Identifying the Major Causes Associated to Rail Intermodal Operation Disruptions Using Causal Machine Learning

Juan Pineda-Jaramillo
Department of Engineering
University of Luxembourg, Esch-sur-Alzette, Luxembourg, 4365
Email: juan.pineda@uni.lu

William McDonald
Luxembourg Centre for Logistics and Supply Chain Management
University of Luxembourg, Luxembourg City, Luxembourg, 1359
Email: william.mcdonald.001@student.uni.lu

Wei Zheng
Luxembourg Centre for Logistics and Supply Chain Management
University of Luxembourg, Luxembourg City, Luxembourg, 1359
Email: wei.zheng.001@student.uni.lu

Francesco Viti
Department of Engineering
University of Luxembourg, Esch-sur-Alzette, Luxembourg, 4365
Email: francesco.viti@uni.lu

Word Count: 6,160 words + 4 tables (250 words per table) = 7,160 words

Submitted [July 29, 2021]

ABSTRACT

Intermodal rail operations represent a complex stochastic system that is impacted by disruptions from diverse causes like extreme weather events, planned and unplanned upstream network delays, equipment failures, labor actions, and intra-railyard inefficiency. Understanding and predicting the occurrence of these disruptions holds the potential to limit their system-wide schedule impact through early-warning prompting mitigating actions.

This paper presents the results of a set of machine learning models trained to predict disruptions in rail intermodal operations, and the most suitable model in terms of the evaluation metrics (e.g., AUC, recall, and F1-score) was used to explore the major predictors of the disruptions and their subsequent delays. The supporting dataset includes intermodal rail journeys with origin the central station of the freight rail network of the National Railway Company of Luxembourg in Bettembourg, connecting several EU countries terminals.

Results show that a gradient boosting machine model, using the CatBoost implementation, outperforms other ML models in terms of the selected evaluation metrics. Additionally, results suggest that the train weight, train length, number of wagons composing the train, weight per wagon, and the month of operation are the major predictors that cause the disruptions in the intermodal operations in the studied rail network.

The outcome of the study suggests that a better distribution of freight weight across the wagons will reduce the probability of a delayed trip, and this insight can be used to optimize the intermodal operations of the National Railway Company of Luxembourg.

Keywords: Rail operation disruptions, Machine learning, Freight transport, Logistics operations, Gradient boosting.

INTRODUCTION

Rail is increasingly chosen for freight transport since it often provides the best set of trade-offs with respect to operational costs, reliability and efficiency, and its utilization increasingly occurs within the intermodal context. Moreover, rail transport is much safer for both operator staff and the public compared to its competitors due to aspects such as advanced control systems that reduce human errors (*1*). Considering that rail intermodal operations is one of the principal actions in the freight transport sector, it is essential to optimize all aspects including the operational use of the rail infrastructure. Disruptions, and their associated delays in rail intermodal operations may occur for many reasons (e.g., disturbance in the flow of operations, accidents, technical failures, lower-than-planned travel speeds, construction and repair works, and extreme weather conditions (*2*, *3*)).

Delays represent positive deviations between realized and scheduled times of events; they are often classified in two groups: those that are caused directly by the variability of process times preparing the train for departure and those caused by the variability in the actual operation of the train along its journey (4). Additionally, delays may be categorized as "primary" delays that originate unexpectedly as the result of extensions of the planned times of individual scheduled processes and "secondary" delays that directly result from the occurrence of a primary delay (5, 6).

Once a disruption occurs, train dispatchers must assess the severity of its impact on the overall schedule and try to reduce losses by taking actions during the operation, in order to reduce the chain of delays that could affect the entire system operation (*7*, *8*). Train disruption prediction, with the aim of optimizing the rail operation and reducing subsequent delays, has been widely studied by different authors (*9–11*); where it is possible to find various approaches, from stochastic methods (*12*, *13*) to machine learning (ML) models (*14–18*). Despite the significant body of research, train disruption prediction models continue to struggle to predict delays and guide mitigating actions in operational environments. Specifically, they often fail to identify the underlying causes of delay and the expected impact to operations, which significantly limits the efficacy of mitigation actions.

Advances in artificial intelligence (AI) have shown promise in addressing the limitations of conventional models. For instance, ML and more specifically deep learning techniques can be applied to process and detect connections between nonlinear, high-dimensional and sequential/time series data (*19–22*), being successfully used to find causality, rather than just correlation, in rail operations (*23*, *24*). Identifying disruptions and their subsequent delay causality in rail intermodal operations is an integral part of quality operational and strategic decision-making and, at the same time, gaps in causal understanding is directly related to ML's challenges of generalizability, interpretability, explainability, and ultimately ability to build robust operational models (*25–27*).

Overall, the existing literature has focused mainly on examining the direct correlations between predictors and disruption occurrence and/or severity in rail operations using traditional statistical methods or ML models, in order to find the best model in terms of prediction accuracy without explaining the causality of disruptions (*5*, *10–17*, *28*). Some studies have examined causal interrelationships between predictors and train disruptions, but they have tended to be highly theoretical, rather than inferring/learning causal links from real-world operational data (6). Furthermore, previous studies dealing with the application of ML models to predict rail disruptions have focused mainly on passenger trains, while studies carried out on freight trains have focused on applying ML models to analyze the impact of design of the rail network rather than the impact of operational features without explaining the contribution of each feature to the disruptions and their subsequent delays (*5*, *29*) . To address a number of these issues, this study uses ML to develop a structural causal network from operational data, exploring the causal relationships between rail operating features and disruptions. The causal network is used to develop a predictive model capable of predicting the risk of disruptions and their subsequent delay causality in rail intermodal operations.

More specifically, this study addresses the effect of different features on disruptions in rail intermodal operations using a supporting dataset of 7,969 trips occurring between November 2019 and April 2021 between the central terminal of the freight rail network of the National Railway Company of Luxembourg (CFL multimodal) in Bettembourg (Luxembourg) and connecting with several countries within the EU. After evaluating a set of ML models, the most suitable approach is employed to identify the major predictors behind the disruptions. Within this framework, the purposes of this study and main contributions of the paper are:

- The examination of different ML models for predicting disruptions in rail intermodal operations based on operations data. We consider different sources of data available for rail intermodal operations in a network connecting various EU countries. To the best of our knowledge, and on the basis of the found literature, this

is the first time that operational data from an extensive network linking various countries has been used for the prediction of disruptions in rail intermodal operations and their subsequent delays.

- The use of ML models is investigated to develop a structural causal network from data to explore the causal interrelationships between operational rail predictors with the aim of discovering the interactions between those predictors that lead to disruptions in rail intermodal operations.

METHODS

This section describes the process to develop a predictive model capable of predicting a disruption that causes a delayed train, and then of exploring the predictors of disruptions in rail intermodal operations applying causal ML approaches. **Figure 1** illustrates the flowchart of the methodology implemented in this study, from data collection to examination of the major key predictors that most influence disruptions in rail intermodal operations. All steps presented in the flowchart are discussed below.

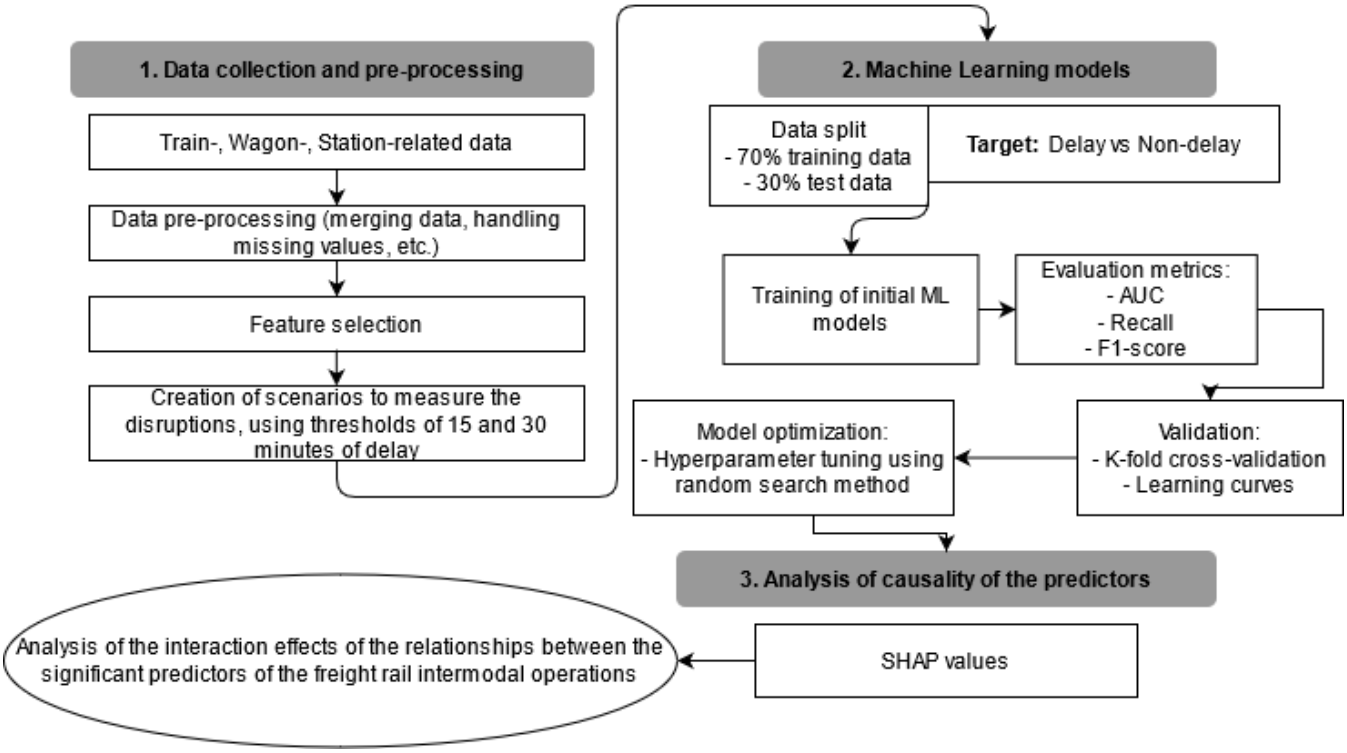

Figure 1 Flowchart of the methodology implemented in this study

Data collection and pre-processing

Data of rail intermodal operations carried out by *CFL multimodal* were provided by the company itself. The CFL data is primarily divided into three datasets regarding (a) the trains (e.g., features concerning the origin and destination of the trains such as station, country, planned and actual times of departure, planned and actual times of arrival, maximum TEU (Twenty-foot equivalent unit), incoterm, TEU count, max length, train length in meters, etc.); (b) the wagons (e.g., information on each wagon comprising the trains of the previous dataset: wagon model, order or wagons in the composition of the train, model max speed, tare weight of the wagon model, tare weight of the actual wagon, etc.); and (c) the stations (e.g., information on the stations through which the trains pass: order, name, city, country, planned and actual times of departure, planned and actual times of arrival, geographical locations of intermediary stations during the trip, etc.). The datasets include the operations performed by CFL multimodal from November 2019 to April 2021.

A unique dataset is created after joining and merging the available datasets using common identifiers (IDs). Then, the unique dataset is filtered to extract the subset of observations related to operations originating in

Bettembourg. In addition, each row is organized in such a way that it presents the trips between a pair of control stations, which are the stations that the train crosses between the station of origin and the station of destination.

After obtaining an unique dataset that presents the trips originating in Bettembourg, the dataset was pre-processed using traditional methods developed in data mining processes such as removing or filling missing values following specific criteria such as the use of the median (numerical features) or the most common values (categorical features), transforming the categorical features using dummy variables, removing redundant features by analyzing the multicollinearity between predictors and identifying the predictive power score, among others (*20*). Furthermore, the z-score normalization method was applied to the dataset to rescale the values of the numerical features without distorting differences in the ranges of values or losing information in order to enhance the results of the models (*30–32*). The z-score normalization method is calculated as presented in **Equation (1)**, where z is the new value calculated using x as the previous value, μ as the mean, and σ as the standard deviation for each column in the dataset.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

The feature to be predicted in this study is the deviation between the scheduled arrival time and the actual arrival time in the trips between a pair of control stations (i.e., the arrival delay for each station pair). We propose a specific threshold in order to identify whether the trip is delayed or not due to the occurrence of a disruption, so the underlying problem can be defined as a binary classification problem (delayed or non-delayed). This approach was chosen because it allows to achieve more general results allowing to explore the causal interrelationships between operational rail predictors with the aim of discovering the interactions between those predictors that lead to disruptions and their subsequent delays in rail intermodal operations. Additionally, in order to evaluate the robustness of the ML models and the significance of the predictors, we treated the binary classification problem using two scenarios, using a threshold of (a) 15 minutes, and (b) 30 minutes. After pre-processing the dataset, we identified a total of 7,969 trips between control stations.

Machine Learning models

Data were randomly split into a training set and a test set with a ratio of 70%-30%, respectively, where each subset was composed of the target feature to predict (*arrived*) and the remaining independent input features following the same proportion, making several tests including different combinations of input features in order to improve the results. This division is carried out with the objective of using the training set to train the ML models, and then using the test set to evaluate their performance (*24*). Then, a set of ML models was trained to predict whether the trip is delayed or not and those models were chosen because they have been widely and successfully implemented in classification problems in different fields (see **Table 1**). Furthermore, the *AUC*, *recall* and *F1-score* evaluation metrics were used as loss functions.

In order to evaluate the performance of the ML models and select the best performing one, the stratified *K*-fold cross-validation method was used to reduce any bias generated by the model (*45*, *46*). This method divides the training set into K subsets, and for each subset the classes are characterised in roughly the same numbers as the entire training set and the incumbent model is applied to the other $K - 1$ subsets, and then the *AUC* metric is selected to evaluate the performance of the model in the holdout/test subset. Furthermore, the process of optimizing a ML model implies the need to tune a set of parameters in order to improve its performance, and the most widely used method to carry out this process is the random search method, which allows evaluating the parameter values that have a greater impact on the performance of the ML model (*47*).

After choosing the best ML model for predicting whether the trip is delayed or not, the learning curve method was implemented in order to identify whether the model has *overfitting* or *underfitting* problems. This method allows analyzing the behavior of the model as a greater number of observations are used in the training process (*48*). The processing time for the training and validation of the ML models is negligible, taking only a few minutes using Python 3.8.5 in an Intel Core i9-10885H CPU @ 2.40 GHz with a memory ram of 32 GB, DDR4, a Hard Disk SSD 1TB NVMe class 40, and a GPU NVIDIA Quadro P620 DDR5.

TABLE 1 ML models used in this study

Model	Description	Literature
Logistic Regression	This model uses a logistic function to model the probability that an observation belongs to a class.	(*33*)
Linear Discriminant Analysis	This model takes advantage of powerful dimensionality reduction methods by locating an efficient linear transformation where the original high-dimensional space data are transformed to a lower dimensional space, and it assumes an identical covariance for all classes.	(*34–36*)
Quadratic Discriminant Analysis	Similar to Linear Discriminant Analysis, but it does not assume that the covariance of each of the classes is identical.	(*36*)
Naïve Bayes	It calculates class probabilities using Bayes theorem while assuming that the features are independent.	(*37–39*)
Gaussian Process Classification	This model assumes some prior distribution on the underlying probability densities and then determines the final classification providing a good fit for the observed data and guaranteeing smoothness.	(*40*)
Multi-Layer Perceptron	This model is an Artificial Neural Network that creates a set of outputs from a set of inputs and is characterized by several layers of input nodes connected as a directed graph between the input and output layers.	(*20*, *24*, *38*)
Support Vector Machine (Radial Kernel)	It classifies observations by projecting the independent features into a high-dimensional feature space, where the classes are linearly separable.	(*14*, *20*, *37–39*, *41*)
K-Nearest Neighbors	It predicts the class of the test sample according to the k training samples, which are the nearest neighbors to the test sample.	(*36*)
Extra Trees	Ensemble method that randomly combines the predictions from many decision trees with the aim of minimizing the variance of the prediction results.	(*41*, *42*)
Adaptive Boosting	Ensemble method that combines the output of different learning models to create a weighted sum that represents the final output of the classifier.	(*39*, *41*)
Random Forest	Ensemble method that uses feature randomness when training many decision trees in parallel in order to create an uncorrelated forest of trees with major accuracy.	(*24*, *37*, *38*)
Gradient Boosting	Ensemble method that builds a sequence of decision trees, where each successive tree aims to improve the previously wrong classifications of the preceding trees.	(*24*, *38*)
CatBoost	Same Gradient Boosting model but using a different open-source library.	(*43*, *44*)

Analysis of causality of the predictors

The Shapley Additive Explanation technique (commonly known as *SHAP* method) is implemented to extract the feature importance and the direct impact of each feature in order to better interpret the output of the ML model, with numerous studies having taken advantage of this method (*49–52*).

The *SHAP* method is based on game theory and offers an insightful method for estimating the contribution of each feature in the output of a model by averaging the differences in predictions over all possible orderings of all input features based on precise solutions to create *SHAP values* (*43*, *53*). This method allows obtaining the magnitude of the contributions of the predictors in the prediction of delayed trips,

In order to obtain the *SHAP values*, suppose an ML model where a group N (with n features) is used to predict an output N, then the contribution ϕ_i of the feature i on the model output $v(N)$ is assigned based on its marginal contribution. Based on numerous axioms to help assign the contribution of each variable, *SHAP values* are calculated as presented in **Equation (2)**, where a linear function of binary variables g is defined based on the additive variable attribution method presented in **Equation (3)**, where $z' \in \{0,1\}^M$ is 0 when a variable is not observed, otherwise is 1, M represents the number of simplified input variables, and $\phi_i \in \mathbb{R}$ (*53*).

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{2}$$

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i Z_i' \tag{3}$$

RESULTS

In this section we present the results of the trained ML models in order to identify the major predictors of disruptions and their subsequent delays in the rail intermodal operations. This section involves two stages: (a) the results of the trained ML models in order to achieve the best performing predictive model based on the evaluation metrics: *AUC*, *recall*, and *F1-score*; and (b) the identification of major predictors of disruptions in rail intermodal operations using the *SHAP* method, in order to analyze the causality of the predictors in the output of the best ML model.

ML models

Several combinations of features were tested in order to identify the combination that achieves better results in the ML models, where the final composition of the dataset is presented in **Table 2**.

TABLE 2 Composition of the final dataset

Feature (type)	Description	Categories	Distribution of the feature in the dataset
TARGET: arrived	Delayed or non-delayed (thresholds: (a) 15 min, (b) 30 min).	2	(a) delayed: 47.8%; non-delayed: 52.2%. (b) delayed: 39.0%; non-delayed: 61.0%.
month_arrival	Month of arrival	12	min: May: 4.4%; max: March:15.6%
Train_Length	Train length [m]	numerical	range: 136.0 - 704.0; median: 544.0
Train_Weight	Train weight, including train tare and freight weight [t]	numerical	range: 272.7 - 2121.0; median: 1278.1
wagon_count	Number of wagons	numerical	range: 4 - 34; median: 16
weight_wagon	Average weight per wagon [t]	numerical	range: 20.8 - 100.7; median: 76.7
Train_Distance_KM	Distance of the TOTAL trip [km]	numerical	range: 84.5 - 1197.9; median: 648.6

Number of observations: 7,969.

Then, after implementing the ML models previously mentioned using the two scenarios (thresholds of 15 and 30 minutes) to define whether the trip between a pair of control stations is delayed or not using the stratified *K*-fold cross-validation method (with *K*=10), we obtained the initial results presented in **Table 3**, where the evaluation metrics are presented. Then, considering that the Random Forest (RF), the Extra Trees (ET), and the CatBoost models have the best performance in terms of the evaluation metrics for both scenarios, the random search method was implemented in order to tune their parameters and improve their performance. The results of the evaluation metrics for each class of the best ML models after tuning their parameters are presented in **Table 4** for initial and tuned models, and their learning curves are presented in **Figure 2**, where it is possible to observe a convergence trend between both curves (training and cross-validation scores) across the full range of models, suggesting that adding more observations to train these models is likely enhance their performance, reducing the risks of *overfitting* and *underfitting* problems.

TABLE 3 Initial results

	15-minutes (delayed trips: 47.8%)			30-minutes (delayed trips: 39.0%)		
Model	**AUC**	**Recall**	**F1-score**	**AUC**	**Recall**	**F1-score**
Random Forest (RF)	**0.771**	0.691	**0.691**	**0.800**	0.671	**0.666**
Extra Trees (ET)	0.758	0.629	0.669	0.785	0.604	0.639
CatBoost	0.729	0.644	0.649	0.766	0.662	0.641
K-Nearest Neighbors (KNN)	0.705	0.637	0.639	0.744	0.661	0.626
Gradient Boosting (GB)	0.684	0.620	0.615	0.716	0.615	0.596
Ada Boost (AB)	0.648	0.578	0.580	0.674	0.617	0.561
Gaussian Process (GP)	0.637	0.637	0.596	0.672	0.664	0.573
MLP Classifier	0.626	0.605	0.582	0.669	0.646	0.568
SVM - Radial Kernel	0.594	0.706	0.607	0.618	0.659	0.538
Logistic Regression (LR)	0.579	0.622	0.569	0.606	0.643	0.534
Linear Discriminant Analysis (LDA)	0.579	0.623	0.569	0.606	0.645	0.535
Naïve Bayes (NB)	0.574	**0.737**	0.609	0.594	**0.713**	0.548
Quadratic Discriminant Analysis (QDA)	0.525	0.675	0.535	0.512	0.672	0.459

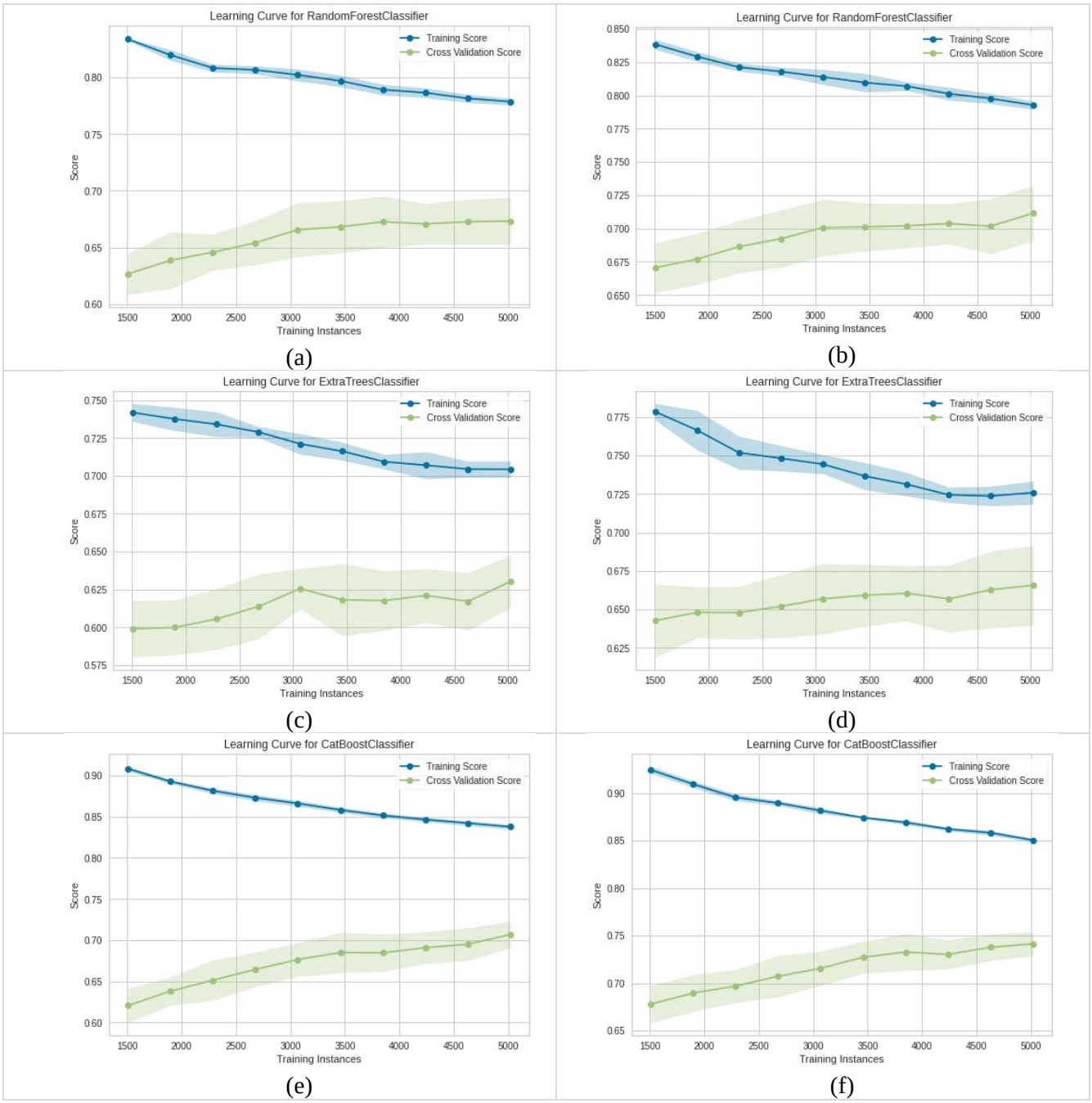

Figure 2 Learning curves of (a) tuned RF (threshold 15-min), (b) tuned RF (threshold 30-min, (c) tuned ET (threshold 15-min), (d) tuned ET (threshold 30-min), (e) tuned CatBoost (threshold 15-min), (f) tuned CatBoost (threshold 30-min)

TABLE 4 Results of the best ML models after tuning their parameters using the random search method

Model	Class	AUC	Recall	F1-score	AUC	Recall	F1-score
Initial RF	delayed train	0.790	0.708	0.710	0.780	0.659	0.644
	non-delayed train	0.790	0.737	0.735	0.780	0.752	0.763
Tuned RF	delayed train	0.760	0.673	0.678	0.760	0.658	0.636
	non-delayed train	0.760	0.717	0.711	0.760	0.737	0.753
Initial ET	delayed train	0.780	0.644	0.689	0.770	0.578	0.613
	non-delayed train	0.780	**0.793**	**0.749**	0.770	**0.803**	**0.775**
Tuned ET	delayed train	0.700	0.701	0.655	0.710	0.622	0.586
	non-delayed train	0.700	0.599	0.640	0.710	0.679	0.707
Initial CatBoost	delayed train	0.760	0.669	0.673	0.740	**0.660**	0.626
	non-delayed train	0.760	0.708	0.704	0.740	0.713	0.739
Tuned CatBoost	delayed train	**0.800**	**0.714**	**0.714**	**0.780**	0.656	**0.647**
	non-delayed train	**0.800**	0.738	0.738	**0.780**	0.762	0.769

From **Table 4** and **Figure 2** it is possible to see that the tuned CatBoost model (a specific framework of the gradient boosting model) outperforms in terms of the evaluation metrics for predicting if the trip in rail intermodal operations is delayed or not considering both scenarios. Therefore, this model is selected to evaluate the impact of the input features on the output of the model, and, furthermore, to analyze causality of disruptions and their subsequent delays.

The gradient boosting model is a tree-based ensemble model that incorporates many weak-learner-models in order to establish a strong-learner model. Hence, a single weak-learner model might not achieve high accuracy for the entire dataset, but it can perform well enough for a subset of the dataset, which means that each weak-learner model improves the performance of the entire model (*54*, *55*). On the other hand, the CatBoost model is a high-performance open-source library for gradient boosting models that has become popular due to demonstrated superiority in performance in different attributes (e.g., minor probability of overfitting, native handling for categorical features and speed of execution) compared to other gradient boosting implementations (*44*, *56*).

To the best of our knowledge, this is the first study in which the gradient boosting model has been used in rail intermodal operation research, more specifically the CatBoost implementation. Other studies that have obtained acceptable results using different ML models in rail operations, but mainly for passenger trains and without studying gradient boosting machines—for instance, implementing stochastic model to predict the propagation of train delays based on Bayesian networks (*11*), and using a support vector machine model to examine the relationship between passenger train delays and some characteristics of railway systems (*14*). Moreover, some studies have explored the use of artificial neural networks to predict delays in passenger trains (*15*, *57–59*) and in freight trains (*29*), including some explorations using deep neural networks for large-scale railway networks (*5*, *60*).

Major predictors of rail intermodal operation disruptions and their subsequent delays

Figure 3 provides the SHAP summary plot that represents an ordered list of the most important features for identifying delayed trips for both scenarios. Although the order of importance of the features for both thresholds varies somewhat, in essence we can observe that the composition of the train affects the results of the ML model. From **Figure 3 (a)** it is possible to observe that higher values of train weight and lower number of wagons correspond to a higher probability of a delayed trip, while the behavior of train distance, weight per wagon, and train length does not allow to visibly identify a trend in their values. On the other hand, **Figure 3 (b)** shows that higher values of train length and weight per wagon correspond to a higher probability of a delayed trip, while it is not possible to visibly recognize a trend in the remaining features, because the high and low values of them are completely mixed. The results also suggest that the operations carried out in some months have a great incidence in operational delays for both scenarios.

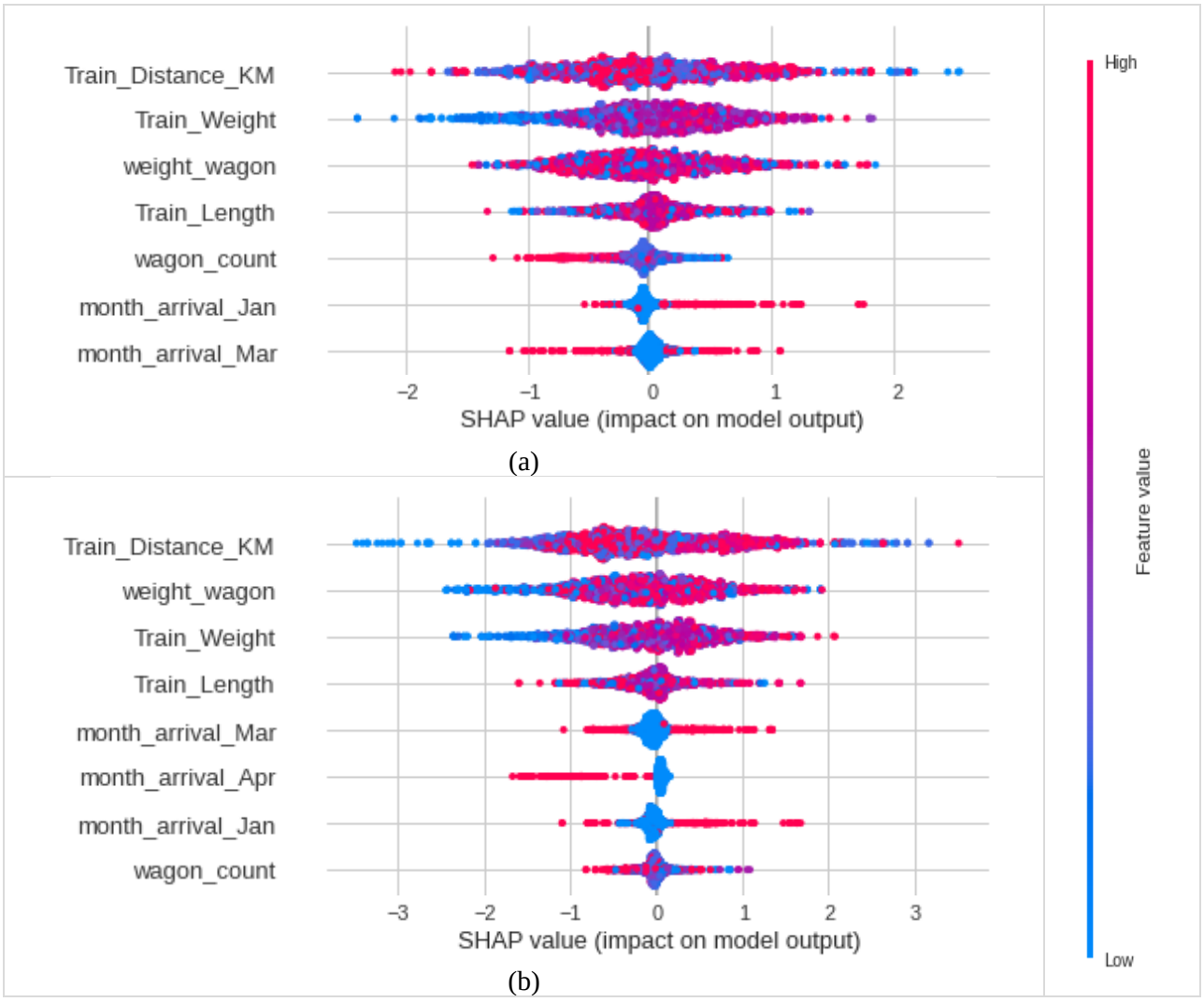

Figure 3 SHAP summary plot for both scenarios, using thresholds of (a) 15 minutes, and (b) 30 minutes

DISCUSSION

The SHAP values allow the identification of train- and operational-related features (e.g., the train weight, train length, number of wagons, weight per wagon, and the month of operation) that explain the disruptions and their subsequent delays in the intermodal operations of the studied rail network. Overall, the greater the train weight, length and weight per wagon, and the lower the number of wagons, the greater the probability that the trip will be delayed, and this delay probability is dependent on the month of the year (**Figure 4**). These interaction plots suggest that if the weight of the freight carried by the train is distributed over a larger number of wagons in an averaged way, the probability that the trip will be delayed will decrease.

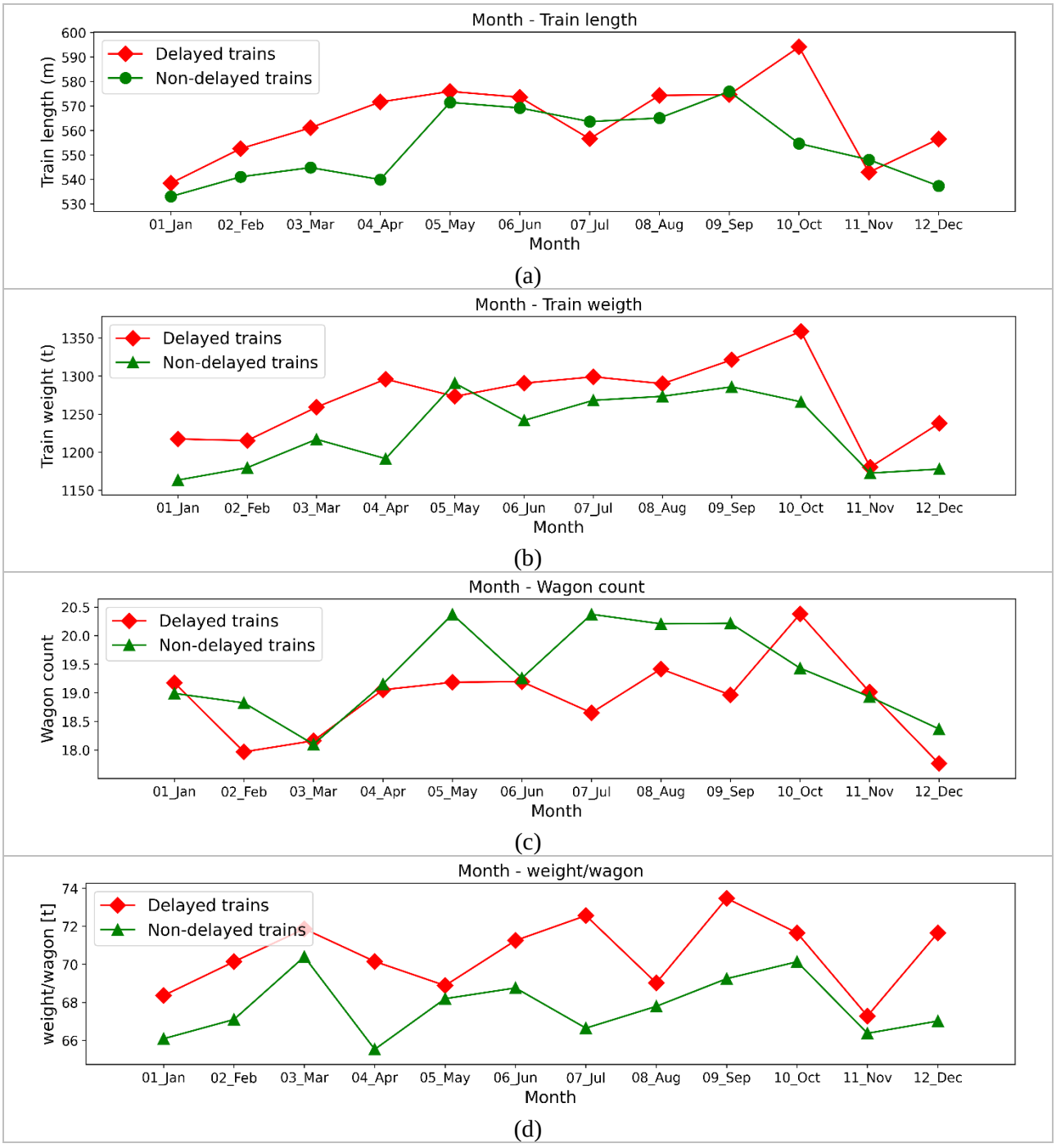

Figure 4 Interaction effects between the major predictors of the rail intermodal operations: (a) month and train length; (b) month and train weight; (c) month and wagon number; (d) month and weight per wagon

Figure 5 presents an in-depth analysis of the relationship between the weight per wagon and the month of operation in the delays in intermodal operations on the studied rail network. In general, most of the intermodal operations carried out by trains with less weight per wagon (<60 t) tend to be non-delayed trips, but as the wagons are heavier, the percentage of delayed trips increases. Besides, the results suggest that there is a greater difference between the number of delayed and non-delayed trips according to the month of operation, where the months of April to August tend to be more efficient months in terms of the number of non-delayed trips, while months like January, February, October and November tend to be more inefficient. At this point it is important to highlight

that the data used in the study comprise 17 months, where the months between April and October have less data than the others. Hence, to draw more trustworthy conclusions on this feature, additional data is needed.

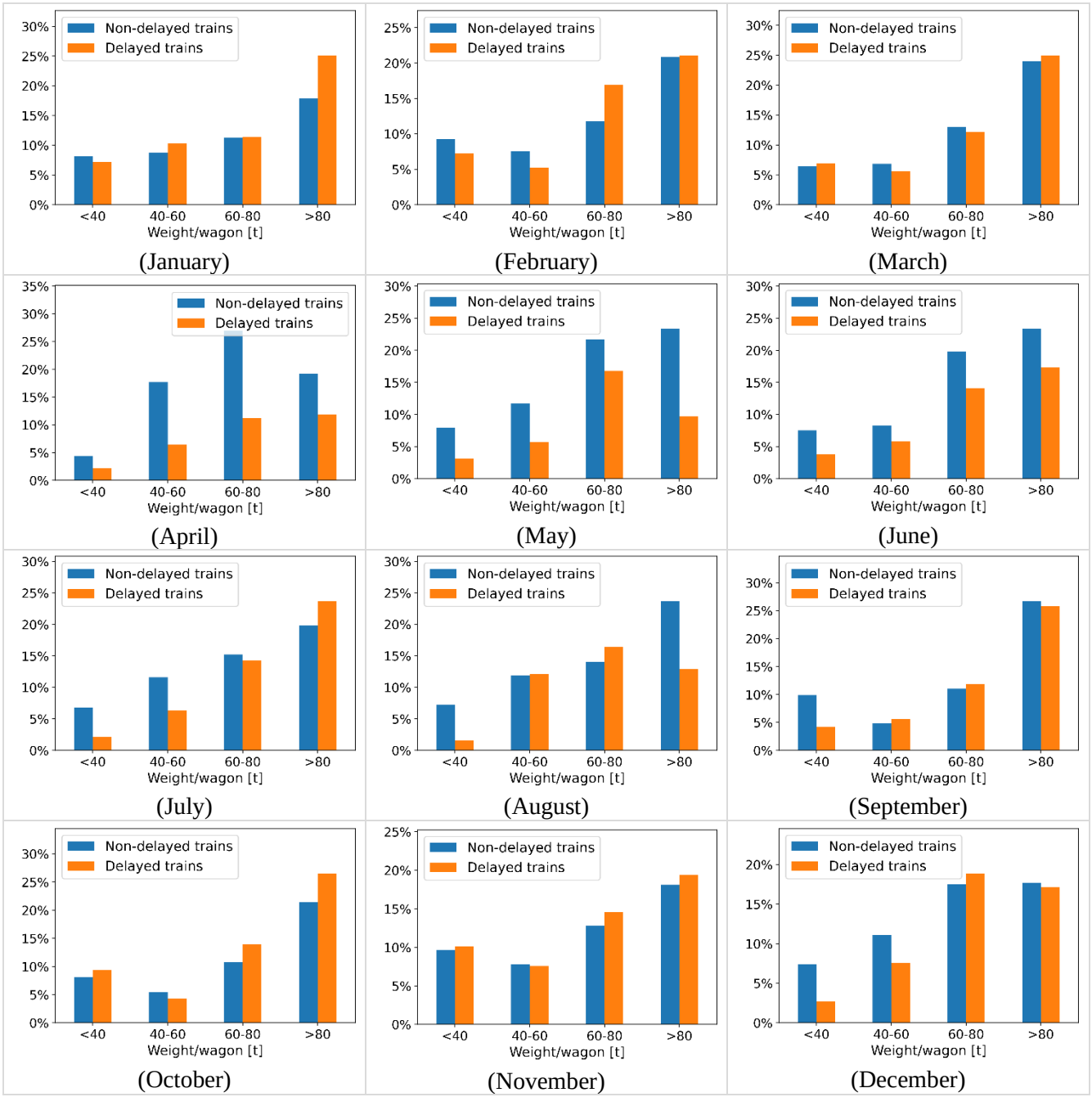

Figure 5 Delayed and non-delayed trains considering their weight per wagon for each month

This study analyzes the relationship between the train and the operational features with the disruptions and their subsequent delays in intermodal rail operations, and the main contributions of this study can be summarized as follows:

- The analysis of the effect of train- operational- and station-related features on the disruptions in the intermodal operations of the studied rail network, from which a reliable predictive ML model was developed, demonstrating for the first time in rail intermodal operations research that the CatBoost model outperforms other ML models.

- The use of the SHAP method to understand the causality of the ML model developed, examining the contribution of each feature in the output of the model in order to identify the major predictors in the disruptions and their subsequent delays of the intermodal operations of the studied rail network. In addition, an in-depth analysis of the relationship of the major predictors was carried out to comprehend the interaction effects between them.
- The predictive model developed in this study can be used as a tool by the National Railway Company of Luxembourg to evaluate future operational interventions with the aim of reducing disruptions and their subsequent delays in its intermodal operations. For instance, a better distribution of the weight of the freight carried by the trains will likely to reduce the probability of a delayed trip.

CONCLUSIONS

This paper has presented an approach to examine the effect of train- operational- and station-related features on disruptions and their subsequent delays in rail intermodal operations. Previous studies including models to predict rail disruptions have been focused on passenger trains, while studies focused on freight trains are scarce and have focused on analyzing the impact at the network design level, rather than train-, operational- and station features. Most critically, these studies have failed to analyze causal relationships between these features within a complex network.

After training a set of ML models for predicting the disruptions and their subsequent delays in the intermodal operations of the studied rail network, a gradient boosting machine model using the CatBoost implementation has been shown to be the most suitable approach to develop a predictive ML model, since it outperforms the other ML models in terms of the predefined evaluation metrics (AUC, recall and F1-score).

The train- and operational- related features that most impact the disruptions in the intermodal operations of the studied rail network are the train weight, train length, number of wagons, weight per wagon and the month of operation, where, in general, the greater the train's weight, length and weight per wagon, and the lower the number of wagons, the greater the probability that the trip will be delayed, suggesting that a better distribution of the weight of the freight carried by the trains will likely to reduce the probability of a delayed trip.

The limitations of the present study include the use of data from only 17 months, which be inadequate to analyze the behavior reflected in the greater difference between the number of delayed and non-delayed trips according to the month of operation. Similarly, some underlying effects due to the global crisis that COVID-19 has unleashed since the first months of 2020 (*61*) may be hidden within the data, which could be identified with novel records for the analysis dataset. In any case, considering the learning curve presented in **Figure 2**, adding new data to the model should improve its performance.

ACKNOWLEDGEMENTS

This study was possible thanks to the collaboration agreement signed between the University of Luxembourg and CFL Multimodal, and funding obtained by the Luxembourg National Research Fund FNR, through the project "ANticipatory Train Optimization with Intelligent maNagEment (ANTOINE)", under grant BRIDGES2020/MS/14767177/ANTOINE. Special thanks to Nathalie Stef and Michael Maraldi from CFL for sharing the data used in this study.

AUTHOR CONTRIBUTIONS

Juan Pineda-Jaramillo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing- Original draft preparation, Visualization. **William McDonald:** Data cleaning, Document editing. **Wei Zheng:** Conceptualization, Investigation. **Francesco Viti:** Conceptualization, Validation, Supervision, Writing- Reviewing and Editing, Funding acquisition.

REFERENCES

1. Cacchiani, V., A. Caprara, and P. Toth. Scheduling Extra Freight Trains on Railway Networks. *Transportation Research Part B: Methodological*, Vol. 44, No. 2, 2010, pp. 215–231. https://doi.org/10.1016/j.trb.2009.07.007.
2. Berger, A., A. Gebhardt, M. Müller-Hannemann, and M. Ostrowski. Stochastic Delay Prediction in Large Train Networks. *OpenAccess Series in Informatics*, Vol. 20, No. January, 2011, pp. 100–111. https://doi.org/10.4230/OASIcs.ATMOS.2011.100.
3. Goverde, R. M. P. A Delay Propagation Algorithm for Large-Scale Railway Traffic Networks. *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 3, 2010, pp. 269–287. https://doi.org/10.1016/j.trc.2010.01.002.
4. Goverde, R. M. P., and I. A. Hansen. Performance Indicators for Railway Timetables. 2013.
5. Huang, P., C. Wen, L. Fu, J. Lessan, C. Jiang, Q. Peng, and X. Xu. Modeling Train Operation as Sequences: A Study of Delay Prediction with Operation and Weather Data. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 141, No. May, 2020, p. 102022. https://doi.org/10.1016/j.tre.2020.102022.
6. Cerreto, F., O. A. Nielsen, S. Harrod, and B. F. Nielsen. Causal Analysis of Railway Running Delays. 2016.
7. Bešinović, N., R. M. P. Goverde, E. Quaglietta, and R. Roberti. An Integrated Micro–Macro Approach to Robust Railway Timetabling. *Transportation Research Part B: Methodological*, Vol. 87, 2016, pp. 14–32. https://doi.org/10.1016/j.trb.2016.02.004.
8. Goverde, R. M. P., N. Bešinović, A. Binder, V. Cacchiani, E. Quaglietta, R. Roberti, and P. Toth. A Three-Level Framework for Performance-Based Railway Timetabling. *Transportation Research Part C: Emerging Technologies*, Vol. 67, 2016, pp. 62–83. https://doi.org/10.1016/j.trc.2016.02.004.
9. Cacchiani, V., D. Huisman, M. Kidd, L. Kroon, P. Toth, L. Veelenturf, and J. Wagenaar. An Overview of Recovery Models and Algorithms for Real-Time Railway Rescheduling. *Transportation Research Part B: Methodological*, Vol. 63, 2014, pp. 15–37. https://doi.org/10.1016/j.trb.2014.01.009.
10. Wen, C., P. Huang, Z. Li, J. Lessan, L. Fu, C. Jiang, and X. Xu. Train Dispatching Management With Data-Driven Approaches: A Comprehensive Review and Appraisal. *IEEE Access*, Vol. 7, 2019, pp. 114547–114571. https://doi.org/10.1109/ACCESS.2019.2935106.
11. Corman, F., and P. Kecman. Stochastic Prediction of Train Delays in Real-Time Using Bayesian Networks. *Transportation Research Part C: Emerging Technologies*, Vol. 95, 2018, pp. 599–615. https://doi.org/10.1016/j.trc.2018.08.003.
12. Lessan, J., L. Fu, C. Wen, P. Huang, and C. Jiang. Stochastic Model of Train Running Time and Arrival Delay: A Case Study of Wuhan–Guangzhou High-Speed Rail. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2672, No. 10, 2018, pp. 215–223. https://doi.org/10.1177/0361198118780830.
13. Yuan, J., R. Goverde, and I. Hansen. Propagation of Train Delays in Stations. In *Computers in Railways VIII* (J. Allan, R. J. Hill, C. A. Brebbia, G. Sciutto, and S. Sone, eds.), WIT Press, Southhampton, UK, pp. 975–984.
14. Marković, N., S. Milinković, K. S. Tikhonov, and P. Schonfeld. Analyzing Passenger Train Arrival Delays with Support Vector Regression. *Transportation Research Part C: Emerging Technologies*, Vol. 56, 2015, pp. 251–262. https://doi.org/10.1016/j.trc.2015.04.004.
15. Yaghini, M., M. M. Khoshraftar, and M. Seyedabadi. Railway Passenger Train Delay Prediction via Neural Network Model. *Journal of Advanced Transportation*, Vol. 47, No. 3, 2013, pp. 355–368. https://doi.org/10.1002/atr.193.
16. Nair, R., T. L. Hoang, M. Laumanns, B. Chen, R. Cogill, J. Szabó, and T. Walter. An Ensemble Prediction Model for Train Delays. *Transportation Research Part C: Emerging Technologies*, Vol. 104, No. May, 2019, pp. 196–209. https://doi.org/10.1016/j.trc.2019.04.026.
17. Milinković, S., M. Marković, S. Vesković, M. Ivić, and N. Pavlović. A Fuzzy Petri Net Model to Estimate Train Delays. *Simulation Modelling Practice and Theory*, Vol. 33, 2013, pp. 144–157. https://doi.org/10.1016/j.simpat.2012.12.005.
18. Ghofrani, F., Q. He, R. M. P. Goverde, and X. Liu. Recent Applications of Big Data Analytics in Railway

Transportation Systems: A Survey. *Transportation Research Part C: Emerging Technologies*, Vol. 90, 2018, pp. 226–246. https://doi.org/10.1016/j.trc.2018.03.010.

19. Bhavsar, P., I. Safro, N. Bouaynaya, R. Polikar, and D. Dera. Machine Learning in Transportation Data Analytics. In *Data Analytics for Intelligent Transportation Systems* (M. Chowdhury, A. Apon, and K. Dey, eds.), Elsevier, Amsterdam, pp. 283–307.
20. Pineda-Jaramillo, J. A Review of Machine Learning (ML) Algorithms Used for Modeling Travel Mode Choice. *Dyna*, Vol. 86, No. 211, 2019, pp. 32–41. https://doi.org/10.15446/dyna.v86n211.79743.
21. Pineda-Jaramillo, J., R. Insa-Franco, and P. Martínez-Fernández. Modeling the Energy Consumption of Trains by Applying Neural Networks. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, Vol. 232, No. 3, 2018, pp. 816–823. https://doi.org/10.1177/0954409717694522.
22. Allah Bukhsh, Z., A. Saeed, I. Stipanovic, and A. G. Doree. Predictive Maintenance Using Tree-Based Classification Techniques: A Case of Railway Switches. *Transportation Research Part C: Emerging Technologies*, Vol. 101, 2019, pp. 35–54. https://doi.org/10.1016/j.trc.2019.02.001.
23. De Martinis, V., and F. Corman. Data-Driven Perspectives for Energy Efficient Operations in Railway Systems: Current Practices and Future Opportunities. *Transportation Research Part C: Emerging Technologies*, Vol. 95, No. August, 2018, pp. 679–697. https://doi.org/10.1016/j.trc.2018.08.008.
24. Pineda-Jaramillo, J., P. Martínez-Fernández, I. Villalba-Sanchis, P. Salvador-Zuriaga, and R. Insa-Franco. Predicting the Traction Power of Metropolitan Railway Lines Using Different Machine Learning Models. *International Journal of Rail Transportation*, 2020. https://doi.org/10.1080/23248378.2020.1829513.
25. Schölkopf, B. Causality for Machine Learning. *arXiv*, 2019, pp. 1–20.
26. Spirtes, P. Introduction to Causal Inference. *Journal of Machine Learning Research*, 2010.
27. Truong, D. Using Causal Machine Learning for Predicting the Risk of Flight Delays in Air Transportation. *Journal of Air Transport Management*, Vol. 91, No. May 2020, 2021, p. 101993. https://doi.org/10.1016/j.jairtraman.2020.101993.
28. Dollevoet, T., D. Huisman, L. Kroon, M. Schmidt, and A. Schöbel. Delay Management Including Capacities of Stations. *Transportation Science*, Vol. 49, No. 2, 2015, pp. 185–203. https://doi.org/10.1287/trsc.2013.0506.
29. Wen, C., W. Mou, P. Huang, and Z. Li. A Predictive Model of Train Delays on a Railway Line. *Journal of Forecasting*, No. February 2019, 2019, pp. 470–488. https://doi.org/10.1002/for.2639.
30. Pineda-Jaramillo, J. A Shallow Neural Network Approach for Identifying the Leading Causes Associated to Pedestrian Deaths in Medellín. *Journal of Transport & Health*, Vol. 19, 2020, p. 100912. https://doi.org/10.1016/j.jth.2020.100912.
31. Bollegala, D. Dynamic Feature Scaling for Online Learning of Binary Classifiers. *Knowledge-Based Systems*, Vol. 129, No. 1, 2017, pp. 97–105. https://doi.org/10.1016/j.knosys.2017.05.010.
32. Shyamala Devi, M., R. M. Mathew, and R. Suguna. Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes Using Machine Learning. *International Journal of Engineering and Advanced Technology*, Vol. 8, No. 6, 2019, pp. 952–956. https://doi.org/10.35940/ijeat.F8255.088619.
33. Menard, S. Six Approaches to Calculating Standardized Logistic Regression Coefficients. *The American Statistician*, Vol. 58, No. 3, 2004, pp. 218–223. https://doi.org/10.1198/000313004X946.
34. Ba, Y., W. Zhang, Q. Wang, R. Zhou, and C. Ren. Crash Prediction with Behavioral and Physiological Features for Advanced Vehicle Collision Avoidance System. *Transportation Research Part C: Emerging Technologies*, Vol. 74, 2017, pp. 22–33. https://doi.org/10.1016/j.trc.2016.11.009.
35. Ching, W. K., D. Chu, L. Z. Liao, and X. Wang. Regularized Orthogonal Linear Discriminant Analysis. *Pattern Recognition*, Vol. 45, No. 7, 2012, pp. 2719–2732. https://doi.org/10.1016/j.patcog.2012.01.007.
36. Kim, K. S., H. H. Choi, C. S. Moon, and C. W. Mun. Comparison of K-Nearest Neighbor, Quadratic Discriminant and Linear Discriminant Analysis in Classification of Electromyogram Signals Based on the Wrist-Motion Directions. *Current Applied Physics*, Vol. 11, No. 3, 2011, pp. 740–745. https://doi.org/10.1016/j.cap.2010.11.051.
37. Zhao, X., X. Yan, A. Yu, and P. Van Hentenryck. Prediction and Behavioral Analysis of Travel Mode Choice: A Comparison of Machine Learning and Logit Models. *Travel Behaviour and Society*, Vol. 20, 2020, pp. 22–35. https://doi.org/10.1016/j.tbs.2020.02.003.

38. Hagenauer, J., and M. Helbich. A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications*, Vol. 78, 2017, pp. 273–282. https://doi.org/10.1016/j.eswa.2017.01.057.
39. Shafiq, M., Z. Tian, A. K. Bashir, A. Jolfaei, and X. Yu. Data Mining and Machine Learning Methods for Sustainable Smart Cities Traffic Classification: A Survey. *Sustainable Cities and Society*, Vol. 60, 2020, p. 102177. https://doi.org/10.1016/j.scs.2020.102177.
40. Xiao, G., Q. Cheng, and C. Zhang. Detecting Travel Modes Using Rule-Based Classification System and Gaussian Process Classifier. *IEEE Access*, Vol. 7, 2019, pp. 116741–116752. https://doi.org/10.1109/ACCESS.2019.2936443.
41. Servos, N., X. Liu, M. Teucke, and M. Freitag. Travel Time Prediction in a Multimodal Freight Transport Relation Using Machine Learning Algorithms. *Logistics*, Vol. 4, No. 1, 2019, p. 1. https://doi.org/10.3390/logistics4010001.
42. Seyyedattar, M., M. M. Ghiasi, S. Zendehboudi, and S. Butt. Determination of Bubble Point Pressure and Oil Formation Volume Factor: Extra Trees Compared with LSSVM-CSA Hybrid and ANFIS Models. *Fuel*, Vol. 269, 2020, p. 116834. https://doi.org/10.1016/j.fuel.2019.116834.
43. Pineda-Jaramillo, J. Travel Time, Trip Frequency and Motorised-Vehicle Ownership: A Case Study of Travel Behaviour of People with Reduced Mobility in Medellín. *Journal of Transport & Health*, Vol. 22, No. April, 2021, p. 101110. https://doi.org/10.1016/j.jth.2021.101110.
44. Yandex. CatBoost. https://catboost.ai/. Accessed Apr. 18, 2021.
45. Young, S. D., W. Wang, and B. Chakravarthy. Crowdsourced Traffic Data as an Emerging Tool to Monitor Car Crashes. *JAMA Surgery*, Vol. 154, No. 8, 2019, pp. 777–778. https://doi.org/10.1001/jamasurg.2019.1167.
46. Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2016.
47. Bergstra, J., and Y. Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, Vol. 13, No. 10, 2012, pp. 281–305.
48. Meek, C., B. Thiesson, and D. Heckerman. The Learning-Curve Sampling Method Applied to Model-Based Clustering. *Journal of Machine Learning Research*, 2002. https://doi.org/10.1162/153244302760200678.
49. Ribeiro, M. T., S. Singh, and C. Guestrin. "Why Should I Trust You?" 2016.
50. Štrumbelj, E., and I. Kononenko. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*, Vol. 41, No. 3, 2014, pp. 647–665. https://doi.org/10.1007/s10115-013-0679-x.
51. Zhou, F., X. J. Yang, and J. C. F. F. de Winter. Using Eye-Tracking Data to Predict Situation Awareness in Real Time During Takeover Transitions in Conditionally Automated Driving. *IEEE Transactions on Intelligent Transportation Systems*, 2021, pp. 1–12. https://doi.org/10.1109/TITS.2021.3069776.
52. Xu, J., A. Wang, N. Schmidt, M. Adams, and M. Hatzopoulou. A Gradient Boost Approach for Predicting Near-Road Ultrafine Particle Concentrations Using Detailed Traffic Characterization. *Environmental Pollution*, Vol. 265, 2020, p. 114777. https://doi.org/10.1016/j.envpol.2020.114777.
53. Lundberg, S. M., and S. I. Lee. A Unified Approach to Interpreting Model Predictions. 2017.
54. Wang, J., B. Liu, T. Fu, S. Liu, and J. Stipancic. Modeling When and Where a Secondary Accident Occurs. *Accident Analysis and Prevention*, Vol. 130, 2019, pp. 160–166. https://doi.org/10.1016/j.aap.2018.01.024.
55. Nti, I. K., A. F. Adekoya, and B. A. Weyori. A Comprehensive Evaluation of Ensemble Learning for Stock-Market Prediction. *Journal of Big Data*, 2020. https://doi.org/10.1186/s40537-020-00299-5.
56. Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. CatBoost: Unbiased Boosting with Categorical Features. 2017.
57. Malavasi, G., and S. Ricci. Simulation of Stochastic Elements in Railway Systems Using Self-Learning Processes. *European Journal of Operational Research*, Vol. 131, No. 2, 2001, pp. 262–272. https://doi.org/10.1016/S0377-2217(00)00126-0.
58. Peters, J., B. Emig, M. Jung, and S. Schmidt. Prediction of Delays in Public Transportation Using Neural Networks. No. 2, 2005, pp. 92–97.
59. Pongnumkul, S., T. Pechprasarn, N. Kunaseth, and K. Chaipah. Improving Arrival Time Prediction of Thailand's Passenger Trains Using Historical Travel Times. 2014.

60. Oneto, L., E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, and D. Anguita. Train Delay Prediction Systems: A Big Data Analytics Perspective. *Big Data Research*, Vol. 11, 2018, pp. 54–64. https://doi.org/10.1016/j.bdr.2017.05.002.
61. Gössling, S., D. Scott, and C. M. Hall. Pandemics, Tourism and Global Change: A Rapid Assessment of COVID-19. *Journal of Sustainable Tourism*, Vol. 29, No. 1, 2021, pp. 1–20. https://doi.org/10.1080/09669582.2020.1758708.