

Article

Machine Learning Methods Benchmarking for Predicting Flight Delays: An Efficiency Meta-Analysis

Hélio da Silva Queiróz Júnior ^{1,2,*} , Viviane Falcão ² , Francisco Gildemir Ferreira da Silva ³ ,
Izabelle Marie Trindade Bezerra ⁴ and Joab Kleber Lucena Machado ⁴

- ¹ Department of Transportation Engineering and Geodesy, Federal University of Bahia, UFBA, R. Prof. Aristides Novis, 2-Federação, Salvador 40110-902, Brazil
- ² Department of Civil and Environmental Engineering, Center of Technologies and Geosciences, Federal University of Pernambuco, UFPE, Avenida da Engenharia, s/n-Cidade Universitária, Recife 50670-420, Brazil; viviane.afalcao@ufpe.br
- ³ Economy Graduate Program, Federal University of Ceará, UFC, Fortaleza 60020-181, Brazil; gildemir@ufc.br
- ⁴ Civil Engineering Department, Federal University of Campina Grande, UFCG, Campina Grande 58429-900, Brazil; izabelle.marie@professor.ufcg.edu.br (I.M.T.B.)
- * Correspondence: helio.junior@ufpe.br

Abstract

Predicting delays in commercial flights is an increasing challenge due to rising air traffic demand, which generates additional costs and operational complexity. This study synthesizes and evaluates machine learning approaches for flight delay predictions, aiming to identify the most accurate prediction logic and assess the role of sample size in model performance. A systematic literature review was conducted, followed by a meta-analysis of 1077 studies published between 2015 and 2025. The studies were classified by prediction logic (binary classification or regression) and evaluated in terms of model effectiveness using Data Envelopment Analysis and Tobit regression to determine the influence of explanatory variables. The results show that binary classification approaches achieved higher average accuracy than regression models did, with confidence intervals validating their relative effectiveness. Furthermore, findings indicate that the use of more complex models does not guarantee improved predictive performance, suggesting that researchers should prioritize robust variable selection rather than constantly adopting increasingly complex methods. This work provides a comprehensive overview of machine learning methods for flight delay predictions and highlights implications for optimizing airport operations and enhancing passenger experience through the adoption of more reliable predictive strategies.

Keywords: flight delay; prediction; meta-analysis; machine learning; DEA



Academic Editors: Yi Li, Meng Meng and Ziyue Yuan

Received: 19 September 2025

Revised: 17 October 2025

Accepted: 23 October 2025

Published: 5 November 2025

Citation: Queiróz Júnior, H.d.S.; Falcão, V.; da Silva, F.G.F.; Bezerra, I.M.T.; Machado, J.K.L. Machine Learning Methods Benchmarking for Predicting Flight Delays: An Efficiency Meta-Analysis. *Sustainability* **2025**, *17*, 9887. <https://doi.org/10.3390/su17219887>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Commercial flight delays are a common problem at airports around the world [1]. The growing demand for air transportation makes these delays increasingly recurrent, imposing additional costs and requiring continuous adjustments to flight management [2].

Accurate prediction of flight delays is paramount importance for the optimization of airport and airline operations. Companies with forecasting capabilities can implement various mitigation strategies, including flight schedule adjustments, the allocation of additional resources, and flight rescheduling [3]. These measures offer the dual benefit of enhancing the passenger experience by reducing uncertainty and waiting times, while concurrently achieving reductions in operating costs and an improvement in the overall efficiency of the

airport system. A reduction in delays is directly associated with decreased maintenance costs, fuel costs, and costs linked to airport services [4]. Furthermore, minimizing delays contributes to the image and reliability of airlines with their customers.

Nevertheless, adequate delay prediction requires implementing sophisticated models, and the use of learning methods is recurrent [5]. While substantial research has applied learning methods to delay prediction, most studies focus on comparing the predictive effectiveness (accuracy) of different machine learning models. To the best of our knowledge, the existing literature lacks a comprehensive analysis that benchmarks these methods based on their operational efficiency, which accounts for the resources (i.e., data and variables) required to achieve the predictive outcome [6]. This study addresses this critical gap by employing Data Envelopment Analysis (DEA) to compare algorithms based on efficiency, and Tobit regression to identify the key determinants influencing that efficiency, thereby providing a novel framework for selecting robust and practical models for flight delay prediction.

Extensive scholarly work has attempted to use learning algorithms for delay anticipation. However, most of these studies focus on comparing specific subsets of methods or different method groups. This selective focus is also common in other areas of transportation and logistics, where machine learning plays a significant role in predicting disruptions, such as in the automotive and textile supply chains [6,7]. While these studies highlight ML efficacy in limited use cases, a necessary overall, cross-method efficiency threshold for flight delay prediction that balances predictive accuracy, model complexity, and data requirements has yet to be established. This study aims to remedy this lack.

The heterogeneity of machine learning predictive methods and their adopted configurations—such as the number of expected outputs, the applied sample size, or the number of explanatory variables—contributes to increased model complexity and differentiates the analysis of prediction capacity [8].

Accurately predicting flight delays is a challenge that necessitates implementing sophisticated models, frequently found on machine learning methodologies [5]. The extant literature suggests that the diversity of predictive methods, each study's specific configuration, the size of the data sample used, and the number of explanatory variables considered play a crucial role in the model's effectiveness [8].

In this context, ref. [4] reported that ensemble storage methods exhibited superior prediction efficiency compared to tree-based methods for delays at Beijing airport. Ref. [9] found that ensemble models classifying flights as delayed or not (binary) were more effective than models using regression logic for delay times in the US airline network. In a similar vein, ref. [10] incorporated the joint use of ensemble prediction methods with Long-Range Short-Term Memory (LSTM), proposing an effective model for Chinese air traffic time data that utilizes a variant of recurrent neural networks (RNNs). The combined use of deep learning methods enables the analysis of large amounts of data [11] and, after numerous processing layers, yields accurate results even in non-binary classificatory prediction models [12].

Given these considerations, the objective of this research is to synthesize and analyze the various machine learning methods employed in studies predicting flight delays. The primary aim is to identify the most accurate method and prediction logic. Furthermore, we examine the potential for combining machine learning methods to increase prediction accuracy. Finally, we analyze the influence of sample size on the performance of the prediction model. The objective is also to establish a unified standard for comparing the results obtained for each proposed method. In addition, we seek to establish an initial analysis parameter applicable in future studies. Ultimately, the study aims to improve the accuracy and robustness of flight delay prediction models.

2. Methodology

Two research methods and two quantitative methods were selected to achieve the objectives defined. A systematic literature review was performed to provide a non-biased summary and comprehensive understanding of studies that apply machine learning methods to predict flight delays and their effects. In addition, a meta-analysis was developed according to the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) framework defined by ref. [13] and applied by ref. [14]. The studies were then ranked in terms of the efficiency of the binary prediction (classification) and regression models using the Data Envelopment Analysis (DEA) Meta-frontier model proposed by ref. [15], which enables the comparison of technical efficiencies between different groups.

2.1. Systematic Literature Review Analysis

A systematic literature review was conducted to provide a comprehensive and unbiased summary of the use of machine learning techniques in flight delay prediction and assessment of their effectiveness.

This review followed the model adopted by refs. [6,15]. The studies were collected and analyzed to identify trends in covered topics in recent years and directions for future research. Following analysis, the data was processed according to the following research strategy.

The literature review reveals that the heterogeneity of prediction results is fundamentally influenced by factors such as the study's time, region, and technical conditions [8]. Region, for instance, affects the relevance of input variables, as evidenced by studies focusing on airport capacity versus airline networks. The overall logic behind the prediction methods varies based on the output required: Binary classification is typically employed for immediate decision-making (delay or no delay), while Regression models are used for granularity (predicting delay duration or propagation). Our DEA and Tobit analysis explicitly address this complexity by evaluating how these intrinsic factors, alongside model configuration, ultimately impact the operational efficiency of the resulting predictions.

2.2. Research Strategy

A comprehensive search was conducted on two prominent databases, Scopus and ScienceDirect, to identify relevant studies. A search was conducted using the keywords "delay," "airport," "prediction," and "Machine Learning," and their respective synonyms. The selection of these terms was informed by prior review research, particularly refs. [6,15], which examined the application of machine learning methods in predicting flight delays. Furthermore, search filters were utilized to restrict the results to peer-reviewed scientific articles. The time range of publications considered was from 2015 to 2025 (early publications), and no restrictions were imposed on the language of the publications.

The research database was constructed in two phases: initially in October 2020 and again in October 2024. Only original studies published between January 2015 and October 2024 were considered.

The studies were then classified based on an analysis of the abstracts and the presentation of information on the capacity of the predictive models used. Consequently, data was collected from three specific groups of variables on the structure and composition employed in the studies for the construction and comparison of the proposed flight delay prediction models (see Table 1). The groups were formed based on the analyses of ref. [8].

Table 1. Data collected from studies on predicting flight delays using machine learning methods.

Variable Group		Description	Possible Values	Variable Type
Author(s)	Publication information	Identifying studies	-	Nominal
Journal		Journal of publication origin		Nominal
Publication Year		Year of study publication in journal	2015–2025	Ordinal
Journal Impact Factor		Impact factor of the journal of origin of the publication (value obtained when the study was collected)	-	Continuous Numerical
Number of Citations in the Study		Number of citations of the study up to October 2024.	-	Discrete numerical
Method Employed	Information on the prediction model(s) used	Machine learning method employed in the study	-	Nominal
Total Sample		The size of the universal sample used in the flight delay prediction model(s)	Values ranging from tens to millions of sample units tested.	Discrete numerical
Test Sample		Portion of the total sample reserved for test analyses of the prediction model(s) tested.	Values ranging from tens to millions of sample units tested	Discrete numerical
Accuracy of the Model(s)		Accuracy of the machine learning model(s) analyzed in the study	-	Continuous numerical
Correct Forecast Units		Units of the test sample correctly predicted by the model(s) applied.	-	Discrete numerical
Predictive Logic		Forecasting logic adopted by the forecasting model(s) analyzed	Regression or Binary	Binary
Study Area		Size of the study area	Airline or route; Airport	Nominal
Independent Variables: Aircraft Rotation Data	Information on the groups of independent variables used in the prediction models	Binary use of variables with aircraft rotation information as independent variables in the model(s) analyzed	They include planned and actual rotations, used to track the propagation of delays and reconstruct primary delays. Characteristics of each flight, such as flight duration, temporal patterns of airline operations (such as year, season, month, day of the week and time of day), and spatial characteristics of departure and arrival airports. Information about the airline's fleet, such as aircraft model, seat configuration and capacity, airport base and aircraft age.	Binary
Independent Variables: Flight Data		Binary use of variables with flight data information as independent variables in the model(s) analyzed		Binary
Independent Variables: Aircraft Data		Binary use of variables with flight data information as independent variables in the model(s) analyzed		Binary

Table 1. *Cont.*

Variable Group	Description	Possible Values	Variable Type
Independent Variables: Passenger Data	Binary use of variables with aircraft data information as independent variables in the model(s) analyzed	Information such as the number of passengers boarding, passenger boarding/disembarking times or information on passenger security and dispatch.	Binary
Independent Variables: Meteorological Data	Binary use of variables with passenger data information as independent variables in the model(s) analyzed	Weather conditions that impact operational procedures, such as temperature, visibility, wind direction and speed, gusts, altitude and humidity.	Binary
Independent Variables: Traffic Data	Binary use of variables with meteorological data information as independent variables in the model(s) analyzed	Information on flight traffic and airport capacity, modeled using a dynamic and stochastic queuing model to estimate traffic-related delays.	Binary
Independent Variables: Airport Capacity Data	Binary use of variables with air traffic data information as independent variables in the model(s) analyzed	Information such as the peak-hour capacity of the passenger terminal (TPS), or information on the runway capacity (PPD) at the airports where the flights originate and end	Binary
Independent Variables: Data on Previous Delays	Binary use of variables with airport capacity data information as independent variables in the model(s) analyzed	Information on flight delays or stages prior to the flight analyzed.	Binary

The initial group pertains to the information contained within each publication, including the identification of studies, such as the authors and the journal of origin. It also encompasses the systematic analysis of the models evaluated, such as the year of publication or the impact factor. In research behavior analyses, such as systematic reviews and meta-analyses, ref. [16] elucidate that the impact factor of a journal can influence confidence in the results of a study, especially in relation to publication bias. In addition, the authors suggest that the inclusion of unpublished data in peer-reviewed journals or journals without an impact factor can affect the estimates, suggesting that consideration of publication status is crucial to the validity of the results.

The second and third groups refer to information on the prediction models developed in the studies, which use machine learning. This data delineates specific characteristics of the structure applied in each study, from the type of machine learning model to the prediction logic used. The ability to predict can be facilitated by employing binary classification logic, where “0” denotes the absence of delay and “1” denotes its presence, or by regression, wherein delay is categorized based on its anticipated duration.

Finally, the groups of independent variables represent the main data applied in the predictive models evaluated, considering that the study area or location can influence the model’s performance. The groups were formed based on the analyses of ref. [8].

2.3. Applying Meta-Analysis to Evaluate Machine Learning Methods

A meta-analysis was subsequently conducted to combine the results of multiple studies. This process provides a more precise estimate of the confidence interval for the accuracy of the machine learning methods applied.

Meta-analysis is of crucial importance in prediction studies, as it facilitates the synthesis of data from multiple studies, thereby providing a comprehensive evaluation of the techniques and models employed. The systematic review was originally developed within the field of biomedical sciences [17–19]. However, it is increasingly applied in the context of artificial intelligence (AI), as evidenced by refs. [20,21]. In addition, ref. [20] proposed that the utilization of meta-analysis in machine learning algorithms facilitates the enhancement of the models constructed, despite the incorporation of data from disparate sources. This process is deemed to be instrumental in enhancing the precision of the forecasts obtained.

As ref. [22] applied meta-analysis to identify research gaps, especially in relation to data sources and data processing techniques, the selection of appropriate machine learning models and optimization techniques, allowing the generation of indices covering various aspects, such as study areas, continents, types of flood models and evaluation metrics, allows a clear visualization of the data.

In the field of transportation studies, ref. [23] applied meta-analysis to ascertain the most accurate machine learning method, in this case for the analysis of energy management for electric vehicles. In alignment with ref. [14], the authors employ the PRISMA meta-analysis methodology, as proposed by ref. [13], ensuring a systematic, transparent, and reproducible review process.

In such cases, reviews consistently observe a common trend of heterogeneity between studies in terms of the type of machine learning methods employed, models, and datasets. This heterogeneity can have a consequential impact on the effectiveness of forecasts. In view of this, ref. [24] elucidate that heterogeneity is significant as it reflects the diversity in modeling approaches and the variables considered.

Consequently, random effects meta-analysis constitutes a statistical approach that facilitates the amalgamation of results from disparate studies, thereby yielding a more precise estimate of an effect, whilst accounting for the variability between studies. This methodology is especially relevant in contexts where machine learning (ML) is used, as the results of different models can vary significantly due to factors such as the quality of the data and the choice of algorithms [25].

Also, ref. [26] states that random-effects meta-analysis, in contrast to the fixed-effects model, which assumes that all studies share a single true effect size, acknowledges that different studies may possess differing effect sizes due to variables such as participant characteristics or intervention implementations. Furthermore, the division of complex studies into groups can significantly enhance the performance of the meta-analysis, thereby facilitating a more detailed and specific analysis of the data.

Considering the points, ref. [8] identifies the area analyzed as one of the divisions to be observed between delay forecasting studies. This is because studies on airports focus on the capacity and demand of facilities, while the analysis of airlines focuses on internal and external factors that can affect the performance of a single company. These differences reflect the complexity of the air transportation system, where both airports and airlines play critical roles, but with different focuses in their analysis.

Subsequently, the size effect analysis was applied between the studies in the groups, as defined by ref. [26]. The effect size is a quantitative measure of the strength of a relationship or the magnitude of an effect in a study. It facilitates the comparison and synthesis of results from different studies, irrespective of the type of data utilized (e.g., continuous or binary data). The study effect is calculated for continuous data as the difference in means

between two groups, using the formula for Cohen's standardized difference in means (d), as outlined in Equation (1).

$$d = (M1 - M2)/(SD_{Pooled}) \quad (1)$$

In this equation, $M1$ and $M2$ represent the means of the two proposed groups, regression and binary, and SD_{Pooled} (Equation (2)) is the combined standard deviation of the groups.

$$SD_{Pooled} = (((n1 - 1) \times S_1^2 + (n2 - 1) \times S_2^2)/(n1 + n2 - 2))^{(1/2)} \quad (2)$$

where S_1^2 and S_2^2 are the standard deviations of the two samples, and $n1$ and $n2$ are the corresponding sample sizes.

In addition, ref. [26] indicates the joint analysis of the variance of the size effect, referring to the measure of dispersion or uncertainty associated with estimating the size effect. A low variance indicates that the effect estimate is accurate, while a high variance suggests greater uncertainty. The calculation of the Size Effect Variance can be calculated using Equation (3), in which $n1$ and $n2$ are the sample sizes for the studies of the two defined analysis groups.

$$Var(d) = ((m1 + n2)/(n1 \times n2)) + (d^2/(2 \times (n1 + n2))) \quad (3)$$

Consequently, the meta-analysis is constructed based on the heterogeneous behavior postulated in studies employing disparate machine learning (ML) methods. The Hedges and ref. [27] method is adopted, which estimates variance adjusted for heterogeneity. The ratio between observed events and expected events (O/E Ratio, or Odds/Expected Ratio) is employed in the Hedges model, thereby enabling a comparison of the observed odds ratio (odds) referring to the units correctly predicted by machine learning (ML) methods, with the expected odds ratio (expected odds) referring to the sample tested by the method. This configuration enables the assessment of the predictive effectiveness of each machine learning (ML) method in predicting flight delays, considering the proportion of correct predictions in the sample tested.

2.4. Ranking of Machine Learning Methods with Data Envelopment Analysis (DEA)

To investigate the heterogeneity of transportation studies, ref. [28] built upon previous meta-analysis studies by incorporating data envelopment analysis (DEA) to compare European and Asian airlines.

Therefore, utilizing Data Envelopment Analysis (DEA) alongside meta-analysis is substantiated by the necessity to assess efficiency across disparate groups operating within analogous technological contexts. According to ref. [29], the combination of DEA and meta-analysis is imperative to assess the divergence in technological progress between specific groups, thereby enriching the efficiency analysis of studies employing disparate methods, such as machine learning (ML).

Moreover, the integration of these approaches has the potential to circumvent the limitations associated with individual methodologies, including the variability in input and output metrics across studies, as evidenced by ref. [30].

As ref. [31] introduced the definition of a joint DEA methodology and systematic analysis through meta-analysis in air transport studies, the authors conducted a route-based performance evaluation, considering the variety of operational characteristics and different route characteristics. Utilizing meta-analysis, the authors were able to implement key variables to assess the efficiency-based performance of the routes examined.

Accordingly, the efficiency under scrutiny is defined as the capacity to circumvent resource expenditure during the production process [32]. In this context, the studies were evaluated using DEA to compare model accuracy (verified via meta-analysis), serving as an efficiency indicator.

While the effectiveness of a model is related to what is produced, without considering the resources needed to do so [33], DEA focuses on efficiency, evaluating performance as a function of the resources employed.

To operationalize the DEA meta-frontier model, core Input and Output components were explicitly defined. The model adopts two inputs (representing resources and complexity): (1) the total sample size (reflecting acquisition and processing costs), and (2) the number of independent variables. The output (representing the desired result) is the Observed/Expected Ratio (O/E Ratio), derived from the meta-analysis, which represents the proportion of correctly predicted instances in the test sample, serving as the benchmark for model effectiveness.

The DEA model selected is the Meta-frontier DEA proposed by ref. [15], which allows for a comparative evaluation of the efficiency of studies that use machine learning to predict flight delays. This approach enables the systematic ranking of studies based on their success rate. This is achieved by conducting a comprehensive analysis of the observed odds ratio, which corresponds to the units correctly predicted by machine learning (ML) models, in relation to the expected odds ratio. The expected odds ratio corresponds to the test sample utilized in the model. Therefore, the DEA Meta-frontier offers a robust structure for classifying studies, allowing the efficiency of each ML method to be assessed individually and ranked according to the groups of methods proposed, which is a factor not previously assessed in review studies.

To calculate the efficiency frontiers, each study had its ML methods organized into clusters, divided between binary forecasting models and regression forecasting models. The methods were positioned according to their effectiveness and efficiency and were allocated in one of the reference quadrants. The definition of these quadrants is derived from the axis that represents the 95th percentile of the meta frontier efficiencies, in accordance with the established average 95% confidence interval (Average CI) obtained through meta-analysis for each group of applied prediction logic. Each method is hierarchized within the group corresponding to its prediction logic. However, the DEA meta-frontier allows the analysis of the excellence of each applied method, regardless of the logic applied.

2.5. Censored Tobit Regression Analysis for Identifying Efficiency Determinants

The censored Tobit regression model is employed to delineate the impact of explanatory variables on the efficiency response of forecasting methods in the effort dimension of a performance measurement model [34], including machine learning methods. According to ref. [35], the Tobit model integrates elements of linear regression models with the analysis of censored data, thereby enabling the estimation of relationships between variables even when the dependent variable is partially observed.

The integration of DEA and Tobit regression constitutes a robust two-stage methodology, designed to first measure efficiency and subsequently identify its key determinants. In the first stage (DEA), we established the efficiency score of each study (DMU) by comparing its output (Efficacy/O/E Ratio) against the utilized inputs (Sample Size and Number of Variables). This DEA efficiency score, representing the dependent variable in the second stage, is intrinsically a technical measure of how effectively resources were converted into results. The second stage (Tobit regression) uses this efficiency score to determine the influence of exogenous and endogenous factors (the explanatory covariates) on that efficiency, addressing why certain studies performed better than others. This sequential

approach ensures that DEA quantifies what happened (conversion efficiency), while Tobit explains why it happened (the structural and data drivers).

The selection of covariates is based on empirical arguments that bridge the efficiency construct (dependent variable, derived from DEA) with the associated data costs and structural characteristics of the studies. Specifically, the covariates are structured into three sub-groups: (1) Publication and Prediction Metrics, including accuracy, sample size, and citation count, which quantify the perceived reliability and data effort; (2) Model Structure, represented by dummy variables for prediction logic (Binary/Regression) and study scope (Airline/Airport); and (3) Data Groups, where variables from Table 1 (e.g., Meteorological Data, Aircraft Data) serve as proxies for the marginal complexity and operational cost of data acquisition, directly influencing the efficiency frontier. This rigorous specification allows us to determine which structural and data-related factors maximize efficiency.

In the context of studies that employ machine learning methodologies to predict flight delays, the financial implications of data usage stem from the progressive escalation in the sample size, which poses a substantial challenge [36]. Many models operate with increasingly specific data or data that also necessitates predictions, such as future weather conditions or dynamic operational information. The diversity and quantity of data groups used often result in complex sets that do not guarantee the achievement of good results and compromise the practical feasibility of applying the models.

The necessity of substantial computing power and the temporal demands of effective data processing impede the ability to make predictions in a timely manner, thereby constraining the applicability of models in real-world scenarios. These factors underscore the necessity to achieve a balance between the volume of data and the operational viability of models. In numerous instances, the computational cost and processing effort do not yield commensurate practical benefits.

For example, ref. [37] posited that variable selection should be formulated as a decision problem that incorporates both the costs of predictors and their predictive adequacy. Consequently, the utilization of a Tobit model facilitates the assessment of the impact among the variables present in the third group of variables (Table 1) employed in the predicted model.

The adopted configuration entails the censoring of the dependent variable, efficiency, at the minimum efficiency limit (efficiency = 0) and the maximum benchmark limit (efficiency = 1), according to the predictive capacity of the method. Efficiency is defined as the model's performance in the results dimension in relation to the selected inputs. The set of variables applied in each forecasting study aims to maximize the generalization capacity of the chosen method. However, the presence of excess variables has the potential to introduce biases or compromise the predictive capacity of the applied model, thereby undermining its effectiveness in real-world settings.

In addition, the impact of the predictive model's structure, including the reliability of the results, as evidenced by the impact factor of the journal of origin for each study [38], and the logic and scope of the analysis employed, can directly influence the model's prediction efficiency.

Therefore, the Tobit model was estimated to have standard errors based on Hessian, and collinearity was measured by the Variance Inflation Factor (VIF). The normality of the residuals was verified through the distribution test and Q-Q plot of the residuals.

3. Results and Analysis

The data collection identified a total of 931 studies in the Scopus database and 146 in ScienceDirect that employ machine learning to predict flight delays. Subsequent processing of these data and the application of a select array of analysis methods enabled a more nu-

anced evaluation of the trends and characteristics associated with publications addressing this subject.

3.1. Systematic Analysis

The analysis of publication trends over time (Figure 1) reveals an almost exponential growth in the number of studies in the Scopus database (a) since 2020. A similar pattern is observed in the ScienceDirect database (b), suggesting a growing trend of studies on delay prediction with machine learning methods.

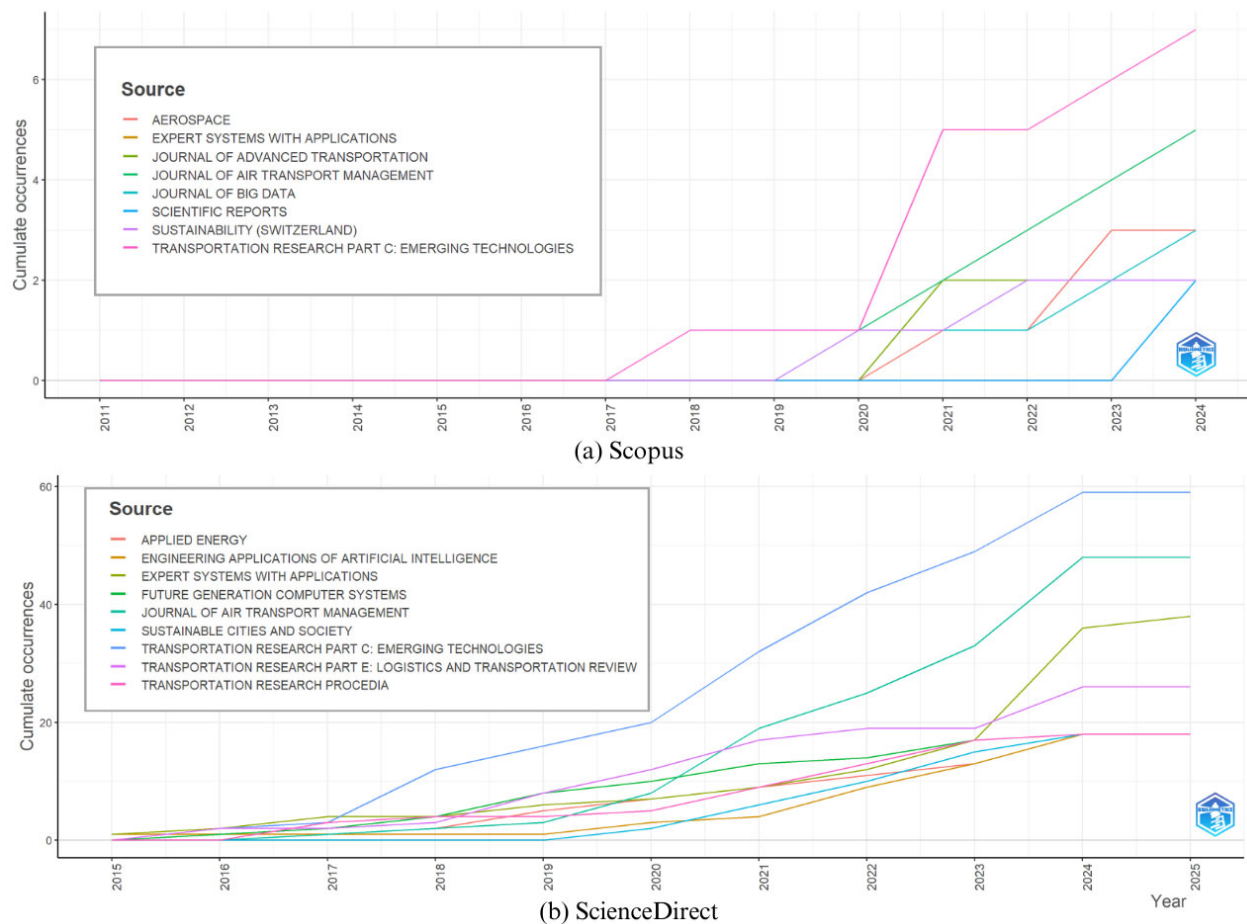


Figure 1. Production of research on delay prediction using machine learning methods over time.

Though the variation in their respective publication dimensions, both bases designate those journals that employ peer review as “main journals.” Illustrative examples of this practice include the Journal of Air Transportation Management and Transportation Research Part C: Emerging Technologies.

However, although peer review is a process that aims to ensure the quality and validity of research, the study by ref. [16] suggests that the reputation of the journal can have a greater impact on the citation of an article in the journal than the intrinsic quality of the study itself.

The researchers found that the original journal’s impact factor was the primary predictor of annual citations, eclipsing other factors such as study methodology and perceived quality. Regarding the trends in the topics addressed, a review of the main terms identified in the titles, abstracts and keywords of the studies, as shown in Figure 2, offers a view of the evolution of how research has centralized the ideas of the analyses carried out.

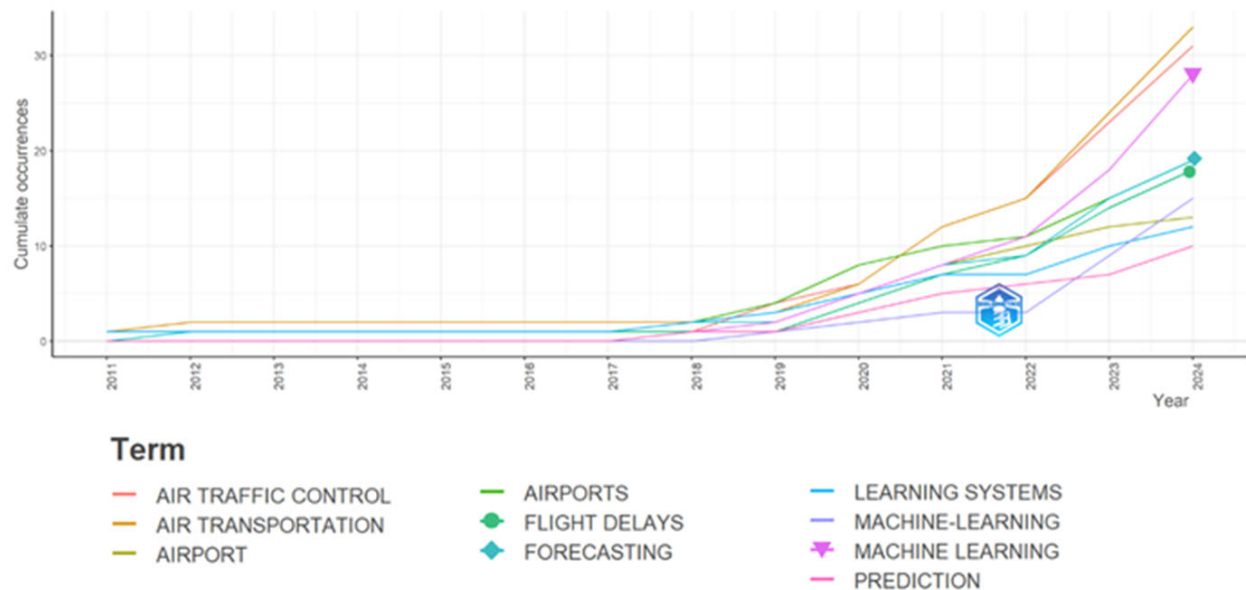


Figure 2. Frequency of the main terms observed in titles, abstracts and keywords of the studies obtained. Temporal analysis from 2011 to 2024.

Future studies in this field are expected to show a substantial rise in the application of machine learning methods [39]. This observation is corroborated by ref. [40], who further posits that as the volume of data increases, the application of machine learning methods becomes increasingly pertinent, particularly for predicting root delays.

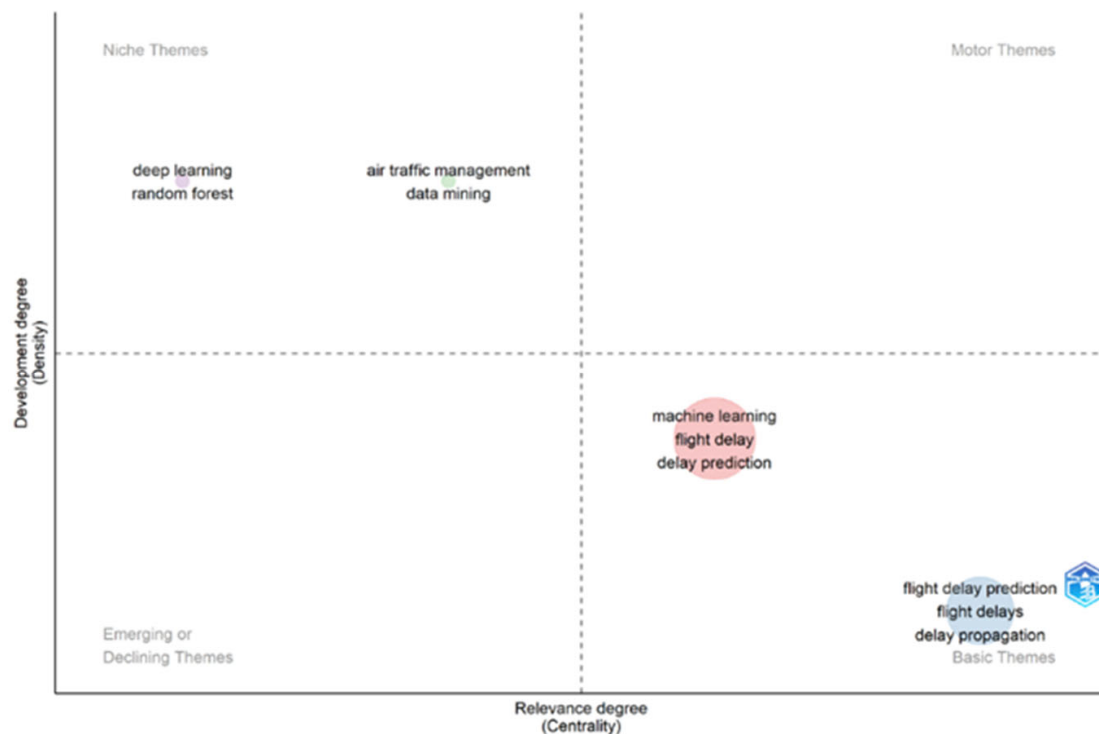
An alternative approach to analyzing trends involves the construction of strategic maps for each journal database (Figure 3). The strategic map is constructed through a thematic analysis of the co-occurrence of keywords or terms extracted from titles and abstracts. The strategic map utilizes two primary dimensions: centrality, which quantifies the strength of connections between a given topic and other topics within the field, and density, which measures the strength of connections within a specific topic.

In the analysis of the Scopus database, the term “niche themes” is characterized as dense but less central topics, which are defined as specialized or limited in scope. Examples include “deep learning” and “random forest,” which, while being highly developed, are of low relevance. This suggests that these topics are well developed but not extensively connected to other themes within the field. Conversely, Basic Themes exhibit a lower degree of density and interconnectedness, often occupying more peripheral positions. In this case, Basic Themes encompass subjects such as “flight delay prediction,” “flight delays,” and “delay propagation,” which, while highly relevant, remain underdeveloped, thereby establishing a foundation for future research endeavors.

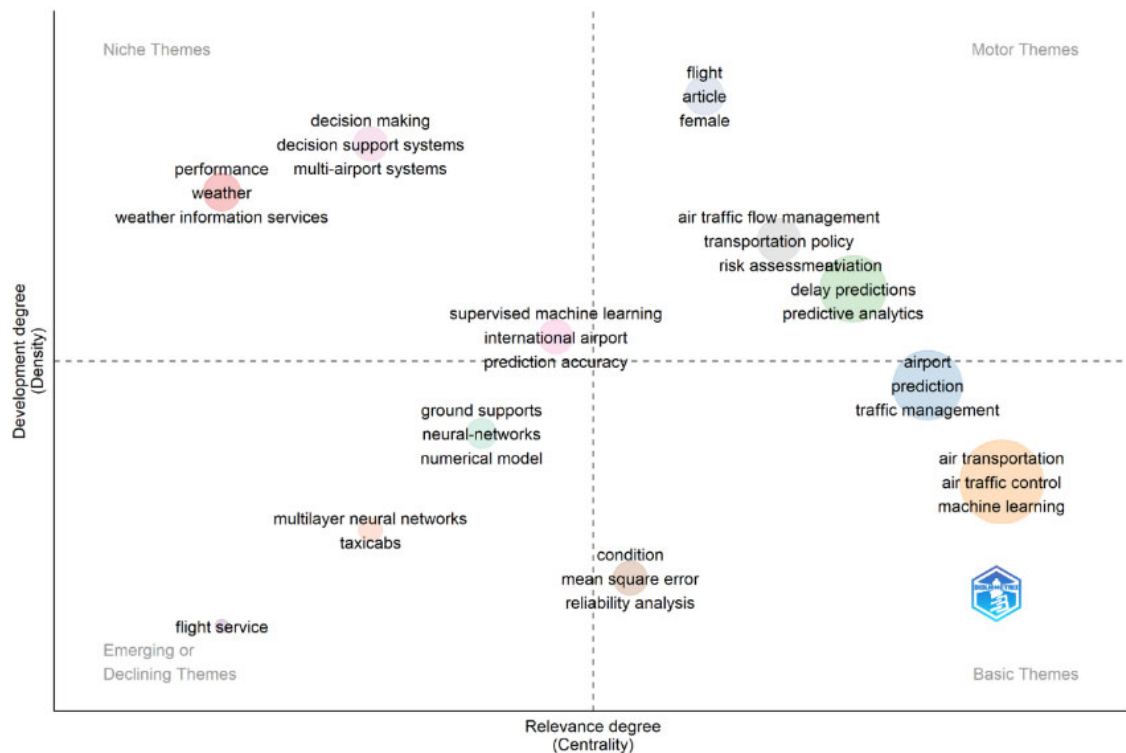
As illustrated in Figure 3b of the ScienceDirect strategy map, the themes pertaining to delay prediction by machine learning methods are categorized into four distinct classifications. The niche themes, which encompass “decision making,” “weather,” and “multi-airport systems,” among others, exhibit a high degree of development yet remain largely disconnected from other themes.

Motor Themes are associated with well-established subjects that are thoroughly integrated into the field. These themes are characterized as highly central and dense, exemplified by “air traffic flow management,” “delay predictions,” and “predictive analytics.” These themes are of significant relevance, have been extensively developed, and propel the research domain. Basic themes include “air transportation,” “air traffic control,” and “machine learning.” These themes are fundamental yet less developed. Conversely, Emerging or Declining Themes are associated with central themes, though exhibiting reduced

density. These themes encompass both emerging and declining subjects, such as “flight service” and “multilayer neural networks,” which are characterized by minimal relevance and development, suggesting potential growth or decline within the field.



(a) Scopus



(b) ScienceDirect

Figure 3. Strategic trend maps of flight delay prediction studies using machine learning method. Data from Scopus (a) and ScienceDirect (b).

A thorough examination of the available data reveals that, while Scopus places a stronger emphasis on machine learning techniques, such as “deep learning” and “random forest,” which are situated within specialized subject areas, ScienceDirect offers a more comprehensive approach by addressing the operational and decision support context. A distinguishing feature of ScienceDirect is the clear delineation of its guiding themes, a feature that is not as prominently exhibited in the Scopus map. A close examination of both maps reveals a noteworthy emphasis on machine learning and delay prediction, albeit with divergent positioning of these subjects with respect to relevance and development. The diversity of specialized subjects found on ScienceDirect indicates a more comprehensive emphasis on operational elements and decision-making processes.

Consequently, there is an increasing emphasis on models that consider delay propagation and cancelation analysis. This focus is guided by the need to understand how delays propagate through transport networks. This perspective is supported by research of refs. [6,39].

Research Strategy: Data Selection and Analysis

Using the data selection and analysis method (PRISMA), Figure 4, 146 and 931 articles were obtained from Scopus and ScienceDirect, respectively, for a total of 1077 studies collected, with 42 duplicate articles and 576 excluded due to title and abstract analysis. The exclusion of the 576 records during the title and abstract screening phase primarily targeted studies that (a) did not focus on flight delay prediction, (b) did not employ a machine learning methodology, or (c) addressed logistics or transportation unrelated to commercial aviation. Regarding the review of the body of the text, as shown in Table 1, 134 articles did not present the information necessary for the collection, while 107 studies were outside the study topic analyzed, and 159 had different study types. This category includes articles that were classified as systematic reviews, literature reviews, editorial materials, or were publications of conference proceedings, as the inclusion criteria required original research articles published in peer-reviewed journals and 29 articles did not apply at least one machine learning method to perform delay prediction.

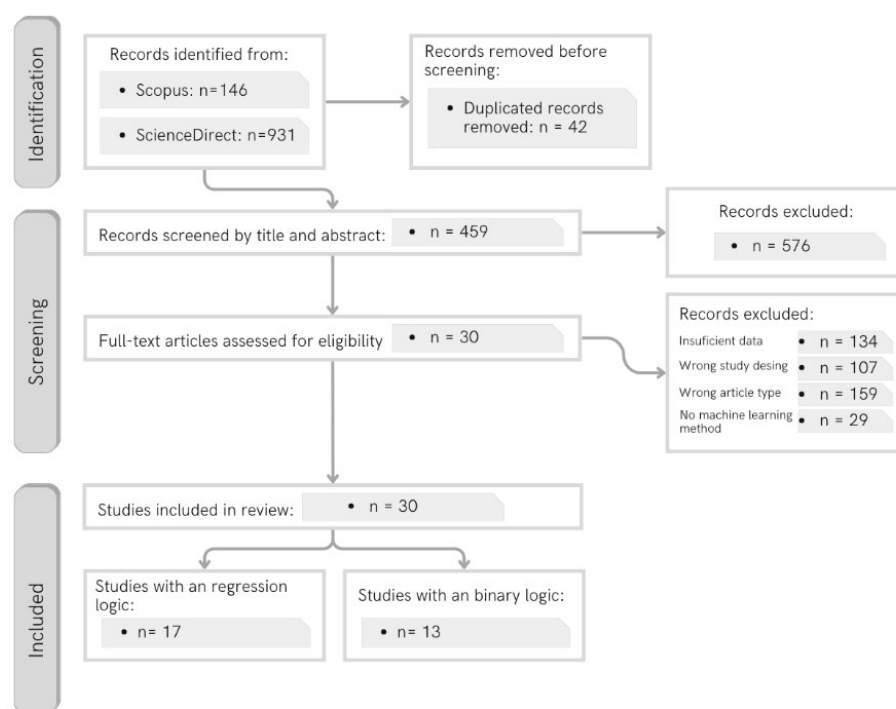


Figure 4. Data collection and study selection flowchart (PRISMA).

Thus, 30 studies were included as eligible for the proposed analyses. To improve the performance of the meta-analysis, the studies were divided into two groups according to the prediction logic used, with 17 studies using regression prediction logic and 13 using binary logic.

Sixty different machine learning methods were applied in the selected articles. Following the classifications presented by [41], the methods were classified into seven main groups according to the similarity of their characteristics and specific applications, Figure 5. This grouping was based on the primary function and structural architecture of the algorithms: Groups A and B aggregate ensemble and tree-based models (known for robustness), Group D aggregates deep learning (focused on generalization and sequential data), Group C aggregates vector machines (focused on classification boundaries), and Group E aggregates neural networks specialized in graph-format data.

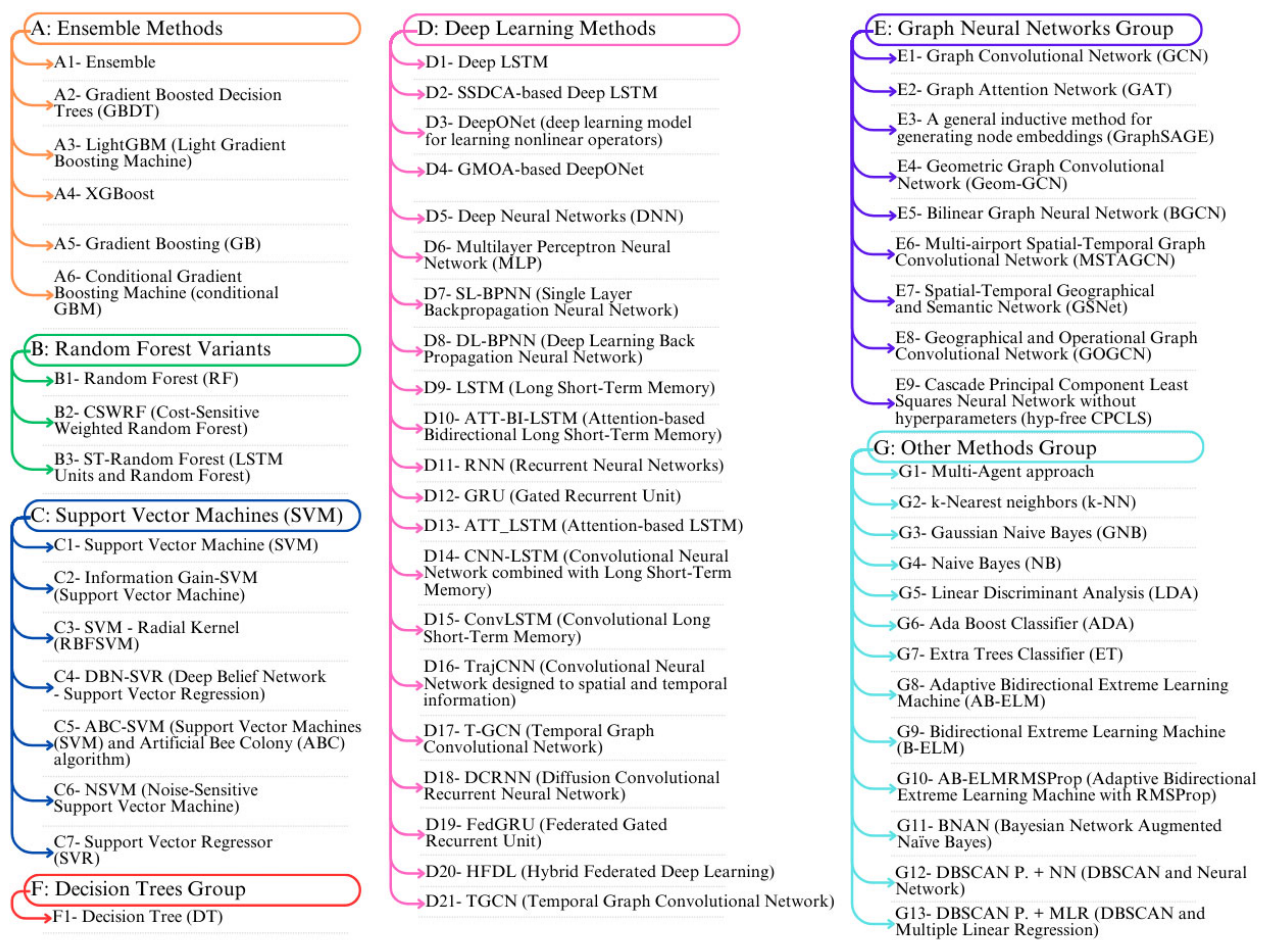


Figure 5. Groups of machine learning methods applied in selected delay prediction studies.

According to [41], ensemble methods such as Gradient Boosting and Random Forest combine multiple models to improve prediction accuracy, reduce variance and increase robustness. The Random Forest variant group, for example, uses multiple decision trees for classification and has specific variants for unbalanced data, such as CSWRF (Cost-Sensitive Weighted Random Forest), and for sequential data, such as ST-Random Forest, which incorporates LSTM. These methods are widely used because of their ability to capture complex patterns and adapt to different challenges, as discussed in the Machine Learning section.

On the other hand, other approaches such as SVMs (Support Vector Machines) and deep learning are also highlighted. SVMs, with variants such as radial kernel and NSVM

(noise-sensitive SVM), are effective in classification and regression tasks [42]. Deep learning methods, such as LSTM and CNN networks, are essential for recognizing patterns in sequential data, such as images and geographic information. Graph neural networks (GCN and GAT) are mentioned as an emerging technique for graph-format data [40,42].

Therefore, although the evolution of delay prediction studies indicates that classical methods, such as decision trees, k-NN, and Naive Bayes (see Table 2), remain relevant due to their simplicity and wide applicability, the emphasis still falls on the use of deep learning methods, which vary according to the configuration or specificity of the models applied.

In 2016, ref. [43] proposed a neural network that optimizes the use of nominal variables, thereby overcoming the multicollinearity problem that is introduced by traditional methods. This approach paved the way for the incorporation of more complex variables in delay prediction, as evidenced by ref. [44], who integrated multiple data sources. Nevertheless, the integration of a more extensive array of variables, despite the augmented generalization capacity of deep learning methodologies, necessitates the implementation of preprocessing techniques such as SMOTE for class balancing.

This approach builds upon ref. [45] by introducing a clustering of trajectory data, followed by a convolutional network (MCNN) to predict estimated time of arrival (ETA) based on air traffic patterns. Concurrently, ref. [36] substituted macroeconomic variables with microeconomic ones, such as airport crowding, to more effectively capture specific factors that directly influence delays. This approach is replicated by [10] that incorporates complex network theory into the detailed analysis of temporal and meteorological correlations of the data.

Conversely, studies such as that of ref. [46] propose a causal approach to preliminary analysis of relationships between data, underscoring the significance of mapping complex relationships between variables to comprehend how events influence each other, thereby contributing to a more explanatory model. In their 2021 study, ref. [47] expanded the scope of research by predicting delays in origin-destination pairs and at airports. They proposed an adapted variable selection scheme.

This adapted selection underwent further evolution in subsequent years, as evidenced by the study by ref. [48], which examined sequence data from ground support services and underscored the significance of critical nodes associated with departure delays. Ref. [49] adopts an explainable artificial intelligence (xAI) approach to integrate varied data, including geographic information and weather conditions, thereby enabling more accurate and understandable predictions.

The present study explores the integration of complex methods, as exemplified by ref. [50], which discusses the application of an integrated hierarchical model in predicting delay status and series duration, thereby reducing ambiguities in operation decisions. In a similar vein, ref. [11] explored comparisons between machine learning algorithms, determining that neural networks and ensemble methods presented the best accuracies when combined.

In 2024, ref. [12] advanced the development of combined techniques, proposing a parallel-series model capable of predicting not only the delay status, but also the reasons and durations of each type of delay coded by IATA. This model is an important differentiator for operational decisions. Concurrently, ref. [51] integrated learning techniques with support vector machines (SVMs) and parameter optimization with the artificial bee colony (ABC) algorithm, enabling comprehensive analysis of the factors that cause delays.

However, spatial and temporal integration models became a central focus from 2023 onwards. Among the studies, ref. [52] propose the inclusion of geographic and operational interactions in airport networks.

In the same line of research, in 2024, ref. [53] developed a model based on a Deep Residual Neural Network (DRN) that combines extrinsic variables and spatial and temporal correlations, facing spatial complexities that previous approaches neglected. Ref. [54] divided the airport network into sub-regions to facilitate predictions at a more granular level while respecting data privacy. In addition, ref. [55] focused on three aspects: real-time prediction, delay prediction and model explainability, to provide more transparency.

Refs. [56–58] focused on capturing spatial and temporal dependencies in more detail. Ref. [59] used the FAST-CA model to overcome traditional limitations, while ref. [56] implemented a causality-driven adjacency matrix to predict propagated delays in the airport network.

These innovations signify significant milestones in the direction of enhanced accuracy and detail in the prediction of flight delays, achieved by the integration of novel factors and analytical frameworks that expand the potential for the application of machine learning in the aviation sector.

Table 2. Summary of selected studies.

DMU	Forecast Logic	Methods Code	Analysis Area	Evaluation Metric	Sample Size	Impact Factor	Citation
Bisandu & Moulitsas (2024) [54]	R ¹	C4, A5, C2, G1, D1, D2, D3, D4	N ⁴	4.23%	3442	7.5	1
Cai et al. (2023) [53]	R ¹	B1, E1, E2, E3, E4, E5, E6, E7, E8	N ⁴	30.96%	1,800,000	5.3	14
Chen, Whang & Zhou (2021) [51]	R ¹	A1, F1, C1	Ap ⁵	74.00%	338,251	6.3	13
Falque, Mazure & Tabia (2024) [59]	R ¹	A3	Ap ⁵	71.47%	10,633,920	2.7	1
Khan et al. (2021) [11]	R ¹	E9, B2, C6, C6	Al ³	67.18%	19,109	7.6	44
Khan et al. (2024) [12]	R ¹	G10	Al ³	58.83%	21,987	3.9	5
Khanmohammadi, Tutun & Kucuk (2016) [44]	R ¹	D6	Ap ⁵	13.66%	1099	0	69
Li et al. (2024) [56]	R ¹	B1	N ⁴	97.00%	5,426,150	2.8	0
Schultz, Reitmann & Alam (2021) [58]	R ¹	D9, D11, D14, D15	Ap ⁵	63.35%	45,900	7.6	32
Shao et al. (2022) [60]	R ¹	D6, C7, A3, D16	Ap ⁵	40.79	27,500	5.5	30
Shen, Chen & Yan (2024) [55]	R ¹	D9, D17, D18, D19, D20	N ⁴	81.80%	12,339,775	7.2	0
Sun et al. (2024) [61]	R ¹	D21	N ⁴	74%	1,048,575	7.5	1
Wang, Liang & Delahaye (2018) [46]	R ¹	G12, G13	Ap ⁵	91.50%	8574	7.6	81
Yang et al. (2023) [49]	R ¹	B1, F1, A4, E1, E2, E3, E1 + E2	Ap ⁵	50%	124,453	7.6	3
Yu et al. (2019) [36]	R ¹	C4, G2, C1	Ap ⁵	89%	528,471	8.3	189
Henriques & Freitas (2018) [45]	B ²	F1, B1, D6	Ap ⁵	84.32%	248,956	0	15

Table 2. Cont.

DMU	Forecast Logic	Methods Code	Analysis Area	Evaluation Metric	Sample Size	Impact Factor	Citation
Khan et al. (2021) [11]	B ²	D7, D8, C1, E9, A1, B1, A2, A4	AI ³	63.73%	19,105	7.6	44
Khan et al. (2024) [12]	B ²	G8, G9, B1, A4, D5, G3, C1, G10	AI ³	64.94%	21,987	3.9	5
Lambelho et al. (2020) [62]	B ²	A3, D6, B1	Ap ⁵	77.63%	2,300,000	3.9	57
Li & Jing (2022) [10]	B ²	B3	N ⁴	90.50%	762,415	7.5	29
Li et al. (2022) [10]	B ²	D14	N ⁴	92.39%	5,426,150	7.5	26
Mamdouh, Ezzat & Hefny (2024) [63]	B ²	D10, D11, D9, D12, D13	Ap ⁵	76.10%	1,000,000	7.5	7
Mokhtarimousavi & Mehrabi (2023) [52]	B ²	C1, C5	Ap ⁵	90.40%	21,298	4.3	11
Pineda-Jaramillo et al. (2024) [50]	B ²	G2, F1, B1, A5, C3, G4, G5, G6, G7	Ap ⁵	63.38%	67,730	4.1	0
Schultz, Reitmann & Alam (2021) [58]	B ²	D9, D11, D14, D15	N ⁴	89.56%	45,900	7.6	32
Tenorio, Marques & Cadarso (2021) [48]	B ²	F1, B1, A5, D6, E1, D9	Ap ⁵	93.11%	7,000,000	0	1
Truong (2021) [47]	B ²	G11	N ⁴	91.97%	1058	3.9	41
Yanying, Mo & Haifeng (2019) [64]	B ²	F1, G4, C1	N ⁴	76.67%	5,635,978	0	9
Birolini & Jacquillat (2023) [65]	B ²	A1	N ⁴	77.30%	15,244	6	6

Forecast Logic: ¹ Regression; ² Binary. Analysis Area: ³ Airline; ⁴ Air Network; ⁵ Airport.

3.2. Meta-Analysis

The grouping of each study was conducted in accordance with the applied prediction logic, as illustrated in Figure 4. Subsequently, the values of the effect coefficient (d) and its variance were calculated. The findings suggest that, with a value of (d) set at 0.1, there is evidence of a small effect between the two groups. As ref. [26] describes, values of (d) close to 0 suggest little or no difference between groups, while larger values indicate a more significant effect. Consequently, a (d) equal to 0.1 suggests that, although negligible, a distinction can be perceived between continuous and binary prediction methods.

Upon examining the variance of (d) between the proposed groups, a minimal value was obtained (2.27×10^{-3}), indicating minimal variability in the effect size and suggesting that the estimate of (d) is highly precise. This finding suggests that the observed difference between the groups is consistent, with minimal uncertainty surrounding the estimate. Consequently, the division of the studies for the meta-analysis into two groups, according to the prediction logic adopted, is valid and significant.

Consequently, the meta-analyses were constructed under the assumption of heterogeneity in prediction studies, as indicated by ref. [23]. The configuration of the meta-analyses, presented in Table 3, identifies each study by means of the authors, year of publication and method analyzed (a study may include more than one machine learning method). The ratio (O/E Ratio) between the observed units (correctly predicted units) and the expected units to be correctly predicted by the model (sample tested by the studies) was used as a measure.

Table 3. Settings adopted by meta-analyses.

Model Components	Adopted Variable
Number of participants	Total sample
Number of events observed	Units of correct predictions
Number of events expected	Test sample
Study identifier	Authors + Year + Method

The random distribution method employed in this study is the ref. [61], which assumes that the variation between studies is of the Normal/Log type. Consequently, the findings of the meta-analyses, as illustrated in Figure 6, demonstrate a 95% confidence interval (CI) with an average value of 0.74 for studies utilizing binary prediction logic (a), exhibiting a variation between [0.70, 0.78]. In contrast, studies employing regression/continuous prediction logic exhibit an average CI of 0.6462, with a variation [0.5258, 0.794].

**Figure 6.** Results of meta-analyses for the group of articles that used (a) binary and (b) regression prediction logic [10–12,29,36,44–56,58–62,65,66].

The discrepancy between the CIs is indicative of the anticipated behavior when comparing groups.

A granular analysis of performance differences across machine learning methods is visually detailed in Figure 6, which plots the Observed/Expected Ratio for each specific model. Furthermore, when compared the Efficacy (accuracy) against Meta-frontier Efficiency, illustrating that while Deep Learning methods (Group D) achieve high efficacy in both logics, simpler methods (Groups A and B) also reach top-tier accuracy in the binary classification group. This variation shows that model grouping, rather than prediction logic alone, determines the highest predictive ceiling.

In a group-specific analysis, the results of the meta-analysis for the binary studies (see Figure 6a) indicate a positive observed effect in the studies analyzed. This is indicated by all Observed-Expected Ratio results (central interval below each figure) being above 0 and below 0.9.

In relation to the confidence interval (CI) ranging from 0.705 to 0.78, this indicates a 95% confidence level that the true effect of binary prediction is confined to this interval. As the range does not include 1.0, it can be deduced that the observed effect is statistically significant, thereby indicating the efficacy of binary prediction. The mean value of 0.741 indicates that, whilst there is a positive effect, it is not of a significant magnitude.

The results of the meta-analysis for the group of studies employing regression-type prediction logic (see Figure 6b) indicate that the observed-expected ratio results are all below 1 and above 0. This finding suggests that, on average, prediction studies utilizing regression logic demonstrate inferior performance in comparison to expectations, thereby implying that prediction may not attain the same level of effectiveness as binary prediction.

A value greater than 0 indicates a positive effect, whilst a value less than 1 suggests that the effect is limited and that the model's performance is deemed unsatisfactory or that there is scope for improvement, as discussed by ref. [26].

To validate the publication biases between the groups, the Egger's tests were applied. The outcomes of this analysis, as illustrated in Figure 7, reveal asymmetry, thereby confirming the existence of a publication bias. This finding challenges the null hypothesis, which posits the absence of such bias.

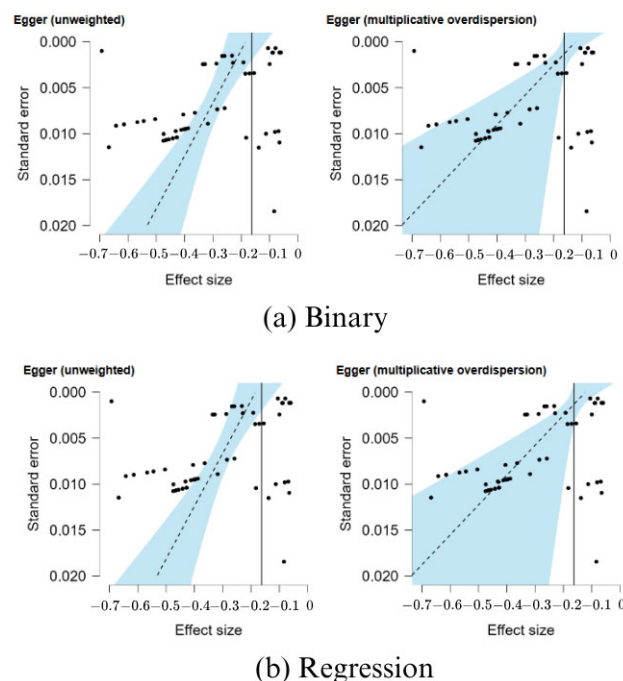


Figure 7. Funnel Plot and Egger tests for the study group with binary logic (a) and regression (b).

The biased characteristic of the analyzed publications is common to studies with heterogeneity, as described by ref. [14].

3.3. Data Envelopment Analysis (DEA)

The efficiency results were analyzed through a joint assessment of the meta frontier, a methodological approach that enabled the hierarchization of the applied methods, irrespective of the forecasting approach employed (see Figure 8).

As demonstrated in Figure 8a, studies that employ binary prediction logic exhibit enhanced efficiency responses when implementing more sophisticated machine learning methodologies, such as those classified in Group D. The DEA shows that the integration approach of machine learning methods proposed by refs. [10,11] and by ref. [51] between the methods of Groups D and C is a common characteristic of both efficient and inefficient models. However, the presence of simpler methods among the efficient data management units (DMUs) in the group, such as the machine learning (ML) methods in groups A and B, suggests that the efficiency of the method is not solely dependent on its complexity.

The group of studies that employ regression-type forecasting logic, as illustrated in Figure 8b, exhibits lower efficiency in comparison to Group (a). The performance of these models is generally lower due to the complexity of the output variables and the need for larger samples.

Furthermore, these models frequently necessitate forecasts with a more extended anticipation horizon compared to the models in the binary group. In this context, the three benchmarks belonging to the regression group undertook a thorough preliminary analysis of the problem to be predicted.

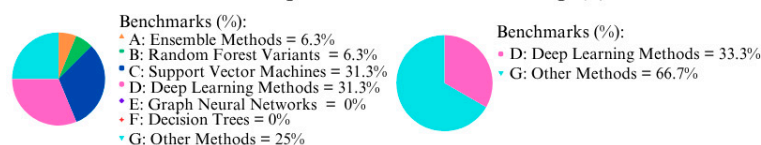
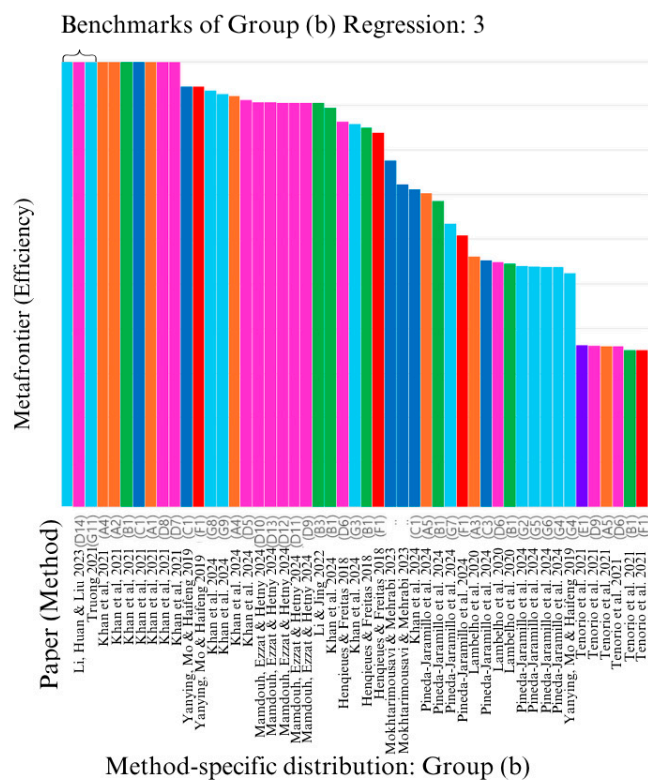
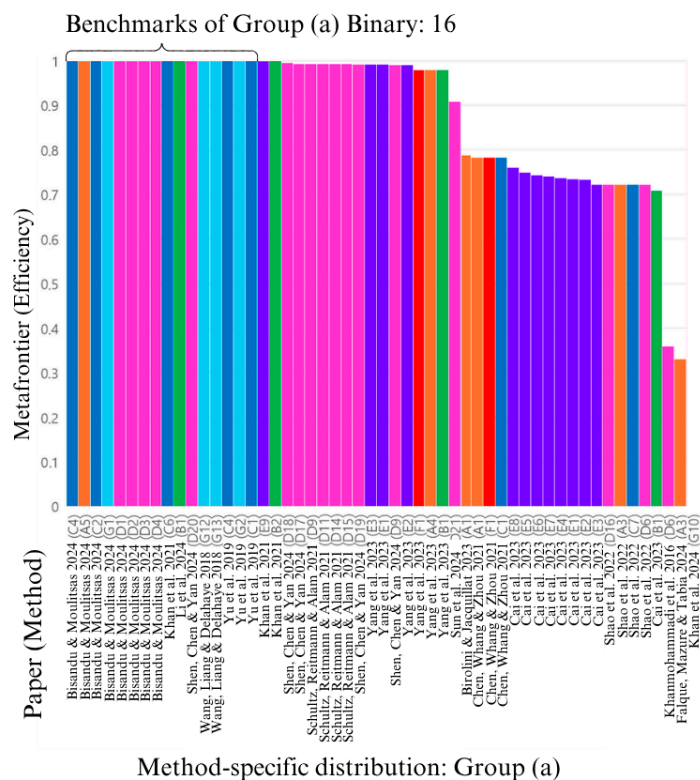
Consequently, when assessing the efficacy of machine learning (ML) algorithms in predicting flight delays, meta-analyses provide validation intervals that quantify the effectiveness of each method. Additionally, the efficiency evaluated by the data envelopment analysis (DEA) method introduces a novel dimension by assessing the individual capacity of each study to transform selected variables into the desired response.

In this manner, the methods are compared by the relationship between efficiency and effectiveness, as illustrated in Figure 9.

Regardless of the prediction logic adopted, it is observed that the balance between the predictive capacity and the efficiency of the sample in generating responses is predominant among the methods in group D.

As a general tendency, graphical neural networks exhibit diminished production capacity for binary logic, as illustrated in Figure 9a. Conversely, the presence of simpler methods in the high-effectiveness and high-efficiency quadrant serves to reinforce the hypothesis that the balance in the performance of prediction models does not necessarily depend on the complexity of the applied machine learning (ML) methods.

The greater dispersion in the production analysis of machine learning (ML) methods that use regression logic, as shown in Figure 9b, is consistent with the lower means of effectiveness and efficiency compared to group (a). The variability in the production scale of the methods in the regression group can be explained by the differences in the outputs, which, although they aim to predict flight delays, vary in dimension (quantity), space (analyzed area), and object (classification to be generalized by the predictive model).



Distribution of Methods Classes between Benchmarks

Figure 8. Meta frontier efficiencies: joint analysis of (a) binary and (b) regression forecast logic groups, [10–12,29,36,44–56,58–63,65,66].

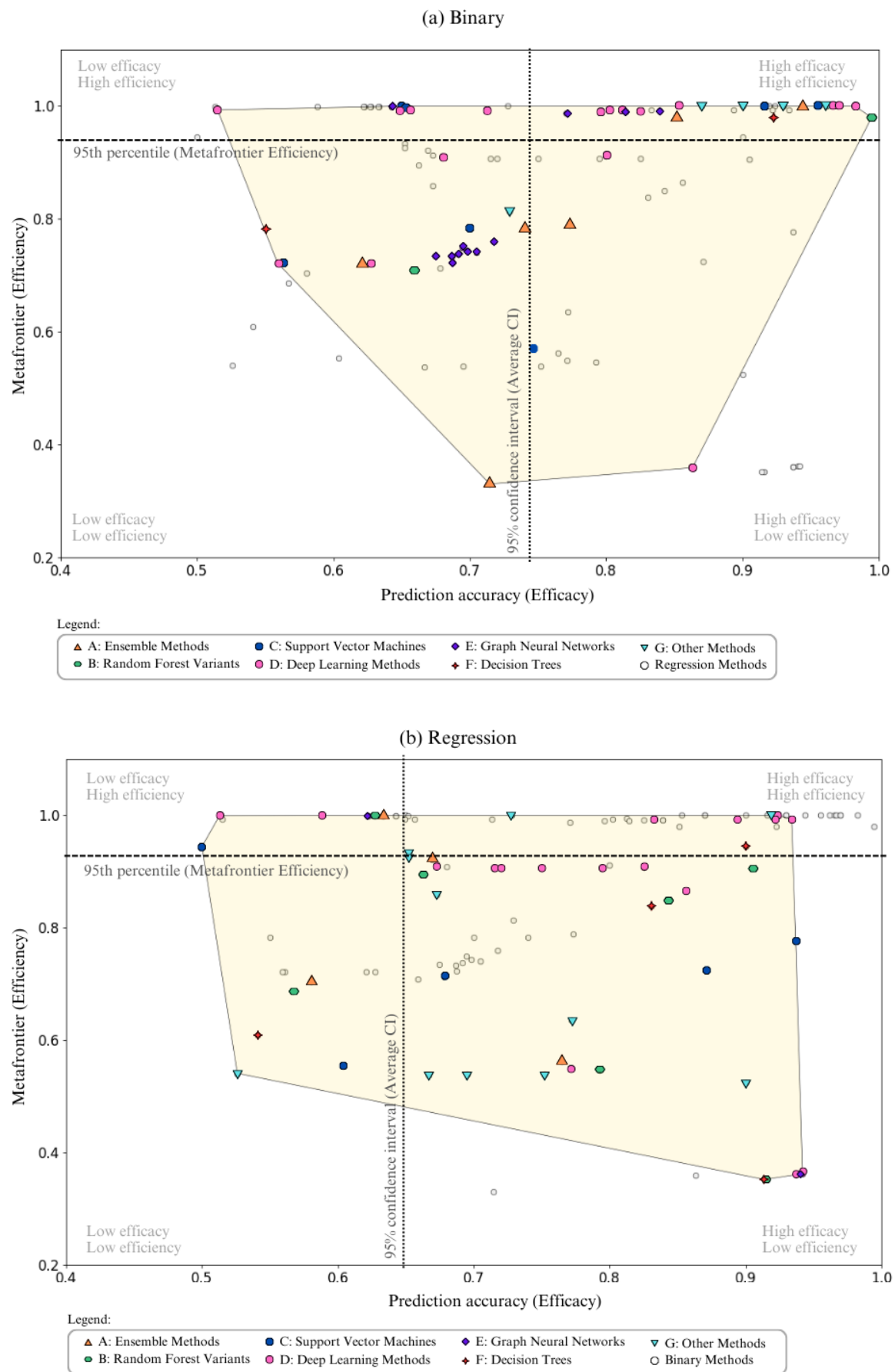


Figure 9. Comparison of Effectiveness vs. Efficiency of ML methods. Groups of methods with (a) binary and (b) regression prediction logic.

3.4. Censored Tobit Regression Model

To ascertain the impact of the groups of variables presented in Table 1 on the efficiency of delay prediction methods by machine learning, two censored regression analysis models of the Tobit type were constructed. In these models, the dependent variable is constrained to non-negative values. In this instance, the efficiency of the meta frontier is measured by DEA.

The models constructed in Table 4 focus on the response of the efficiency of the methods to factors exogenous and endogenous to the applied method.

Table 4. The following compares the proposed censored regression models.

Variable	Tobit Models—Factors Influencing Efficiency			
	Model 1	<i>p</i> -Value	Model 2	<i>p</i> -Value
Constant	0.4151	3.32×10^{-16} ***	0.1883	0.0162 **
Dummy (Regression)	−0.0404	0.0322 **	−	−
Prediction Accuracy	0.4068	1.89×10^{-17} ***	0.4547	8.72×10^{-30} ***
Total Sample	-7.477×10^{-8}	2.02×10^{-28} ***	-7.144×10^{-8}	1.18×10^{-32} ***
Correctly Predicted Instances	2.603×10^{-7}	6.27×10^{-14} ***	2.18701×10^{-7}	3.18×10^{-12} ***
Dummy (Airline)	−0.0468	0.0154 **	−	−
Dummy Var. (Aircraft)	0.0711	0.0004 ***	−	−
Dummy Var. (PAX)—Passengers	−0.1812	1.89×10^{-8} ***	−	−
Dummy Var. (Meteor.)—Meteorology/Weather	0.1035	2.44×10^{-6} ***	0.1450	2.49×10^{-15} ***
Dummy Var. (Airport Capacity.)	0.1722	4.49×10^{-16} ***	0.1631	1.41×10^{-21} ***
Dummy Var. (Flight data)	−	−	0.2362	0.0009 ***
Dummy Var. (Air Routes)	−	−	−0.1618	1.39×10^{-19} ***
Dummy Var. (Air Traffic)	−	−	0.0774	8.16×10^{-5} ***
Dummy Var. (Previous Delays)	−	−	0.0405	0.0224 **
Impact Factor	−0.0097	0.0148 **	−0.0354	3.86×10^{-21} ***
Quotes	−	−	0.0007	0.0008 ***
n (number of observations)	1–109		1–109	
Chi-square (10)	565,488		780,543	
Log Likelihood	127,572		142,655	

Legend: * significant at the 10 percent level; ** significant at the 5 percent level; *** significant at the 1 percent level.

The regression analysis of model 1 aims to identify the method that will maximize the performance of the predictive model, ensuring that the selected inputs are aligned with the specificities of the model and its ability to produce robust results.

The negative value of the Dummy (Regression) variable and the Total Sample coefficient (also negative) indicate a reduction in efficiency in these contexts, despite positive coefficients for accuracy and correctly predicted instances [41]. Conversely, Model 1 demonstrates, with a highly significant *p*-value (0.0154), that analyses centered on a specific airport (as opposed to an airline) exhibit enhanced efficiency [16].

Regarding the group of variables, both Tobit models indicate an increase in efficiency with the use of meteorological information and data on airport capacity. The analysis also reveals that the coefficients associated with the enhancement in efficiency are comparatively diminished for additional variables, with certain variables, such as Aircraft Rotation Data and Passenger data, even exhibiting negative values. The number of citations demonstrates a positive relationship with the efficiency of the predictive models, as indicated in model 2, while the Impact Factor shows a negative relationship.

In consideration of the limitations inherent in the defined models, the normality tests of the residuals indicate that the hypothesis of a normal distribution of the errors is to be

rejected. This may have ramifications for the validity of the significance tests; however, as previously mentioned in other contexts, the Tobit model may still be applicable even in cases where normality is not met.

4. Discussion: Implications for Efficiency and Variable Selection

The comprehensive analysis (combining meta-analysis, DEA, and Tobit regression) yields three primary findings that challenge the conventional pursuit of model complexity in flight delay prediction. First, binary classification achieves superior average efficacy compared to regression models. Second, DEA establishes that model complexity is not synonymous with operational efficiency. Third, Tobit regression identifies that the selective use of contextual data is a positive driver of efficiency, while reliance on large samples can be detrimental.

These findings prompt critical discussion regarding the underlying mechanisms of predictive success in the aviation sector:

4.1. *The Efficacy and Efficiency of Binary Classification*

The results of the meta-analysis establish that binary classification approaches achieve higher average efficacy compared to regression models. This finding, where simpler logic outperforms more resource-intensive approaches, warrants closer examination. The superior efficacy of binary classifiers can be primarily attributed to the defined scope of the prediction task. By focusing solely on classifying the presence or absence of a delay (“0” or “1”), the model’s generalization capacity is maximized, reducing the need for the extensive computational complexity required by regression models [27], which predict a continuous duration of delay.

Furthermore, regression logic studies have been shown to predominantly use a greater number of advanced methods [57], incorporating the largest number of variables, and consequently, larger samples [64].

From an operational standpoint, a binary outcome aligns directly with a fundamental need of the airline industry: the necessity to make immediate operational decisions. For instance, decisions regarding flight schedule adjustments, resource reallocation [67], and initial crew adjustments require only a simple binary signal, rather than the exact duration. Consequently, the effectiveness of these models aligns directly with a fundamental practical need of the industry, suggesting that the effectiveness of these models aligns directly with a fundamental practical need of the industry, thereby enhancing the perceived utility and effectiveness of binary classification [68].

4.2. *The Value of Parsimony over Complexity*

The results from the DEA meta-frontier analysis reveal that model complexity is not synonymous with operational efficiency. This phenomenon is empirically validated by the presence of simpler methods, such as those from Ensemble Methods (Group A) and Random Forest Variants (Group B), achieving high effectiveness and efficiency benchmarks. This observation is critical because it supports the hypothesis that the efficiency of a method does not depend exclusively on its generalization potential or iteration capacity.

The value of parsimony arises because efficiency is maximized by a precise and limited set of explanatory variables. In contrast, models relying on extensive data often face the challenge of redundancy and noise, which can compromise the predictive capacity of the applied model. This inability to generalize responses ultimately gives rise to a second problem, namely that a large mass of data presents a considerable amount of redundant information. Consequently, the performance of complex methods is hin-

dered when the computational cost and processing effort do not yield commensurate practical benefits.

Furthermore, for regression models seeking a higher degree of granularity and continuous output, the lower efficiency found by the DEA underscores the necessity of a thorough preliminary analysis. The high variability in the production scale of regression methods can be mitigated when studies undertake a causal approach to comprehend the interrelationships among the data. Mapping complex relationships between variables should be the starting point in analyses based on regression logic to ensure the efficacy of the model. Thus, parsimony, when achieved through robust variable selection, leads to optimization that ensures results are achieved with greater precision and less operational complexity.

4.3. Balancing Predictive Power with Operational Cost

The combined insights from the DEA and Tobit regression models allow this study to establish a comprehensive framework for balancing predictive power with the operational cost inherent in data collection and complexity. The Censored Tobit regression model acts as a crucial analytical bridge, quantifying how specific structural decisions and data clusters influence the efficiency derived from DEA.

The analysis confirms that the pursuit of prediction accuracy must be tempered by a cost-sensitive variable selection approach. The negative coefficient observed for the Total Sample size and the Regression Logic dummy variable indicates that, despite their potential for generalization, the marginal cost of implementing these extensive inputs often outweighs the marginal gain in efficiency. This constraint aligns with the necessity of making predictions in a timely manner, where the temporal demands of effective data processing impede the ability to make timely predictions [37].

Conversely, the model highlights specific variables whose information value justifies their acquisition cost. The positive and significant coefficients associated with Meteorological Information and Airport Capacity Data demonstrate that these data clusters are direct drivers of efficiency. These variables are directly associated with the common causes of delays in the area under analysis [52], proving that when the variables are directly associated with the core cause of the delay, they enhance the reliability of the study [11].

However, the analysis of additional operational variables, such as Aircraft Rotation Data and Passenger Data, reveals a distinct trade-off: their coefficients exhibit negative or diminished values. The findings suggest that the employment of these variables can, in numerous instances, become negligible or even contribute as noise, thereby diminishing the efficacy and efficiency of the models. This complexity is further underscored by the fact that efficiency is enhanced in analyses centered on a specific airport (instead of an airline), suggesting that the operational control mechanisms and reduced variability in data [44,51] contribute to enhanced reliability of the findings.

Therefore, the ultimate strategy for achieving robust predictive performance lies in identifying and prioritizing a parsimonious cluster of high-impact variables (like those related to Meteorology and Airport Capacity) while resisting the temptation to incorporate highly complex or high-cost data inputs that reduce operational efficiency.

5. Conclusions

This study conducted a comprehensive meta-analysis on the application of machine learning methods in flight delay prediction, revealing a significant growth in research output since 2020, with a total of 1077 studies collected. The findings of the study indicated that binary prediction methods exhibited a higher mean accuracy in comparison to regression models. This observation is supported by a 95% confidence interval, which signifies the efficacy of binary methods. Furthermore, data envelopment analysis (DEA) demonstrated

that the efficiency of complex models does not necessarily translate into better predictive results, thereby highlighting the importance of careful variable selection.

The analyses conducted are pivotal for the advancement of research in flight delay prediction, as they provide a solid foundation for the selection of more effective machine learning methods. Moreover, the analysis of heterogeneity across studies underscores the necessity for more integrated and adaptive approaches in flight delay research.

Notwithstanding its substantial contributions, this study is not without its limitations. The heterogeneity of the methods and the variability in the samples used in the studies analyzed may have influenced the results of the meta-analysis, making direct comparisons between different approaches difficult. Furthermore, the utilization of published data may introduce publication bias, thereby limiting the generalizability of the findings.

The DEA meta-frontier model, while robust, assumes that all decision-making units operate at an optimal scale and is sensitive to the number of studies and variables included in the input/output set, potentially overstating the efficiency of certain DMUs. Furthermore, while we utilized the Variance Inflation Factor (VIF) to monitor multicollinearity in the Tobit regression, the non-normal distribution of the residuals, though acceptable for Tobit analysis in some contexts, may affect the validity of the significance tests. This necessitates a cautious interpretation of the determinants of efficiency.

In the context of future research endeavors, the findings establish a clear agenda focused on efficiency and resource allocation. Future research lines should include: (1) Creating a formal framework for cost-sensitive variable selection; (2) Researching hybrid predictive pipelines; and (3) Applying the combined DEA and Tobit methodology to other efficiency-accuracy trade-off sensitive fields.

In summary, the present study offers a comprehensive overview of contemporary practices in the domain of flight delay prediction, thereby establishing a benchmark for future research in this field. The practical application of the findings could lead to significant improvements in airline operational efficiency and passenger experience, thereby underscoring the continued relevance of machine learning in the aviation industry. The adoption of more effective methods and the consideration of contextual variables are essential steps in addressing the growing challenges of modern aviation.

Author Contributions: Conceptualization, H.d.S.Q.J. and V.F.; methodology, H.d.S.Q.J., V.F. and F.G.F.d.S.; software, H.d.S.Q.J. and V.F.; validation, V.F. and F.G.F.d.S.; formal analysis, H.d.S.Q.J.; investigation, H.d.S.Q.J., V.F. and F.G.F.d.S.; resources, I.M.T.B. and J.K.L.M.; data curation, H.d.S.Q.J.; writing—original draft preparation, H.d.S.Q.J.; writing—review and editing, I.M.T.B. and J.K.L.M.; visualization, V.F. and F.G.F.d.S.; supervision, V.F. and F.G.F.d.S.; project administration, V.F.; funding acquisition, I.M.T.B. and J.K.L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external public funding. The cost of publication was financially supported by Zenite Engenharia e Consultoria.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because access is restricted for some of the publications. Requests to access the datasets should be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, C.; Wang, X. Airport congestion delays and airline networks. *Transp. Res. Part E* **2019**, *122*, 328–349. [\[CrossRef\]](#)
2. Liu, Y.; Yin, M.; Hansen, M. Economic costs of air cargo flight delays related to late package deliveries. *Transp. Res. Part E* **2019**, *125*, 388–401. [\[CrossRef\]](#)
3. de Almeida, E.E.; Oliveira, A.V.M. An econometric analysis for the determinants of flight speed in the air transport of passengers. *Sci. Rep.* **2023**, *13*, 4573. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Wang, Z.; Liao, C.; Hang, X.; Li, L.; Delahaye, D.; Hansen, M. Distribution prediction of strategic flight delays via machine learning methods. *Sustainability* **2022**, *14*, 15180. [\[CrossRef\]](#)
5. Chu, A.-M. *A Bayesian Network Approach to Identify Influential Factors for Flight Delays*; Massachusetts Institute of Technology: Cambridge, UK, 2012.
6. Jebbor, I.; Hachimi, H.; Benmamoun, Z. Artificial Intelligence in Predicting Automotive Supply Chain Disruptions: A Literature Review. In *Advances in Intelligent Systems and Digital Applications*; Springer: Cham, Switzerland, 2025; pp. 1–13. [\[CrossRef\]](#)
7. Jebbor, I.; Benmamoun, Z.; Hachimi, H. Forecasting supply chain disruptions in the textile industry using machine learning: A case study. *Ain Shams Eng. J.* **2024**, *15*, 103116. [\[CrossRef\]](#)
8. Carvalho, L.; Sternberg, A.; Gonçalves, L.M.; Cruz, B.; Soares, J.A.; Brandão, D.; Carvalho, D.; Ogasawara, E. On the relevance of data science for flight delay research: A systematic review. *Transp. Rev.* **2020**, *41*, 852–872. [\[CrossRef\]](#)
9. Rebollo, J.J.; Balakrishnan, H. Characterization and prediction of air traffic delays. *Transp. Res. Part C* **2014**, *44*, 231–241. [\[CrossRef\]](#)
10. Li, Q.; Jing, R. Flight delay prediction from spatial and temporal perspective. *Expert Syst. Appl.* **2022**, *201*, 117662. [\[CrossRef\]](#)
11. Khan, W.A.; Ma, H.-L.; Chung, S.-H.; Wen, X. Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. *Transp. Res. Part C* **2021**, *129*, 103225. [\[CrossRef\]](#)
12. Khan, W.A.; Chung, S.-H.; Eltoukhy, A.E.E.; Khurshid, F. A novel parallel series data-driven model for IATA-coded flight delays prediction and features analysis. *J. Air Transp. Manag.* **2024**, *114*, 102488. [\[CrossRef\]](#)
13. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#)
14. Okraszewska, R.; Romanowska, A.; Laetsch, D.C.; Gobis, A.; Reisch, L.A.; Kamphuis, C.B.; Lakerveld, J.; Krajewski, P.; Banik, A.; Braver, N.R.D.; et al. Interventions reducing car usage: Systematic review and meta-analysis. *Transp. Res. D Transp. Environ.* **2024**, *131*, 104217. [\[CrossRef\]](#)
15. Christopher, J.O.; Prasada Rao, D.S.; George, E.B. Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. *Empir. Econ.* **2008**, *34*, 231–255. [\[CrossRef\]](#)
16. Rudolpho, C.W.; Katz, I.M.; Lavigne, K.N.; Zacher, H. Job crafting: A meta-analysis of relationships with individual differences, job characteristics, and work outcomes. *J. Vocat. Behav.* **2017**, *102*, 112–138. [\[CrossRef\]](#)
17. Reljic, T.; Sehovic, M.; Lancet, J.; Kim, J.; Ali, N.A.; Djulbegovic, B.; Extermann, M. Benchmarking treatment effects for patients over 70 with acute myeloid leukemia: A systematic review and meta-analysis. *J. Geriatr. Oncol.* **2020**, *11*, 1293–1308. [\[CrossRef\]](#)
18. Reiss, R.; Gaylor, D. Use of benchmark dose and meta-analysis to determine the most sensitive endpoint for risk assessment for dimethoate. *Regul. Toxicol. Pharmacol.* **2005**, *42*, 55–65. [\[CrossRef\]](#)
19. Khoja, L.; Atenafu, E.G.; Suciu, S.; Leyvraz, S.; Sato, T.; Marshall, E.; Keilholz, U.; Zimmer, L.; Patel, S.P.; Piperno-Neumann, S.; et al. Meta-analysis in metastatic uveal melanoma to determine progression free and overall survival benchmarks: An international rare cancers initiative (IRCI) ocular melanoma study. *Ann. Oncol.* **2019**, *30*, 1370–1380. [\[CrossRef\]](#)
20. Papalia, G.F.; Brigato, P.; Sisca, L.; Maltese, G.; Faiella, E.; Santucci, D.; Pantano, F.; Vincenzi, B.; Tonini, G.; Papalia, R.; et al. Artificial Intelligence in Detection, Management, and Prognosis of Bone Metastasis: A Systematic Review. *Cancers* **2024**, *16*, 2700. [\[CrossRef\]](#)
21. Jaiteh, M.; Phalane, E.; Shiferaw, Y.A.; Voet, K.A.; Phaswana-Mafuya, R.N. Utilization of Machine Learning Algorithms for the Strengthening of HIV Testing: A Systematic Review. *Algorithms* **2024**, *17*, 362. [\[CrossRef\]](#)
22. Akinsoji, A.H.; Adelodun, B.; Adeyi, Q.; Salau, R.A.; Odey, G.; Choi, K.S. Integrating Machine Learning Models with Comprehensive Data Strategies and Optimization Techniques to Enhance Flood Prediction Accuracy: A Review. *Water Resour. Manag.* **2024**, *38*, 4735–4761. [\[CrossRef\]](#)
23. Arévalo, P.; Ochoa-Correa, D.; Villa-Ávila, E. A Systematic Review on the Integration of Artificial Intelligence into Energy Management Systems for Electric Vehicles: Recent Advances and Future Perspectives. *World Electr. Veh. J.* **2024**, *15*, 364. [\[CrossRef\]](#)
24. Utami, W.; Sugiyanto, C.; Rahardjo, N. Artificial intelligence in land use prediction modeling: A review. *IAES Int. J. Artif. Intell.* **2024**, *13*, 2514–2523. [\[CrossRef\]](#)
25. Yaghoubi, E.; Yaghoubi, E.; Khamees, A.; Razmi, D.; Lu, T. A systematic review and meta-analysis of machine learning, deep learning, and ensemble learning approaches in predicting EV charging behavior. *Eng. Appl. Artif. Intell.* **2024**, *135*, 108789. [\[CrossRef\]](#)

26. Borenstein, M.; Hedges, L.V.; Higgins, J.P.T.; Rothstein, H.R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* **2010**, *1*, 97–111. [CrossRef] [PubMed]
27. Hedges, L.V.; Olkin, I. *Statistical Methods for Meta-Analysis*; Academic Press, Inc.: Orlando, FL, USA, 1985.
28. Arjomandi, A.; Dakpo, K.; Seufert, J. Have Asian airlines caught up with European airlines? A by-production efficiency analysis. *Transp. Res. Part A* **2018**, *111*, 389–403. [CrossRef]
29. Shah, W.U.H.; Hao, G.; Yan, H.; Shen, J.; Yasmeen, R. Forestry Resource Efficiency, Total Factor Productivity Change, and Regional Technological Heterogeneity in China. *Forests* **2024**, *15*, 152. [CrossRef]
30. Zubir, M.Z.; Noor, A.A.; Rizal, A.M.M.; Harith, A.A.; Abas, M.I.; Zakaria, Z.; Bakar, A.F.A. Approach in inputs & outputs selection of Data Envelopment Analysis (DEA) efficiency measurement in hospitals: A systematic review. *PLoS ONE* **2024**, *19*, e0293694.
31. Yen, B.; Li, J.-S. Route-based performance evaluation for airlines—A metafrontier data envelopment analysis approach. *Transp. Res. Part E* **2022**, *161*, 102706. [CrossRef]
32. Barbosa, F.C.; Fuchigami, H.Y. *Análise Envoltória de Dados: Teoria e Aplicações Práticas*; ULBRA: Itumbiara, Brazil, 2018.
33. Meza, L.A.; Gomes, E.G.; Neto, L.B. Curso de Análise de Envoltória de Dados. XXXVII Simpósio Brasileiro de Pesquisa Operacional. 2005; pp. 2520–2547. Available online: <http://www.din.uem.br/~ademir/sbpo/sbpo2005/pdf/arq0289.pdf> (accessed on 14 July 2021).
34. Greene, W.H. *Econometric Analysis*, 5th ed.; Prentice Hall: Hoboken, NJ, USA, 2003; pp. 145–198.
35. Wooldridge, J.M. *Introdução à Econometria*; Editora Thomson; Cengage: San Francisco, CA, USA, 2006; pp. 295–328.
36. Yu, B.; Guo, Z.; Asian, S.; Wang, H.; Chen, G. Flight delay prediction for commercial air transport: A deep learning approach. *Transp. Res. Part E* **2019**, *125*, 203–221. [CrossRef]
37. Miyawaki, K.; MacEachern, S.N. Economic variable selection. *Can. J. Stat.* **2023**, *51*, 19–37. [CrossRef]
38. Halpern, N.; Mwesiumo, D.; Budd, T.; Suau-Sanchez, P.; Bråthen, S. Segmentation of passenger preferences for using digital technologies at airports in Norway. *J. Air Transp. Manag.* **2021**, *91*, 102005. [CrossRef]
39. Hatipoğlu, I.; Tosun, Ö. Predictive Modeling of Flight Delays at an Airport Using Machine Learning Methods. *Appl. Sci.* **2024**, *14*, 5472. [CrossRef]
40. Callahan, M.; Wears, R.L.; Weber, E. Factors associated with postpublication citation. *JAMA* **2002**, *287*, 2847–2850. [CrossRef] [PubMed]
41. Sternberg, A.; Soares, J.; Carvalho, D.; Ogasawara, E. A Review on Flight Delay Prediction. *arXiv* **2022**, arXiv:1703.06118.
42. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Science + Business Media, LLC: Berlin/Heidelberg, Germany, 2006; pp. 1–729.
43. Zhou, Z.-H. *Machine Learning*; Springer: Singapore, 2021; pp. 1–536.
44. Khanmohammadi, S.; Tutun, S.; Kucuk, Y. A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. *Procedia Comput. Sci.* **2016**, *95*, 237–244. [CrossRef]
45. Henriques, R.; Feiteira, I. Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport. *Procedia Comput. Sci.* **2018**, *138*, 638–645. [CrossRef]
46. Wang, Z.; Liang, M.; Delahaye, D. A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area. *Transp. Res. Part C* **2018**, *95*, 280–294. [CrossRef]
47. Truong, D. Using causal machine learning for predicting the risk of flight delays in air transportation. *J. Air Transp. Manag.* **2021**, *91*, 101993. [CrossRef]
48. Tenorio, V.M.; Marques, A.G.; Cadarso, L. Signal processing and machine learning for air traffic delay prediction. *Transp. Res. Procedia* **2021**, *58*, 463–470. [CrossRef]
49. Yang, Z.; Chen, Y.; Hu, J.; Song, Y.; Mao, Y. Departure delay prediction and analysis based on node sequence data of ground support services for transit flights. *Transp. Res. Part C* **2023**, *153*, 104217. [CrossRef]
50. Pineda-Jaramillo, J.; Munoz, C.; Mesa-Arango, R.; Gonzalez-Calderon, C.; Lange, A. Integrating multiple data sources for improved flight delay prediction using explainable machine learning. *Res. Transp. Bus. Manag.* **2024**, *56*, 101161. [CrossRef]
51. Chen, Z.; Wang, Y.; Zhou, L. Predicting weather-induced delays of high-speed rail and aviation in China. *Transp. Policy* **2021**, *101*, 14–22. [CrossRef]
52. Mokhtarimousavi, S.; Mehrabi, A. Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. *Int. J. Transp. Sci. Technol.* **2023**, *12*, 230–244. [CrossRef]
53. Cai, K.; Li, Y.; Zhu, Y.; Fang, Q.; Yang, Y.; Du, W. A geographical and operational deep graph convolutional approach for flight delay prediction. *Chin. J. Aeronaut.* **2023**, *36*, 17–31. [CrossRef]
54. Bisandu, D.B.; Moulitsas, I. Prediction of flight delay using deep operator network with gradient-mayfly optimisation algorithm. *Expert Syst. Appl.* **2024**, *247*, 123306. [CrossRef]
55. Shen, X.; Chen, J.; Yan, R. A spatial-temporal model for network-wide flight delay prediction based on federated learning. *Appl. Soft Comput.* **2024**, *154*, 111380. [CrossRef]

56. Li, C.; Mao, J.; Li, L.; Wu, J.; Zhang, L.; Zhu, J.; Pan, Z. Flight delay propagation modeling: Data, Methods, and Future opportunities. *Transp. Res. E Logist. Transp. Rev.* **2024**, *185*, 103525. [\[CrossRef\]](#)
57. Qu, J.; Wu, S.; Zhang, J. Flight Delay Propagation Prediction Based on Deep Learning. *Mathematics* **2023**, *11*, 494. [\[CrossRef\]](#)
58. Schultz, M.; Reitmann, S.; Alam, S. Predictive classification and understanding of weather impact on airport performance through machine learning. *Transp. Res. C Emerg. Technol.* **2021**, *131*, 103119. [\[CrossRef\]](#)
59. Falque, T.; Mazure, B.; Tabia, K. Machine learning for predicting off-block delays: A case study at Paris—Charles de Gaulle International Airport. *Data Knowl. Eng.* **2024**, *152*, 102303. [\[CrossRef\]](#)
60. Shao, W.; Prabowo, A.; Zhao, S.; Koniusz, P.; Salim, F.D. Predicting flight delay with spatio-temporal trajectory convolutional network and airport situational awareness map. *Neurocomputing* **2022**, *472*, 280–293. [\[CrossRef\]](#)
61. Sun, M.; Tian, Y.; Wang, X.; Huang, X.; Li, Q.; Li, Z.; Li, J. Transport causality knowledge-guided GCN for propagated delay prediction in airport delay propagation networks. *Expert Syst. Appl.* **2024**, *240*, 122426. [\[CrossRef\]](#)
62. Lambelho, M.; Mitici, M.; Pickup, S.; Marsden, A. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *J. Air Transp. Manag.* **2020**, *82*, 101737. [\[CrossRef\]](#)
63. Mamdouh, M.; Ezzat, M.; Hefny, H. Improving flight delays prediction by developing attention-based bidirectional LSTM network. *Expert Syst. Appl.* **2024**, *238*, 121747. [\[CrossRef\]](#)
64. Dong, X.; Zhu, X.; Hu, M.; Bao, J. A Methodology for Predicting Ground Delay Program Incidence through Machine Learning. *Sustainability* **2023**, *15*, 6883. [\[CrossRef\]](#)
65. Birolini, S.; Jacquillat, A. Day-ahead aircraft routing with data-driven primary delay predictions. *Eur. J. Oper. Res.* **2023**, *310*, 379–396. [\[CrossRef\]](#)
66. Yu, Y.; Mo, H.; Li, H. A Classification Prediction Analysis of Flight Cancellation Based on Spark. *Procedia Comput. Sci.* **2019**, *162*, 480–486. [\[CrossRef\]](#)
67. Chang, B.R.; Tsai, H.-F.; Mo, H.-Y. Ensemble Meta-Learning-Based Robust Chipping Prediction for Wafer Dicing. *Electronics* **2024**, *13*, 1802. [\[CrossRef\]](#)
68. Zheng, Z.; Zou, B.; Wei, W.; Tian, W. A Data-Light and Trajectory-Based Machine Learning Approach for the Online Prediction of Flight Time of Arrival. *Aerospace* **2023**, *10*, 675. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.