Research in Transportation Business & Management 56 (2024) 101161

Contents lists available at ScienceDirect

Research in Transportation Business & Management

journal homepage: www.elsevier.com/locate/rtbm

Integrating multiple data sources for improved flight delay prediction using explainable machine learning

Juan Pineda-Jaramillo [a], Claudia Munoz [b,c,*], Rodrigo Mesa-Arango [d], Carlos Gonzalez-Calderon [e], Anne Lange [f]

[a] Universidad Politécnica de Valencia, Spain
[b] Universidad de Antioquia, Colombia
[c] Universidad Nacional de Colombia, Colombia
[d] Florida Institute of Technology, USA
[e] Universidad Nacional de Colombia, Colombia
[f] Frankfurt University of Applied Sciences, Germany

ARTICLE INFO

Keywords:
Flight delays
Explainable machine learning
Shap method
Sobol method

ABSTRACT

Flight delays negatively impact costs, customer satisfaction, and revenue in the aviation industry. As a result, it is critical to identify the factors that cause flight delays for each airport, as they can vary depending on various attributes associated with their operations.

This study proposes an explainable artificial intelligence (xAI) methodology for identifying the features that affect airport delays by integrating data from multiple sources and implementing explainable artificial intelligence. The methodology incorporates operational data, airport information, geographic data, and weather data combined and used to train a series of machine learning models. Furthermore, the SHAP and Sobol techniques are used to thoroughly analyze the features that influence flight delays for the specific case of the airport in Santiago, Chile.

The results show that a linear discriminant analysis model is best suited for predicting flight delays in this specific case study, and the features that have the most significant impact on delays are the international flight status, average temperature at the destination airport, wind speed, and average temperature at Santiago airport.

The proposed methodology could be applied by airlines that can collect data from multiple sources and conduct similar investigations, leading to the development of a decision support system to make better-informed decisions and reduce the impact of flight delays.

1. Introduction

Air transportation has become an indispensable part of modern life, as people travel for various reasons, including business and leisure. Globalization, economic growth, and advancements in air travel technology have all contributed to the increased air-transportation utilization. As global trade and multinational companies expand, air-transportation utilization grows due to the agility and efficiency of this mode (IATA, 2019). Air travel has become a critical factor in global trade and commerce, enabling individuals and goods to be transported to different parts of the world in just a few hours.

Beyond merely facilitating movement, the evolution of air travel technology has been pivotal in broadening the scope and capacity of air transportation. This evolution includes not only the introduction of aircraft with higher fuel efficiency, but also the incorporation of advanced technologies aimed at streamlining operations and improving the passenger experience. Next-generation air traffic management systems, digital booking and boarding processes, and improvements in aircraft design and materials have all contributed to the industry's growth. These technological advancements have not only improved the efficiency and accessibility of air travel, but have also significantly increased its convenience and safety, making it more appealing to a broader audience (FAA, 2021).

The aviation industry is critical to global connectivity for individuals

* Corresponding author at: Calle 65 No. 78 – 28 M1, Medellín, Colombia.
E-mail addresses: juapija1@upv.es (J. Pineda-Jaramillo), claudia.munoz1@udea.edu.co, chmunozh@unal.edu.co (C. Munoz), rmesaarango@fit.edu (R. Mesa-Arango), cagonza0@unal.edu.co (C. Gonzalez-Calderon), anne.lange@fb3.fra-uas.de (A. Lange).

https://doi.org/10.1016/j.rtbm.2024.101161
Received 26 September 2023; Received in revised form 4 June 2024; Accepted 17 June 2024
Available online 22 June 2024

and businesses, enabling worldwide travel and commerce. Despite significant advancements in air travel technology, flight delays continue to be a challenge for airlines due to weather and operational issues, resulting in inconvenience, time loss, and increased expenses for airlines and passengers (Gui et al., 2020; Stone, 2018). Since flight delays have far-reaching consequences for airports and airspace, it is critical to address such issues thoroughly. In addition, flight delays upset passengers, which causes adverse economic effects to themselves, as well as airlines and airports (Britto, Dresner, & Voltes, 2012; Oza, Sharma, Sangoi, Raut, & Kotak, 2015). Accurately predicting flight delays is a critical step that airlines can take to improve operational efficiency and provide better service to their passengers (Bubalo & Gaggero, 2021; Degas et al., 2022).

In our study, a 'flight delay' is defined as the difference between scheduled and actual flight departures. Multiple operational, weather-related and space weather factors have been identified as potential contributors of flight delays. Operational factors such as airport congestion, aircraft turnaround time, crew scheduling and availability, air traffic control issues, and maintenance issues can all cause delays (See, Ülkü, Forsyth, & Niemeier, 2023). According to the Federal Aviation Administration (FAA), weather is a significant cause of flight delays, with elements such as thunderstorms, snow, and high winds leading to disruptions (FAA, 2021; Khaksar & Sheikholeslami, 2019). Visibility, cloud cover, and extreme temperatures further impact flights (Allan, Gaddy, & Evans, 2001; Algarin Ballesteros & Hitchens, 2018; Borsky & Unterberger, 2019). Moreover, space weather events, particularly solar storms and solar maximums, present additional challenges. These events can cause HF communication blackouts, failures in GNSS-based navigation and surveillance systems, and increased cosmic radiation, posing a significant risk to aviation safety and operations (David et al., 2023; Xue, Yang, Liu, & Cong, 2023; Zinke, 2023). Airline-specific issues such as overbooking, gate availability, and scheduling practices can all contribute to delays. Understanding the intricate interplay of these various factors is critical for accurately forecasting and mitigating flight delays. Precise flight delay prediction is not only a profitable practice for the aviation industry but also enhances passenger travel experiences (Gholami & Khashe, 2022).

According to Chakrabarty (2019), the increasing availability of data has prompted the use of artificial intelligence (AI) and machine learning (ML) to predict flight delays in recent years. However, traditional ML algorithms can be challenging to interpret, making it difficult to identify the underlying causes of flight delays. A promising alternative is explainable AI (xAI), i.e., a subfield of AI that prioritizes transparency and interpretability in decision-making processes (Ribeiro, Singh, & Guestrin, 2016). With xAI algorithms it is possible to explain the factors contributing to flight delays, allowing airlines to make better-informed decisions by learning more about the causes of such delays. Furthermore, airlines can improve their accuracy in predicting flight delays by integrating data from multiple sources into xAI algorithms (Kaewunruen, Sresakoolchai, & Xiang, 2021).

The proposed study aims to fill a gap in the literature by offering a methodology to identify the factors contributing to flight delays at a specific airport using xAI algorithms. The process considers flight-related information such as the origin and destination airports, scheduled departure and arrival times, and weather conditions at both the origin and destination airports based on the flight schedule. By incorporating these data into the flight delay prediction model, airlines can achieve more accurate and reliable predictions, allowing them to make informed decisions and take proactive measures to minimize the impact of flight delays on passengers and business operations. Furthermore, by constantly adding more data from various sources, the xAI algorithms can improve their predictive performance and adapt to changing circumstances, making them even more effective at predicting flight delays and providing their passengers with the best travel experience.

The structure of this paper has been organized to facilitate a comprehensive understanding of our research.: Section 2 reviews the relevant literature, establishing the foundation upon which our study is built. Section 3 details the datasets and the methodology behind constructing an xAI-based model for predicting airline flight delays with an emphasis on weather features. Section 4 presents the findings from the models employed to predict commercial flight delays. Following this, Section 5 delves into the implications of our findings, discussing their relevance and impact on the aviation industry and related fields. Lastly, Section 6 concludes the paper, summarizing the principal outcomes and proposing directions for future research in this area.

2. Literature review

2.1. Methodologies for predicting flight delays

Flight delays constitute a significant issue in the aviation industry, affecting both passengers and airlines. The use of data and analytics to predict flight delays has received much attention in recent years. This practice has demonstrated reductions in the impacts of flight delays and improvements in air travel efficiency (Malighetti, Morlotti, Redondi, & Paleari, 2023).

There are two methods for modeling and predicting flight delays: statistical models and machine learning models. Traditional statistical methods, such as regression analyses and time series analyses, are used in statistical models to predict flight delays. These models determine the relationship between various factors and the likelihood of a delay. For example, Mokhtarimousavi and Mehrabi (2023) investigated the relationship between weather conditions, flight distance, and aircraft type in predicting flight delays using regression analysis.

ML models, on the other hand, involve training algorithms on historical data to identify patterns and expect future flight delays. These algorithms can consider various factors contributing to delays, such as weather conditions, aircraft performance, and flight routes. However, the sample size must be large enough to apply ML to flight delay analysis (Gholami & Khashe, 2022).

ML techniques such as artificial neural networks, decision trees, random forests, and support vector machines have previously been used successfully for flight delay prediction (Carvalho et al., 2021; Singh, Jayaprakash, & Agarwal, 2022). For example, Gui et al. (2020) investigated operational and weather-induced airline delays using various ML techniques such as decision trees, random forests, and K nearest neighbors. Khaksar and Sheikholeslami (2019) used Bayesian modeling, decision trees, cluster classification, random forests, and hybrid methods to predict delays for US flights, revealing that visibility, wind, and departure time were the predictable parameters of delay in the airline network. Using regression models and neural networks, Sridhar, Wang, Klein, and Jehlen (2011) predicted weather-related flight delays and cancellations at the national, regional, and airport levels.

Often, these studies employ black-box ML models that fail to explain the factors contributing to flight delays. To overcome that limitation, the research presented in this paper proposes a method for predicting flight delays using xAI incorporating data from multiple sources. The proposed methodology can identify the features that impact flight delays and explain the contributing factors by integrating air operations data, airport information, geographical information, and weather data with procedures that extract valuable insights from the models.

2.2. Methodologies to measure feature impact

In prediction tasks, features in a prediction model frequently collaborate, indicating the presence of feature interaction. There have been several methods proposed for measuring feature interaction. Initially, statistical methods were presented, with Hastie and Tibshirani (2017) offering the analysis of variance (ANOVA) test to measure feature interaction, with the corresponding p-value for each pair of features computed following the ANOVA test to measure feature interaction. This method, however, involves a significant amount of

computation time. The χ^2 test was used by Loh (2002) and Lou, Caruana, Gehrke, and Hooker (2013) to investigate pairwise feature interactions. Models based on decision trees, such as random forests, use a tree structure to measure feature interaction, in which two features, f1, and f2, are said to interact if they are on the same path of a decision tree (Deng, 2019; Sorokina, Caruana, Riedewald, & Fink, 2008; Wright, Ziegler, & König, 2016). Recently, partial dependency-based methods such as Friedman and Popescu's H-statistics method have been proposed.

Another popular method to measure feature interaction is the partial dependency plot, which depicts the minor impact of one or two features on the predicted outcome of the machine-learning model (Friedman, 2001; Zhao & Hastie, 2021). A partial dependence plot can reveal whether the target-feature relationship is linear, monotonic, or complex.

Another critical methodology for feature interaction is the SHapley Additive exPlanation (SHAP) method. The SHAP method is used to determine the importance of each feature and its direct impact on the model result, allowing for accurate interpretation of the best model's output and identification of the elements with the most significant impact on predicting commercial flight delays (Xu, Wang, Schmidt, Adams, & Hatzopoulou, 2020).

The SHAP method computes the impact of each input feature (ϕ_i for feature i) on the model output $v(N)$ as shown in Eq. (1), where a linear function of binary features is defined based on the additive feature attribution shown in Eq. (2), where $z' \in \{0,1\}^M$ is 0 if a feature is not observed, and is 1 otherwise, and M defines the number of input features (Lundberg & Lee, 2017).

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S\bigcup\{i\}) - v(S)] \tag{1}$$

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2}$$

2.3. Variance-based sensitivity analysis

Sensitivity analysis is essential in machine learning because it quantifies the contribution of input features to model output. The goal of variance-based sensitivity analysis is to determine how much of the uncertainty in the model output can be attributed to the fate of the input feature. It is a powerful tool for testing model comprehension and identifying non-influential parameters.

Several techniques have been proposed to improve the computational efficiency of variance-based sensitivity analysis (Plischke, Borgonovo, & Smith, 2013). The Morris method is a low-cost method for calculating sensitivity indices requiring several models runs. It works well with models with many input features (Morris, 1991). Another popular method for calculating sensitivity indices is the Fourier amplitude sensitivity test (FAST), which uses a Fourier series to represent the input parameters and is especially useful for models with high-dimensional input spaces.

Delta Moment-Independent Measure (Delta-MIM) is a variance-based sensitivity method based on the Delta test that can efficiently compute sensitivity indices for models with many input features. The Monte Carlo Filter (MCF) is a recently proposed method that uses a filtering approach to estimate the sensitivity indices and is particularly useful for computationally expensive models.

Sobol's variance-based sensitivity method is one of the most popular approaches for measuring the model output variance attributed to each input feature or interaction. This method systematically examines the contribution of individual input features to the output variance. It divides the total variance into components representing the individual and joint contributions of the input variables. This method requires many model evaluations, which can be computationally costly.

On the other hand, Sobol's variance-based sensitivity method assesses how much of the model output uncertainty can be attributed to uncertainty in the input features used to train the model (Saltelli et al., 2010). This method investigates the best ML model's performance across various scenarios. It uses a factor-fixing approach to determine which input feature can be fixed (that is, assigned any value within its range) without affecting the model output. Sobol's variance-based sensitivity method estimates sensitivity indices independently (that is, potential non-additivity in the model does not affect its sensitivity index) and allows for quantification of interaction effects on model behavior (Sobol, 1993); Mesa-Arango et al., 2023. Ou, Liu, Dong, and Wang (2014) used genetic algorithms where they extended and applied the SOBOL method to analyze the effect of uncertainties on the stability of the flight control system. The authors used the technique to create a new clearance framework due to its suitability for systems with strong nonlinearity, input factors varying in large intervals, and input factors subjecting to random distributions.

Following the construction of a model, it is applied to a random sample to define the ranges of values for each input feature. Then, the Sobol method calculates the total order indexes for all input features in the model. This process assists in determining how much each input feature contributes to the model's output uncertainty. It thus aids in identifying the critical input features that require more attention during model training.

The total order index of feature i measures the contribution to the output variance caused by a model input, including the contribution to the output variance caused by the single input feature i alone, as well as the contribution to the variance of the input generated by the interaction of the input feature i with the other input features. The formulation of the total order index is shown in Eq. (3), where $V[E(Y|X_{\sim i})]$ is the conditional variance of the expected value of Y after fixing for all features except X_i, averaged over all X_i values, and $V(Y)$ is the unconditional variance of the model's output. Previous studies (Saltelli et al., 2010, 2007) can help readers understand the critical equations used in the Sobol method.

$$S_{Ti} = 1 - \frac{V[E(Y|X_{\sim i})]}{V(Y)} \tag{3}$$

In predicting flight delays, variance-based sensitivity analysis can provide insights into the importance of various features in predicting flight delays, such as departure time, weather conditions, airline carrier, and airport location. Understanding the relative importance of these features allows airlines and airports to make data-driven decisions to improve operations and reduce delays.

Table 1 presents a review of the literature on the use of machine learning (ML) approaches (models such as random forest, support vector regression, and neural networks, among others) for predicting flight delays, feature impact analysis (including partial dependence plots, SHAP values, and so on), and variance-based sensitivity analysis (such as Sobol's method, Morris's method, and the Delta Moment-Independent Measure (Delta-MIM). Overall, Table 1 provides a thorough overview of the methods used to predict flight delays and how they have been applied in the literature.

2.4. Contributions to previous literature

This study contributes to the aviation field by proposing a methodology for predicting flight delays using xAI data from multiple sources, which has not been approached yet. By integrating air operations data, airport information, geographical information, and weather data, the proposed methodology can identify the features that impact flight delays and provide clear explanations of their contributing factors. This method differs from traditional methods, which typically rely solely on operational data and can produce inaccurate predictions. The proposed methodology uses data from multiple sources to improve the predictive performance of xAI algorithms. In addition, using xAI models ensures that predictions are transparent and interpretable. Airlines must

Table 1
Summary of data-driven methods for flight delay prediction.

Study	Objective	Main Contribution	Data source	Approach		AFI	VBSA
				SM	ML		
Dai (2024)	Propose a clustering-based model for decomposing the flight delay prediction	A combination of ANOVA and the Forward Sequential Feature Selection (FSFS) algorithm is used to determine the most influential indicators on flight delays	Arriving flights at JFK International Airport		X		
Abdelghany, Guzhva, and Abdelghany (2023)	Predict flight block time and evaluate airlines' block time padding strategies.	Development of machine learning models for predicting flight block time and evaluating airlines' block time padding strategies can improve airline scheduling and reduce flight delays.	US domestic flights		X	FI	
Qu, Xiao, Yang, and Xie (2023)	Propose a flight delay prediction method based on Att-Conv-LSTM	The model can predict the specific delay time using spatio-temporal neural network and considering meteorological information	Four China domestic airports		X		
Wang, Mao, Li, Li, and Tu (2023)	Prediction of estimated time of arrival for multi-airport systems via "Bubble" mechanism.	Introduces a novel "Bubble" mechanism for accurate medium-term ETA prediction in MAS, employing ARIMA for out-MAS travel time and a novel MSB-LSTM model for in-MAS travel time prediction based on new spatio-temporal features.	MAS in China, Guangdong–Hong Kong–Macao Greater Bay Area		X		
Li, Guan, and Liu (2023)	A CNN-LSTM framework for flight delay prediction.	Proposes a CNN-LSTM framework to address spatial-temporal correlations and extrinsic factors in flight delay prediction, showcasing significant accuracy improvements over benchmark models. Highlights the framework's utility for airport regulators in enhancing on-time performance through advanced management strategies.	U.S. domestic flights		X		
Ayaydin and Akcayol (2022)	Reducing the adverse effects of cancellation, delay and diversion events within the aviation ecosystem	Use ML for preventing financial and moral losses that may occur as a result of delays in flights and to take necessary precautions by predicting the flight delay in advance	Turkey domestic flights		X		
Li and Jing (2022)	Predict flight delays from spatial and temporal perspectives	A real-time monitoring and high-accuracy prediction system to alleviate flight delays	China domestic flights		X		
Gholami and Khashe (2022)	To develop a flight delay prediction system by analyzing data from domestic flights inside the US	The proposed models take different inputs from the user through Alexa and then predict flight delays	US domestic flights		X		
Guimaraes, Soares, and Ventura (2022)	Predict if passengers in a connecting flight will lose their connection	Predict missed flight connections in an airline's hub airport using historical data on flights and passengers, and analyze the factors that contribute additively to the predicted outcome for each decision horizon	Portugal airport flight		X	SHAP	
Dalmau, Ballerini, Naessens, Belkoura and Wangnick (2021)	Predict the take-off time of individual flights.	The model allows for improved take-off time predictions as early as the initial flight plan is received.	EUROCONTROL Maastricht Upper Area Control Centre		X	SHAP FI	
Zhu and Li (2021)	To develop a novel spatial weighted recurrent neural network model to provide better flight time predictions	With the improved flight time prediction, fuel loading can be optimized, resulting in reduced fuel consumption.	METAR data and airline records from Hong Kong	X	X		
Yi, Zhang, Liu, Zhong, and Li (2021)	To prove that the Stacking algorithm has advantages in airport flight delay prediction	A flight delay prediction classification method based on the Stacking algorithm at Boston Logan International Airport	Boston Logan International Airport flight		X		
Tang (2021)	Prediction of the occurrence of flight delays	Performed a prediction of the occurrence of flight delays by adapting it into an ML problem. Due to the imbalanced nature of the data set, evaluation measures were weighted to eliminate the dominant effect of non--delayed flights over delayed flights	John F. Kennedy International Airport departure flights	X	X		
Alharbi and Prince (2020)	Predict the flight delay more precisely by applying both the tools and deep learning.	Predict delay using a hybrid approach of the Saudi Airlines' flights	Saudi Airlines flights		X		
Zhang and Ma (2020)	Establish a flight delay prediction model based on an ML algorithm to predict the departure delay at an airport considering both flight information and weather conditions at the airport	The results show that the scheduled departure time is the most essential feature and is positively related to the departure delay.	Newark Liberty International Airport departure flights		X	SHAP	
Lambelho, Mitici, Pickup, and Marsden (2020)	Predict arrival/departure flight delays and cancellations	A generic assessment of strategic flight schedules, using Key Performance Indicators derived from predictions on flight delays and cancellations	Flight schedules in the period 2013–2018 at London Heathrow Airport		X	SHAP	

(*continued on next page*)

Table 1 (*continued*)

Study	Objective	Main Contribution	Data source	Approach		AFI	VBSA
				SM	ML		
Khaksar and Sheikholeslami (2019)	To identify parameters that enabled practical estimation of delays in US and Iranian flights	The hybrid approach exhibited a performance superior to those of the other methods. Parameters such as fleet age and aircraft type exerts potent effects on flight delays in the Iranian network, whereas weather conditions strongly influence flight delays in the US network	US and Iranian flight dataset		X		
Choi, Kim, Briceno, and Mavris (2016)	Predict airline delays caused by inclement weather conditions using data mining and supervised ML algorithms	The proposed prediction model enabled to classify airline delays caused by rough weather conditions using historical weather and traffic data of individual OD pairs by utilizing ML algorithms.	US domestic flight data		X		
Belcastro, Marozzo, Talia, and Trunfio (2017)	To implement a predictor of the arrival delay of a scheduled flight due to weather conditions	The model predicted arrival delays considering both flight information (origin airport, destination airport, scheduled departure and arrival time) and weather conditions at the origin airport and destination airport according to the flight timetable. It identifies a handy pattern of flight delay that may help airlines in reducing delays.	US flights by major air carriers	X	X		
Ou et al. (2014)	To validate the robust stability of the flight control system of a hypersonic flight vehicle	Proposed a new clearance framework for robust stability clearance of the control system of hypersonic flight vehicle based on theory and global uncertainty SA (GUSA)	–				SOBOL
Our study	Flight delay predictor (binary)	By implementing this weather-based predictor, airlines, airports, and flight-booking websites can improve their overall efficiency and customer satisfaction by providing more accurate flight information and reducing the impact of flight delays due to weather conditions.	Flight operations from Santiago de Chile International Airport		X	SHAP	SOBOL

* Analysis of feature impact (AFI), Variance-based sensitivity analysis (VBSA), Statistics methods (SM), Machine Learning (ML), Features Interaction (FI), SHapley Additive exPlanation (SHAP).

understand the root causes of flight delays and take proactive measures to mitigate their impact. Overall, the findings of this study contribute to a better understanding of the factors that cause flight delays, which can assist airlines in making better decisions and improving the passenger experience.

3. Dataset and models

This section outlines the procedure for creating a predictive model for airline flight delays using xAI. Fig. 1 illustrates the steps taken in the methodology, described in detail below. All the procedures outlined in this section were conducted using Python 3.8, with computational resources sufficient to support the processing and analysis demands of the study.

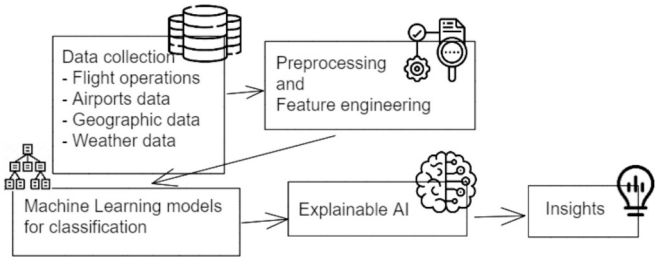

Fig. 1. Flow diagram of the methodology followed in this study.

3.1. Data collection

To develop a predictive model for airline flight delays using an xAI approach, our study utilized a comprehensive dataset encompassing flight operations from Santiago de Chile International Airport (SCL) to 63 domestic and international destinations throughout the year 2017 (January 1 to December 31). This comprehensive dataset integrates flight operation records, airport information, and geographical data, offering a holistic view of the factors influencing flight delays. The weather data, an essential component of our analysis, was sourced from the National Oceanic and Atmospheric Administration (NOAA), a scientific agency within the US Department of Commerce, employing the Meteostat Python package (Meteostat, 2022) to incorporate hourly surface data pertinent to the airports and times corresponding to the departures and arrivals of flights from Santiago de Chile. Table 2 outlines the detailed composition of our dataset, while section 3.2 delves into the data processing and feature engineering approaches employed. The dataset was provided through collaboration with a major airline operating within the region, ensuring a rich and detailed basis for our analysis.

In the analysis of flight delays at Santiago de Chile International Airport (SCL) for the year 2017, our focus is on understanding the operational and external factors contributing to delays. To provide a more targeted analysis, we have detailed the distribution of destinations by continent in our dataset as follows: South America accounts for 88.3% of destinations, North America 8.4%, Europe 2.5%, and Oceania 0.7%. This distribution highlights the geographic scope of SCL's flight operations and its relevance to our study on flight delays. For additional statistics, including the frequency of flights and other relevant operational data, readers are directed to Table 3 in Section 3.2.

Table 2
Characteristics of the data sources.

Feature	Description	Data source
date_i	Scheduled date and time of the flight	Flight operations
flight_i	Scheduled flight number	Flight operations
ori_i	Scheduled origin city code	Flight operations
des_i	Scheduled destination city code	Flight operations
comp_i	Scheduled flight airline code	Flight operations
date_o	Date and time of flight operation	Flight operations
flight_o	Operated flight number	Flight operations
ori_o	Code of the city of origin of the operation (Santiago de Chile)	Flight operations
des_o	Code of the city of destination of the operation	Flight operations
comp_o	Airline code of operated flight	Flight operations
day	Day of the month of flight operation	Flight operations
month	Number of months of flight operation	Flight operations
day_name	Day of the week of flight operation	Flight operations
flight_type	Type of flight: national or international	Flight operations
company_name	Name of the airline that operates	Flight operations
city_ori	Origin city name	Flight operations
city_dest	Destination city name	Flight operations
airport_type	Destination airport type	Airport information
latitude_dest	Latitude at the destination airport	Geographical data
longitude_dest	Longitude at the destination airport	Geographical data
elevation_dest	Elevation at destination airport [m]	Geographical data
continent_dest	Continent at the destination airport	Airport information
country_dest	Country at the destination airport	Airport information
tavg_dest	Average temperature at destination airport at landing [C]	Weather information
prcp_dest	Precipitation at destination airport at landing [mm]	Weather information
wspd_dest	Wind speed at destination airport at landing [km/h]	Weather information
pres_dest	Air pressure at destination airport at landing [hPa]	Weather information
tavg_santiago	Average temperature at Santiago airport at takeoff [C]	Weather information
prcp_santiago	Precipitation at Santiago airport at takeoff [mm]	Weather information
wspd_santiago	Wind speed at Santiago airport at takeoff [km/h]	Weather information
pres_santiago	Air pressure at Santiago airport at takeoff [hPa]	Weather information

Table 3
Descriptive statistics of the explanatory features in the final dataset.

Feature id	Number of Categories	Feature distribution and statistics
diff_15 (target feature)	2	Yes: 11.7%, No: 88.3%
day_name	7	Thursday: 15.0%, Friday: 14.8%, Monday: 14.7%, Sunday: 14.5%, Wednesday: 14.3%, Tuesday: 14.3%, Saturday: 12.4%
flight_type	2	National: 55.2%, International: 44.8%
company_name	23	Grupo LATAM: 60.6%, Sky Airline: 21.0%, Copa Air: 2.9%, Others: 15.5%
city_dest	60	Buenos Aires: 9.0%, Antofagasta: 8.7%, Calama: 7.8%, Lima: 7.5%, Others: 67.0%
period_day	3	morning: 37.8%, afternoon: 37.2%, night: 25.0%
airport_type_dest	3	medium_airport: 64.6%, large_airport: 35.1%, small_airport: 0.4%
continent_dest	4	South America: 88.3%, North America: 8.4%, Europe: 2.5%, Oceania: 0.7%
month	12	January: 9.1%, December: 9.0%, November: 8.9%, October: 8.6%, August: 8.5%, February: 8.4%, March: 8.3%, September: 8.3%, July: 8.1%, May: 7.8%, April: 7.6%, June: 7.3%
day	–	mean: 15.7277, median: 16.0, min: 1.0, max: 31.0
high_season	–	mean: 0.3324, median: 0.0, min: 0.0, max: 1.0
elevation_dest	–	mean: 399.9424, median: 89.61, min: 2.44, max: 4070.6
diff_flight	–	mean: 0.4963, median: 0.0, min: 0.0, max: 1.0
diff_company	–	mean: 0.2736, median: 0.0, min: 0.0, max: 1.0
distance	–	mean: 2157.5520, median: 1225.4, min: 195.82, max: 11890.46
tavg_dest	–	mean: 17.1891, median: 17.0, min: −10.1, max: 32.8
wspd_dest	–	mean: 12.6466, median: 11.6, min: 0.0, max: 56.2
tavg_santiago	–	mean: 15.7462, median: 15.8, min: 3.5, max: 27.5
prcp_santiago	–	mean: 0.4470, median: 0.0, min: 0.0, max: 20.8
wspd_santiago	–	mean: 10.4202, median: 10.5, min: 3.9, max: 17.8
pres_santiago	–	mean: 1016.1829, median: 1015.8, min: 1009.9, max: 1028.5

3.2. Data pre-processing and feature engineering

In this phase, data was combined to ensure that each row accurately represents a distinct flight operation from Santiago de Chile airport to various destinations. This process included integrating flight characteristics, destination airport details, and weather conditions experienced during both takeoff in Santiago and landing at the arrival destination. The final dataset, following data imputation to address missing values, facilitated the development of features crucial for our analysis. These features, each capturing different aspects of flight operations, are defined as follows:

- Diff_15 (Target feature): This feature indicates quantifies flight delays by assessing if a flight's actual landing time exceeds the scheduled landing time by more than 15 min. Specifically, we define 'arrival delay' based on the moment the aircraft touches down on the runway, rather than the in-gate time or the conclusion of taxi-in. This approach aligns with aviation research standards where a 15-min threshold is widely recognized for classifying a flight as delayed. Such a metric is endorsed by regulatory bodies, including the Federal Aviation Administration (FAA), and is corroborated by studies from Guleria, Cai, Alam, and Li (2019); Belcastro et al. (2017); Prince and Simon (2015) as a significant benchmark for delay analysis.
- High_season: A binary indicator reflecting whether the flight occurred during peak travel periods, as defined by specific date ranges throughout the year: between December 15 and March 3, July 15 and July 31, or September 11 and September 30.
- Period_day: Categorizes the departure time of day into morning, afternoon, or night, based on the local time of departure: morning (5 h–12 h), afternoon (12 h–19 h), or night (19 h–5 h).
- Diff_flight, Diff_company: Dummy variables indicating changes in scheduled flight number, and operating airline, respectively.
- Distance: The geographical distance between Santiago de Chile Airport and the destination, calculated using the Haversine formula.

Each developed feature is derived at the individual flight operation level, rather than aggregated over daily, monthly, or annual periods, to preserve the granularity needed for precise delay prediction and analysis.

After feature engineering, the dataset underwent a series of preprocessing steps to ensure the integrity and usability of the data for model development. Outliers were identified and removed using the interquartile range method, specifically targeting the target feature (Diff_15), to mitigate their potential impact on model accuracy. Given the dataset's composition, with many parts being categorical, these were transformed into dummy features through one-hot encoding, facilitating

their use in predictive modeling. Numeric features were rescaled using z-score normalization to standardize their distribution (Bollegala, 2017; Pineda-Jaramillo, Martínez-Fernández, Villalba-Sanchis, Salvador-Zuriaga, & Insa-Franco, 2021), ensuring that model performance was not skewed by variable scales. To streamline the dataset further and reduce redundancy, a correlation analysis using the Pearson method was performed, enabling the selection of the most relevant features for training the models.

The dataset initially comprised 68,206 flight operations for the entire year of 2017, from January 1 to December 31. After comprehensive data cleaning, which involved merging with weather datasets and applying imputation techniques to fill missing values, the final dataset was refined to include 67,730 flight operations. Missing values were addressed using median imputation for numeric features and mode imputation for categorical attributes, based on the distribution of the data within the same month and destination signature to preserve the dataset's original characteristics as closely as possible. This approach ensured that the data used in our models was as complete and representative as possible.

In addition, it is important to note that the final dataset is unbalanced,11.7% of flights experiencing delays of more than 15 min. This imbalance presents a challenge for predictive modeling, as there are significantly more observations for the 'not delayed' class than the 'delayed' class. Such disparities can lead to models that are biased towards predicting the majority class more accurately, diminishing their effectiveness in identifying delayed flights - a common issue in classification problems. To counteract this, we employed specific techniques designed to handle imbalanced datasets, enhancing our model's ability to generalize across both classes. The composition and distribution of the final dataset's features, both categorical and numerical, are detailed in Table 3, providing insight into the data foundation upon which our analysis is built.

3.3. Machine learning models

In the initial stage, the dataset was divided into training (68.5%), test (28.5%), and holdout (5%) sets, aiming to use the training set for model training, the test set for model evaluation, and the holdout set for final validation to confirm model generalizability. Given the binary categorical nature of the target feature (delayed or not delayed), classification models were deemed suitable for this study. It was observed that the dataset exhibited a class imbalance, with delayed flights being significantly underrepresented compared to non-delayed flights. To address this issue, we explored several class imbalance techniques, including Random Under Sampling, SMOTE (Synthetic Minority Over-sampling Technique), Random Over Sampling, and SMOTE-ENN (SMOTE with Edited Nearest Neighbors). After preliminary testing with a sample of the data and evaluating metrics such as accuracy, precision, recall, and AUC score, SMOTE-ENN emerged as the most effective method for this specific dataset.

SMOTE-ENN is a hybrid approach that synthesizes new examples from the minority class (delayed flights) using SMOTE, then refines the dataset by removing synthetic examples that introduce ambiguity, as performed by the ENN algorithm. This technique is well-acknowledged for its effectiveness in addressing class imbalance issues in datasets, ensuring that the models developed are not biased towards the majority class, and it has been widely recognized and validated in the literature for its effectiveness in handling imbalanced datasets, ensuring that our models accurately reflect the complexities of flight delay prediction (Islam, Abdel-Aty, Cai, & Yuan, 2021; N. Kim & Hong, 2021; Kumar et al., 2022; Muntasir Nishat et al., 2022; B. Wang et al., 2022).

Following that, ten ML models were trained that had previously been used in classification problems with similar dataset size and input features. The goal was to determine which model best predicted whether a commercial flight would be delayed, as shown in Table 4. The hyperparameter tuning process can improve classification models by combining the values of the various parameters. Each model's performance is measured using accuracy, recall, and ROC_AUC (Bergstra & Bengio, 2012).

Table 4
Models implemented in this study.

ML Model	Description
Logistic Regression	Logistic regression uses a logistic function to determine the class to which each observation belongs (Menard, 2004).
Decision Trees	Decision trees recursively divide the feature space to perform classification using a tree-like structure (Hagenauer & Helbich, 2017).
Naive Bayes	Naive Bayes classifiers are probabilistic models based on Bayes' theorem, with strong (naive) independence assumptions between the features. They are particularly known for their simplicity and efficiency in high-dimensional datasets (Hagenauer & Helbich, 2017; Shafiq, Tian, Bashir, Jolfaei, & Yu, 2020; X. Zhao, Yan, Yu, & Van Hentenryck, 2020).
K Nearest Neighbors	K Nearest Neighbors uses the closest neighbors of each data point to determine the classes (Kim, Choi, Moon, & Mun, 2011).
SVM - Radial Kernel	The Radial Basis Function (RBF) kernel is a popular SVM variant used for classification, utilizing a radial basis function as its kernel to handle non-linear data separation effectively (Patle & Chouhan, 2013)
Linear Discriminant Analysis	Linear Discriminant Analysis is a dimensionality reduction technique used in the machine learning classification process, aiming to find a linear combination of features that best separate two or more classes of objects (Ba, Zhang, Wang, Zhou, & Ren, 2017; Ching, Chu, Liao, & Wang, 2012; K. S. Kim et al., 2011; Pineda-Jaramillo, Barrera-Jiménez, & Mesa-Arango, 2022).
Ada Boost Classifier	AdaBoost, short for Adaptive Boosting, is an ensemble learning method that combines multiple weak classifiers to create a strong classifier, by adjusting the weights of incorrectly classified instances (Servos, Liu, Teucke, & Freitag, 2019; Shafiq et al., 2020).
Extra Trees Classifier	Extra Trees Classifier is an ensemble learning method that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve predictive accuracy and control overfitting (Servos et al., 2019; Seyyedattar, Ghiasi, Zendehboudi, & Butt, 2020).
Random Forest	Random Forest uses feature randomness to simultaneously train an array of decision trees (ensemble method) to produce a more effective forest of trees (Hagenauer & Helbich, 2017; Pineda-Jaramillo et al., 2022).
Gradient Boosting	Gradient Boosting is an ensemble method that successively trains several decision trees, with each new tree correcting a previous tree's incorrect classification (Hagenauer & Helbich, 2017; Mesa-Arango et al., 2023; Pineda-Jaramillo, 2021).

The k-fold cross-validation method is used to validate the training of the machine learning model. This method involves dividing the movement set into K groups and ensuring that each group contains an equal proportion of delayed and non-delayed flights. The model is trained using K-1 subsets in turns, and the ROC_AUC metric is used to evaluate the model's performance in the K subset that was not used for training. The final ROC_AUC score is calculated by averaging the ROC_AUC values from all folds. The k-fold cross-validation method is widely used to avoid underfitting and overfitting issues (Witten, Frank, Hall, & Pal, 2016; Young, Wang, & Chakravarthy, 2019).

3.4. Explainable machine learning

Models must demonstrate a high level of learning performance and be amenable to scrutiny, interpretation, and decision-making based on the outcomes they produce to be considered part of the Explainable Machine Learning paradigm. As a result, after selecting the most effective ML model, two techniques, the SHAP method and Sobol's variance-based sensitivity method are used to investigate and explain the factors causing flight delays departing from Santiago de Chile airport. These

approaches can guide decision-making and improve the transparency and accountability of the model's outputs by providing valuable insights into the importance of individual features and their interactions in predicting flight delays.

4. Results

This section presents the results and analysis of the xAI-based models used for flight delay forecasting. The area is divided into two parts: (i) the performance metrics of the models used to achieve a robust and reliable model, such as accuracy, recall, and ROC_AUC, and (ii) the influence of features associated with delays in flights departing from the Santiago de Chile airport. These two components comprise the methodology for predicting flight delays using xAI, which uses data from multiple sources.

4.1. Models

The ML models described in Section 3.3 were trained and optimized using the k-fold cross-validation method, and the results are shown in Table 5.

After evaluating a broader range of models as depicted in Table 5, it was observed that while gradient boosting and random forest excel in terms of accuracy, and ROC_AUC scores, our primary focus was on maximizing recall due to its critical importance in identifying delayed flights. Upon comprehensive comparison, Linear Discriminant Analysis (LDA) emerged as the standout model, exhibiting the highest recall among the models tested, alongside robust accuracy and ROC_AUC scores. This makes LDA particularly well-suited for predicting flight delays, given the operational significance of correctly identifying as many delayed flights as possible.

Consequently, we have opted to proceed with a detailed evaluation of the LDA model, implementing optimization techniques to refine its performance further. The effectiveness and generalizability of LDA are demonstrated through its consistent performance across all key metrics, as shown in Table 6. To further illustrate our model's performance and specifically address the valuable insights provided by a confusion matrix, we present this analysis in Fig. 2. This model's selection is grounded in its ability to offer a balanced approach to predicting flight delays at Santiago de Chile International Airport, providing both accuracy and interpretability in understanding the factors contributing to delays.

On the other hand, while precision is an important metric in many ML applications, it was not considered the primary criterion for model selection in our study. Precision measures the proportion of true positive predictions in the pool of all positive predictions made by the model. However, in the context of predicting flight delays, where the primary operational goal is to identify as many actual delays as possible to mitigate their impact, the cost of missing a delayed flight (a false negative) outweighs the cost of falsely predicting a delay (a false positive). Given the significant imbalance in our dataset, with delayed flights being much less frequent than non-delayed flights, a model that excels in recall ensures that fewer delayed flights are missed, even if it means accepting a higher rate of false positives. Consequently, while precision remains a useful metric for assessing model performance, our focus on recall over precision reflects a strategic decision to prioritize the detection of delayed flights, acknowledging the operational realities and the greater negative implications of failing to identify delays accurately.

Table 5
Results.

Model	Accuracy	ROC_AUC	Recall
Logistic Regression (LR)	0.543	0.725	**0.817**
K Neighbors (KNN)	0.580	0.621	0.616
Decision Tree (DT)	0.667	0.589	0.468
Random Forest (RF)	**0.752**	**0.749**	0.547
Gradient Boosting (GB)	**0.772**	**0.763**	0.534
SVM - Radial Kernel (RBFSVM)	0.604	0.676	0.671
Naive Bayes (NB)	0.526	0.618	0.683
Linear Discriminant Analysis (LDA)	0.567	0.725	**0.824**
Ada Boost Classifier (ADA)	0.641	0.649	0.557
Extra Trees Classifier (ET)	0.695	0.642	0.486

Table 6
Results of the LDA model in the three datasets.

Set	Accuracy	ROC_AUC	Recall
Training (using k-fold cv method) (68.5%)	0.567	0.725	0.824
Test (28.5%)	0.544	0.725	0.815
Holdout (5%)	0.544	0.720	0.809

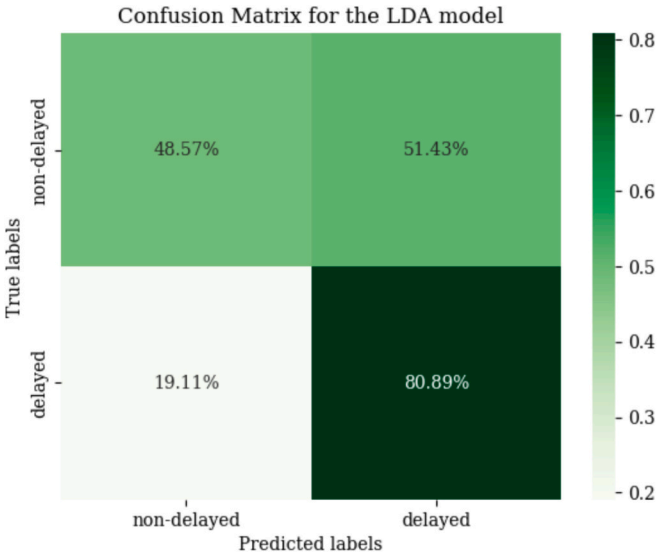

Fig. 2. Confusion Matrix for the test set for the LDA model.

4.2. SHAP method

Figure 3 depicts the significance of the features in the delays of commercial flights departing from the Santiago de Chile airport, as well as the direct impact of each element on the output magnitude for the LDA model. This plot offers an understanding of feature contributions beyond traditional linear interpretations. All features are ordered in descending order based on their global impact on the calculated SHAP values, demonstrating that international flight status, average temperature at the destination airport, wind speed, and average temperature at Santiago airport, among others, are the features that have the most significant influence on commercial flight delays departing from Santiago de Chile airport.

4.3. Sobol's variance-based sensitivity method

Sobol's variance-based sensitivity method was applied to the trained LDA model to assess the impact of input features on the prediction of commercial flight delays departing from the Santiago de Chile airport. Based on the ranges of the original dataset, a random sample of all input features was generated, and the LDA model was applied to this sample to predict flight delays. The Python SALib library calculated the total-order indices for all input features in the model using the resulting values.

The results of these indices for the most influential features are shown in Fig. 4, where we can see that the order of importance differs slightly from the order established by the SHAP method. When estimating delays for commercial flights departing from the Santiago de Chile airport, the most influential features in the LDA model output according to Sobol's variance-based sensitivity method are similar.

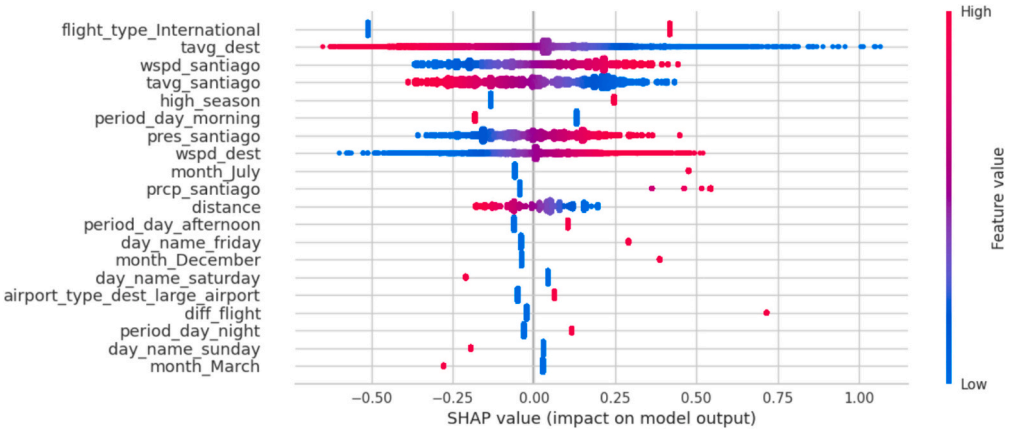

Fig. 3. SHAP plot of the most essential input features in the LDA model, sorted by global importance.

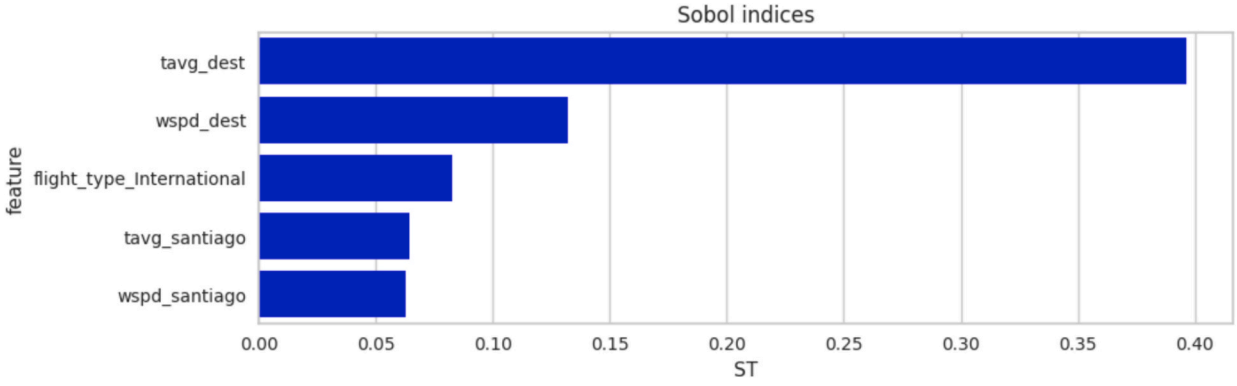

Fig. 4. Total effect indices of the LDA model's input features.

Given that the estimate includes the interaction effects of the same input features, the total-order index can be used to identify a possible non-influential part. Although the fact that the flight's destination differs from the scheduled destination appears not to affect the LDA model's ability to predict delays in commercial flights departing from the Santiago de Chile airport using the SHAP method, the Sobol results show that the interaction between this feature and the other input features is critical for such a task.

5. Discussion and implications

This research introduces an innovative xAI-based methodology for predicting flight delays, leveraging a comprehensive integration of diverse data sources. Unlike traditional regression models, our xAI approach excels in identifying complex, non-linear relationships between various factors and flight delays without relying on a priori assumptions. By analyzing air operations data, airport specifics, geographical details, and weather conditions, we provide clear explanations of the intricate dynamics influencing flight delays (AhmadBeygi, Cohn, Guan, & Belobaba, 2008; AhmadBeygi, Cohn, & Lapp, 2010; Beatty, Hsu, Berry, & Rome, 1999; Brueckner, Czerny, & Gaggero, 2022).

We found that factors such as the time of day, airport type and location, as well as environmental and weather conditions significantly influence flight delays at SCL. Our analysis reveals that international flights and those during peak travel times are particularly prone to delays, intensified by conditions such as high temperatures at takeoff, high wind speeds, and low temperatures at landing (Baumgarten, Malina, & Lange, 2014; Mayer & Sinai, 2003; Yimga, 2017). Our methodology's advantage lies in its ability to identify and quantify the nuanced effects of these factors, providing actionable insights for mitigating their impact.

The application of SHAP values and sensitivity analyses has proven instrumental in our understanding, revealing, for instance, the disproportionate impact of late-day departures on delay probability (Ahmad-Beygi et al., 2008, 2010; Beatty et al., 1999). This insight underscores the necessity for airlines to adjust scheduling and resource allocation strategically to mitigate the risk of cascading delays.

Moreover, our findings regarding the impact of airport characteristics on delay propensity, such as elevation and airport size, highlight the need for a comprehensive approach to delay mitigation that accounts for the diverse nature of these influences (Fan, Wu, & Zhou, 2014; Santos & Robin, 2010). The observed correlation between high temperatures and delay likelihood further suggests the importance of adapting operational strategies to climatic conditions (Coffel, Thompson, & Horton, 2017).

By embedding our methodology within the broader context of aviation operations, we underscore its significance not merely as an academic exercise but as a practical tool for improving the efficiency and reliability of air travel, where the most important implications for practitioners can be summarized as follows:

- Operational Adjustments: Airlines and airports can use these insights to refine scheduling, improve resource allocation, and tailor operational strategies to mitigate identified risk factors effectively. For example, insights into time-of-day or weather-related delays can help to inform more resilient scheduling practices, lowering the risk of cascading delays.
- Strategic Planning: The detailed analysis of delay factors provides a robust basis for long-term strategic planning, emphasizing the importance of flexibility and adaptability in operations to handle the

dynamic nature of flight delays. Understanding the complexities of flight delays allows airlines and airports to develop flexible and adaptable strategies for managing the dynamic nature of air travel, particularly during peak travel times and in adverse weather conditions.

- Policy Formulation: Our findings offer valuable insights for policymakers and regulatory bodies, supporting the development of policies aimed at enhancing air traffic management and minimizing delay incidences. By identifying systemic issues that contribute to delays, targeted interventions can be developed to improve the overall efficiency of the aviation ecosystem.
- Future Research Directions: This study lays the groundwork for future investigations into the application of xAI in aviation, suggesting further exploration of data integration techniques and model enhancements to refine delay predictions.

6. Conclusion and future work

The aviation industry is an essential enabler of global trade and passenger mobility. However, flight delays challenge airlines, resulting in inconvenience, lost time, and increased costs. Predicting flight delays accurately is critical for airlines to improve overall performance and provide better services to passengers. This research contributes to the aviation field by proposing a methodology for predicting flight delays based on xAI, a promising alternative to traditional machine learning methods. XAI can provide a better understanding of the data and transform the results into actionable insights that can guide decision-making processes by incorporating data from multiple sources and utilizing SHAP and Sobol's variance-based sensitivity techniques.

The linear discriminant analysis model is best suited to predicting delayed flights in this specific case of flights at Santiago de Chile airport. The study emphasizes the importance of flight destination, time of day effects, weather-related characteristics, and month effects in predicting flight delays and recommends that xAI algorithms be used to explain the factors contributing to flight delays.

According to an examination of SCL flight schedules, flights are typically delayed later in the day and overnight. International flights are delayed longer due to increased traffic volume and the complexity of international travel, and short trips from SCL have more delays. While medium and large airports offer better connectivity and convenience, they are also more prone to delays due to increased flight volume and complicated airport operations. Regarding meteorological factors, high takeoff temperatures at SCL, high wind speeds, and low temperatures during landing at destination airports can increase the potential for flight delays. Because of these weather-related factors, flight delays at SCL are worse in the winter than in the summer. Furthermore, flights to lower-elevation airports are more likely to be delayed due to warm weather and low air density, which affect aircraft performance. Finally, on average, Saturdays have more SCL flight delays.

Overall, the findings suggest that airlines, airports, and passengers can use the proposed model's predictions as a recommender system to make better-informed decisions and reduce the impact of flight delays. Our study travels an initial exploration into predicting flight delays using an xAI approach. While our findings provide valuable insights within the confines of our dataset, we acknowledge that expanding the scope of variables could enhance the robustness and applicability of our model. Future research could significantly benefit from testing the proposed model across multiple airports, which would allow for a broader evaluation of its effectiveness. Additionally, exploring the impact of a wider range of factors, such as flight frequency, destination popularity, and other obtainable features—including specific aircraft data—could offer a more comprehensive understanding of the dynamics influencing flight delays. By incorporating these variables, subsequent studies can build on our groundwork to develop more nuanced models that offer to the complexities of air travel and flight delay prediction.

While our study provides valuable insights into predicting flight delays using an xAI approach, several limitations should be acknowledged:

- Methodological Constraints: Our primary objective is to demonstrate how explainable AI (xAI) techniques, specifically SHAP values and Sobol's variance-based sensitivity analysis, can provide actionable insights into the determinants of flight delays. While these methods are effective in revealing complex, non-linear interactions between variables, they may not capture the magnitude of relationships as clearly as traditional regression methods. Future research could benefit from combining these xAI techniques with traditional regression models to enhance interpretability.
- Generalizability: Our findings are based on data from Santiago de Chile International Airport (SCL). While the methodology can be applied to other airports, specific results and feature importance may differ based on local conditions and operational characteristics. Future research should consider testing the model across multiple airports to evaluate its broader applicability.
- Temporal Data: The dataset used is from 2017, and changes in airport operations and influencing factors over time may impact the applicability of our findings. Future research should use more recent data to validate and refine our model, ensuring that it reflects current conditions and practices.

By addressing these limitations, future studies can build on our groundwork to develop more robust and generalizable models for predicting flight delays, ultimately enhancing the reliability and efficiency of air travel.

Funding

This research received no specific grant from public, commercial, or not-for-profit funding agencies.

CRediT authorship contribution statement

Juan Pineda-Jaramillo: Software, Methodology, Data curation. **Claudia Munoz:** Writing – review & editing, Conceptualization. **Rodrigo Mesa-Arango:** Visualization, Investigation. **Carlos Gonzalez-Calderon:** Writing – original draft, Methodology, Data curation, Conceptualization. **Anne Lange:** Visualization, Supervision.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Data availability

Data will be made available on request.

References

Abdelghany, A., Guzhva, V. S., & Abdelghany, K. (2023). The limitation of machine-learning based models in predicting airline flight block time. *Journal of Air Transport Management, 107*, Article 102339. https://doi.org/10.1016/j.jairtraman.2022.102339

AhmadBeygi, S., Cohn, A., Guan, Y., & Belobaba, P. (2008). Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management, 14*(5), 221–236. https://doi.org/10.1016/j.jairtraman.2008.04.010

AhmadBeygi, S., Cohn, A., & Lapp, M. (2010). Decreasing airline delay propagation by re-allocating scheduled slack. *IIE Transactions, 42*(7), 478–489. https://doi.org/10.1080/07408170903468605

Algarin Ballesteros, J. A., & Hitchens, N. M. (2018). Meteorological factors affecting airport operations during the winter season in the Midwest. *Weather, Climate, and Society, 10*(2), 307–322. https://doi.org/10.1175/WCAS-D-17-0054.1

Alharbi, B., & Prince, M. (2020). A hybrid artificial intelligence approach to predict flight delay. *International Journal of Engineering Research and Technology, 13*(4), 814. https://doi.org/10.37624/IJERT/13.4.2020.814-822

Allan, S., Gaddy, S., & Evans, J. (2001). *Delay causality and reduction at the new York City airports using terminal weather information systems*. Mass, USA: Lincoln Laboratory, Massachusetts Institute of Technology Cambridge.

Ayaydin, A., & Akcayol, M. A. (2022). Derin Öğrenme Tabanlı Havacılık Uçuş Verilerinde Gecikme Durumunun Tahmin Edilmesi. *Bilişim Teknolojileri Dergisi, 15*(3), 239–249. https://doi.org/10.17671/gazibtd.1060646

Ba, Y., Zhang, W., Wang, Q., Zhou, R., & Ren, C. (2017). Crash prediction with behavioral and physiological features for advanced vehicle collision avoidance system. *Transportation Research Part C: Emerging Technologies, 74*, 22–33. https://doi.org/10.1016/j.trc.2016.11.009

Baumgarten, P., Malina, R., & Lange, A. (2014). The impact of hubbing concentration on flight delays within airline networks: An empirical analysis of the US domestic market. *Transportation Research Part E: Logistics and Transportation Review, 66*, 103–114. https://doi.org/10.1016/j.tre.2014.03.007

Beatty, R., Hsu, R., Berry, L., & Rome, J. (1999). Preliminary evaluation of flight delay propagation through an airline schedule. *Air Traffic Control Quarterly, 7*(4), 259–270. https://doi.org/10.2514/atcq.7.4.259

Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. (2017). Using scalable data Mining for Predicting Flight Delays. *ACM Transactions on Intelligent Systems and Technology, 8*(1), 1–20. https://doi.org/10.1145/2888402

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(10), 281–305.

Bollegala, D. (2017). Dynamic feature scaling for online learning of binary classifiers. *Knowledge-Based Systems, 129*(1), 97–105. https://doi.org/10.1016/j.knosys.2017.05.010

Borsky, S., & Unterberger, C. (2019). Bad weather and flight delays: The impact of sudden and slow onset weather events. *Economics of Transportation, 18*, 10–26. https://doi.org/10.1016/j.ecotra.2019.02.002

Britto, R., Dresner, M., & Voltes, A. (2012). The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review, 48*(2), 460–469. https://doi.org/10.1016/j.tre.2011.10.009

Brueckner, J. K., Czerny, A. I., & Gaggero, A. A. (2022). Airline delay propagation: A simple method for measuring its extent and determinants. *Transportation Research Part B: Methodological, 162*, 55–71. https://doi.org/10.1016/j.trb.2022.05.003

Bubalo, B., & Gaggero, A. A. (2021). Flight delays in European airline networks. *Research in Transportation Business & Management, 41*, Article 100631. https://doi.org/10.1016/j.rtbm.2021.100631

Carvalho, L., Sternberg, A., Maia Gonçalves, L., Beatriz Cruz, A., Soares, J. A., Brandão, D., … Ogasawara, E. (2021). On the relevance of data science for flight delay research: A systematic review. *Transport Reviews, 41*(4), 499–528. https://doi.org/10.1080/01441647.2020.1861123

Chakrabarty, N. (2019). A data mining approach to flight arrival delay prediction for American Airlines. In *2019 9th Annual information technology, electromechanical engineering and microelectronics conference (IEMECON)* (pp. 102–107). https://doi.org/10.1109/IEMECONX.2019.8876970

Ching, W. K., Chu, D., Liao, L. Z., & Wang, X. (2012). Regularized orthogonal linear discriminant analysis. *Pattern Recognition, 45*(7), 2719–2732. https://doi.org/10.1016/j.patcog.2012.01.007

Choi, S., Kim, Y. J., Briceno, S., & Mavris, D. (2016). Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th digital avionics systems conference (DASC)* (pp. 1–6). https://doi.org/10.1109/DASC.2016.7777956

Coffel, E. D., Thompson, T. R., & Horton, R. M. (2017). The impacts of rising temperatures on aircraft takeoff performance. *Climatic Change, 144*(2), 381–388. https://doi.org/10.1007/s10584-017-2018-9

Dai, M. (2024). A hybrid machine learning-based model for predicting flight delay through aviation big data. *Scientific Reports, 14*(1), 4603. https://doi.org/10.1038/s41598-024-55217-z

Dalmau, R., Ballerini, F., Naessens, H., Belkoura, S., & Wangnick, S. (2021). An explainable machine learning approach to improve take-off time predictions. *Journal of Air Transport Management, 95*, Article 102090. https://doi.org/10.1016/j.jairtraman.2021.102090

David, P., Kriegel, M., Berdermann, J., Kauristie, K., Jacobsen, K. S., Fabbro, V., … Keil, R. (2023). Performance indicator development addressing mitigation of the space weather impacts on GNSS. *Journal of Space Safety Engineering, 10*(3), 324–330. https://doi.org/10.1016/j.jsse.2023.07.004

Degas, A., Islam, M. R., Hurter, C., Barua, S., Rahman, H., Poudel, M., … Aricó, P. (2022). A survey on artificial intelligence (AI) and eXplainable AI in air traffic management: current trends and development with future research trajectory. *Applied Sciences, 12*(3), 1295. https://doi.org/10.3390/app12031295

Deng, H. (2019). Interpreting tree ensembles with in trees. *International Journal of Data Science and Analytics, 7*(4), 277–287. https://doi.org/10.1007/s41060-018-0144-8

FAA. (2021). *The future of air transportation: Technology and innovation*. Federal Aviation Administration.

Fan, L. W., Wu, F., & Zhou, P. (2014). Efficiency measurement of Chinese airports with flight delays by directional distance function. *Journal of Air Transport Management, 34*, 140–145. https://doi.org/10.1016/j.jairtraman.2013.10.002

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Gholami, S., & Khashe, S. (2022). Flight delay prediction using deep learning and conversational voice-based agents. *Journal for Engineering, Technology, and Sciences, 89*(1), 60–72.

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology, 69*(1), 140–150. https://doi.org/10.1109/TVT.2019.2954094

Guimaraes, M., Soares, C., & Ventura, R. (2022). Decision support models for predicting and explaining airport passenger connectivity from data. *IEEE Transactions on Intelligent Transportation Systems, 23*(9), 16005–16015. https://doi.org/10.1109/TITS.2022.3147155

Guleria, Y., Cai, Q., Alam, S., & Li, L. (2019). A multi-agent approach for reactionary delay prediction of flights. *IEEE Access, 7*, 181565–181579. https://doi.org/10.1109/ACCESS.2019.2957874

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications, 78*, 273–282. https://doi.org/10.1016/j.eswa.2017.01.057

Hastie, T. J., & Tibshirani, R. J. (2017). *Generalized additive models*. Routledge. https://doi.org/10.1201/9780203753781

IATA. (2019). *Globalization and the air transport industry*. International Air Transport Association.

Islam, Z., Abdel-Aty, M., Cai, Q., & Yuan, J. (2021). Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention, 151*(December 2020), Article 105950. https://doi.org/10.1016/j.aap.2020.105950

Kaewunruen, S., Sresakoolchai, J., & Xiang, Y. (2021). Identification of weather influences on flight punctuality using machine learning approach. *Climate, 9*(8), 127. https://doi.org/10.3390/cli9080127

Khaksar, H., & Sheikholeslami, A. (2019). Airline delay prediction by machine learning algorithms. *Scientia Iranica, 26*(5 A), 2689–2702. https://doi.org/10.24200/sci.2017.20020

Kim, K. S., Choi, H. H., Moon, C. S., & Mun, C. W. (2011). Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current Applied Physics, 11*(3), 740–745. https://doi.org/10.1016/j.cap.2010.11.051

Kim, N., & Hong, S. (2021). Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city. *Information Processing & Management, 58*(1), Article 102410. https://doi.org/10.1016/j.ipm.2020.102410

Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., … Uddin, M. (2022). Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. *Healthcare, 10*(7), 1293. https://doi.org/10.3390/healthcare10071293

Lambelho, M., Mitici, M., Pickup, S., & Marsden, A. (2020). Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management, 82*, Article 101737. https://doi.org/10.1016/j.jairtraman.2019.101737

Li, Q., Guan, X., & Liu, J. (2023). A CNN-LSTM framework for flight delay prediction. *Expert Systems with Applications, 227*, Article 120287. https://doi.org/10.1016/j.eswa.2023.120287

Li, Q., & Jing, R. (2022). Flight delay prediction from spatial and temporal perspective. *Expert Systems with Applications, 205*, Article 117662. https://doi.org/10.1016/j.eswa.2022.117662

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12*, 361–386.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 623–631). https://doi.org/10.1145/2487575.2487579

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.

Malighetti, P., Morlotti, C., Redondi, R., & Paleari, S. (2023). The turnaround tactic and on-time performance: Implications for airlines' efficiency. *Research in Transportation Business & Management, 46*, Article 100874. https://doi.org/10.1016/j.rtbm.2022.100874

Mayer, C., & Sinai, T. (2003). *Why do airlines systematically schedule their flights to arrive late?* Wharton School, University of Pennsylvania.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58*(3), 218–223. https://doi.org/10.1198/000313004X946

Mesa-Arango, R., Pineda-Jaramillo, J., Araujo, D. S. A., Bi, J., Basva, M., & Viti, F. (2023). Missions and factors determining the demand for affordable mass space tourism in the United States: A machine learning approach. *Acta Astronautica, 204*(September 2022), 307–320. https://doi.org/10.1016/j.actaastro.2023.01.006

Meteostat. (2022). Meteostat Python Package, Version 1.6.5. Retrieved from https://pypi.org/project/meteostat/, 20 de febrero de 2023, 12:44.

Mokhtarimousavi, S., & Mehrabi, A. (2023). Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. *International Journal of Transportation Science and Technology, 12*(1), 230–244. https://doi.org/10.1016/j.ijtst.2022.01.007

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics, 33*(2), 161–174. https://doi.org/10.1080/00401706.1991.10484804

Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., … Khan, M. R. H. (2022). A comprehensive investigation of the

performances of different machine learning classifiers with SMOTE-ENN oversampling technique and Hyperparameter optimization for imbalanced heart failure dataset. *Scientific Programming, 2022*, 1–17. https://doi.org/10.1155/2022/3649406

Ou, L., Liu, L., Dong, S., & Wang, Y. (2014). Robust stability clearance of flight control law based on global sensitivity analysis. *Journal of Applied Mathematics, 2014*, 1–10. https://doi.org/10.1155/2014/153602

Oza, S., Sharma, S., Sangoi, H., Raut, R., & Kotak, V. C. (2015). Flight delay prediction system using weighted multiple linear regression. *International Journal of Engineering and Computer Science, 4*(4), 11668–11676. ISSN:2319-7242.

Patle, A., & Chouhan, D. S. (2013). SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)* (pp. 1–9). https://doi.org/10.1109/ICAdTE.2013.6524743

Pineda-Jaramillo, J. (2021). Travel time, trip frequency and motorised-vehicle ownership: A case study of travel behaviour of people with reduced mobility in Medellín. *Journal of Transport & Health, 22*(April), Article 101110. https://doi.org/10.1016/j.jth.2021.101110

Pineda-Jaramillo, J., Barrera-Jiménez, H., & Mesa-Arango, R. (2022). Unveiling the relevance of traffic enforcement cameras on the severity of vehicle–pedestrian collisions in an urban environment with machine learning models. *Journal of Safety Research*. https://doi.org/10.1016/j.jsr.2022.02.014

Pineda-Jaramillo, J., Martínez-Fernández, P., Villalba-Sanchis, I., Salvador-Zuriaga, P., & Insa-Franco, R. (2021). Predicting the traction power of metropolitan railway lines using different machine learning models. *International Journal of Rail Transportation, 9*(5), 461–478. https://doi.org/10.1080/23248378.2020.1829513

Plischke, E., Borgonovo, E., & Smith, C. L. (2013). Global sensitivity measures from given data. *European Journal of Operational Research, 226*(3), 536–550. https://doi.org/10.1016/j.ejor.2012.11.047

Prince, J. T., & Simon, D. H. (2015). Do incumbents improve service quality in response to entry? Evidence from Airlines' on-time performance. *Management Science, 61*(2), 372–390. https://doi.org/10.1287/mnsc.2014.1918

Qu, J., Xiao, M., Yang, L., & Xie, W. (2023). Flight delay regression prediction model based on Att-conv-LSTM. *Entropy, 25*(5), 770. https://doi.org/10.3390/e25050770

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135–1144. Doi:10.48550/arXiv.1602.04938.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications, 181*(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., … Tarantola, S. (2007). *Global sensitivity analysis. The Primer*. John Wiley & Sons, Ltd.. https://doi.org/10.1002/9780470725184

Santos, G., & Robin, M. (2010). Determinants of delays at European airports. *Transportation Research Part B: Methodological, 44*(3), 392–403. https://doi.org/10.1016/j.trb.2009.10.007

See, K. F., Ülkü, T., Forsyth, P., & Niemeier, H.-M. (2023). Twenty years of airport efficiency and productivity studies: A machine learning bibliometric analysis. *Research in Transportation Business & Management, 46*, Article 100771. https://doi.org/10.1016/j.rtbm.2021.100771

Servos, N., Liu, X., Teucke, M., & Freitag, M. (2019). Travel time prediction in a multimodal freight transport relation using machine learning algorithms. *Logistics, 4*(1), 1. https://doi.org/10.3390/logistics4010001

Seyyedattar, M., Ghiasi, M. M., Zendehboudi, S., & Butt, S. (2020). Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models. *Fuel, 269*, Article 116834. https://doi.org/10.1016/j.fuel.2019.116834

Shafiq, M., Tian, Z., Bashir, A. K., Jolfaei, A., & Yu, X. (2020). Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustainable Cities and Society, 60*, Article 102177. https://doi.org/10.1016/j.scs.2020.102177

Singh, J., Jayaprakash, M. D., & Agarwal, R. (2022). Flight delay prediction for Indian air carriers with explainable artificial intelligence. In *Third international conference on smart Technologies in Computing, electrical and electronics (ICSTCEE)* (pp. 1–6). https://doi.org/10.1109/ICSTCEE56972.2022.10099797

Sobol, I. M. (1993). **Sensitivity analysis for nonlinear mathematical models**. *Mathematical Modelling Computational Experiments, 1*, 407–414.

Sorokina, D., Caruana, R., Riedewald, M., & Fink, D. (2008). Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on machine learning - ICML '08* (pp. 1000–1007). https://doi.org/10.1145/1390156.1390282

Sridhar, B., Wang, Y., Klein, A., & Jehlen, R. (2011). Modeling flight delays and cancellations at the national, regional and airport levels in the United States. In *Proceedings of the 9th air traffic management Research and Development seminar, Berlin, Germany*.

Stone, M. J. (2018). Impact of delays and cancellations on travel from small community airports. *Tourism and Hospitality Research, 18*(2), 214–228. https://doi.org/10.1177/1467358416637252

Tang, Y. (2021). Airline flight delay prediction using machine learning models. In *ACM international conference proceeding series (pp. 151–154)*. https://doi.org/10.1145/3497701.3497725

Wang, B., Zhang, C., Wong, Y. D., Hou, L., Zhang, M., & Xiang, Y. (2022). Comparing resampling algorithms and classifiers for modeling traffic risk prediction. *International Journal of Environmental Research and Public Health, 19*(20), 13693. https://doi.org/10.3390/ijerph192013693

Wang, L., Mao, J., Li, L., Li, X., & Tu, Y. (2023). Prediction of estimated time of arrival for multi-airport systems via "bubble" mechanism. *Transportation Research Part C: Emerging Technologies, 149*, Article 104065. https://doi.org/10.1016/j.trc.2023.104065

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques. In *Data mining: Practical machine learning tools and techniques (4th ed.)*.

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics, 17*(1), 145. https://doi.org/10.1186/s12859-016-0995-8

Xu, J., Wang, A., Schmidt, N., Adams, M., & Hatzopoulou, M. (2020). A gradient boost approach for predicting near-road ultrafine particle concentrations using detailed traffic characterization. *Environmental Pollution, 265*, Article 114777. https://doi.org/10.1016/j.envpol.2020.114777

Xue, D., Yang, J., Liu, Z., & Cong, W. (2023). Forward-looking study of solar maximum impact in 2025: Effects of satellite navigation failure on aviation network operation in the Greater Bay Area, China. *Space Weather, 21*(12). https://doi.org/10.1029/2023SW003678

Yi, J., Zhang, H., Liu, H., Zhong, G., & Li, G. (2021). Flight delay classification prediction based on stacking algorithm. *Journal of Advanced Transportation, 2021*, 1–10. https://doi.org/10.1155/2021/4292778

Yimga, J. O. (2017). Airline code-sharing and its effects on on-time performance. *Journal of Air Transport Management, 58*, 76–90. https://doi.org/10.1016/j.jairtraman.2016.10.001

Young, S. D., Wang, W., & Chakravarthy, B. (2019). Crowdsourced traffic data as an emerging tool to monitor car crashes. *JAMA Surgery, 154*(8), 777–778. https://doi.org/10.1001/jamasurg.2019.1167

Zhang, B., & Ma, D. (2020). Flight delay prediciton at an airport using maching learning. In *2020 5th international conference on electromechanical control technology and transportation (ICECTT)* (pp. 557–560). https://doi.org/10.1109/ICECTT50890.2020.00128

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics, 39*(1), 272–281. https://doi.org/10.1080/07350015.2019.1624293

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society, 20*, 22–35. https://doi.org/10.1016/j.tbs.2020.02.003

Zhu, X., & Li, L. (2021). Flight time prediction for fuel loading decisions with a deep learning approach. *Transportation Research Part C: Emerging Technologies, 128*, Article 103179. https://doi.org/10.1016/j.trc.2021.103179

Zinke, L. (2023). Halloween-like solar storm impacts. *Nature Reviews Earth and Environment, 4*(11), 735. https://doi.org/10.1038/s43017-023-00496-9