Predicting Reactionary Delays in a Hub-Spoke Network using Graph Attention Neural Networks

Constança Veiga, Marta Ribeiro
Faculty of Aerospace Engineering
Delft University of Technology
Delft, Netherlands

Marie Carré
Operations Research & Air Traffic Management
Swiss International Air Lines (SWISS)
Zurich, Switzerland

***Abstract*—Reactionary delays are a critical challenge in airline operations, especially within hub-spoke networks, where disruptions at spoke airports propagate and amplify throughout the fleet. Accurate prediction of these delays is essential for effective network planning, as errors can lead to flight cancellations, missed connections, and curfew infringements. However, current state-of-the-art delay prediction models do not fully integrate all elements that cause reactionary delays and affect subsequent operations. This study aims to close this gap by using a Graph Attention Network (GAT) model to predict reactionary delay distributions within a fleet network and identify the most critical flights through the analysis of attention weights. Using operational data from Swiss International Air Lines' short-haul fleet, the GAT model integrates node-level features, such as flight-specific parameters, and edge-level features, including rotational dependencies and passenger connections, to capture the spatial-temporal dynamics of delay propagation. The GAT model achieved reliable predictive accuracy, particularly on medium-delay days, of a root mean squared error of 15.59 minutes and a mean absolute error of 10.50 minutes. The results further reveal that the model comprehends the ripple effects caused by rotation delays. Furthermore, its attention weights confirm its capability to identify critical flights and connections, enabling the airline to allocate resources more effectively.**

***Keywords*—Reactionary Delays; Delay Propagation; Graph Attention Network; Airline Operations; Critical Flights**

I. INTRODUCTION

Airline punctuality is critical not only for operational efficiency but also for passenger satisfaction and cost control. Delays in air traffic operations can be classified as primary delays and reactionary delays. Primary delays are initial disruptions caused by factors such as technical issues, weather conditions, or air traffic control restrictions. Reactionary delays, on the other hand, occur when earlier disruptions affect later flights and can significantly amplify the impact of the initial delay [1]. In the European air traffic network, reactionary delays account for approximately 45% of total delays, sometimes exceeding primary delays in their cumulative effect [2].

Reactionary delays present a unique challenge, especially within hub-spoke airline networks, where spoke airports can introduce unpredictable complexities due to arrival delays [3]. Spoke airports often have limited resources and operational constraints, making them more susceptible to delays that can propagate through the network. Hub-spoke networks suffer more from reactionary delays than point-to-point networks because delays originating at spoke airports can quickly spread to hub airports, which are central nodes connecting multiple routes [4]. This interconnectedness means that a single delay in a spoke airport can have a ripple effect, causing widespread disruptions across the network.

Current state-of-the-art approaches typically focus on estimating departure delays at specific airports, often using the previous flight's delay as an input [5]. However, because actual delays for later flights are unknown earlier in the day, prediction accuracy decreases over longer time horizons. To mitigate the effects of reactionary delays, airlines need predictions that allow sufficient time for operational adjustments. Moreover, many studies lack passenger data, which is essential for assessing the impact of reactionary delays on connecting flights. To address these challenges, this study analyzes the entire short-haul fleet of Swiss International Air Lines (SWISS), comprising of approximately 350 flights per day connecting Zurich Airport, the central hub, to numerous spoke airports across Europe. The dataset includes detailed flight information such as scheduled and actual arrival and departure times, minimum ground times, and confidential connecting passenger data.

This research explores whether a Graph Attention Network (GAT) model can accurately determine reactionary delay distributions within a fleet network, specifically examining the role that spoke airports play in delay propagation. By using a dynamic, graph-based model and real operational data from SWISS, the model can address the complex, real-time nature of delay propagation, especially in hub-spoke networks where spoke airports play a critical role. Furthermore, the attention weights within the GAT model highlight the most critical flights, offering insights into which connections have the greatest impact on delay propagation. This interpretability is important for operational decision-making, as it allows SWISS to identify and prioritize interventions on key flights that could mitigate widespread disruptions and improve overall network performance.

This paper is structured as follows. Section II highlights the current state-of-the-art and the research gap covered by this work. Section III describes the methodology applied in this study. The results of the GAT model are then presented and validated in Sections IV and V, respectively. Finally, Sections VI and VII discuss and conclude this work.

II. Related Work

Delay propagation in air traffic networks has been studied through a variety of methods, ranging from mathematical and statistical models to machine learning techniques, that aim to capture the complexity of how delays spread over interconnected flights. Early mathematical approaches introduced concepts such as the 'delay multiplier', which quantifies how an initial delay amplifies throughout the network [6]. Other techniques employed Monte Carlo simulations, where statistical distributions capture variability in processes such as ground operations [3]. More sophisticated models, including Delay Propagation Trees and Bayesian Networks, explicitly account for the non-independent nature of delays. For instance, a Delay Propagation Tree with a Bayesian Network (DPT-BN) approach modeled interlinked factors (i.e., aircraft, crew, and passenger connections) using conditional probability distributions [7]. Although valuable for revealing causal relationships, many of these mathematical and statistical models are typically validated on a limited number of datasets or restricted operational scenarios, which hinders their generalizability.

Statistical methods such as regression analysis and time-series forecasting have provided additional insights into delay patterns. For example, linear regression models have been used to link early delays to subsequent disruptions at airports [8], while Granger causality and its refinements enable the construction of Delay Causality Networks (DCNs) that track how delays flow between airports [9], [10]. However, these methods frequently overlook real-time and dynamic factors such as evolving weather conditions or peak operational periods, limiting their practical applicability.

With increased computational power, machine learning (ML) methods have gained traction. Gradient Boosting Decision Trees (GBDTs) and Random Forests excel at capturing non-linearities in large datasets [11], and have been used by organizations like EUROCONTROL to improve arrival time predictions [5]. Although these approaches often show strong performance close to the departure time, their accuracy typically diminishes for longer prediction horizons. Deep learning models, including LSTM, Recurrent Neural Networks (RNNs) [12], and Graph Neural Networks (GNNs) [13], further exploit spatial-temporal dependencies. For example, LSTM-based methods forecast delays for specific look-ahead intervals, while graph-based architectures such as the Spatial-Temporal Gated Multi-Attention Graph Network can anticipate delays across wide airport networks [14]. Despite these advancements, most deep learning models are still primarily validated on historical datasets and rarely incorporate real-time data streams or airline-specific resource constraints, such as crew or passenger connections.

As a result, several gaps persist in literature. Access to detailed passenger connection data is frequently restricted, limiting the ability to fully capture how delays ripple across flights. Many studies focus on U.S. air traffic data, leaving European or other regional networks underrepresented. Moreover, although multiple causal factors, such as aircraft, crew, and passengers, are known to contribute to delays, they are rarely modeled together. Lastly, a persistent challenge lies in dynamically updating predictions in real-world operations, where conditions like weather and congestion fluctuate quickly.

This research addresses these limitations by adopting a GAT model capable of dynamically capturing both the importance and evolving relationships of interconnected flights. In contrast to traditional Graph Convolutional Networks, which treat all connections with static weights, a GAT model continuously adjusts the relevance of flight-to-flight connections and can more effectively leverage edge features. This approach enables the model to better reflect changing conditions within the network and accurately predict how delays propagate. By identifying the flights and connections that have the greatest impact on cascading delays, the proposed research aims to improve the resilience and efficiency of airline operations.

III. Methodology

GATs are inherently complex, making it crucial to thoroughly understand their architecture in order to effectively comprehend their functionality and identify appropriate applications. To present the methodology used in this study, Section III-A introduces the GAT model, and its architecture is discussed in Section III-B. A description of the model training procedure, along with the tuned hyperparameters, are included in Sections III-C and III-D, respectively.

A. Problem Definition and Graph Representation

The problem of predicting flight delay propagation is formulated as a graph-based problem where each flight is modeled as a node, and the relationships between flights (such as connections and rotations) are represented as edges. Let $G = (V, E)$ be a directed graph where:

- $V = \{v_1, v_2, ..., v_n\}$ represents the set of flights, with each node v_i corresponding to a unique flight.
- $E = \{e_{ij}\}$ represents the set of edges, where an edge e_{ij} exists if flight v_i is connected to flight v_j through passenger connections, crew rotations or aircraft rotations.

This is visually represented in Figure 1. Each node v_i is associated with a feature vector $\mathbf{h}_i$ that includes both categorical and numerical attributes related to the flight. Similarly, each edge e_{ij} is associated with a feature vector $\mathbf{e}_{ij}$, capturing the relationship between connected flights.

Each component of the GAT model utilizes specific node, edge, and graph features to capture complex relationships

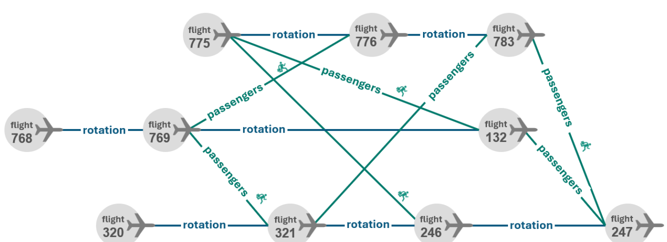

Figure 1: Illustration of the flight network, showing nodes (flights), edges (connections), and their interdependencies.

within the network. Table I displays the features associated with each node-level feature vector $\mathbf{h}_i$. Table II lists the features associated with each edge-level feature vector $\mathbf{e}ij$ for connections between nodes i and j. Graph-level features, shown in Table III, characterize the entire graph. It was decided to include the day and month as graph features to help the model capture seasonal patterns and daily trends affecting flight delays.

TABLE I. Features for Flight Nodes

Features	Type
Aircraft type	Categorical
Departure airport	Categorical
Arrival airport	Categorical
Delay codes	Categorical
Estimated time of departure (ETD)	Numerical (time)
Scheduled departure time	Numerical (time)
Scheduled arrival time	Numerical (time)
Actual departure time	Numerical (time)
Actual arrival time	Numerical (time)
Flight distance	Numerical
Number of passengers	Numerical

TABLE II. Features for Flight Edges

Features	Type
Connection type	Categorical
Flight number	Numerical
Number of connecting passengers	Numerical
Destination	Categorical
Airport capacity	Numerical
Number of connections per class	Numerical
Connecting time	Numerical (minutes)
Minimum connecting time	Numerical (minutes)
Minimum ground time	Numerical (minutes)
Number of HON and Senator members	Numerical
Number of wheelchair passengers	Numerical

TABLE III. Graph-Level Temporal Features

Features	Type
Day	Numerical (time)
Month	Numerical (time)

B. Model Architecture

The GAT model leverages multi-layer attention-based graph convolutions to predict arrival times by learning from the relationships and dependencies among flights. The model is implemented using the PyTorch Geometric library, with three attention layers that progressively refine the node-level embeddings through neighborhood aggregation. The architecture of the model has two primary components.

1) Attention Mechanism: The first layer in the model applies the graph attention mechanism, which assigns attention coefficients α_{ij} to each edge e_{ij} between a node v_i (the target node) and its neighboring node v_j. These coefficients quantify the importance of neighboring nodes in determining the representation of a target node. The attention coefficients are computed as follows:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top\left[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top\left[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k\right]\right)\right)} \tag{1}$$

where a is a weight vector, W is a learnable weight matrix, $\mathcal{N}(i)$ denotes the neighbors of node i, and $\|$ represents concatenation. These coefficients determine the influence of neighboring nodes on the target node during the aggregation process. The LeakyReLU activation function introduces a small, non-zero gradient for negative input values, allowing the model to retain information from these inputs rather than setting them to zero.

2) Layers: Following the attention mechanism, two Graph Attention Convolutional (GATv2Conv) layers are used to further refine the node embeddings. These layers continue aggregating information from neighboring flights using the learned attention weights, improving the expressiveness of each flight's representation:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right) \tag{2}$$

where l indicates the layer number, σ is a non-linear activation function, and α_{ij} are the attention weights from the previous layer. The final layer of the network is a linear layer that maps each node's final embedding to a predicted arrival time $\hat{t}_{\text{arr},i}$ for each flight v_i.

C. Training

The GAT model is trained using historical flight data from SWISS, covering the period from January 1st to September 30th 2024. This dataset includes detailed records of flight schedules, actual departure and arrival times, delay codes, and various other flight-related features. Based on this historical data, the model is tasked with learning to predict delay propagation. Furthermore, the dataset is pre-processed to encode temporal features using trigonometric functions to capture their cyclical nature. Both node and edge features are normalized to improve training stability.

D. Hyperparameter Optimization

A three-layer GAT architecture was found to provide good prediction abilities whilst not overfitting. Each layer has progressively fewer attention heads in each subsequent layer to refine relational patterns. Moving from two to three layers improved the model's ability to capture complex dependencies, while a fourth layer led to overfitting. Weight decay was omitted to enable more flexible parameter tuning, and a dropout rate of 0.1 was used to mitigate overfitting without restricting the model's adaptability. The final hyperparameters, shown in Table IV, achieved the best balance of performance and generalization.

TABLE IV. Hyper-parameter Fine-tuning

Hyper-parameter	Value(s)	Optimum
Batch Size	1 – 100	10
Epoch	100 – 1000	500
Drop-out	0.1 – 0.6	0.2
Learning Rate	0.001 – 0.1	0.002
Number of Layers	1, 2, 3, 4	3

IV. Results

This section presents the model's training and testing performance, along with evaluation metrics such as RMSE and MAE to assess prediction accuracy. Through an analysis of loss curves (Subsection IV-A), RMSE, MAE, and MAPE (Subsection IV-B), the model's generalization capability and its effectiveness across different delay scenarios is assessed. Finally, in Subsection IV-C, feature importance is analyzed.

A. Training and Testing Performance

The training loss, in Figure 2, decreases rapidly and stabilizes at 0.0702 by epoch 1000, indicating effective pattern capture without overfitting. The test loss, in Figure 2, follows a similar trend and stabilizes at 0.0071, which is lower than the training loss, suggesting robust generalization and potentially lower test data complexity. Overall, these results confirm that the chosen architecture and hyperparameters are effective and reliable for predictions on new data.

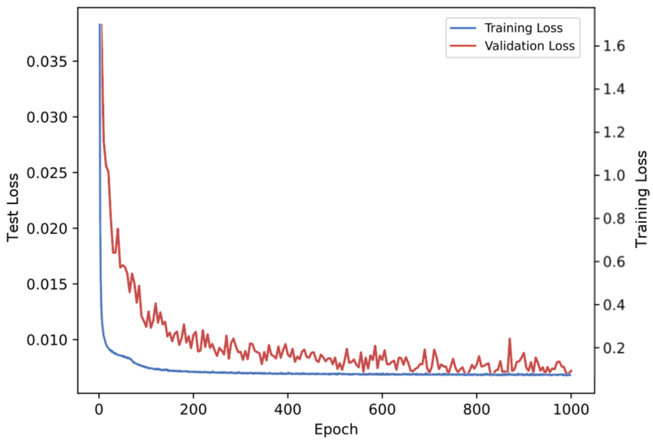

Figure 2: Loss curves

B. Performance Measures

The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) metrics are used to assess model performance. For the validation dataset, these metrics indicate prediction errors across all flights, representing overall accuracy for the entire day before each flight day begins. Both RMSE and MAE are approximately 16.35 minutes, suggesting a consistent error margin. The low MAPE of 2.61% implies a strong alignment between predicted and actual arrival times.

Table V displays the performance of the GAT model across various delay day categories. The model achieves the highest accuracy of 85.7% on medium delay days, likely due to the large number of medium delay examples in the training dataset. In contrast, low delay days exhibit lower accuracy as a result of their scarcity, while high delay days remain inherently challenging to predict, exhibiting the highest errors with an accuracy of 31.7%. Moreover, the RMSE consistently exceeds the MAE across all categories, highlighting the influence of outliers on performance metrics.

TABLE V. Performance Metrics by Delay Day Category

Test Days	MAE [min]	RMSE [min]	Accuracy
High delay days	27.39	37.56	31.7%
Medium delay days	10.50	15.59	85.7%
Low delay days	13.45	23.90	72.3%
All test days	16.30	26.52	65.3%

C. Feature Importance

GNN Explainer provides interpretability for GNNs by identifying important substructures and features that contribute most to a model's predictions [15]. Figure 3 displays the average importance of each node feature across samples. The ground time and total number of passengers are identified as having the highest influence on arrival predictions. This aligns with operational patterns where increased passenger counts and ground handling requirements are known contributors to delays. Additionally, the departure airport also scores highly, indicating that certain airports, such as high-traffic hubs, may be more susceptible to delays. In turn, scheduled time represented by the encoded values for both scheduled arrival and departure times hold moderate importance. These features capture daily and weekly cycles, such as peak travel hours, that can influence delay likelihood.

Regarding the edge features in Figure 4, rotation connections show notably higher importance than passenger connections, reflecting the impact of consecutive flights sharing the same aircraft: if an earlier flight is delayed, the subsequent flight using that aircraft is likely delayed as well. This strong dependency underscores the importance of aircraft

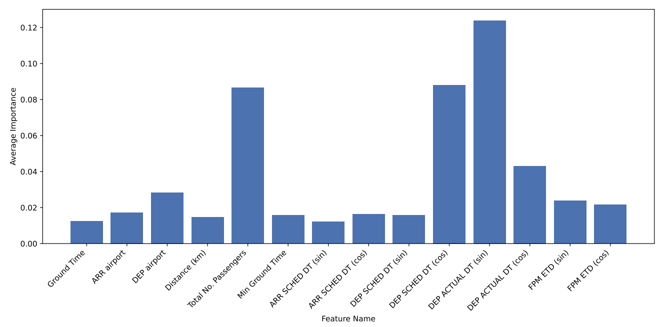

Figure 3: Average node feature importance across samples

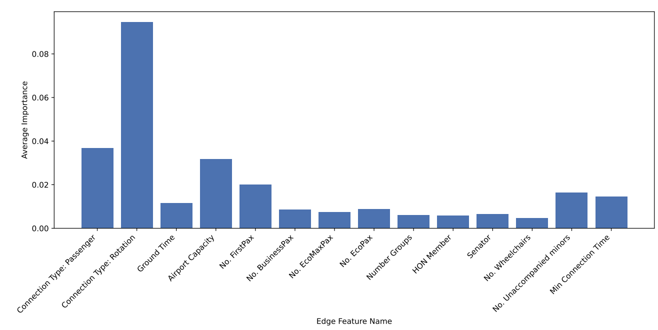

Figure 4: Average edge features importance across samples

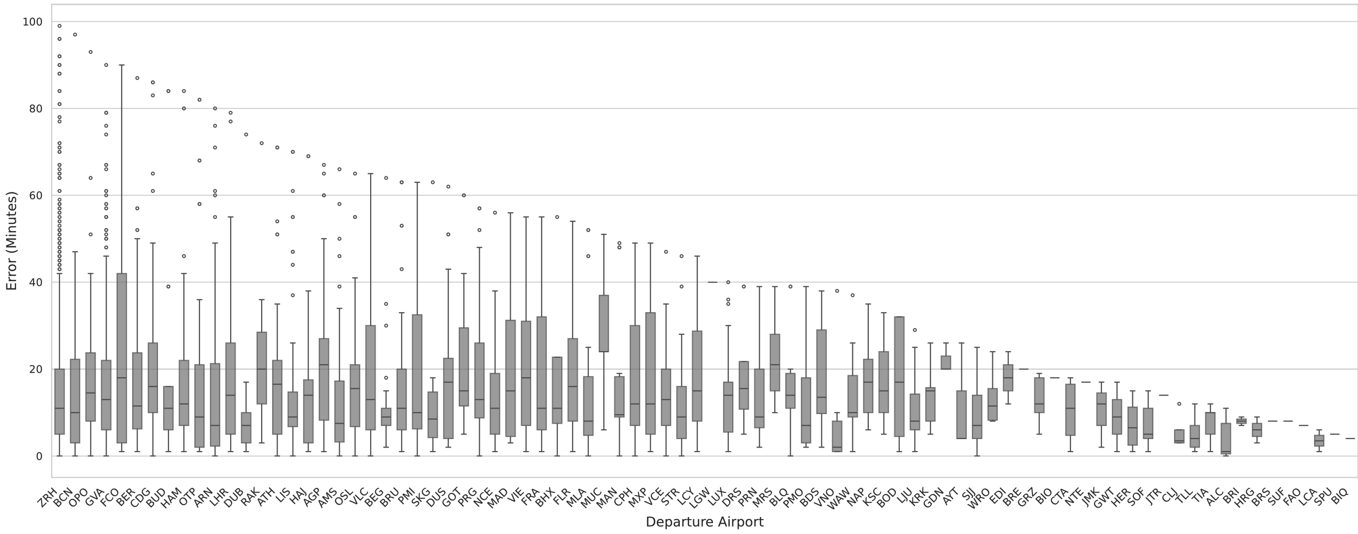

Figure 5: Error vs Departure airport

rotation, as disruptions in one segment can cascade into subsequent flights. According to the GAT model, passenger-related features, such as the number of economy passengers (i.e., EcoMaxPax, EcoPax) and group sizes, demonstrate moderate importance in delay predictions. Interestingly, the model assigns relatively high significance to the number of first-class passengers, possibly reflecting operational factors linked to premium travelers.

Special service features, such as wheelchairs and unaccompanied minors, also appear to have low importance. However, this does not necessarily imply irrelevance for delay predictions; their limited frequency (around 10% for wheelchairs and 5% for unaccompanied minors) may constrain the model's ability to capture their effects on network-wide delays. Finally, minimum connection time demonstrates moderate importance, indicating that short connection windows raise the likelihood of delays if disruptions occur at any point in the network.

V. Validation

This section further elaborates on the results previously presented to assess the quality of the model. Section V-A presents the error per airport, followed by Section V-B which investigates the model's understanding of delay propagation. Section V-C compares the results obtained with a baseline model, where XGBoost is used to predict arrival times on a per-flight basis before the start of the day. Section V-D presents unforeseen cases that the model cannot predict and Section V-E evaluates the attention weights generated by the model. It is important to note that, to protect confidentiality, all flight numbers and airport codes (except ZRH, the SWISS hub) have been altered and assigned random values.

A. Error vs Departure Airport

An analysis of error versus departure airport, as seen in Figure 5, illustrates the geographical influence on model predictions. Each airport reflects a unique operational environment, with variability in delays likely influenced by factors such as airport size, congestion levels, and connectivity. For instance, larger international hubs, such as CDG, exhibit higher error margins, potentially due to the greater complexity associated with larger volumes of connecting passengers and intricate scheduling interdependencies. This further strengthens the idea that airport-specific features are crucial to understanding the model.

When comparing to the historical delay average per airport, it was observed that the model has a tendency to overpredict delays. The model performs with varying accuracy across the airports, reflecting their unique delay patterns. For example, FCO exhibits a median error of approximately 20 minutes, with whiskers extending up to 90 minutes, reflecting high variability in prediction accuracy. This variability aligns with FCO's historical delays, which are generally low but can span a wide range. Occasional overpredictions appear to contribute to FCO's unpredictability in delay increases.

In contrast, AMS demonstrates a clear pattern of adaptability. With a median error of approximately 10 minutes, its stable historical delay profile, characterized by low median delays and narrow variability, is clearly reflected in the prediction error. This performance indicates that the model operates efficiently in environments with predictable delays while consistently achieving low error rates without frequent overestimates. Overall, airports with consistent delay patterns, such as AMS and BHX, demonstrate lower prediction errors, whereas airports with more unpredictable delays, like FCO and CDG, experience a higher frequency of overpredictions.

B. Propagation Delay

The GAT model attempts to detect how a delay in one flight can ripple through subsequent connections. By predicting arrival times and comparing the model's output with actual delay propagation, this section examines whether the GAT model can effectively predict these cascading delay effects. The delay propagation curves provide a side-by-side comparison of real and predicted delays for individual aircraft, with red indicating actual delay propagation and blue representing the model's predictions. These predictions are generated by running the model at 02:00 UTC.

Certain aircraft, such as Aircraft A (Figure 6), demonstrate a strong alignment between predicted and actual delay patterns, indicating that the GAT model effectively captures propagation in these cases. The model's predicted curve closely follows actual fluctuations throughout the day, suggesting it has learned the sequence of dependencies between flights. Similarly, for Aircraft B (Figure 7), the predicted delay curve follows the real trajectory, underscoring the model's ability to track delay fluctuations and identify dependency patterns in a structured network. Although the GAT captures the overall propagation trend, it does not fully match the magnitude of the actual delay, illustrating a limitation while still reinforcing the model's capability to represent critical relationships driving delay propagation.

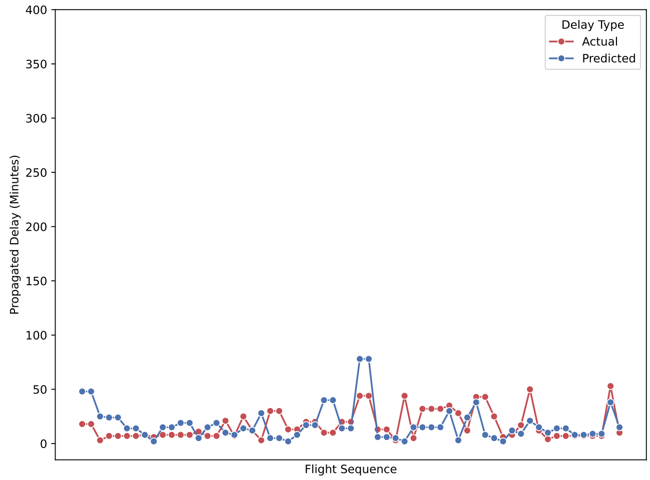

Figure 6: Aircraft A

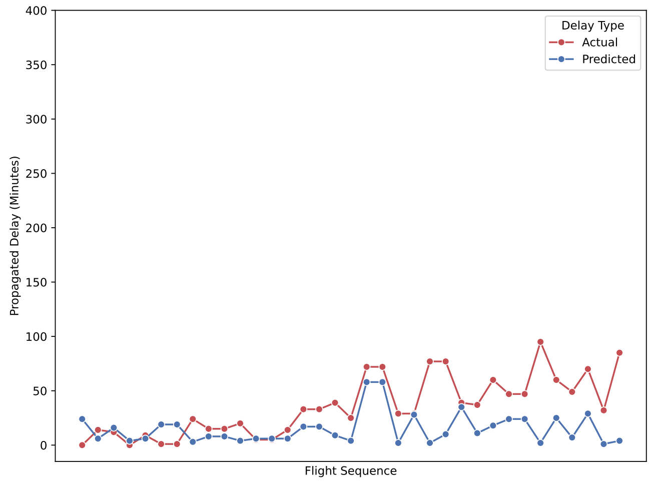

Figure 7: Aircraft B

However, not all cases display such alignment. In Figure 9, the GAT model's predictions remain relatively flat compared to the pronounced peaks in the real data, suggesting complexities in Aircraft D's flight connections or rotations that the model cannot fully capture. Similarly, for Aircraft C in Figure 8, the predicted delays do not closely track the real propagation pattern, particularly at the beginning. This gap could imply that the model's graph structure struggles to generalize to aircraft with less predictable or atypical patterns of delay propagation.

These inconsistencies reveal that while the GAT model may perform well in certain structured network scenarios, it may lack the versatility required to accurately model propagation for more irregular cases, potentially due to limited information or insufficient representation of key operational features within the dataset.

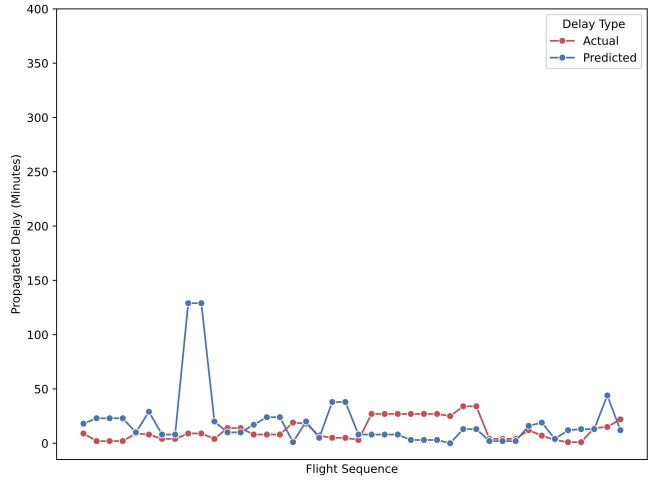

Figure 8: Aircraft C

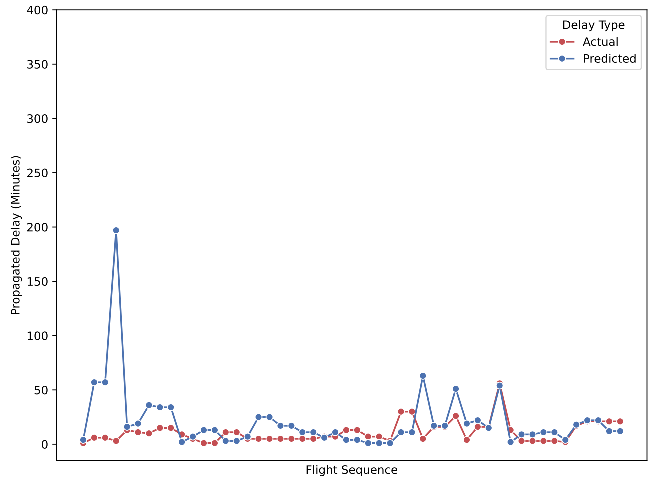

Figure 9: Aircraft D

C. Comparison with Baseline Model

The GAT model is evaluated against the existing approach used by SWISS for arrival time prediction. The current model employs XGBoost, a gradient-boosting framework optimized for efficiency and performance, to predict arrival times per flight before the start of the day. XGBoost uses a set of features primarily related to each flight's structural characteristics, as detailed in Table VI, and relies on a more limited feature set compared to the GAT model. This analysis investigates how the additional features used by the GAT model and its graph-based approach lead to potential performance improvements. The models also differ in feature handling; the GAT model does not include average ground time or maximum ground time, and it does not explicitly represent features such as ground time before a flight or the number of flights without a break as node or edge attributes. Instead, the graph structure implicitly captures these elements through the connections and interactions between flights, effectively leveraging relational and temporal dependencies.

The baseline model's feature importance, in Table VI, has a strong reliance on temporal and structural characteristics. Departure time of day ranks highest, likely reflecting congestion patterns. Departure airport and previous ground time also score highly, emphasizing the influence of origin-specific factors and prior schedules. Additional features, such as departure month, city pair, and turnaround metrics, capture seasonal, route-specific, and operational considerations. Comparing the GAT and SWISS models reveals that both assign high importance to scheduled and actual departure times, along with the departure airport. However, the GAT model also emphasizes total passenger count, a feature absent in the SWISS model. Furthermore, while the SWISS model treats departure month and day as categorical variables, the GAT model integrates these time-based factors within its graph structure.

Table VII shows that both the SWISS and GAT models yield nearly identical RMSE performance in predicting arrival times, with minimal day-to-day differences. The RMSE values indicate similar accuracy levels, with each model occasionally showing a marginal advantage, and the consistent standard deviation suggests comparable variability in prediction errors. These results imply that the additional complexity of the GAT model's graph-based features does not translate into a significant performance improvement over the XGBoost model, suggesting that the current feature set already captures the essential dynamics for accurate prediction.

The comparison between the GAT and SWISS models reveals differences in error evolution throughout the day. As shown in Figure 10, the SWISS model exhibits lower error during earlier time blocks but shows increasing error as the day progresses, suggesting that accumulating delays or growing network complexity impact its performance. In contrast, the GAT model maintains a more stable error across all time blocks, indicating consistent handling of various operational conditions. On some days, one model consistently does better throughout the day, or their performance converges, indicating that external factors (e.g., traffic patterns, network congestion) impact their effectiveness.

TABLE VI. Features for the Baseline Model

Features	**Type**
Departure time of day (min)	Numerical (time)
Minimum ground time	Numerical
Ground time before flight	Numerical
Maximum ground time	Numerical
Average ground time	Numerical
Departure month	Categorical
Number of flights without break	Numerical
Off-block to on-block time (scheduled)	Numerical (time)
Departure weekday	Categorical
City pair	Categorical
Departure airport IATA code (scheduled)	Categorical

TABLE VII. RMSE Comparison between Baseline Model and GAT Model by Day

Date	**SWISS Model** (RMSE ± SD)	**GAT Model** (RMSE ± SD)
22-10-24	17.52 ± 12.06	17.18 ± 10.50
23-10-24	18.28 ± 13.00	16.98 ± 11.63
24-10-24	18.03 ± 13.18	17.77 ± 11.94
25-10-24	18.72 ± 12.76	17.76 ± 11.95
26-10-24	27.88 ± 22.39	18.95 ± 12.65
27-10-24	22.77 ± 16.43	22.93 ± 15.01
28-10-24	20.55 ± 14.69	21.15 ± 13.34
29-10-24	18.03 ± 13.12	22.78 ± 14.08
30-10-24	19.35 ± 13.76	21.98 ± 13.72
31-10-24	16.30 ± 10.05	19.98 ± 13.17

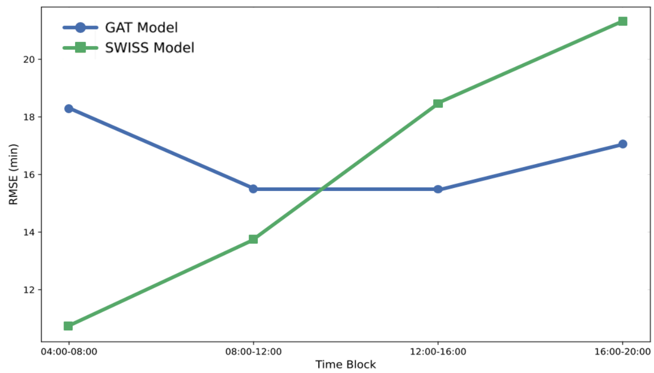

Figure 10: Comparison of SWISS (Baseline model) and GAT Model on 22-10-2024

Examining Table VII alongside the average daily delay and its variability reveals a correlation with the performance differences between the SWISS and GAT models. On days with higher average delays and greater variability, the GAT model outperforms the SWISS model significantly, likely owing to its ability to capture complex interdependencies through graph-based features. On days with lower average delays, the SWISS model performs better, suggesting that its simpler structure is more effective under stable, low-delay conditions. When delays are moderate and variability is low, both models achieve comparable performance. These findings underscore the influence of delay distribution characteristics on model effectiveness.

D. Unforeseen Events

Predictive models, such as GATs, offer potential solutions by forecasting delays based on historical data and flight interdependencies. However, irregular and unforeseen events, such as Aircraft on Ground (AOG) incidents, pose significant challenges. Over a 21-day period, 392 heavily delayed flights (delays greater than 45 minutes) were recorded, including 40 AOG cases—averaging 1.9 per day.

One such incident involved a technical failure at Zurich airport, where a grounded aircraft caused the cancellation of its rotation and delays across three downstream flights. Despite the disruption, the model underpredicted the resulting delays, suggesting its difficulty in accounting for such abrupt, pattern-breaking events. On average, each AOG incident impacts approximately 3.5 other flights, amplifying the disruption within the flight network.

Delays can be attributed to various causes beyond AOG incidents. As shown in Table VIII, these include technical failures, operational disruptions, and external factors. For example, a baggage sorting system failure combined with adverse weather at the destination led to a departure delay in Zurich, which in turn caused a rotational delay for a subsequent flight. Such events, including equipment malfunctions not addressed by routine maintenance and unscheduled crew shortages, present significant challenges to the model's predictive accuracy. External factors such as weather, strikes, and ground handling also contribute to sporadic and difficult to foresee disruptions. As these events require immediate operational responses, the GAT model's reliance on historical patterns may prove insufficient, limiting its predictive power in real-time, random disruption scenarios.

TABLE VIII. Distribution of Delay Reasons

Delay Reason	Number of Flights
AOG (Aircraft on Ground)	40
Technical Issues (Code 89)	50
Flight Operations & Crew (Code 81)	25
Immigration, Customs, Health (Code 84)	15
Weather (Codes 63f, 13)	10
Industrial Action (Code 16c)	5
Ground Handling (Codes 65, 83)	5
Air Traffic Control (Code 39b)	3
Others	14
Total	**132**

This analysis underscores the limitations of the GAT model in forecasting delays caused by unforeseen and irregular operations. In particular, AOG incidents and technical issues accounted for a notable portion of high-delay flights, yet the model consistently underpredicted the delays for these cases. Across the 40 AOG flights, the model's predicted delays were, on average, 36 minutes lower than actual delays. These factors significantly contribute to overall flight delays, yet their random, sudden nature makes them difficult to anticipate. Understanding the frequency and impact of these events is, therefore, critical for enhancing operational resilience and refining flight delay prediction models.

E. Attention Weights

The GAT model's attention weights provide insight into its internal prioritization by identifying significant connections within the flight network. Although some patterns are clear, others remain less straightforward. A recurring theme across datasets is the high attention weights assigned to 'first flights of the day'. These flights initiate the day's rotation cycle, and because early delays can propagate through subsequent connections, a phenomenon known as the snowball effect, the model inherently recognizes their operational importance.

To further analyze how the model evaluates flight importance, the "first flights of the day" are excluded, and the subsequent flights with the highest attention weights are examined. In Table IX, these flights are compared with the priority flights designated by SWISS based on connectivity, passenger load, and operational impact. This comparison measures the extent to which the model's prioritization aligns with the decisions from an operational perspective. For example, Flight 6081 receives a high attention weight, likely because of its numerous connections or its proximity to high-passenger hubs. Delays on such flights may disrupt direct connections and lead to misconnections for many passengers, emphasizing their operational importance. The attention mechanism also identifies flights operating within tightly packed rotations. For instance, Flight 9123 (Figure 11) follows a dense schedule with minimal ground time, leaving little margin for delay recovery. These flights, particularly those with short turnaround times at busy airports, are highly sensitive to even minor disruptions. The model assigns them higher attention due to their tight schedules and the potential impact on overall network performance.

Some flights, such as 24264 and 3142 (Table IX), receive high attention weights even though they do not appear immediately critical. This observation raises questions about the model's internal criteria. These flights may have historical patterns of delays or incidents that result in elevated attention weights. Alternatively, their roles within specific rotation cycles may render them more sensitive to disruptions than initially apparent. For example, a flight connecting a secondary hub or providing a special service may have a delay history that propagates disruptions, leading the model to prioritize it. Additionally, flights such as 24264 may have fewer connections yet remain significant because of their geographical locations. The model may recognize dependencies

TABLE IX. Highest Attention Weights for 25/10 (Excluding Initial Flights of the Day)

Flight	Attention	Criticality	Origin	Destination	Edges
24264	0.999896	Moderate	KXP	GVA	2
6081	0.999404	High	HFR	ZRH	17
16072	0.999269	High	DXB	ZRH	18
3142	0.999115	Moderate	MQT	GVA	6
810	0.999090	High	TRN	ZRH	22
5179	0.998873	Moderate	ZRH	FZY	2
16632	0.998140	Low	GVA	JMC	1

TABLE X. Lowest Attention Weights for 04/11 per node

Flight	Attention	Criticality	Origin	Destination	Edges
6912	0.055620	Low	ZRH	WBT	8
822	0.068699	Low	ZRH	TRN	2
19166	0.104608	Moderate	ZRH	XTR	2
3118	0.106501	Moderate	ZRH	MQT	7
2885	0.112978	Low	ZRH	MQF	3
2084	0.115062	Moderate	ZRH	LHT	4
2893	0.116590	Moderate	MQF	ZRH	15

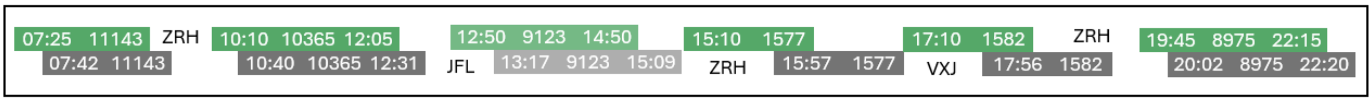

Figure 11: Rotation of critical Flight 9123 (green represents scheduled and gray actual times)

on these flights, especially if delays on these routes affect subsequent rotations involving more critical hubs.

The flights with the lowest attention weights in Table X demonstrate the GAT model's focus on connectivity over other operational features. For instance, the 2084-2091 rotation was flagged as a priority by SWISS due to Flight 2091's international group connections. However, the GAT model assigns it low importance because it does not account for the critical nature of large passenger volumes connecting to international flights at the end of the day. Flight 19166 is classified as semi-critical because its paired flight has five international connections with short layovers, necessitating a high-speed rotation. Similarly, flight 822, also semi-critical, operates within a tightly packed schedule with minimal ground-time buffers. These examples further highlight the GAT model's emphasis on the number of connections rather than on passenger metrics or schedule tightness.

In conclusion, the model's high-attention nodes reveal a notable trend: flights with high connectivity (i.e., multiple connections or groups of connecting passengers) frequently receive elevated attention scores. For instance, flights with numerous connections to major hubs (e.g., ZRH) or closely packed rotations often appear in the top ranks. This focus aligns with operational intuition, as these flights represent points in the network where disruptions could cascade, impacting large numbers of passengers and subsequent flights. Additionally, flights identified as 'Priority flights' by SWISS generally appear within the high-attention subset, indicating a strong alignment between the model's attention outputs and established criticality metrics.

VI. Discussion

The GAT model has shown notable strengths in predicting delay propagation within typical operational conditions, particularly for flights with high connectivity and consistent scheduling patterns. By effectively capturing network dependencies, the model leverages relational data to anticipate how delays may affect connected flights. The model's use of attention weights is effective in identifying flights essential for maintaining network stability, such as those that have numerous connections. This capability to discover key flights highlights the potential of the GAT model in detecting flights that could significantly impact the fleet network if delayed.

However, the GAT model encounters limitations when tasked with predicting irregular delays caused by unexpected factors, such as technical issues, crew shortages, or sudden weather changes. Such cases often lack the predictability the model depends on, revealing a gap in its ability to generalize to less predictable events. Although the model captures dependencies within the flight network, it relies heavily on historical data patterns, which limits its adaptability in real-time, unexpected scenarios.

When compared to SWISS's XGBoost model, the GAT model exhibited comparable RMSE values across test days, with both models performing well under typical operational conditions. However, the GAT model showed a distinct advantage in scenarios involving cascading delays, achieving an RMSE of 18.95 minutes compared to the SWISS model's 27.88 minutes on a particularly challenging operational day. This highlights the advantage of using a graph-based approach to capture interdependencies between connected flights more effectively than purely gradient-boosting models. Conversely, the SWISS model outperformed the GAT model on days with lower average delays, achieving an RMSE of 16.30 minutes compared to the GAT model's 19.98 minutes on an operationally stable day. This suggests that while the SWISS model remains efficient for typical day-to-day operations, the GAT model performs better in predicting networked delays that propagate across multiple connections.

Additionally, the GAT model's performance aligns with and surpasses other models from existing research used in delay prediction for long-term prediction (before day of operations begins). For instance, Random Forest models applied to Colombia's airport network achieved an RMSE of 33.8 minutes [11], and MSTAGCN models on the U.S. airport network yielded 30.7 minutes [14]. By comparison, the GAT model's RMSE of 15.59 minutes on medium-delay days underscores its ability to adapt to the complex dynamics of European airline networks. However, the model does not outperform specialized sequential prediction models, such as LSTMs, which achieve RMSE values between 6.31 and 7.73 minutes for short look-ahead predictions [12]. However, these models lack the GAT model's comprehensive understanding of the network structure of flights and, hence, are not appropriate for modeling delay propagation across interconnected flights.

The attention mechanism embedded in the GAT model has proven insightful for identifying flights with high operational impact. The model often assigns higher attention weights to flights that influence network stability, such as those initiating daily operations or involving significant passenger connections. This alignment with operational priorities underscores the relevance of the model's attention outputs in highlighting flights where delays may propagate widely through the network. Moreover, Zurich and Geneva emerge as critical nodes within the network, reinforcing the central role of these hubs in maintaining operational efficiency. Outliers, meaning non-critical flights receiving unexpectedly high attention weights, suggest that the model might identify unquantified operational risks or factors not currently captured by the priority flights.

Lastly, the model's performance is influenced by the quality and scope of available data, including the absence of real-time operational data, such as maintenance records and updated crew rotations. The lack of this data may restrict the model's ability to adjust dynamically to evolving operational conditions. Consequently, this limitation may partly explain discrepancies between predicted and actual delays in certain flights or aircraft types with atypical schedules. Expanding the model's data inputs to incorporate weather forecasts and maintenance data could enhance its predictive accuracy and allow for more adaptive responses in future applications.

A. Future Work

Building on the GAT model's strong performance, a key area for improvement is integrating real-time operational data, such as maintenance checks, crew rotations, and detailed weather forecasts. This would allow the model to update dynamically rather than relying solely on historical patterns, reducing prediction gaps in rapidly changing conditions. Adding airport-specific features, such as runway capacity, ground congestion, and metrics like transfer passenger volumes, would further refine predictions by capturing each airport's unique operational context. Incorporating maintenance details, like fleet age and the frequency of technical checks, could also enhance reliability, especially for older aircraft that are more prone to delays.

Moreover, a deeper examination of propagation dynamics, specifically identifying which connections (e.g., international vs. domestic) are most prone to cascading delays, could offer targeted strategies to mitigate knock-on effects. Future work might explore global attention mechanisms or hybrid architectures (e.g., combining GAT with LSTM) to account for these extended dependencies. By addressing these areas, the GAT model can become an even more robust tool for strengthening resilience and efficiency across a fleet network.

VII. Conclusion

Reactionary delays currently account for approximately 45% of total delays. This research demonstrates the potential of a GAT model to predict reactionary delay distributions within a hub-spoke airline network, with particular attention to the role of spoke airports. By leveraging both node-level features (representing individual flights) and edge-level attributes (capturing dependencies among flights), the model effectively identifies how delays propagate through critical connections, including rotational links, high-volume passenger transfers, and key spoke-hub routes.

Performance metrics underline the model's accuracy, especially on moderate-delay days (RMSE of 15.59 minutes, MAE of 10.50 minutes), and suggest a sensitivity to more extreme disruptions (RMSE around 37.56 minutes on high-delay days). The difference between RMSE and MAE indicates that, while most predictions are reasonably accurate, outliers significantly inflate the RMSE. The GAT's attention weights provide deeper insight into which nodes (flights) and edges (connections) are most influential, thereby enabling more precise operational interventions.

Despite its strong performance, the GAT model sometimes struggles with irregular, extreme disruptions, emphasizing the need for richer inputs such as weather forecasts, maintenance schedules, and crew availability data. Additionally, while the GAT excels at modeling localized delay propagation, it may overlook longer-range effects in highly interconnected networks.

Overall, this research showcases the GAT model's potential to enhance delay forecasting and operational decision-making, laying a foundation for more resilient airline network management. For future work, including features on airport capacity, ground congestion, or runway availability has the potential to increase performance. Additionally, expanding the training dataset beyond one year would capture greater seasonal and operational variability, reducing the model's vulnerability to outliers or rare events.

Acknowledgment

We would like to thank David Graber for his invaluable insights and technical guidance on the development of the Graph Attention Network model used in this work.

References

[1] EUROCONTROL, "CODA Digest: Delays to Air Transport in Europe—Annual 2013," 2014.
[2] EUROCONTROL, "All-Causes Delays to Air Transport in Europe - Quarter 2 2023" 2023.
[3] H. Fricke and M. Schultz, "Delay Impacts onto Turnaround Performance Optimal Time Buffering for Minimizing Delay Propagation," in 8th USA/Europe Air Traffic Management Research and Development Seminar, 2009.
[4] V. M. Eguíluz, E. Hernández-García, F. Plo and J. J. Ramasco, "Data-driven modelling of the Tree of Reactionary Delays," in 6th International Conference on Research in Air Transportation, 2014.
[5] R. Dalmau, A. Trzmiel and S. Kirby, "PETA: Combining Machine Learning Models to Improve Estimated Time of Arrival Predictions," in SESAR Innovation Days, 2023.
[6] R. Beatty, R. Hsu, L. Berry and J. Rome, "Preliminary Evaluation of Flight Delay Propagation through an Airline Schedule," Air Traffic Control Quarterly, vol. 7, 1999.
[7] C. Wu and K. Law, "Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model," Transportation Research, n. 122, p. 62-77, 2019.
[8] C. Wu and K. Law, "Flight Delay Propagation Impact on Strategic Air Traffic Flow Management," Transportation Research Record, n. 2177, p. 105-113, 2010.
[9] W. Du, M. Zhang, Y. Zhang, X. Cao and J. Zhang, "Delay causality network in air transport systems," Transportation Research Part E, vol. 118, p. 466-476, 2018.
[10] Z. Jia and X. Cai, "Delay propagation network in air transport systems based on refined nonlinear Granger causality," Transportmetrica B, vol. 10, n. 4, p. 1-13, 2022.
[11] JG. Muros Anguita and O. Diaz Olariaga, "Prediction of departure flight delays through the use of predictive tools based on machine learning/deep learning algorithms," The Aeronautical Journal, n. 128, p. 111-133, 2024.
[12] J. Sun, T.L.E. Dijkstra, K. Aristodemou, V.S. Buzetelu, T. Falat, T.G. Hogenelst, N. Prins and B.C. Slijper, "Designing Recurrent and Graph Neural Networks to Predict Airport and Air Traffic Network Delays," in 10th International Conference for Research in Air Transportation FAA & Eurocontrol, 2022.
[13] K. Cai, Y. Li, Y. Fang and Y. Zhu, "A Deep Learning Approach for Flight Delay Prediction through Time-Evolving Graphs," IEEE Transactions on Intelligent Transportation Systems, 2021.
[14] H. Zheng, Z. Wang, C. Zheng, Y. Wang, X. Fan, W. Cong and M. Hu, "A graph multi-attention network for predicting airport delays," Transportation Research Part E, 2024.
[15] R. Ying, D. Bourgeois, J. You, M. Zitnik and J. Leskovec, "GNNExplainer: Generating Explanations for Graph Neural Networks," in 33rd Conference on Neural Information Processing Systems, 2019.