INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)

e-ISSN : 3139-065X

www.ijemss.com

editor@ijemss.com

Vol. 01, Issue 4, August 2026, pp:423-428

Predicting Flight Disruptions Using Machine Learning

[1]Akash Krishna A [2]Anisha S [3]Rukshana Begam

[1]Assistant Professor, Acharya Institute Of graduate Studies Bengaluru, Karnataka, India, akashkrishna@acharya.ac.in

[2]Customer Service Agent, Airport Operations and customer Service, Trichy, Tamil Nadu, India

[3]Student, BBA (Aviation), Bharathidasan university, Trichy, Tamil Nadu, India

DOI:https://doi.org/10.5281/zenodo.21833403

ABSTRACT

Flight disruptions, including delays and cancellations, impose substantial economic and operational costs on airlines, airports, and passengers worldwide, making accurate prediction a priority for the aviation industry. This study undertakes a secondary analysis of existing peer-reviewed literature on the application of machine learning to flight disruption prediction, with the purpose of consolidating fragmented findings into a coherent picture of current capability and limitation. Rather than collecting fresh flight data, the paper synthesises reported methodologies, model choices, feature sets, and performance metrics across a range of published studies covering ensemble methods, boosting algorithms, and hybrid architectures. The synthesis shows that tree-based ensemble models, particularly Random Forest and Gradient Boosting variants, consistently outperform simple linear and distance-based classifiers, with reported accuracies often exceeding ninety percent when weather, carrier, and historical delay-propagation features are included. Class imbalance between delayed and on-time flights emerges repeatedly as a major methodological obstacle, addressed unevenly across studies through resampling techniques. The paper concludes that while predictive accuracy has improved steadily, inconsistent evaluation protocols and limited real-time deployment studies remain significant gaps that future research should address.

Keywords: *flight delay prediction; machine learning; aviation analytics; random forest; feature engineering; class imbalance*

1. INTRODUCTION

Air transport is one of the most schedule-sensitive industries in the world, and even small disturbances to a single flight can propagate through an airline's network for hours. A late-arriving aircraft, a crew that times out of duty, or a sudden change in weather at a hub airport can each trigger a chain of knock-on delays that touch thousands of passengers before the day is over. Industry estimates over the past decade have repeatedly placed the direct and indirect cost of flight delays in the tens of billions of dollars annually when fuel, crew overtime, missed connections, and passenger compensation are taken together. For passengers the cost is measured differently, in missed meetings, lost holidays, and the general erosion of trust in air travel as a dependable mode of transport.

For much of the history of commercial aviation, disruption management has been reactive: airlines responded to delays as they occurred rather than anticipating them. The growth of digital flight-tracking infrastructure, open government datasets such as those maintained by national bureaus of transportation statistics, and low-cost cloud computing has changed this picture considerably. It is now possible to combine historical flight schedules, weather observations, air traffic control data, and airport congestion indicators into a single dataset large enough to train predictive models. Machine learning, with its ability to detect non-linear relationships among many interacting variables, has become the dominant analytical approach used to convert this data into forward-looking predictions of delay or cancellation.

A substantial and growing body of published research has applied classification and regression algorithms, ranging from logistic regression and decision trees to gradient boosting and deep neural networks, to the flight disruption problem. Reported results vary widely depending on airport, time horizon, feature set, and evaluation metric, which makes it difficult for a newcomer to the field to judge which approaches are genuinely more effective and which apparent gains are artefacts of a particular dataset or evaluation choice. This paper responds to that difficulty by stepping back from any single dataset and instead examining, side by side, what a cross-section of recent studies report about model choice, feature engineering, and predictive performance. In doing so it aims to give a grounded, evidence-based answer to the question of what machine learning can currently offer to the problem of predicting flight disruptions, and where its limits lie.

2. NATURE AND SCOPE OF THE STUDY

This is a descriptive, desk-based secondary analysis rather than an empirical study built on newly collected flight data. The unit of analysis is the published research article: each source paper is treated as a data point describing a modelling exercise, and the present study aggregates and compares these exercises rather than repeating them. This design was chosen deliberately. Flight operations data of the scale needed for a credible primary study, spanning multiple airports, seasons, and years, is not readily available to an independent researcher without institutional access to airline or regulatory systems. A synthesis of already-published, peer-reviewed results makes it possible to draw conclusions that are grounded in a wider range of airports and time periods than any single primary dataset could offer.

The scope of the review is bounded in three ways. First, thematically, the study restricts itself to supervised machine learning approaches applied to two related outcomes: binary or multi-class delay classification (on-time versus delayed, or delay-severity bands) and continuous delay-duration regression. Studies focused purely on air traffic flow management, crew scheduling optimisation, or airspace capacity planning are outside the boundary of this review unless they explicitly report a disruption-prediction model. Second, temporally, the review concentrates on studies published within roughly the last six years, since this is the period in which ensemble and boosting methods (Random Forest, XGBoost, LightGBM, CatBoost) became the dominant tools in this space, superseding earlier work built mainly around basic decision trees and linear models. Third, geographically, the underlying studies draw predominantly on data from major hub airports in the United States, supplemented by a smaller number of studies from other regions, so conclusions about absolute accuracy figures should be read as most directly applicable to large, high-traffic hub environments rather than smaller regional airports.

Within these boundaries, the variables examined across the reviewed studies include flight-level features (scheduled departure and arrival time, carrier, origin-destination pair, aircraft type), operational features (previous-flight delay, turnaround time, gate assignment), and exogenous features (weather conditions, air traffic congestion, holiday or seasonal indicators). The rationale for including this particular mix of variable types is that it mirrors what the reviewed literature itself treats as the standard feature space for this problem, allowing like-for-like comparison across studies.

3. REVIEW OF LITERATURE (Font-Cambria, Bold, Font Size -12)

3.1 Early classification approaches

Work applying machine learning to individual flight delay prediction using supervised classification established the basic template that later studies would refine: a labelled historical dataset of flights, a set of schedule and carrier features, and a comparison of several off-the-shelf classifiers against standard accuracy-based metrics. One such comparative study evaluated several classification algorithms on flight records departing from a major New York hub airport and reported that a Decision Tree model produced the highest accuracy, precision, recall, and F1-score among the algorithms tested, each close to or above ninety-seven percent (Chakrabarty, 2019, as reported in comparative studies of JFK departure data). This result is notable both for the very high accuracy reported and for how it illustrates a recurring pattern in the literature: simple, interpretable models occasionally outperform more complex ones on narrowly-scoped, single-airport datasets, even though this advantage does not always generalise across airports.

3.2 Ensemble and boosting methods

A second cluster of studies moved beyond single-algorithm comparisons toward ensemble and boosting techniques applied to larger, multi-airport datasets. A hybrid clustering-based model combining density-based clustering with an optimised random forest architecture, evaluated on aviation big data, reported an average prediction accuracy of approximately ninety-seven percent, an improvement of nearly five percentage points over prior comparable efforts, alongside a substantial gain in processing speed achieved by decomposing the prediction task into smaller sub-problems handled by cluster-specific models. This line of work highlights a second recurring

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
e-ISSN : 3139-065X
editor@ijemss.com
Vol. 01, Issue 4, August 2026, pp:423-428

theme in the literature: performance gains increasingly come not from a single stronger algorithm but from architectural choices, such as clustering flights into more homogeneous subgroups before applying a predictive model, or optimising ensemble weights through metaheuristic search.

A related contribution proposed a purpose-built algorithm for predicting individual flight delay minutes using regression rather than classification, incorporating newly engineered features capturing prior-flight-delay and following-flight-delay relationships, and tested this approach across three separate forecast horizons of two, six, and twelve hours using ten different machine and deep learning models. This regression-oriented strand of the literature is important because binary classification (delayed versus on-time) can obscure operationally important differences between a five-minute delay and a five-hour delay; regression-based studies attempt to close this gap, though at the cost of typically lower headline accuracy figures than classification studies report.

3.3 Weather and exogenous factors

A further group of studies incorporated external weather data explicitly rather than treating it implicitly through seasonal indicators. One such study combined flight data from a major hub with weather observations from a national environmental information centre and compared linear regression, support vector regression, decision tree regression, random forest regression, and regularised regression methods, generally finding that tree-based regressors handled the non-linear interaction between weather severity and delay duration more effectively than linear approaches. This supports a broader observation across the literature that weather variables, while individually modest in predictive power, interact strongly with airport-specific congestion effects, and models capable of capturing such interactions tend to outperform additive linear models.

3.4 Addressing class imbalance

Because on-time flights heavily outnumber delayed flights in most operational datasets, a recurring methodological concern in the literature is class imbalance. A systematic evaluation of six classifiers, including Decision Tree, Random Forest, Support Vector Classifier, Logistic Regression, K-Nearest Neighbours, and Naive Bayes, applied resampling techniques such as Random Oversampling, the Synthetic Minority Oversampling Technique, and Adaptive Synthetic Sampling to correct skewed class distributions, and validated results using stratified cross-validation together with metrics less sensitive to imbalance, such as the Matthews Correlation Coefficient and ROC-AUC, rather than relying on raw accuracy alone. This same body of work reports that a random forest regression model achieved an R-squared value of approximately 0.94 and a root-mean-square error of about 12.5 minutes, and identifies late aircraft arrival, carrier-attributable delay, weather, and National Air System congestion as the dominant contributing factors, a finding echoed by a separate statistical and machine-learning study that identified temperature, prior delay rate, month, and weekday as significant predictors and achieved its best result with a Gradient Boosting classifier.

3.5 Comparative and synthesis studies

A recent broad comparative study traced the evolution of machine learning use in flight delay prediction back to its earliest documented application in 2014, noting a marked rise in research activity from 2016 onward as an increasingly diverse array of algorithms was applied to the problem, including Logistic Regression, Naive Bayes, Neural Networks, XGBoost, Random Forest, CatBoost, and LightGBM; across the studies it reviewed, Random Forest emerged as the single most frequently used and generally best-performing technique. A separate synthesis paper similarly surveyed multiple prior studies and reported that one evaluation using JFK airport data achieved an accuracy of 97.78 percent with a Decision Tree model, with Random Forest and Gradient Boosted Tree models following closely behind at 92.40 percent and 93.34 percent respectively, while a different study using a Support Vector Regressor approach addressed the practical challenge of processing very large datasets by sampling and organising records on a month-by-month basis and using CatBoost to select the fifteen most informative features prior to model training.

Taken together, the literature reviewed here shows a clear trajectory: from single-algorithm, single-airport classification exercises toward larger, multi-source, ensemble-driven, and imbalance-aware modelling pipelines. What remains comparatively underexplored, however, is direct, standardised comparison across studies. Because

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)

e-ISSN : 3139-065X

www.ijemss.com

editor@ijemss.com

Vol. 01, Issue 4, August 2026, pp:423-428

each paper uses its own dataset, feature set, and evaluation metric, headline accuracy figures such as 97.78 percent or 97.2 percent cannot be safely read as directly comparable measures of one model being superior to another; they are, more precisely, measures of how well a particular model fits a particular airport's data under a particular set of conditions. This gap between the appearance of comparability and its practical difficulty is itself an important finding of the present review and is returned to in the discussion section below.

4. OBJECTIVES OF THE STUDY

The specific objectives guiding this review are as follows:

1. To identify and categorise the machine learning algorithms most commonly applied to flight disruption prediction in the recent published literature.
2. To compare reported predictive performance across studies in terms of the features used, the evaluation metrics applied, and the resulting accuracy or error figures, while noting where such comparisons are and are not methodologically sound.
3. To examine how the reviewed studies address the practical challenge of class imbalance between delayed and on-time flights, and to assess which mitigation techniques appear most effective.
4. To identify the exogenous and operational factors (weather, carrier, prior-flight delay propagation, air traffic congestion) that the literature most consistently associates with flight disruption.
5. To surface gaps in the existing body of research, particularly around standardised evaluation and real-world deployment, that can guide future primary research in this area.

5. DISCUSSION OF RESULTS

The synthesis of findings across the reviewed studies supports several conclusions that speak directly to the objectives set out above. With respect to algorithm choice, tree-based ensemble methods, particularly Random Forest and gradient-boosting variants such as XGBoost, LightGBM, and CatBoost, appear far more frequently among the best-performing models than linear or distance-based methods such as logistic regression or k-nearest neighbours. This is consistent with the underlying nature of the prediction problem: flight disruption is driven by the interaction of many variables, for example a weather event combined with an already-congested hub combined with a tight aircraft turnaround, rather than by any single variable acting in isolation, and tree-based ensembles are structurally well suited to capturing exactly this kind of interaction effect without requiring the researcher to specify it in advance.

At the same time, the very high accuracy figures reported by some single-airport, single-algorithm studies (in excess of ninety-seven percent in more than one case) should be interpreted cautiously rather than taken at face value as evidence of a solved problem. Several of the reviewed papers achieving such figures work with a single airport over a single time window, which tends to inflate apparent accuracy relative to what a model would achieve if deployed across a more heterogeneous, multi-airport, multi-season operational setting. The studies that evaluate performance using metrics less sensitive to class imbalance, such as the Matthews Correlation Coefficient and the area under the ROC curve, alongside stratified cross-validation, generally report more moderate and arguably more trustworthy performance figures than studies relying on raw accuracy alone, which can appear artificially high simply because on-time flights vastly outnumber delayed ones in most datasets.

The findings on contributing factors are more consistent across the literature than the findings on absolute accuracy. Late aircraft arrival (the propagation of a delay from an earlier flight using the same aircraft), carrier-attributable delay, adverse weather, and air traffic congestion recur across nearly every study reviewed as the dominant drivers of disruption, regardless of which specific algorithm was used to model them. This convergence suggests that the underlying causal structure of flight disruption is reasonably well understood, even where the specific predictive architecture used to model it varies considerably from one study to the next. This matters practically: it implies that an airline seeking to reduce disruption does not necessarily need the single most accurate published model, but does need reliable, timely data feeding these known factors, since the marginal gain from better real-time weather and upstream-delay data appears, across the literature, to matter more than the marginal gain from switching between competing tree-based algorithms.

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)

e-ISSN : 3139-065X

www.ijemss.com

editor@ijemss.com

Vol. 01, Issue 4, August 2026, pp:423-428

The regression-oriented studies reviewed here, which predict delay duration rather than a simple delayed/on-time label, generally report lower headline performance than classification studies, but arguably answer a more operationally useful question, since an airline rebooking passengers or reassigning crew needs to know how long a delay is likely to be, not merely whether one will occur. The comparative rarity of well-validated regression studies relative to classification studies, and the near-absence of studies evaluating live, real-time deployment of these models rather than offline historical validation, stand out as the two clearest gaps in the current literature. Addressing these two gaps, through more regression-focused research and through field trials of deployed prediction systems, would represent the most valuable direction for future work in this area, more so than continuing to add incremental algorithmic variations to an already crowded space of classification studies evaluated on similar historical datasets.

In summary, the evidence reviewed here indicates that machine learning, and ensemble tree-based methods in particular, can predict flight disruption with a meaningful degree of accuracy when supplied with weather, carrier, and delay-propagation features, but that reported accuracy figures vary too much across incompatible evaluation setups to be compared directly, and that the field would benefit from greater standardisation of datasets and metrics, more attention to delay-duration regression, and empirical study of real-world deployment rather than offline validation alone.

Table 1: Summary of Reviewed Studies

Study focus	Method(s) used	Key reported result	Data scope
Single-airport classification	Decision Tree, Random Forest, Gradient Boosted Tree	~97.8% accuracy (Decision Tree)	JFK, one year
Hybrid clustering + ensemble	DBSCAN clustering + optimised Random Forest	~97.2% accuracy; +2.49% via clustering	Multi-airport big data
Delay-minute regression	FDPP-ML, 10 ML/DL models	Improved MAE/RMSE at 2, 6, 12 hr horizons	US flight network
Weather-integrated prediction	Linear, SVR, Decision Tree, Random Forest, Ridge, Lasso	Tree-based regressors outperform linear models	JFK + weather data
Imbalance-aware evaluation	6 classifiers + SMOTE/ADASYN/ROS	$R^2 \approx 0.94$, RMSE ≈ 12.5 min (Random Forest)	Multi-source dataset
Comparative survey	Cross-study review, 2014–present	Random Forest most frequently top-performing	Literature synthesis

REFERENCES(Font-Cambria, Bold, Font Size -12)

1. [1] Chakrabarty, N. (2019). A data mining approach to flight arrival delay prediction for American airlines. Proceedings of the 9th Annual Information Technology, Electronics and Mobile Communication Conference. https://ieeexplore.ieee.org/document/9243339/
2. [2] Alsolami, D., Mehmood, R., & Albeshri, A. (2021). Airline flight delay prediction using machine learning models. Proceedings of the 5th International Conference on Future Networks and Distributed Systems. https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725
3. [3] Esmaeilzadeh, E., & Mokhtarimousavi, S. (2023). A novel intelligent approach for flight delay prediction. Journal of Big Data, 10, Article 179.

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)

e-ISSN : 3139-065X

www.ijemss.com

editor@ijemss.com

Vol. 01, Issue 4, August 2026, pp:423-428

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00854-w

4. [4] Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2024). A hybrid machine learning-based model for predicting flight delay through aviation big data. Scientific Reports, 14, Article 4665. https://www.nature.com/articles/s41598-024-55217-z
5. [5] Gopalakrishnan, K., & Balakrishnan, H. (2023). Flight delay prediction using machine learning. Proceedings of the IEEE International Conference on Recent Trends in Electronics and Communication. https://ieeexplore.ieee.org/document/10126220/
6. [6] Dogru, N., & Subasi, A. (2024). Predictive modeling of flight delays at an airport using machine learning methods. Applied Sciences, 14(13), 5472. https://www.mdpi.com/2076-3417/14/13/5472
7. [7] Singh, V., Meel, P., & Vishwakarma, D. K. (2025). Flight delay prediction: Evaluating machine learning algorithms for enhanced accuracy. PeerJ Computer Science / PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12685205/
8. [8] Rao, A., & Suresh, K. (2024). Deciphering air travel disruptions: A machine learning approach. arXiv preprint. https://arxiv.org/pdf/2408.02802