A hybrid ensemble learning-based prediction model to minimize delay in air cargo transport using bagging and stacking

Rosalin Sahoo[a], Ajit Kumar Pasayat[b], Bhaskar Bhowmick[c],

Kiran Fernandes[d] and Manoj Kumar Tiwari[e]*

[a,b,c]*Indian Institute of Technology, Kharagpur, India;* [d]*Durham University Business School, Durham, UK;* [e]*National Institute of Industrial Engineering (NITIE), Mumbai, India*

*corresponding author

Manoj Kumar Tiwari,

mkt09@hotmail.com

Director Office, National Institute of Industrial Engineering (NITIE), Mumbai, Maharashtra, 400087, India

A hybrid ensemble learning-based prediction model to minimize air cargo transport risk using bagging and stacking

Manufacturing productivity is inextricably linked to air freight handling for the global delivery of finished and semi-finished goods. In this article, our focus is to capture the transport risk associated with air freight which is the difference between the actual and the planned time of arrival of a shipment. To mitigate the time-related uncertainties, it is essential to predict the delays with adequate precision. Initially data from a case study in the transportation and logistics sector were pre-processed and divided into categories based on the duration of the delays in various legs. Existing datasets are transformed into a series of features, followed by extracting important features using a decision tree-based algorithm. To predict the delay with maximum accuracy, we used an improved hybrid ensemble learning-based prediction model with bagging and stacking enabled by characteristics like time, flight schedule, and transport legs. We also calculated the dependency of accuracy on the point in time during business process execution is examined while predicting. Our results show all predictive methods consistently have a precision of at least 70 percent, provided a lead-time of half the duration of the process. Consistently, the proposed model provides strategic and sustainable insights to decision-makers for cargo handling.

Keywords: air cargo handling, predictive analysis, feature selection, Ensemble learning, machine learning

1. Introduction

The publishers of International Journal of Production Research (IJPR), one of the most important and cited journals in the production research sector, have taken the initiative to publish some of the best articles in cutting-edge areas related to production management, manufacturing, and logistics to address some of the industry's most pressing issues on the occasion of 60th volume anniversary of the journal. IJPR has published some promising contributions related to logistics, cargo handling, and freight transport through various modes of transportation such as air, sea, and land over the last few decades. Precisely, there is a plethora of literature available covering various aspects of air freight handling.

However, cargo risk management in terms of handling delays in shipping goods across several legs is less explored, leaving a considerable gap that presents an opportunity to investigate the delay scenario in air cargo and solving it using innovative approaches. In this paper, we present a novel method of predicting delays in cargo handling by air using a hybrid ensemble learning approach.

The air cargo industry has become an indispensable part of the global economy, holding an important niche in transporting high-valued commodities. Continuously increasing demand from the industries and the end-users have encapsulated rapid development in the cargo industry over the past decades. Thus, the worldwide cargo market is considered to be an essential entity to boost the global economy. Freights which are perishable, time-sensitive, or used in just-in-time supply networks are mostly delivered by air cargo to the consignees worldwide at competitive prices (Wen et al., 2019). The timely delivery of shipments to their intended destination makes transit time one of the most important variables in logistics. The entire time it takes for items to be picked up and delivered has a big influence on industrial and manufacturing productivity (Du et al., 2018)). To maintain a synonymous economic cycle, manufacturing companies typically rely on air cargo to get the inventory with minimum lead time, and production going as fast as possible to satisfy the rebounding demand for goods. Shipment delivery time is also important for buyers who need to use the product being delivered to satisfy various needs at the end user and it also affects the transportation cost incurred by the buyer. Therefore, longer times in delivering the shipments are directly proportional with high logistic expenses. Both suppliers and customers rely on collaborative order management to avoid typical last-mile delivery problems (Zhang et al., 2021). Just like production and warehousing delays, transportation mayhem is inevitable sometimes. There exist certain uncontrollable reasons responsible for delay like weather, traffic, and disruptions are sometimes brought on by poor communication, time management, and decision making. Understanding possible delays ahead of time may lead to significant savings and avoid having to pay extra for air freight in an emergency, which is a frequent (and expensive) alternative for car and electronics providers, as well as time-sensitive loads like pharmaceutical industries (Chih et al., 2021). Instead of being compelled to choose the fastest option regardless of cost, providers may employ digital technology to identify the most financially sensible method to satisfy customer needs. Flight delays have a significant impact on the global economy as it impacts the industries and manufacturing sectors that have strong ties to the aviation industry. As reported by the International air transport Association (IATA), the annual worth of goods transported via air cargo is over $6.4 trillion (IATA, 2017a). Research reveals that improved strategic management enabling cutting-edge technologies in air cargo might enhance the market compatibility and customer demand (Dolgui et al., 2018) and (Zheng et al., 2020).

In the past decade, globalization of the world economy has positively impacted the risk of supply chain disruptions. The whole chain is becoming longer, more complicated as more partners are involved (Yoon et al., 2020), (Hendricks & Singhal, 2005) and (Tang, 2006). In cargo logistics, transport risk is one of the key performance indicators, quantified as the time difference between the actual and planned arrival time of shipment. The two most essential characteristics addressed, as earliness and tardiness, are very much undesirable by customers and freight forwarders (Wen et al., 2019). Earliness is a measure of finishing operations before due time, and tardiness is a measure of a delay in executing certain operations. Tardiness is considered the primary cause behind delays in production and delivery to all downstream customers, and earliness induces additional storage and handling costs (Metzger, Leitner, Ivanovi, et al., 2015) and (Zoutendijk, 2021). Disruption risk in transportation is often considered the extreme risk, which is more than 48-hour delays or more than 24 hours earliness, severely impacting the customers' operations and the freight forwarders. Distinguishing disruption risks from the routine deviations within a day is referred to as recurrent risks (Shang et al., 2017). In air transportation, a leg is the segment of the trip between consecutive flight stops within the boarding point to a destination point. Freight forwarder plays a vital role in global trade which directly impacts the supply chain. Therefore forwarders must adopt risk management strategies for making operations smoother and agile.

Flight delays are unavoidable, and they contribute significantly to airline profits and losses. One of the most important performance metrics for any transportation system is delay (Du et al., 2018). Specifically, commercial aviation stakeholders in freight carriers define delay as the time it takes for a flight to be delayed or rescheduled. As a result, a delay may be defined as the difference between the scheduled and actual departure or arrival timings of a plane. Several factors can cause an initial flight delay, including air carrier difficulties, harsh weather, air traffic control, and so on. However, because of interconnected resources, a propagated delay develops. Therefore delay in air freight can be divided into two groups: controllable delay and uncontrollable delay. Aircraft is the most typical re-source. Because the same aircraft flies many flight legs, a delay in one trip might have an impact on subsequent flights by the same aircraft. Researchers across the globe have recently been actively emphasizing the different problems of successful cargo handling. In recent studies, (Metzger, Leitner, Ivanovic, et al., 2015) and (Shang et al., 2017) concentrated on mitigating air cargo risk by applying empirical evaluation techniques and predictive modeling techniques. While addressing the transport risk directly related to delay in shipment, previous studies show no scientific evidence of the critical features responsible for the delay. Predicting and analyzing the delays and their causes have long been active subjects of research because of their vital importance in air traffic control, airline decision making and ground delay program (B. Yu et al., 2019). The authorities involved in cargo handling have a multitude of indicators related to tolerance thresholds for flight delays. Inflight delay is inevitable and it plays an important role in both profits and loss of the airlines (Thiagarajan et al., 2017). Prediction of such delays is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became cumbersome due to the complexity of the air transportation system, the number of prediction methods, and the deluge of flight data. to analyze event logs in order to predict process delay using a number of process risk indicators such as execution time, waiting time, and resource involvement . A methodological breach remains in adopting advanced techniques to compare and justify the applicability and usefulness of the prediction methods adopted in previous studies.

To address the foregoing gaps, we initially perform data pre-processing to understand the pattern and to remove stale data to avoid faults in making real-time decisions. We analyze the important features responsible for delaying a shipment in each transport leg using a feature selection algorithm. Our evidence suggests that the generated important features are responsible for delaying transporting the shipment in maximum cases. Further, we perform an empirical comparison of results obtained using a few machine learning and deep learning classifiers followed by the ensembling technique to achieve better than state-of-the-art results. While most of the researchers emphasize risk mitigation strategies using random yield distribution, disruption probabilities, and predictive monitoring techniques (Wen et al., 2019) and (Shang et al., 2017), we compared and evaluated the effects of classifiers in terms of prediction accuracy and applicability in a real-world industrial case study in the field of transport and logistics. The purpose of this article is to demonstrate, from the air freight forwarder's point of view, an effective risk management strategy aimed at reducing the risk arising from competitive market conditions. This study provides an in-depth description of the evaluation method to assess the benefits with real data sets. The case study adopted in this research is based on a large international forwarding company involving actual transport and logistics services that covers a global supply chain network. The dataset consists of 3942 business process instances within five months and incorporates 56,802 business activity completion instances.

The successive contribution of our study reveals that the prediction techniques based on machine learning classifiers can predict the time delay violation in an air cargo shipment process. The most critical stages responsible for the delay identified using the feature selection technique, which further enhances classification algorithms' performance. Then, the results obtained through prediction algorithms are compared and ensembled to improve the prediction capability. Our prediction mechanism aims to understand whether the actual delivery duration exceeds the planned duration throughout a complete shipment process execution, which will help assess and forecast the transport risks. The outcome of our study reveals that the adopted models can provide adequate prediction accuracy with feasible lead times; e.g., all predictive models consistently maintain the accuracy of at least 70% provided a lead time of half the process duration.

The remainder of the article is organized as follows. Section 2 illustrates the problematic facets of the air freight logistics business and the summary of how our study advances state-of-the-art in terms of prediction. In Section 3, we explain the overview of the dataset used for this study, followed by a detailed explanation of the methodology adopted in Section 4. Section 5 presents our findings by analyzing the results and further representing a detailed discussion. Section 6 concludes the paper with managerial implications and future research directions.

2. Related Literature

The air cargo sector dominates a significant part of the logistics industry due to the fastest shipping and fewer physical barriers, despite being an expensive mode of transportation. Due to the involvement of multiple factors, the risk associated with this sector is immense, which has been studied by many researchers worldwide to overcome the significant challenges (Rodríguez-Sanz et al., 2019). While ensuring transport risk is the foremost challenge facing the air cargo sector, different researchers (Han et al., 2007; Ross et al., 2010) (Han et al., 2007),(Ross et al., 2010),(Zúñiga et al., 2011),(Chan et al., 2014),(Pathak et al., 2019) and (Tsolakis et al., 2021) closely examine the shipping process to reduce risk factors. Flight delays caused by system congestion and other reasons have been a constant source of business losses and monetary deprivation in the aviation industry in the past few years (Archetti & Peirano, 2019). The concept of various risks in air transport operations has been investigated by, (Choi et al., 2019), and (Neal & Koo, 2020), focusing on the commercial and safety risks. Flight departure delay has been estimated by (Tu et al., 2008) to solve the disparity between the expected time of departure and the real-time of departure. (Metzger, Leitner, Ivanovi, et al., 2015) demonstrated the potential of predictive monitoring in the air cargo industry, and as a solution to increase business profitability and sustainability, a new cloud and services-based collaboration and convergence model have been proposed. The authors also studied risk management in the air cargo sector and evaluated the strategies that will benefit the shippers and the forwarders. A study on perishable air cargo has been conducted by (Azadian et al., 2012), and a framework is proposed to account for the real-time information inaccuracy. (Shang et al., 2017) and (Wen et al., 2019) also focused on air transport risk considering the delay in shipment as a significant factor and solved their models using different solution approaches.

Many researchers and practitioners have introduced data-driven predictive methods to obtain maximum efficiency with better accuracy to address the air cargo sector's significant challenges. (Elbert et al., 2012) evaluated possible risk management tools for air freight forwarders using a balanced approach to pricing and financial hedging. Using real-time details, (Azadian et al., 2012) studied dynamic routing of perishable air cargo, formulating a novel Markov decision model. Experimental findings have shown that the approach suggested could improve delivery reliability and reduce expected costs. A potential predictive monitoring practice was proposed by (Metzger, Leitner, Ivanovi, et al., 2015) in the air cargo logistics sector to mitigate the service-oriented issues, and it was an event-driven architectural approach. Further, the authors extended this study that aims to forecast possible issues in an air cargo shipment during process execution by contrasting three major classes of predictive control strategies focused on machine learning, aggregation, and constraint satisfaction of Quality-of-Service (QoS). The outcome of this study shows that evidence shows that for particular accuracy metrics, certain combinations of techniques can outperform individual techniques. (Shang et al., 2017) developed a model-based method to mitigate the transport risk in air cargo considering a data-driven approach. Use data from international logistics for air freight, and the authors investigated directions of predicting transport threats using the Bayesian non-parametric model and compared the result with a linear model and a flexible mixture model. The effect of the scheduled block time assigned for a flight, a factor managed by airlines, on-time arrival efficiency has been analyzed by (Ambra et al., 2019) using a structural estimation approach from econometrics. In summary, although previous research efforts have concentrated on the development of data-driven predictive modeling for air freight logistics, the active involvement of powerful computational algorithms such as machine learning techniques are less explored.

We proposed a data-driven predictive modeling approach using machine learning-enabled by time, flight plan, and transport legs to forecast the shipping delays to resolve the research gap. Summarizing the related work and addressing the derived literature gaps, this article makes substantial contributions to literature and practice. Firstly, using the existing case study, we deliver an in-depth data analysis and classified the delays into different subgroups according to the duration of earliness and tardiness. Enhancing the model's efficiency in terms of accuracy and reducing the training time, we implemented a feature selection procedure to generate a shipment's critical stages that favourably affect the delay parameter. Different machine learning classifiers have been exploited and compared to obtain maximum accuracy in terms of F-score, precision, recall, and error metrics. Finally, we present ensemble learning methods where different classifiers are strategically combined to solve a particular computational intelligence problem. We also provide evidence that shows that integrating the predictions of several different classifiers into a single robust prediction can boost classification efficiency.

3. Data set overview

Being one of the essential pillars of the global transport chain, the air cargo sector receives significantly less attention due to confidential operating procedures. It is essential to highlight the unique features of the dataset presented in the case study to understand our approach better. Figure 1. shows the transportation and logistics framework covered by the case study where up to three identical shipments from the consignors are consolidated and shipped collectively to the consignees to get some monetary gain with more cargo transport protection.

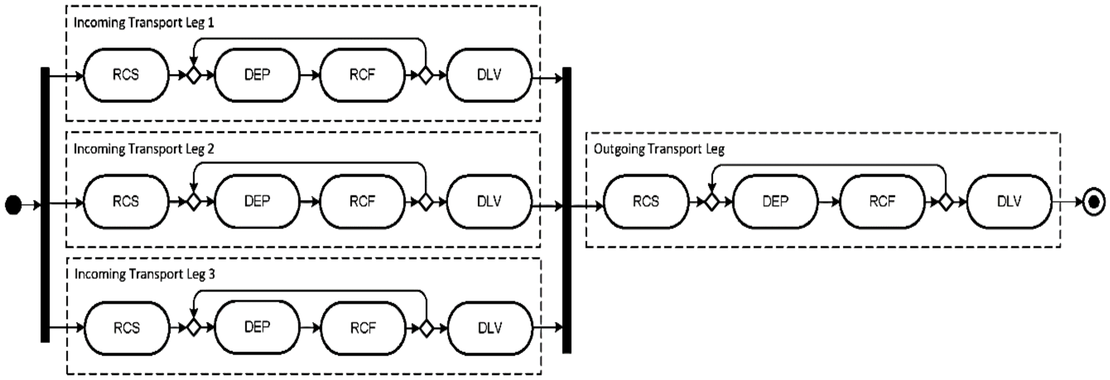


Figure 1. The transportation and logistics framework covered by the case study

The business process includes three incoming legs and one outgoing leg and each transport leg includes the following physical transport services labeled as activities as presented in Figure 1 (citation). One transport leg may contain one or more segments where shipments are shifted to other convenient flights at stopover airports. In such a scenario, RCS refers to shipment checks in at departure airline by producing a receipt; DEP is the departure of aircraft with confirmed goods on board, RCF indicates acceptance of freight at the arrival airport, DLV shows delivery of freight at the destination after receipt of shipment is signed and handed over to the consignee. This data involves the monitoring and tracing activities from a forwarding firm over five months. Cargo 2000 is an IATA project that works on innovative quality control services for the air transport market that maintains remarkable accountability in the supply chain (IATA, 2017b).

The data was reconstructed from the case study with execution traces of 3,942 real business process instances, consisting of 7,932 transport legs and 56,082 invocations of operation within a time span of 5 months. Out of total service invocations, all the segments RCS, DEP, RCF, and DLV have 11874, 16167, 16167, and 11874 instances respectively. For any of the business process services, each execution trace contains scheduled and successful durations (in min) as well as airport codes for the departure and arrival services. The emphasis is on the transfer of tangible items, and the handling in the business process of transport information differs depending on whether the documents are paper-based or electronic, as the data set did not allow anyone to differentiate between the different types of documents. Multiple flight segments may be included in a transport leg (e.g., in case the freight is moved to other flights or airlines at stopover airports). RCF loops back to DEP in our case. The number of segments per leg in this case study varies from one to four.

4. Methodology

This section introduces the background information of all the methods that have been adopted by our study. We have performed an extensive data analysis in Section 4.1 by classifying the data into different groups in the first stage and generated an extended data set from the existing data that has been discussed in Section 3. We also have elaborated the feature engineering and feature selection technique in Section 4.2 to reduce the dimensionality of the dataset and to identify important features by generating a score for each feature.

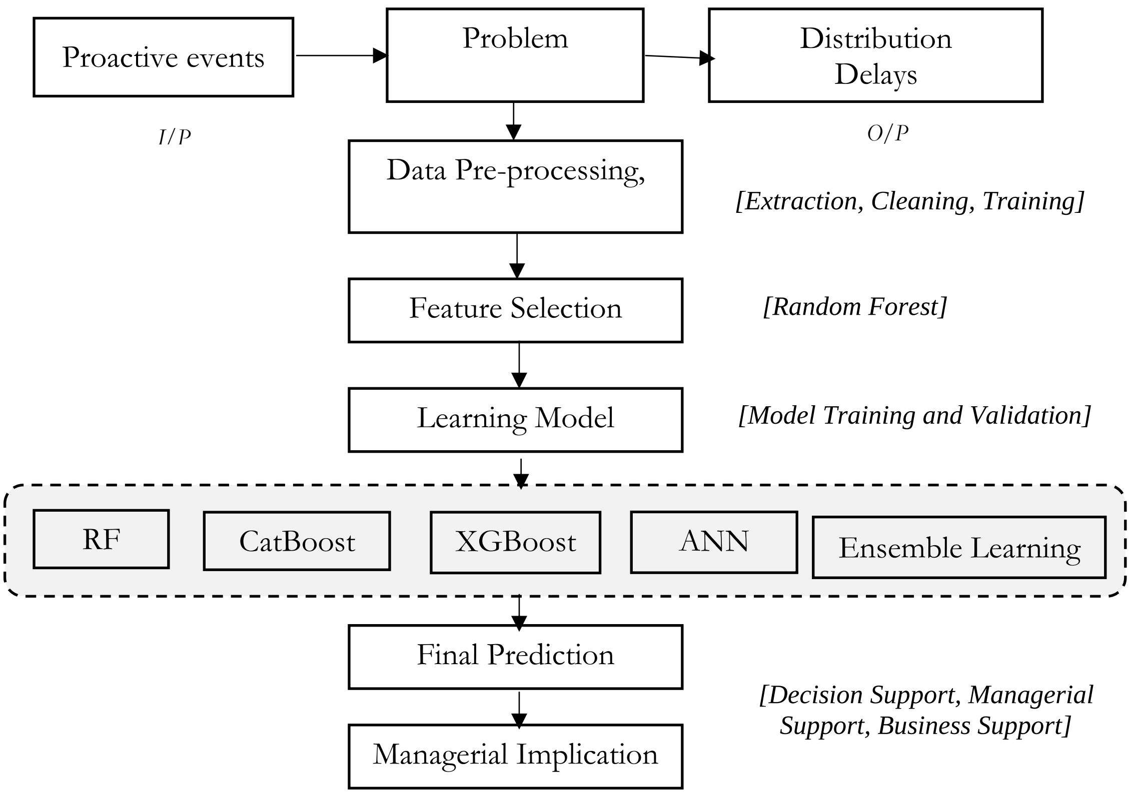


Fig. 2. Predictive Modelling Framework

We also have elaborated the feature engineering and feature selection technique in Section 4.2 to reduce the dimensionality of the dataset and to identify important features by generating a score for each feature. The description of all the machine learning classifiers and ensembling techniques are further discussed in Section 4.3 to provide a basic knowledge of the applicability in the context of our research. Figure 2. exhibits the predictive modeling framework of our study in an elucidate manner.

4.1. Data Pre-processing

Data Pre-processing plays a vital role in constructing models for machine learning. The pooled raw data of our model contains some outliers, out-of-range values, few missing values, and error values. While training the model, ambiguity increases due to redundant and irrelevant information that leads to incorrect results, as data quality has a direct effect on model output. The above issues raise the need for data pre-processing before the model's training. In our analysis, activities such as cleaning and transforming the data, selecting the case, removing the appropriate attributes, and normalizing the data if it is not within the acceptable range were used in the pre-processing data system. Firstly, all missing values are filled by the mean imputation technique; random values are identified and resolved where inconsistent data is present through outlier mining. Further, while performing data transformation, we normalize the attributes and aggregate the logically dependent and similar attributes. In data cleaning, the duplicate entities are deleted as the volume of the data is huge. Using the actual and planned shipment duration, the delay has been calculated across all incoming and outgoing legs. In our scenario, the generated sample with no delays is identified and removed. The original dataset contains 104 attributes divided into four main groups: incoming transport leg 1, incoming transport leg 2, incoming transport leg 3, and outgoing transport leg. The information related to planned duration and actual duration of shipment delivery is recorded for each transport leg. The data contains information of unique id for the process throughout the incoming and outgoing legs. The delay is classified based on earliness and tardiness. Further, data were sectioned into two groups: positive delay and negative delay based on time.

Table 1

Positive and Negative delays obtained from the dataset.

	Time (in hrs)		**Time (in hrs)**
Positive Delay	0-24	Negative delay	0-24
	24-48		24-48
	48-72		48-72
	>72		>72

Classifying the delays into different groups having similar characteristics simplifies the learning process. The empirical distribution of all shipments is examined from the data (almost 56000 transport services), and the classification is done accordingly. Positive delay mostly referred to those deviations where the shipments reach the concerned receiver before the planned arrival time, whereas negative delay depicts the delivery falls beyond the planned arrival time. We divided the delays into two groups as positive and negative delays to better the prediction technique. It is essential to keep track of delays to understand the pattern to accurately forecast the deviation in the freight transport process's execution. The non-symmetric distribution of transport risk with several bumps at positive delays and negative delays can be visualized in Figure 3 across all the transport legs.

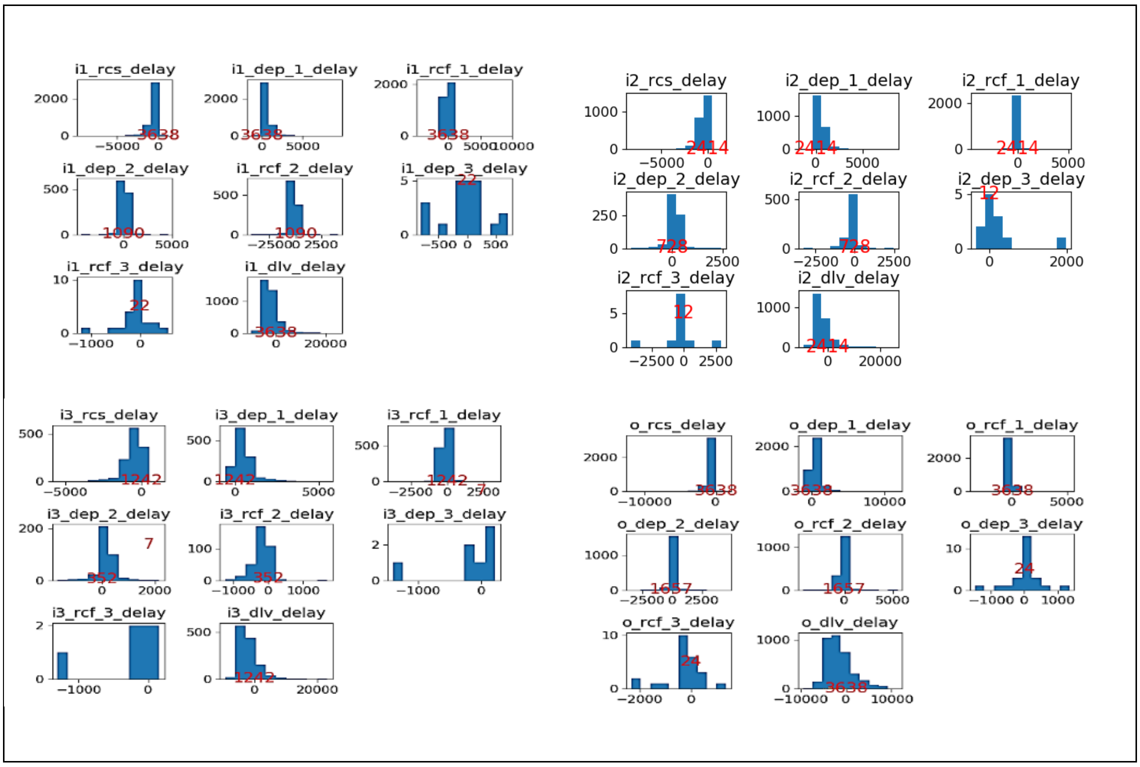


Fig. 3. Distribution of delay across Incoming and outgoing Transport leg

Box plots are a standardized way for showing the distribution of the data by showing the min-max median value and representations of outliers in a dataset and Figure 4. illustrates the box plot of the dataset to get an idea of the ranges of delays. It is an approach of summarizing a set of data measured on an interval scale, which is useful when comparing distributions between many groups or datasets. The delays after 29256 min i.e. 20.31 days and before -26050 min i.e. 18.09 days are considered an outlier. The maximum deviation can be seen from the graphs as a total delay of 5000 min is ranged as >72 hours delay. It is evident that delays of shipments with incoming leg one and leg two have similar delay distribution. Delays are generally having local mean around different days. Delay of shipments with three incoming legs has much more variance that is more prone to delay, and the curve is getting rough evenly.

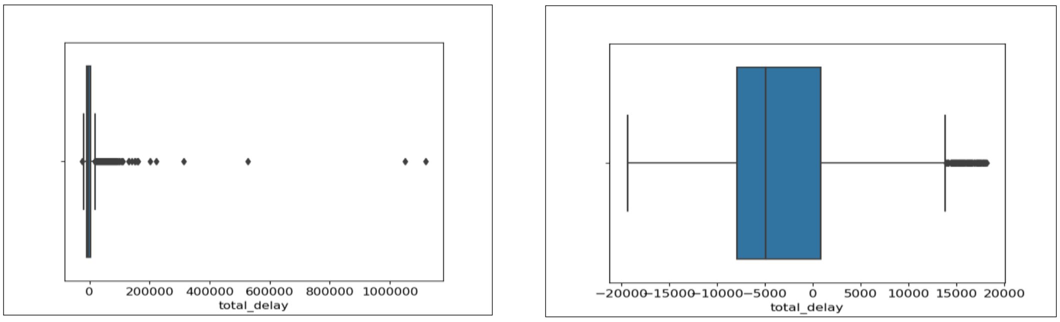


Figure 4. BOX Plot of the dataset to get the idea of the ranges of delays

4.2. Feature Engineering and Feature Selection

The method of converting raw data into features that improve the performance of the predictive model is termed feature engineering. It is used to prepare the correct input dataset, consistent with the machine learning algorithm. To improve the accuracy with reduced time and performance of the model, researchers adopt the feature selection method. Feature selection is a crucial stage of data analysis when dealing with datasets having huge variables. It is applied to select a much smaller set of essential features that are relevant for classification (Bejani & Ghatee, 2018). These features are used to boost the efficiency of machine learning algorithms. In this research, we have identified delays in each leg of a shipment as the critical feature used to improve the prediction algorithm's performance. New attributes have been generated using the existing attributes from the dataset on which the prediction analysis is done. It is an essential preprocessing step in machine learning that has been widely used in many domains ranging from text classification to complex networks. This study considers feature selection using the Random Forest technique to further enhance the algorithm's accuracy and efficiency. On doing feature importance of the dataset using Random Forest, we found the importance of individual features responsible for transportation risk. The proposed method Random forest is based on embedded feature selection type, which is highly accurate, generalized better, and interpretable.

4.2.1. Feature selection using Random Forest (RF)

RF provides an internal measure of feature importance after computation, which is useful to select important features (Lin et al., 2017). RF is a set of decision trees that comprises internal nodes and leaves (L. Breiman, 2001). The important feature can be calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability is evaluated by dividing the number of samples that reach the node by the total number of samples. The higher the value, the more important the feature. The pre-processed data was first converted into a series of features in the feature extraction and selection process based on having zero and non-zero delay values. The importance of a variable a_i for predicting a response variable Y is evaluated by averaging the sum of the weighted reduction in the residual sum of the square (RSS) for all nodes t where a_i is used over the number of regression trees. The variable importance is denoted by

$$VarIMP\left(a_i\right) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(S_t) = a_i} D(t) \Delta k\left(S_t, t\right) \tag{1}$$

where,

$D(t)\Delta k\left(S_t, t\right)$	:	Weighted reduction in the RSS by dividing the internal node t into two child nodes
$D(t) = N_t / N$	:	The proportion of samples of the data at node t
N	:	Total number of samples drawn to create the regression tree
N_t	:	Total number of trees of regression in a random forest
T	:	Regression tree structure
S_t	:	Split at node t
$v(S_t)$	:	The splitting variable for the split S_t

$\Delta k\left(S_t, t\right) = k(t) - D_1 k\left(t_1\right) - D_2 k\left(t_2\right)$: RSS at node t, t_1 and t_2 denote the two child nodes.

4.3. Overview of classification algorithm

This section briefly introduces the prediction mechanism of all the techniques that are experimentally compared in this study. Machine learning is a study of algorithms that automatically allows systems to learn and develop from practice. We have purposely adopted a mixture of the different classes of well-known most relevant techniques such as Random Forest (RF), boosting algorithms, artificial neural network (ANN), and ensemble learning models for the sake of generalizability of our outcomes. Out of the entire dataset, 20% are considered testing data, while the remaining are considered the training data set. Essentially there are specific steps followed while training the machine learning algorithms. Firstly, the prediction algorithm is trained using an existing set of training datasets having historical instances. In this step, the most important features are taken into consideration to generate a more accurate solution. We use a normalized scale to visualize the impact of prediction lead-time on prediction accuracy, which compensates for the fact that the number of service executions varies between individual instances of the process ranging from 20% to 100% provided in Table 6.

4.3.1.Random Forest (RF): Breiman (2001) proposed the RF algorithm concept, a classification approach that builds the predictive model consisting of multiple decision trees. RF has a wide range of applications because of its better stability and generalization (Mursalin et al., 2017). For any model starting from a few hundred to several thousand trees, the number of decision trees varies. This study has developed 100 decision trees using individual samples where each sample is the root node of individual decision trees. One decision tree can be constructed by splitting one node into two child nodes by randomly selecting the variables. One variable (often called a splitting variable) is selected from all other variables to determine the optimal split. The split variable's value is referred to as the cut point. When there are no more instances to split, the splitting process is continued until the stopping criterion is met. RF technique is widely used by many researchers for feature selection and classification problems and the details can be referred from (Pavlov, 2019), (Nair et al., 2019), and (Ribeiro & dos Santos Coelho, 2020)*4.3.2. CatBoost (Gradient boosting with categorical features support)* is an improved variant of the algorithm for gradient boosting decision trees explicitly developed to accommodate categorical features. It uses binary decision trees as base predictors. While training this bosting algorithm, the data is randomly shuffled and reshuffled multiple times to calculate the mean for every object only on its historical data. CatBoost uses oblivious decision trees, which use the same splitting criteria over the entire tree level. Let us consider the dataset $D = \{(X_i, Y_i)\}_{i=1,..,n.}$, a differentiable loss function $L(y, F(x))$, and the number of iterations to be N. the boosting technique aims to search for an approximation to a function that minimizes the loss function. The mapping function is termed as $F_0(x)$ is to minimize the loss function with a constant value is as follows:

$$F_0(x) = \arg\min_{f} \sum_{i=1}^{n} L(y, F(x)) \tag{2}$$

The pseudo-residuals (r_{im}) are computed from iteration 1 to *n* to solve the optimization problem, as shown in Equation 3.

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1(x)}} \quad for \;\; i = 1, ..., n. \tag{3}$$

Further, the base learner $\beta_m(x)$ is fitted to r_{im} using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$. The multiplier γ_m is computed in the next step by solving the following one-dimensional optimization problem as shown in Equation 4.

$$\gamma_m = \arg\min_{f} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma\beta_m(x_i)) \tag{4}$$

Thus the model can be updated using the following equation, 5, to generate the output.

$$F_m(x) = F_{m-1}(x) + \gamma_m \beta_m(x) \tag{5}$$

Equation (3)-(5) demonstrates the categorical gradient boosting method's significant steps where the parameterized function includes splitting variables, cut points, and nodes of individual trees.

4.3.3. XGBoost (Extreme gradient boosting) is originated from the idea of gradient boosting. This algorithm has received the utmost attention from researchers and practitioners due to its excellent performance in predicting and preventing overfitting issues reducing the computational time. This can be achieved by simplifying the objective functions and combining prediction and regularization while retaining maximum computational efficiency (Hughes et al., 2019). The objective function is defined as:

$$O = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{k=1}^{t} R(f_k) + C \tag{6}$$

Here, $R(f_k)$ represents the regularization term at the *k* time iteration, and *C* is a constant term, which can be selectively omitted. It is denoted as:

$$R(f_k) = \alpha H + \frac{1}{2}\eta \sum_{j=1}^{H} \omega_j^2 \tag{7}$$

Here, α defines the complexity of leaves, *H* represents the number of leaves, η is the penalty parameter, ω_j is the output result of each leafy node. In conclusion, the optimization of the objective function is turned into the minimal quadratic function being calculated. In comparison, XGBoost has a more substantial potential to avoid overfitting due to the addition of regularization.

4.3.4. ANN (Artificial Neural Network)

ANN is considered the most popular method of deep learning based on a computational model made of a series of artificial neurons. Every artificial neuron represents a particular output function called the activation function, while the memory of an ANN model represents the weight of each relationship between two neurons(Schmitz et al., 1999). The output of ANN can be expressed as,

$$y_j = \psi\left(\sum_{i=1}^{n} w_{ji} x_i + \Theta_j\right) \tag{8}$$

Where, Θ_j : External threshold, offset or bias, w_{ji} : Synaptic weights, x_i : Input, y_j : Output. An ANN model's performance depends on on-link nodes, weights, and activation features.

4.3.5. Ensemble Learning

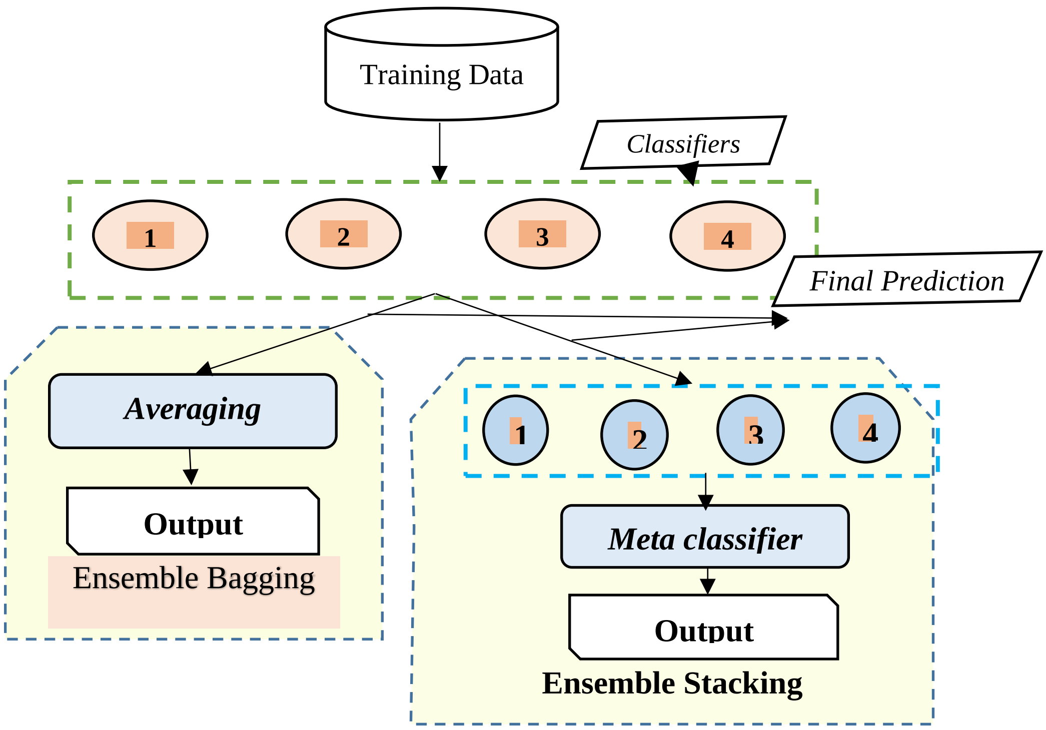
In machine learning, ensemble algorithms have achieved success by merging several weak learners to form one strong learner. Ensemble learning can boost the classification's efficiency by integrating several different classifiers' predictions into a single robust prediction (Porwik et al., 2019). The main objective of ensemble learning is to reduce the probability of selecting a single poorly performing learning algorithm and boost one algorithm's outcome using an intelligent ensemble of many individual algorithms (Zhu et al., 2019) and (J. J. Q. Yu, 2020). In several applications, the usefulness of ensemble learning has been extensively demonstrated.

Table 2

Comparison of different Ensemble Models

Ensemble Method	Objective	Type	Aggregation Method
Bagging	Decrease Variance	Parallel	Averaging
Stacking	Reduce Bias, Decrease Variance	Hybrid	Regression

Bagging (ELB) is an ensemble approach intended to stabilize the base classifier's accuracy (Leo Breiman, 1996). It claims to decrease variation and help to prevent overfitting. It often considers homogeneous weak learners, learns them independently from each other, and integrates them in a deterministic averaging process (Agarwal & Chowdary, 2020). Bagging adopts the most common methods of aggregating simple learners' outputs, i.e., averaging regression and voting for classification problems. The prediction combines multiple base classifiers $E = \{E_1, E_2, ..., E_T\}$ where the number of base learners is T. The classifiers are trained on the data set $D = \{(x_1, y_2),(x_2, y_2),...,(x_m, y_m)\}$ where x is the input vector and y is the class label associated with it.



sFig. 5. The proposed fusion framework of Ensemble learning using bagging ad stacking

Stacking (ELS) often considers heterogeneous weak learners, learns them in parallel, and combines them to generate a new prediction based on the predictions of the various weak models by training a meta-model (Wolpert, 1992) and (Ribeiro & dos Santos Coelho, 2020). The output of the first-level learners serves as input for the meta-learner. The first-level learners are often made up of different and diverse learning algorithms, although it is possible to create stacked ensembles from the same learning algorithms (Bai et al., 2021). Experiments have shown that a stacked ensemble works quite well and, on average better than a single best classifier.

Above mentioned algorithms are implemented in our model to predict the accuracy of the delay that occurred. To the best of our knowledge, this analysis is the first attempt to apply the methodology of feature selection and the combination of machine learning algorithms to improve accuracy in forecasting the delay.

Table 3

Pseudocode of ensemble learning algorithm using Bagging

Ensemble Using Bagging	
Input	$D = \{(x_1, y_2),(x_2, y_2),...,(x_m, y_m)\}$; Base learning algorithm E ; Number of base learners T.
Output	Ensemble Classifier : $H(x) = \sum_{t=1}^{T} h_t(x) = y$
Process	1. ***for*** $t = 1,...,T$: 2. $D_{bs} = (D,\ D_{bs})$ % D_{bs} is the bootstrap distribution 3. $h_t = E(D_{bs})$ 4. ***end***

Table 4

Pseudocode of ensemble learning algorithm using Stacking

Ensemble using stacking	
Input	$D = \{(x_1, y_1),(x_2, y_2),...,(x_m, y_m)\}$; First-level learning algorithms $L_1,...,L_T$; Second-level learning algorithm L.
Output	Ensemble Classifier : $H(x) = h^*(h_1(x),...,h_T(x))$
Process	1. ***for*** $t = 1,...,T$; % Train the first-level learner 2. $h_t = L_t(D)$; % first-level learning algorithm L_t 3. ***end*** 4. $D^* = \phi$; % Generate a new data set

5. ***for*** $i = 1, \ldots, m$;
6. ***for*** $t = 1, \ldots, T$;
7. $z_{it} = h_t(x_i)$;
8. ***end***
9. $D^* = D^* \cup ((z_{i1}, \ldots, z_{iT}), y_i)$;
10. ***end***
11. $h^* = L(D^*)$; % Train the second-level learner *h**
% the second-level learning algorithm *L*
% new data set *D**.

---

5. Experimental Analysis

This section followed the approaches mentioned above to analyze and compare the prediction accuracy using an industry dataset. We first determined and selected important features from our data set, and then the empirical results are presented for all the prediction techniques with a general discussion.

5.1. Feature selection

The process of feature selection plays an essential role in optimizing machine learning algorithms' performance and accuracy by choosing a small segment of more features in the dataset (Barbour et al., 2018). In our work, the invalid data and null values are removed from the extended data set. They were then limited to a regular range with a mean value of 0 and a standard deviation of 1. We implement Random Forest to select features based on feature score in which each feature is having a score based on the weight associated with it. The weight of the feature refers to the number of times it appears in the tree. Feature defines the structure of the tree; therefore, selecting the correct features will provide a better tree structure. After our experiment, out of 37, we got 27 non-zero important features, out of which the three most important features responsible for transportation delay of a shipment across all the legs are o_dlv_delay, i2_dlv_delay, and i1_dlv_delay with scores 0.365961, 0.117987, 0.114570 respectively. Appendix A contains all the features' specifics and is the most dominant characteristic of all the functions. It is inferred from the result that when the planned duration time deviates from the actual duration of freight delivery at the outgoing transport leg, the process gets highly impacted, which leads to maximum delay.

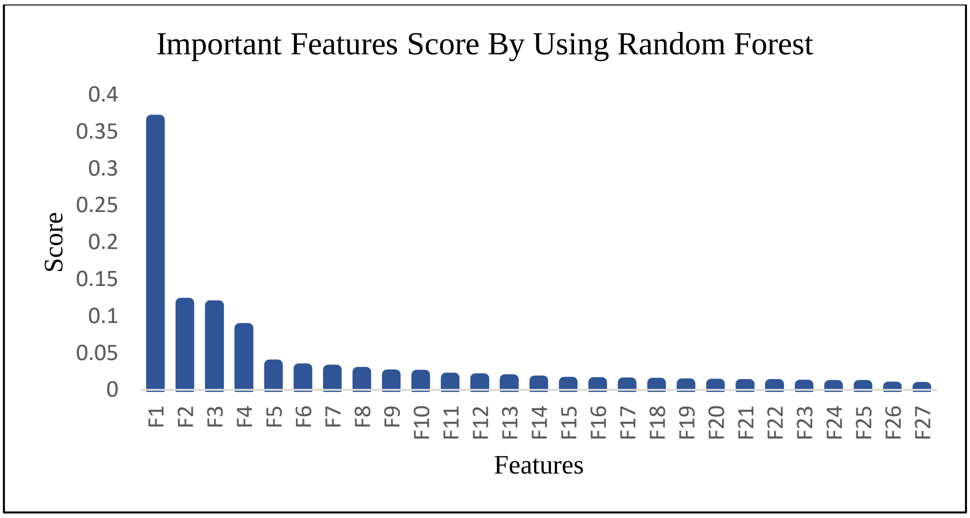


Figure 6.Scores of each feature to generate the importance

5.2. Prediction accuracy and computational efficiency

For evaluating the prediction methodology based on machine learning, we have implemented Random Forest, XGBoost, CatBoost, ANN, and in the Ensemble model, we are using Bagging, Stacking methods. For each algorithm, we loaded the test dataset in Python 3.7.4 and performed it on a computer with Intel (R) Core(TM) i5-8300H CPU with memory equipped with 16 GB RAM. We identify the delays and calculate them as it can be useful to quantify delays for complicated decision-making. After classification, we applied an ensemble machine learning model to forecast the estimated delay accurately. To the best of our knowledge, first classification, feature selection, and then forecasting delays were not explored in previous studies. Various performance measures such as F-score, precision, recall, specificity, error value, and accuracy are used to assess the proposed method's performance (El-dahshan et al., 2010). Confusion matrix, one of the common approaches to measuring the performance of the model. The two classes are known as the positive class and the negative class in a confusion matrix.

To quantify four metrics, each predicted class is compared to its actual class, for instance: True Positives (TP) implies the number of delay instances accurately classified as positive classes, False Positives (FP) states the number of delay instances wrongly classified as positive classes, True Negative (TN) implies the number of delay instances that are appropriately classified as negative classes, the number of delay instances that are wrongly classified as negative classes are termed as false negative (FN). Furthermore, we used F-score, precision, sensitivity, specificity, accuracy, and errors to assess prediction quality. Those performance measures can be extracted following the equations in table 6, based on the confusion matrix.

Table 6

Confusion matrix table obtained after implementing the algorithms with accuracy values

Pos (%)	RF					XGBoost					CatBoost					ANN					ELB					ELS				
	TP	TN	FP	FN	Accuracy	TP	TN	FP	FN	Accuracy	TP	TN	FP	FN	Accuracy	TP	TN	FP	FN	Accuracy	TP	TN	FP	FN	Accuracy	TP	TN	FP	FN	Accuracy
20	70	193	16	58	0.780	65	192	22	58	0.762	68	193	16	60	0.774	69	195	19	54	0.783	64	195	17	61	0.768	75	198	14	50	0.810
40	71	194	15	57	0.786	67	195	19	56	0.777	69	195	19	54	0.783	71	194	15	57	0.786	65	192	22	58	0.762	73	198	14	52	0.804
60	71	193	15	58	0.783	71	194	15	57	0.786	70	193	16	58	0.780	75	197	11	54	0.807	77	200	10	50	0.821	81	204	11	41	0.845
80	73	198	14	52	0.804	73	198	14	52	0.804	77	200	10	50	0.821	78	199	9	51	0.821	82	205	9	41	0.851	88	204	9	36	0.866
100	101	214	7	15	0.934	107	214	9	7	0.952	107	215	6	9	0.955	103	214	13	7	0.940	106	218	3	10	0.961	108	217	5	7	0.964

The number of positive class predictions that currently belong to the positive class is quantified by precision. Precision is the percentage of positive observations predicted accurately to the total positive observations predicted. Sensitivity is the ratio of the number of samples truly classified to the overall number of actual samples. F- Metric is calculated as the weighted average of Precision and Recall. Specificity can be expressed as the proportion of actual negative cases that are correctly identified. The accuracy matrix can be stated as the ratio between correct predictions to the total number of predictions made. We demonstrated the comparison of four classification methods, namely RF, XGBoost, CatBoost, ANN, along with the Ensemble technique using Bagging and stacking to equitably compare the output of different methods, using the same standard parameters to demonstrate the benefit of our methods. The accuracy of the predicted values is given in Table 6.

Table 7

Comparison of accuracy indicators

Model	Accuracy	Precision	Recall	F-score	Specificity	Error	Type-I	Type-II	Training Time (In Secs)
RF	<u>93.47</u>	93.51	87.06	90.16	96.83	0.065	0.1293	0.0316	0.822
XGBoost	95.25	93.85	92.24	93.03	96.83	0.047	0.077	0.0316	0.498
CatBoost	95.54	94.69	92.24	93.44	97.28	0.044	0.077	0.0271	0.888
ANN	94.07	93.64	88.79	91.15	96.83	0.064	0.1121	0.0317	0.915
ELB	*96.14*	97.24	91.37	94.21	93.96	0.038	0.086	0.0135	0.147
ELS	**96.44**	95.58	93.91	94.74	97.75	0.0356	0.0608	0.0225	0.181

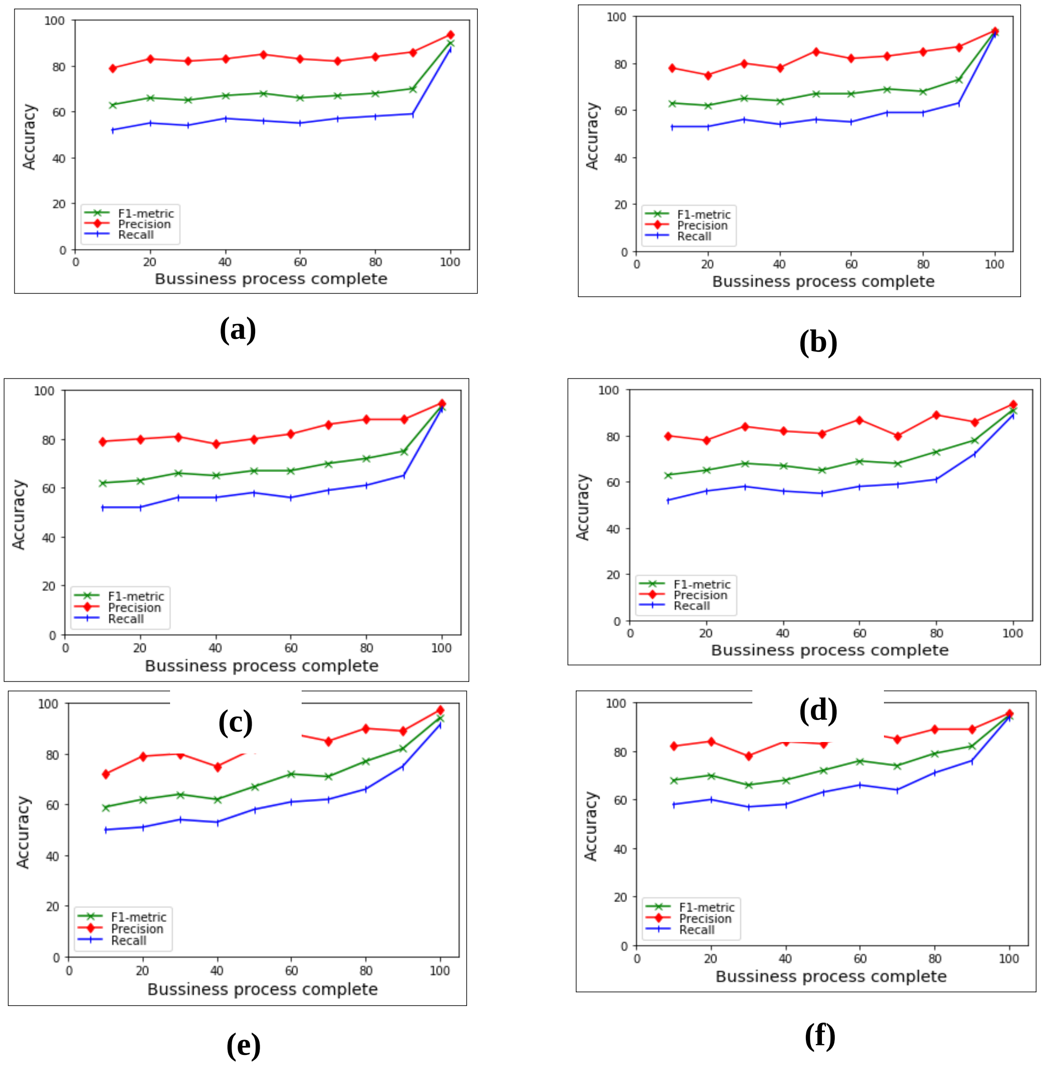


Figure 7. Accuracy indicators graph for (a) Random Forest (b) XGBoost (c) CatBoost (d) ANN (e) Ensemble bagging (f) Ensemble stacking

5.3. Discussion

Implementing the data from the case study, we explored various ways to forecast the accuracy in delay occurrence to avoid transport risk. Tables 6 and Table 7 show the experimental outcomes of all the algorithms in our datasets, where the best result is highlighted in bold, the second-best is italicized, and the least is underlined in Table 7 as per the accuracy values. All the classification algorithms are implemented to generate F-score, precision, recall, specificity, error, and accuracy value to validate the algorithms' significant aspect. Our empirical results show that one important determinant of cargo transport risk is the different delay-related features that drive the entire risk analysis process. The focal findings of the study reveal that selecting the important feature using the Random Forest (RF) algorithm unlocks the primary factors responsible for the delay. To the best of our knowledge, such analysis has not been done before in this particular environment related to risk analysis in the air cargo sector. It is observed that the maximum impact occurs on a transport process when the shipment gets delayed at the outgoing transport leg than the initial incoming legs. The feature selection technique affects the overall phenomenon by reducing the number of input variables when developing a predictive model with better performance. The prediction accuracy and computational efficiency, as mentioned below, reveals the methodological findings generated from the experimental analysis carried out in this study.

Empirical findings obtained from the case study discussed in Section 3 indicate that all predictive methods consistently have an accuracy of at least 70 percent given a lead time of half the duration of the process execution. However, the comparison of our predictive models showed considerable variation in the rate of true/false and positive/negative predictions they provide. Based on these results, predictive monitoring strategies were combined via meta-regressor in such a way that the risk of taking unnecessary steps in the case of false positive predictions or the risk of not taking the required precautions in the case of false-negative predictions was minimized. Results also show that ensemble techniques indeed outperform individual techniques in terms of all the performance indicators. Specifically ensembling the predictions obtained through individual techniques using bagging and stacking improves overall accuracy with the minimum error value. To evaluate the effect of several algorithms, a comparative study was conducted by training our predictive model using RF, XGBoost, CatBoost, and ANN. Further ensembling techniques are used to generate more accurate solutions by combining the predictions of many base estimators to produce one adequate predictive model. As shown in Table 7, accuracy, F-score, and precision values obtained using RF compared with other competitor algorithms fails to exhibit sufficient superiority in our case, although it simplifies transformation and performs well in classification problems. Besides, we found RF requires much computational power and time to build up numerous trees to generate the output with comparatively more error value. In our study, both CatBoost and XGBoost algorithm presents excellent classification performance on extended datasets. The observations are best represented by the performance evaluation metrics used to assess the contingency results. Here, CatBoost outperforms existing boosting algorithms like XGBoost with a minimal margin but the limitation observed while running the CatBoost algorithm is its comparably high computational time as the dataset has many numerical features. In our case, the training time of CatBoost was a little higher than the XGBoost algorithms. In contrast with all other algorithms tested, ANN shows better accuracy, precision, recall, and F-score than other machine learnings algorithms, but this does not provide the best result in our case due to the comparatively maximum error value. The computational time taken by ANN is more compared to other algorithms in our case. All classification algorithms exhibited a comparatively higher computational burden and proneness to overfitting. Our proposed hybrid ensemble technique results better than individual classification algorithms in terms of performance. We observed that ensemble bagging and stacking algorithms generally have a more stable performance during the training and validating process than others. The further threshold changes are observed to have a substantial negative effect on the precision of identification of majority groups, which further impacts prediction accuracy. Among all the techniques ensemble-based stacking methods are the most computationally powerful in predicting accuracy with minimum training time. The classification performance of our proposed algorithms is constrained by several factors, including the size of feature space and the number of base classifiers. To make a more in-depth analysis of the outcomes, Figure 7. visualize prediction accuracy concerning contingency table metrics. We will note from these plots that there is a very limited number of points affecting the prediction accuracy in the execution of the process. At first, accuracy improves significantly after the synchronization point (60 percent mark) and then after the last RCF service of the outgoing transport leg. Initially, accuracy is very low but improves towards the end of the process execution.

The experimental findings help in guiding the development of the optimal classification model, which will be beneficial for the experiments with real-life datasets. The prediction model provided in this article offers an explicit interpretation of delay correlations between time, flight schedule, and previous delay. As delays are directly related to costs, businesses need accurate cargo transport predictions. The model stands to offer strategic and sustainable insights for freight transport planning and other related applications to decision-makers and provides forwarders a roadmap to wisely focus on selecting flights and routes with less transport risk to transfer precious goods. By using Machine Learning efficiently, the decision-makers can easily predict potential delays in a few clicks with minimum stipulated time and adapt their assignment and planning of resources accordingly while simultaneously providing the best customer experience. The techniques used in our study may help understand the air cargo transport industry, lacking a proper platform for data analysis and forecasting. This model has vital application in decision-making from freight forwarders' perspectives. While transporting consolidated shipments through the air with multiple legs, this model will help choose the suitable airlines and routes by which the shipment can be transported. This model is also beneficial in terms of the customer's perspective in choosing the right path.

6. Conclusion

Airfreight forwarders and shippers face threats resulting from competitive market environments, so it is necessary to use risk management strategies to address market shifts. This paper attempted to address the techniques to analyze and predict transport risks using the data from international air cargo logistics. The proposed solution contributes in many ways. Firstly, Cargo 2000 data have been critically analyzed to understand the cause of actual delay that occurs at each transport leg by implementing data engineering. Secondly, electing the most important feature that is responsible for causing delay throughout the shipment process by generating a score of each feature, which further helped in amplifying the performance of classifiers. Thirdly, different classification algorithms are adopted to find out multiple performance indicators such as accuracy, F-score, precision, recall, error values; and lastly a hybrid ensemble integration strategy which includes Bagging and Stacking, is proposed to obtain the final output. The empirical findings indicate that many conventional methods are outperformed by our suggested techniques and achieve more balanced and stable classification outcomes. Our study provides a roadmap to solve practical, real-world problems in air cargo shipment industries by predicting the delay, which will impose a significant advantage on freight forwarders in a competitive environment. The study's key drawback is that it focuses on one network of actors participating in a single air freight handling method. The geographical location, local facilities, and air freight volumes involved in the transaction may also affect the outcomes. This may have consequences for how reflective the situation could be in achieving the study's purpose and objectives. For future research, multiple dimensions such as more transport leg, complex network, and multi-haul flights can be considered. An improved version of machine learning algorithms can be deployed for better results. By implementing advanced algorithms to deal with this problem, our future work will also consider expanding classification approaches and exploring some effective solutions with several other coding strategies.

Appendix

Table A.1

Description of the Features extracted from raw dataset (F= [F1, F2,. . ., F37])

Feature	Description	Feature	Description	Feature	Description	Feature	Description
F1	o_dlv_delay	F11	i1_dep_2_delay	F21	i2_hops	F31	i2_rcf_3_delay
F2	i2_dlv_delay	F12	o_dep_2_delay	F22	o_dep_1_delay	F32	i3_dep_2_delay
F3	i1_dlv_delay	F13	i2_rcf_2_delay	F23	i1_rcs_delay	F33	legs
F4	i3_dlv_delay	F14	i2_rcs_delay	F24	i1_hops	F34	i3_dep_3_delay
F5	o_rcf_2_delay	F15	i1_rcf_2_delay	F25	i1_dep_1_delay	F35	i3_rcf_3_delay
F6	i3_rcf_1_delay	F16	i2_dep_1_delay	F26	i3_dep_1_delay	F36	o_dep_3_delay
F7	i3_rcf_2_delay	F17	o_rcs_delay	F27	i2_dep_2_delay	F37	i2_dep_3_delay
F8	o_rcf_1_delay	F18	o_hops	F28	i1_rcf_3_delay		
F9	i3_hops	F19	i2_rcf_1_delay	F29	i1_dep_3_delay		
F10	i1_rcf_1_delay	F20	i3_rcs_delay	F30	o_rcf_3_delay		

References

Agarwal, S., & Chowdary, C. R. (2020). A-Stacking and A-Bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection. *Expert Systems with Applications*, *146*. https://doi.org/10.1016/j.eswa.2019.113160

Ambra, T., Caris, A., & Macharis, C. (2019). Towards freight transport system unification: reviewing and combining the advancements in the physical internet and synchromodal transport research. *International Journal of Production Research*, *57*(6), 1606–1623. https://doi.org/10.1080/00207543.2018.1494392

Archetti, C., & Peirano, L. (2019). Air intermodal freight transportation: The freight forwarder service problem. *Omega*, 102040. https://doi.org/10.1016/j.omega.2019.02.009

Azadian, F., Murat, A. E., & Chinnam, R. B. (2012). Dynamic routing of time-sensitive air cargo using real-time information. *Transportation Research Part E: Logistics and Transportation Review*, *48*(1), 355–372. https://doi.org/10.1016/j.tre.2011.07.004

Bai, B., Li, G., Wang, S., Wu, Z., & Yan, W. (2021). Time series classification based on multi-feature dictionary representation and ensemble learning. *Expert Systems with Applications*, *169*(September 2020), 114162. https://doi.org/10.1016/j.eswa.2020.114162

Barbour, W., Martinez Mori, J. C., Kuppa, S., & Work, D. B. (2018). Prediction of arrival times of freight traffic on US railroads using support vector regression. *Transportation Research Part C: Emerging Technologies*, *93*(May), 211–227. https://doi.org/10.1016/j.trc.2018.05.019

Bejani, M. M., & Ghatee, M. (2018). A context aware system for driving style evaluation by an ensemble learning on smartphone sensors data. *Transportation Research Part C: Emerging Technologies*, *89*(February), 303–320. https://doi.org/10.1016/j.trc.2018.02.009

Breiman, L. (2001). ST4_Method_Random_Forest. *Machine Learning*, *45*(1), 5–32.
Breiman, Leo. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.3390/risks8030083
Chan, F. T. S., Shekhar, P., Tiwari, M. K., Chan, F. T. S., Shekhar, P., & Dynamic, M. K. T. (2014). Dynamic scheduling of oil tankers with splitting of cargo at pickup and delivery locations : a Multi- objective Ant Colony-based approach. *International Journal of Production Research*, *52*(24), 7392–7409. https://doi.org/10.1080/00207543.2014.932932
Chih, M., Rico, T., Jiana, M., & Wang, F. (2021). What drives freight transportation customer loyalty ? Diverging marketing approaches for the air freight express industry. *Transportation*, *48*(3), 1503–1521. https://doi.org/10.1007/s11116-020-10104-0
Choi, T. M., Wen, X., Sun, X., & Chung, S. H. (2019). The mean-variance approach for global supply chain risk analysis with air logistics in the blockchain technology era. *Transportation Research Part E: Logistics and Transportation Review*, *127*(May), 178–191. https://doi.org/10.1016/j.tre.2019.05.007
Dolgui, A., Tiwari, M. K., Sinjana, Y., Kumar, S. K., & Son, Y. (2018). Optimising integrated inventory policy for perishable items in a multi-stage supply chain. *International Journal of Production Research*, *7543*, 1–24. https://doi.org/10.1080/00207543.2017.1407500
Du, W., Zhang, M., Zhang, Y., Cao, X., & Zhang, J. (2018). Delay causality network in air transport systems. *Transportation Research Part E*, *118*(September), 466–476. https://doi.org/10.1016/j.tre.2018.08.014
El-dahshan, E. A., Hosny, T., & Salem, A. M. (2010). Hybrid MRI techniques for brain image classification Hybrid intelligent techniques for MRI brain images classification. *Digital Signal Processing*, *20*(2), 433–441. https://doi.org/10.1016/j.dsp.2009.07.002
Elbert, R., Bogusch, C., & Özsucu, Ö. (2012). Risk Management for Air Freight Forwarders: Analysis of Flexible Price Agreements and Financial Hedging. *Supply Chain Forum: An International Journal*, *13*(4), 40–50. https://doi.org/10.1080/16258312.2012.11517305
Han, C. P., Knott, K., & Egbelu, P. J. (2007). A heuristic approach to the three-dimensional cargo-loading problem. *International Journal of Production Research*, *7543*. https://doi.org/10.1080/00207548908942585
Hendricks, K. B., & Singhal, V. R. (2005). An empirical analysis of the effect of supply chain disruptions on long-run stock price performance and equity risk of the firm. *Production and Operations Management*, *14*(1), 35–52. https://doi.org/10.1111/j.1937-5956.2005.tb00008.x
Hughes, S., Moreno, S., Yushimito, W. F., & Huerta-Cánepa, G. (2019). Evaluation of machine learning methodologies to predict stop delivery times from GPS data. *Transportation Research Part C: Emerging Technologies*, *109*(October 2018), 289–304. https://doi.org/10.1016/j.trc.2019.10.018
IATA. (2017a). *Air freight volumes post their strongest year of growth since 2010*. *December 2017*, 1–4.
IATA. (2017b). *Cargo 2000 Dataset*. Kaggle. https://www.kaggle.com/crawford/cargo-2000-dataset
Lin, L., Wang, F., Xie, X., & Zhong, S. (2017). Random forests-based extreme learning machine ensemble for multi-regime time series prediction. *Expert Systems with Applications*, *83*, 164–176. https://doi.org/10.1016/j.eswa.2017.04.013
Metzger, A., Leitner, P., Ivanovi, D., Schmieders, E., Franklin, R., Carro, M., Dustdar,

S., & Pohl, K. (2015). Process Monitoring Techniques. *Ieee Transactions on Systems, Man, and Cybernetics: Systems*, *45*(2), 276–290.

Metzger, A., Leitner, P., Ivanovic, D., Schmieders, E., Franklin, R., Carro, M., Dustdar, S., & Pohl, K. (2015). Comparing and combining predictive business process monitoring techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(2), 276–290. https://doi.org/10.1109/TSMC.2014.2347265

Mursalin, M., Zhang, Y., Chen, Y., & Chawla, N. V. (2017). Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing*, *241*, 204–214. https://doi.org/10.1016/j.neucom.2017.02.053

Nair, R., Hoang, T. L., Laumanns, M., Chen, B., Cogill, R., Szabó, J., & Walter, T. (2019). An ensemble prediction model for train delays. *Transportation Research Part C: Emerging Technologies*, *104*(May), 196–209. https://doi.org/10.1016/j.trc.2019.04.026

Neal, C., & Koo, T. T. R. (2020). Demand for cargo airships: An analysis of mode choice decision making in the freight transport industry. *Journal of Air Transport Management*, *83*(October 2019), 101741. https://doi.org/10.1016/j.jairtraman.2019.101741

Pathak, D. K., Thakur, L. S., & Rahman, S. (2019). Performance evaluation framework for sustainable freight transportation systems. *International Journal of Production Research*, *57*(19), 6202–6222. https://doi.org/10.1080/00207543.2019.1602741

Pavlov, Y. L. (2019). Random forests. *Random Forests*, 1–122. https://doi.org/10.1201/9780367816377-11

Porwik, P., Doroz, R., & Wrobel, K. (2019). An ensemble learning approach to lip-based biometric verification, with a dynamic selection of classifiers. *Expert Systems with Applications*, *115*, 673–683. https://doi.org/10.1016/j.eswa.2018.08.037

Ribeiro, M. H. D. M., & dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied Soft Computing Journal*, *86*, 105837. https://doi.org/10.1016/j.asoc.2019.105837

Rodríguez-Sanz, Á., Comendador, F. G., Valdés, R. A., Pérez-Castán, J., Montes, R. B., & Serrano, S. C. (2019). Assessment of airport arrival congestion and delay: Prediction and reliability. *Transportation Research Part C: Emerging Technologies*, *98*(November 2018), 255–283. https://doi.org/10.1016/j.trc.2018.11.015

Ross, A., Jayaraman, V., & Robinson, P. (2010). Optimizing 3PL service delivery using a cost-to- serve and action research framework. *International Journal of Production Research*, *7543*. https://doi.org/10.1080/00207540600603969

Schmitz, G. P. J., Aldrich, C., & Gouws, F. S. (1999). ANN-DT : An Algorithm for Extraction of Decision Trees from Artificial Neural Networks. *IEEE Transactions on Neural Networks*, *10*(6), 1392–1401.

Shang, Y., Dunson, D., & Song, J. S. (2017). Exploiting big data in logistics risk assessment via Bayesian nonparametrics. *Operations Research*, *65*(6), 1574–1588. https://doi.org/10.1287/opre.2017.1612

Tang, C. S. (2006). Perspectives in supply chain risk management. *International Journal of Production Economics*, *103*(2), 451–488. https://doi.org/10.1016/j.ijpe.2005.12.006

Thiagarajan, B., Srinivasan, L., Sharma, A. V., Sreekanthan, D., Vijayaraghavan, V., Student, U., & Foundation, S. (2017). A Machine Learning Approach for

Prediction of On-time Performance of Flights. *IEEE Access*.
Tsolakis, N., Zissis, D., Papaefthimiou, S., & Korfiatis, N. (2021). Towards AI driven environmental sustainability : an application of automated logistics in container port terminals logistics in container port terminals. *International Journal of Production Research*, *0*(0), 1–21. https://doi.org/10.1080/00207543.2021.1914355
Tu, Y., Ball, M. O., & Jank, W. S. (2008). Estimating flight departure delay distributions - A statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, *103*(481), 112–125. https://doi.org/10.1198/016214507000000257
Wen, X., Xu, X., Choi, T.-M., & Chung, S.-H. (2019). Optimal Pricing Decisions of Competing Air-Cargo-Carrier Systems--Impacts of Risk Aversion, Demand, and Cost Uncertainties. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1–15. https://doi.org/10.1109/tsmc.2019.2930725
Wolpert, D. (1992). Stacked Generalization (Stacking). *Neural Networks*, *5*, 241–259.
Yoon, J., Talluri, S., Yildiz, H., & Sheu, C. (2020). The value of Blockchain technology implementation in international trades under demand volatility risk. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2019.1693651
Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport : A deep learning approach. *Transportation Research Part E*, *125*(October 2018), 203–221. https://doi.org/10.1016/j.tre.2019.03.013
Yu, J. J. Q. (2020). Semi-supervised deep ensemble learning for travel mode identification. *Transportation Research Part C: Emerging Technologies*, *112*(July 2019), 120–135. https://doi.org/10.1016/j.trc.2020.01.003
Zhang, P., Member, G. S., & Kuang, H. (2021). An Equilibrium Service Capability Recovery Model for a Port-Hinterland Service Network With Delay or Disruption Risk. *IEEE Access*, *9*, 79033–79050. https://doi.org/10.1109/ACCESS.2021.3084302
Zheng, S., Yang, Z., He, Z., Wang, N., & Chu, C. (2020). Hybrid simulated annealing and reduced variable neighbourhood search for an aircraft scheduling and parking problem. *International Journal of Production Research*, *7543*. https://doi.org/10.1080/00207543.2019.1629663
Zhu, Y., Zhou, L., Xie, C., Wang, G. J., & Nguyen, T. V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, *211*(September 2018), 22–33. https://doi.org/10.1016/j.ijpe.2019.01.032
Zoutendijk, M. (2021). Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem. *Aerospace*, *8*(152).
Zúñiga, C., Piera, M. Á., & Narciso, M. (2011). Revisiting the pallet loading problem using a discrete event system approach to minimise logistic costs. *International Journal of Production Research*, *7543*(May 2010). https://doi.org/10.1080/00207541003702234