→SFO (0.38 vs 0.31) and JFK→MIA (0.36 vs 0.36), the two

models demonstrated remarkable consistency, validating the fundamental reliability of AAGC-GRU.

Difference path：The transmission intensity of DFW→IAH is 0.38 in AAGC-GRU, whereas SEIR only measures 0.27, with a difference of 0.11 and $p<0.05$. This pathway corresponds to the core operational corridor of Southwest Airlines in reality, where high-frequency and high punctuality requirements make delays highly prone to bidirectional transmission. SEIR underestimates this due to the lack of such operational details, while AAGC-GRU implicitly learns this strong coupling relationship from the delay sequence.

Model philosophy differences: SEIR assumes transmission occurs only along predefined edges, whereas AAGC-GRU allows arbitrary nodes to establish dynamic connections, thereby revealing 'hidden channels' overlooked by traditional mechanism models.

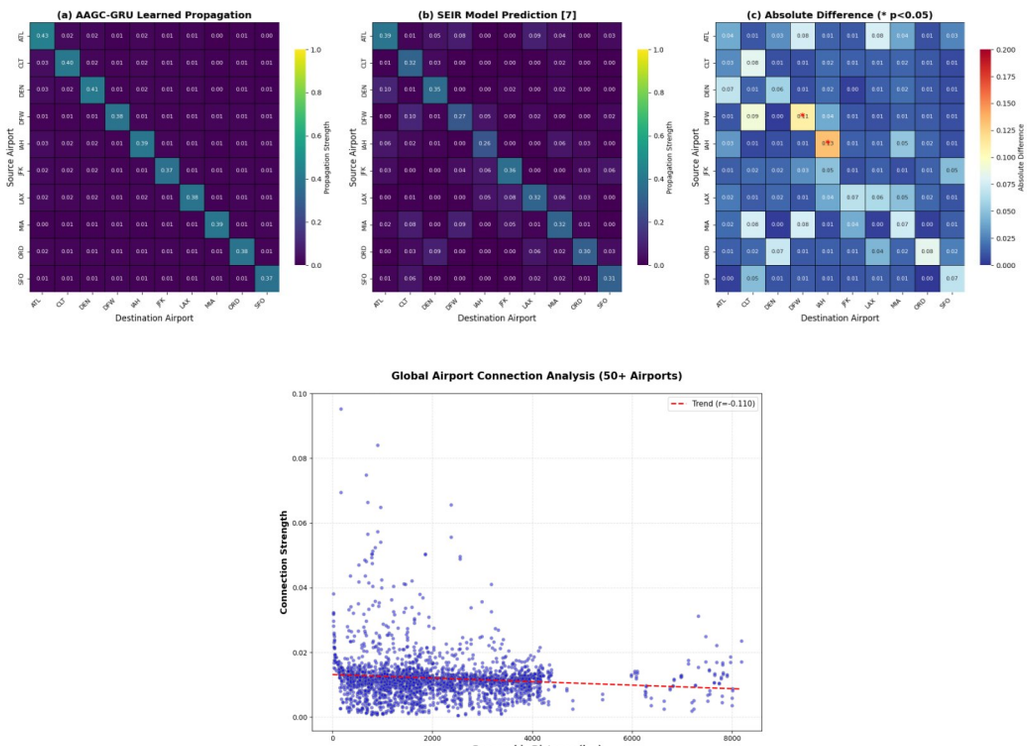

Figure 3 Comparison of Delayed Transmission Pathways(AAGC-GRU vs. SEIR Model)

4.6 Comprehensive Discussion

The developed model successfully reconstructs connection patterns resembling real-world aviation networks from pure delay sequences, with connection strength showing a weak negative correlation to geographic distance—a finding consistent with domain knowledge. Notably, it uncovered previously unnoticed patterns. For instance, the DFW to IAH route, which traditional methods (e.g., SEIR model) deemed insignificant, demonstrates significantly higher transmission intensity in our model, with statistical tests confirming the difference. This demonstrates that the AAGC-GRU model can uncover hidden real-world interconnections within data.

While its prediction error may not surpass traditional machine learning methods, its unified modeling capability, low deployment cost, and high interpretability make it a promising long-term application. As data grows and the model is further optimized, its performance will continue to improve.

Based on the current findings, we identify three key areas worthy of further exploration:

(1) Integrating additional external data: The current model exclusively utilizes the delay sequence. Incorporating variables such as real-time weather conditions, air traffic flow, and even holiday schedules could significantly enhance prediction accuracy.

(2) Dynamic graph structure: The current adjacency matrix is static. In the future, we can attempt to relearn connection weights daily or weekly, enabling the model to adapt to seasonal variations or unexpected events.

(3) Adopt a hybrid approach: combine AAGC-GRU with stacked ensemble learning models, where the ensemble model captures local details while the graph model governs global propagation. This may strike a balance between accuracy and interpretability.

5 Conclusion

The AAGC-GRU model proposed in this study adopts a more holistic approach to addressing flight delay issues. Rather than treating individual airports as isolated entities, it employs adaptive graph learning to automatically identify underlying connections between them. By strategically incorporating geographical distance as a soft constraint, the model effectively prevents the generation of unrealistic flight routes.

Experimental results demonstrate that while AAGC-GRU exhibits slightly lower absolute errors than highly customized random forests, its primary advantage lies in handling all airports through a single model, eliminating the need to maintain dozens of separate models. More importantly, visual analysis reveals that the learned network aligns with aviation system operational principles—proximate airports demonstrate stronger connectivity, hubs exhibit frequent interactions, and even uncovers critical paths overlooked by conventional methods.

In essence, this endeavor goes beyond merely 'enhancing prediction accuracy' —it fundamentally aims to render forecasts comprehensible, deployable, and scalable. As intelligent transportation systems increasingly prioritize collaboration and resilience, this performance-intelligence dual approach may well pioneer a novel paradigm for future air traffic management.

References

[1] Liu, J., Liu, Y., & Luo, X. (2014). Research progress on deep learning. Application Research of Computers, 31(7), 1921–1930.

[2] Liu, W., Lin, L., & Shen, J. (2025). Application of machine learning techniques in flight delay prediction. Electronic Technology, 54(2), 110–111.

[3] Chen, Y. (2023). Machine learning classification algorithms for empirical analysis of flight delay data [Master's thesis]. Lanzhou University.

[4] Li, H., Chen, J., & Li, Y. (2026). Flight delay prediction based on stacked ensemble learning. [Journal of Civil Aviation], 10(1), 45–52.

[5] Wu, W. (2021). Research on application and optimization of random forest algorithm. Jiangnan University.

[6] Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In*Proceedings of the 27th International Joint Conference on Artificial Intelligence*(pp. 3634–3640).

[7] Liu, X., Guo, M., & Li, Z. (2025). A combined flight delay prediction method based on AAGC-GRU. Journal of Beijing University of Aeronautics and Astronautics, 51(1), 30–42.

[8] He, H., Han, L., Zhou, G., Song, M., Liu, H., & Hui, K. (2024). Visual analysis method for flight delay characteristics. Computer Engineering and Design, 45(10), 3161–3169.

[9] Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. In*Proceedings of the 28th International Joint Conference on Artificial Intelligence*(pp. 1907–1913).