TUS

Technological University of the Shannon:
Midlands Midwest
Ollscoil Teicneolaíochta na Sionainne:
Lár Tíre Iarthar Láir

Predicting Shipment Delays in Supply Chain Using Machine Learning

Submitted by: Balaji Kornepati

Student Id: K00302097

Thesis Supervisor: Michelle Ahern

This Research Project is submitted in Partial Fulfilment of the Degree of Master of Science in Data Analytics.

Declaration

I have read the University's code of practice on plagiarism. I hereby certify that this material, which I now submit for assessment on the program of study leading to the award of a Master of Science in Data Analytics is entirely my own work and has not been taken from the work of others, expect to the extent that such work has been cited within the text of my work.

Future students may use the material contained in this thesis provided that the source is acknowledged in full.

Student ID Number: K00302097

Name of the Candidate: Balaji Kornepati

Signature of the Candidate: Balaji Kornepati

Date: 31/08/2025

Acknowledgement

I would like to thank my supervisor for her assistance in guiding me through this research process. I would also like to thank my family for their support during this time.

Table of Contents

List of Tables

List of Figures

Abstract

The last-mile delivery supply chain is a final part of supply chain where supply is carried out to the delivery point, to the customer door. This step can be the most expensive of logistics as multiple individual deliveries are usually served to multiple customers in an environment of individual addresses, traffic jams, and schedule sensitivities. It is also the most problematic stage since delays are obvious on this occasion on grounds of these challenges. This study aims at evaluating machine learning and deep learning algorithm to forecast delivery delays using an integration of structured operational and sequential time series data. The analysis built and tested several different models using a data set of 30,000 Amazon deliveries supplemented with weather data via the Visual Crossing API, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) and ensembled Voting Classifier. Instead of using traditional methods, a new hybrid RF-LSTM model was proposed, whose feature extraction power of Random Forest and temporal modelling power of LSTM networks combine.

The evaluation of the models was carried out on the basis of the accuracy, precision, recall and F1-score. The results showed that the hybrid RF+LSTM model was statistically higher than all the other baseline models with an accuracy of 87.24%, precision of 0.89, recall of 0.92, and an F1-score of 0.90. To increase interpretability of explanations, SHAP (SHapley Additive Explanations) analysis was adopted, with traffic density, the distance of delivery, the features of the agent, and weather being the most significant ones responsible for the presence of delays.

The results show that the hybrid structures can be quite efficient when dealing with the multidimensional complexity of the last-mile logistics. In addition to prediction accuracy, the research underlines the significance of model transparency as a source of stakeholder trust and operational practicality. The present study provides a well-grounded, interpretable model of delay prediction that can be used by operators of logistics technologies, technology developers and policymakers who are interested in improving efficiency and reliability of last-mile delivery systems utilized by them, or others.

List of Abbreviations

Abbreviations	Full Title
RF	Random Forest
LSTM	Long Short-Term Memory
NN	Neural Networks
ML	Machine Learning
ANN	Artificial Neural Networks
GB	Gradient Boosting
SVM	Support Vector Machine
AI	Artificial Intelligence
SHAP	Shapley Addictive Explanations
XAI	Explainable Artificial Intelligence
DSS	Decision Support Systems
API	Application Programming Interface
CNN	Convolutional Neural Network

Chapter 1 Introduction

1.1 Problem Statement

The concept of last-mile logistics, or the transportation of goods between a distribution centre and a final consumer, has become an area of focus when it comes to operational optimisation, as well as research. Along with the growth of e-commerce, which is enjoying globally averaged high-digit annual growth, the expectations of customers regarding faster and more predictable deliveries have increased. As such, the service-level agreements have tightened, distribution networks have become fragmented, and it has become harder to maintain operational efficiency.

Industry reports indicate that the last-mile segment consumes over 50 percent of the overall spend on logistics thus making it the most expensive as well as the resource-intensive segment of the supply chain. The reason why its outlay becomes high is the complicated character of the route planning in crowded urban environments joined with the ambiguity of traffic patterns, the inconsistency of weather effects, and the unevenness of client disposition. The traffic congestion, the limited delivery windows in the urban environment significantly reduce the flexibility of the operations whereas the time boundedness of the delivery routes and the sparse recipient density in the rural environment continue to cause inefficiencies.

Delivery delays form one of the most important operational issues in the modern supply chain. Even small delays in the planned delivery periods can cause significant downstream effects: failure to meet appointments, unwise use of resources related to the relocation of activities and the worsening of reputation. In the perspective of the end consumer, the inability to fulfil delivery expectations is quite often understood as the violation of the provision of trust

specifically in cases when the communication practices are inefficient or not accurate. In very competitive delivery settings, persistent patterns of lateness may also trigger customer turnover and, in the long term, will culminate in quantifiable market share decreases.

The modern machine learning (ML) and artificial intelligence (AI) developments allow using multi-source datasets and make it possible to discover demanding relationships between operational variables, temporal relationships, and environmental variables. This kind of system can provide predictions that dynamically adapt to the dynamic contexts. Despite this, weak interpretability is a serious barrier to the adoption of high performing models into an operational environment. Managers would be unwilling to base inferences on the generated insights when clear reasoning behind the predictions is discarded.

In the current research, the two problems of complex interdependence and temporal dynamics in enterprise data were addressed by designing a hybrid prediction framework that combines Random Forest (RF)- an algorithm designed to work with clear-cut tabular data-with Long Short-Term Memory (LSTM) networks, which are well-known to capture sequential and time-based dependencies. The framework also uses SHAP (SHapley Additive Explanations) to provide interpretable, explainable explanations of the importance of features influencing each prediction. These variables are linked with the operational data, including external data, in particular weather data (through the Visual Crossing API). This allows the model to provide high predictive accuracy without losing interpretability required to inform the subsequent action.

1.2 Motivation

The current research considers practical and theoretical aspects of inefficiency in delivery. In real life, late deliveries will create inefficiencies and hike operation costs and customer satisfaction. Advanced forecasts of these delays allow such mitigation through the re-routing

of vehicles, re-scheduling delivery processes, or delivering a warning on the extent of the delay in advance effectively avoiding the disruption of the service and increasing the viability of the enterprise as a whole, strengthening its position in a markets competitive significance (Al-Saghir, 2022).

Academically, most existing studies have used machine-learning techniques to solve logistic problems, but most adopted a hybrid approach of combining structured operational data with sequential temporal characteristics, with few studies. In addition to that, not many studies have been done concerning explainable AI in this area. Most precise models operate as black boxes, and their internals are opaque, therefore, risking the trust and the ability to utilize in operational conditions. The current thesis fills in this gap by introducing a hybrid RF-LSTM solution that is both accurate and interpretable, thereby matching operational reliability with the predictive performance.

1.3 Research Aim

The aim of the thesis is to train, test, and analyse a hybrid Random Forest Long Short-Term Memory model to effectively predict the occurrence of delivery delays in last-mile logistics processes, and estimate the potential of the model in operational use.

1.4 Research Questions

1. Does Machine Learning models effectively predict delivery delays in logistics operations?
2. How does the predictive performance of the traditional models compare to deep learning and hybrid models?
3. Which operational and environmental features have the best influence on delivery delays predictions?

1.5 Scope of the study

This research is focused on the field of last-mile delivery operation based on the historical delivery data that were enriched by external weather information sources provided by the Visual Crossing API. The data will contain both structured operation attributes [e.g. age of driver, rating, distance to delivery, and traffic conditions) and time series position dependent sequences [e.g. order time and scheduling of pickup). The study is based in one and the same operational situation to keep the feature of data homogeneity, but the limitation of the study is the generalisability of the findings to other areas, industries or delivery models.

1.6 Significance of Study

The research contributes to the rationale behind utilizing hybrid machine learning frameworks on complicated logistic predicting problems. Specifically, it shows that an ensemble of random forest (RF) and the long short-term memory (LSTM) is capable of leveraging tree and sequence-specific patterns in features at the same time, and hence of capturing a more comprehensive representation of underlying problem structure. As part of this hybrid framework, a Shapley additive explanations (SHAP) module is used to meet the interpretability needs without having an adverse effect on the predictivity accuracy.

In terms of operationalization, the resulting RF-LSTM pipeline provides a deployable predictive model that would allow logistics managers to have a superior system of predicting delivery delays to facilitate prevention measures. The interpretability module facilitates understanding to the decision-makers as regards to the reasoning behind the disruptions predicted thus fostering trust and intervention measures towards resolving disruptions. Furthermore, the flexible nature of the model makes the integration with the existing logistics

management frameworks easy with minimal interference and hence increases scalability and can easily be adopted.

1.6 Conclusion

This chapter has set out the context and motivation to research into predictive modelling of delivery delays associated with last-mile logistics. It has presented the purpose, the objectives, questions as well as significance of the research and has placed the research in the context of the existing industry challenges. The proposed research will achieve the intended outcome of a practical predictive framework based on combining Random Forest and LSTM models along with SHAP interpretability, which will be accurate, clear, and at the stage of implementation. The rest of this thesis is organized as follows. Chapter 2 provides a literature review of the topics or questions that have been studied and examined before on last-mile delivery, supply chain delays, and the application of selection of machine learning and hybrid models in the logistics forecasting process, establishing the academic background to the research. In Chapter 3, the methodology used to conduct the research is described and includes data acquisition, preprocessing, feature engineering, model selection and the design of the hybrid RF-LSTM framework. Chapter 4 shows the empirical results, such as comparisons among models and exploratory data analysis, feature interpretability using SHAP. Chapter 5 gives a discussion of the results as regards to the stated goals of the research and literature as well as giving practical implications, suggesting methodological constraints and recommendations on future research.

Chapter 2 Literature Review

2.1 Introduction

Efficient supply chain management is essential for ensuring timely delivery and maintaining customer satisfaction, as disruptions in logistics can lead to substantial financial losses, inventory discrepancies, and reputational harm. Interruptions in the supply chain may trigger monetary losses, imbalances in the stocks, and customer dissatisfactions at all supply chain levels. (Al-saghir, 2022). Harmony is a descriptive tool it provides statistics from the past data and it allows users to get an average of elapsed time between transportation booking and destination delivery by using historical data. However, (Jonquais & Krempl, 2019) and (Viellechner, 2020) argue that such static models are not adequate to handle the rapid changes like the occurrence of severe weather or the presence of congestion at the ports.

Recent research increasingly explores the use of machine-learning (ML) methods to improve logistics forecasting. (Al-saghir, 2022) notes that shipment delays classification using the Random Forest (RF) models still work whether data are complete or not. (Manai, 2018) shows the flexibility of RF when considering risks associated with suppliers and (Wang, 2021) states that Gradient Boosting (GB) models such as XGBoost are effective with structured logistics data and output features that may be helpful in making business decisions.

In spite of these advancements, certain internal constraints remain. According to (Keung, 2021) Support Vector Machines (SVMs) work effectively in binary classification, but they fail to identify the ordering patterns suitable to real-time prediction. (Verma, 2024)further argues that the “black-box” nature of many ML models limits practical implementation because managers require transparency and trust in predictive system.

Some researchers have started adding explainable-artificial-intelligence (XAI) methods like SHAP (SHapley Additive Explanations) to ML pipelines in response to these limitations. As (Jahin, 2025) and (Jauhar, 2024) indicate, integrating SHAP with ML models makes it significantly more accurate in making predictions and make the models easier to interpret. Nonetheless, these advancements have focused mainly on areas that deal with demand prediction and perishable-products management as opposed to shipment-delay prediction.

A clear research gap therefore exists in developing hybrid models that combine real-time data processing, sequential learning, and interpretability specifically for predicting shipment delays. This thesis seeks to fill that gap by proposing a hybrid RF-LSTM architecture integrated with SHAP analysis to deliver both predictive accuracy and actionable insights for proactive logistics decision-making.

2.2 Literature Review and Methods

2.2.1 Inclusion and Exclusion Criteria

Criteria Type	Included	Excluded
Publication Type	Peer-reviewed journals, conference papers, industry whitepapers	Blogs, newspapers, non-academic articles
Time Frame	Studies published from **2011 to 2025**	Studies older than 10 years
Relevance	Focus on **machine learning**, **AI**, **predictive models**, or **risk analysis** in supply chains, Explainable AI	Studies unrelated to delays, supply chain, or machine learning

Criteria Type	Included	Excluded
Data Availability	Articles with clear datasets, experiments, or case studies	Articles lacking methodology or empirical analysis

Table 1: Overview of Literature Review Papers

2.2.2 Thematic and analytical Review of literature

2.2.3 Model Structures and Predictive Capabilities

Several empirical studies have shown that machine learning (ML) models outperform traditional statistical methods when it comes to predicting shipping delays at least in terms of modelling non-linear structural features that are characteristic of shipping time series data. To give some examples, (Al-saghir, 2022) reported that Random Forest (RF) ensemble classifier has solid predictive performance under the observational completeness constraint, a property that can be directly related to the aggregated structure of the former since it helps reduce overfitting and, thus, improve generalization performance. Similar results are outlined by (Jonquais & Krempl, 2019) that Random Forest (RF) might have a significant predictive discrimination in a shipment delay context and should be adopted when it comes to risk-related applications. (Manai, 2018), as an example of complementary literature, has revealed that RF is capable of handling supplier risk assessment activities due to computational soundness and the capacity to combine logistical characteristics with heterogeneity. The works focusing on Support Vector Machines (SVM) provide divergent and complementary evidence. In significant, high-dimensional scenarios, (Al-saghir, 2022) found high scores of discrimination accuracy in a circumstance where SVM was utilized to differentiate between held up and on-time deliveries. These observations are supported based on the work of (Jonquais & Krempl, 2019) which states that SVM optimally identifies a hyperplane, which can be utilized in effective classification of data even as it gets more complex in nature.

Empirical studies of temporal learning and adaptability in the context of logistics often show divergence findings. Random forest (RF) and support vector machines (SVM) are examples of standard formulations that are inherently about Markovian dependencies; that cannot encode sequential relation and dynamically changing patterns. (Keung, 2021)draws a conclusion that SVM may be introduced into use in purely static classification environments, but it does not work well when it comes to modelling time-based relationships and thus cannot be used in real-time forecasting, especially in highly volatile and disruptive conditions. In the attempt to overcome these weaknesses, scholars have been focusing more on the use of deep learning models, more specifically, Long Short-Term Memory (LSTM). (Keung, 2021) illustrates that LSTM models are effective in capturing sequential change in the pattern of shipments, thus making the model more accurate for using in high volatile environments. (Jahin, 2025) confirm this opinion and establish that LSTM networks have enough flexibility in modelling complex logistics time-series and are therefore suitable in dynamic prediction problems. Yet, both studies highlight the great strain that LSTM models require in terms of computing resources, as well as the low level of inherent interpretability that hampers their operationalization.

This thesis aims to address this gap by developing a hybrid RF-LSTM model enhanced with SHAP analysis to provide both accurate forecasts and transparent insights for proactive logistics decision-making.

2.2.4 Explainability and decision-making support

While predictive accuracy is essential for reliable decision-making, explainability is equally critical for real-world applications. Many of the effective models e.g. deep-learning architectures of the form LSTM, and the ensemble methods, including RF, are not interpretable enough. The presence of such constraints hampers their integration into those systems where transparency and trust cannot be negotiable.

On demand and inventory forecasting, (Verma, 2024) and (Jahin, 2025) attempted to answer this problem by letting SHAP values and permutation-based explanations filter through their ML pipelines. A similar course was taken by (Jauhar, 2024), who integrated explainable AI (XAI) techniques into LSTM-based models to directly forecast perishable goods and, therefore, simplify the ambiguity related to its output.

While these examples yield promising outcomes, they mainly focus on demand and inventory prediction. The use of shipment delay prediction usage described by XAI is relatively limited. With the aim of narrowing this gap, This thesis attempts to combine SHAP with a hybrid RF-LSTM system. The proposed framework delivers accurate delay predictions and presents managerial friendly descriptions of the input that make some shipments susceptible to delay.

2.3 Factors Affecting Shipment delays

Delays in shipments occur as a result of a combination of both internal operation deficiencies and external shocks. The subsequent discussion questions the major factors of such delays which are constituted by port congestion, supplier constraints, negative weather phenomena, and systemic shocks included by the COVID-19 pandemic. The current body of literature is evaluated taking into consideration both descriptive correctness and methodological consistency and practical applicability.

2.3.1 Port Congestion

Port congestion is a major cause of shipment delays, particularly along global high-volume corridors. (Jonquais & Krempl, 2019) reveals that increasing volumes of trade and infrastructural constraints based on structure create the significant bottlenecks at critical ports, which in turn, causes shipping time disruptions. In line with this finding, (Fancello, 2011) advises on the use of the Decision Support Systems (DSS) which utilise the neural networks to optimise labour and equipment used in the container terminals. However, these frameworks are, to a large extent, formulated to the context of centralised facilities and do not put the idea of decentralised facilities, or multimodal hubs as identified by (Viellechner, 2020), into consideration. In addition, according to (Yucekaya, 2024), extreme weather events further contribute to the congestion process since it leads to the disruption of trans-shipments in ports, thus having the need to incorporate predictive models that would allow sharing environmental information with the congestion prediction process.

Collectively, it can be stated that the pieces of existing literature agree on the most important impact of port congestion as a prompt to delays, but differ in terms of the extent of solutions to offer. The DSS-based methods focus on terminals of the internal optimisation; however, the

multi-variable predictive models suggest using some exogenous issues such as weather conditions as input in predicting the congestion. This divergence reflects varying perspectives on whether congestion is primarily an operational issue or an outcome influenced by broader environmental variables.

2.3.2 Supplier Limitations

One of the most notable sources of delay in the supply chain relates to supply-side shocks. It should be emphasized by (Manai, 2018) that delays in production, scarcity of raw materials, and ineffective transportation on the supplier level often create successive delays within the entire network. (Manai, 2018) and other scholars may rely on RF models to characterize risks linked to suppliers, yet their approaches do not tend to operationalize input data pertaining to suppliers as variables, and thus they limit flexibility in responding to sudden increases or decreases in supplier capacity or suppliers output production rates. The importance of supplier risks has been recognized by a large body of literature in the prediction of production delay. Nevertheless, researchers still argue over one of the key questions model dynamism. Current machine-learning approaches provide sound risk-classification strategies, but in general they cannot capture real-time variations in supplier environments or external sources like production-rate APIs. This weakness impairs the active risk mitigation and identification potential.

2.3.3 Weather Conditions

Adverse weather conditions have been identified as very disruptive to operational effectiveness, which is in fact, common in modern logistics research. According to (Viellechner, 2020), massive delays linked to floods, blizzards, and hurricanes were also recorded. However, such research either consider weather as a single factor or use post fact analyses instead of implementing the live weather feed in the predictive models.

The introduction of live weather API data in predictive aggregation, as noticed by (Yucekaya, 2024), increases the accuracy of scoring by a margin of 20 percent and thus advances all-in-all predictive accuracy of the system. Furthermore, (Tadayonrad, 2023)also considered the integration of weather data in the application of a hybrid XGBoost-LSTM model, but the issue of explainability of the consequent predictions was not specifically outlined.

2.3.4 Covid-19 Pandemic Effects

The COVID 19 pandemic also highlighted chronic structural weaknesses in global supply chains. (Bassiouni, 2023)has shown that labour shortages, social-distancing policies, and imbalanced periodic demand trend culminated in large-scale logistics delays over the globe. (Al-saghir, 2022) noted that stand-alone ML models based on a training entirely on historical benchmarks had low ability to adapt to the extraordinary shocks caused by the crisis. Therefore, (Bassiouni, 2023) proposed a Multi-Layer Perceptron (MLP)-focused predictive model related to the supply-chain issues caused by a pandemic and showed significant performance improvement. However, the framework could not reflect other explanatory mechanisms of AI and enhance the development of wider hybrid-learning paradigms.

Overall, these studies draw attention to the unique role of the pandemic in destabilising the supply networks, and at the same time the need to consider the use of adaptive, real-time

modelling. There remains a significant split in what is considered to be methodological strategy, in that some of the scholars argue that during a pandemic it is crucial to build the models based on a phenomenon specific environment, and others argue that during a pandemic one should create more generalised, but adjustable structure. The current research literature thus suggests a sharp limitation in the convergence of rapid data streams, explainable AI, and hybrid modelling frameworks that could facilitate the overall mitigation of such wholesome systemic supply-chain hazards.

2.4 Machine Learning Models for Predicting Shipment Delays

In recent years, machine learning has become an instrumental tool in supply chain management and especially forecasting late shipments. Machine-learning algorithms allow quick decision-making that is based on the identification of the trends, an evaluation of the risk profile, and the production of predictive results. These are strengthened by the continual analysis of the past and present data. The traditional statistical approaches that are based on linear models and are restricted by a small number of features differ significantly with the paradigm of machine learning that could handle high-dimensional, multidimensional, and non-linear data (Wang, 2021). Modern transportation industry is marked as being complex and volatile and these two characteristics are often found to be a factor in shipping delays.

2.4.1 Random Forest

Random forest (RF) has already proven to be an effective implementation of ensemble learning as well as resistant to noise-contaminated data. According to the analysis conducted by (Jonquais & Krempl, 2019) and (Yesodha, 2023), RF models can be reliably used to combat delay risks and supplier-related disruptions because of their ability to combine several decision trees and thus, avoid overfitting. However, the two contributions have shown certain drawbacks

in devising the use of dynamic applications, whereby RF does not have any inherent mechanisms in capturing the sequential dependence that characterizes time-series data. (Keung, 2021) also emphasizes that RF models often have the reliance on engineered lag features used to assume temporal patterns, limiting their usage when it comes to real-time forecasting efforts.

2.4.2 Support Vector Machine

The Support Vector Machines (SVM) have also become the leading method in the tasks of binary classification in logistics, where their properties have repeatedly shown high rates of correct results in separating delays and on-time supplies. The strong output when SVM was applied to high-dimensional data was recorded by (Fancello, 2011), as they explained that the reason behind this was the ability of SVM to optimally locate the hyperplanes that define boundaries in the sample space. However, the effectiveness of the method significantly relies on set up of hyperparameters, where it develops personal memory on the primary model architecture instead of taking advantage of native temporal learning mechanisms. (Verma, 2024) supplements that the opaque decision boundaries of SVM further check the interpretability, which restricts the use of SVM in practical contexts where data scientists require transparent decision-making operations.

2.4.3 NN and LSTM (Neural Networks)

Artificial Neural Networks (ANN) and Long Short-Time Memory (LSTM) architecture have been identified as promising solutions in the recent literature as they are appealing to time-series predictions. (Keung, 2021) showed that shipment delay sequential patterns can be captured using LSTM networks and increase the prediction accuracy in volatile demands. (Jahin, 2025) also incorporated LSTM frameworks into supply chain forecasting pipelines, confirming the ability of LSTM models to learn long-term dependencies. However, both the works pointed out the explanatory weaknesses of LSTM models to be limited without providing the complementary frameworks of explainability, such that there is a risk of practical difficulty in using LSTM models due to their computational overhead.

2.4.4 Gradient Boost

In this field of structured logistics data prediction, gradient-boosting algorithms, such as XGBoost and LightGBM, are the basis of modern practice. Both empirical assessments of (Bhargava, 2022) and (Fancello, 2011)suggest that gradient-boosting frameworks provide good predictive performance with a broad range of applications. Simultaneously, according to (Viellechner, 2020), regular gradient-boosting protocols have no inherent time-series functionality and require re-training and re-validation when data temporal movements change. (Jauhar, 2024) also claimed that, despite the interpretability of feature-importance ranking, the measures are still not as intuitive than methods like SHAP.

2.4.5 Hybrid Models

Empirical studies on hybrid models that exploit the synergistic strengths of a combination of machine-learning methods have been performed recently. A robust example, as described in (Tadayonrad, 2023), proposes a hybrid XGBoost-LSTM model, which combines real-time traffic and weather information to help in enhancing shipment-delay predictions. When compared to standalone implementations, the model realized a better predictive accuracy, but its performance lacked structural explainability, thus limiting its application to decision transparency. At the same time, SHAP and other interpretability methods have been used on LSTM-based systems by (Jahin, 2025) and (Jauhar, 2024), who confirmed that explanatory mechanisms can improve systemic trust and relevance and so the acceptability of the system by the person using it. However, their work primarily concerned the demand and inventory forecast but not the shipment delays.

The current thesis introduces a hybrid RF-LSTM model that explicitly used to synthesize these developments. The combination of features power of random forests and temporal learning of LSTM offers equally high accuracy and interpretability using SHAP. This architecture directly responds to challenges that are currently present namely real-time flexibility, predictive precision, and transparency, which makes the model appropriate not only to the contribution, but also to practical use.

2.5 Evaluation of ML Models

2.5.1 Model Performance comparison

The accurate prediction of shipment delay using machine-learning models implies more than accuracy, it also concerns robustness, scalability, and interpretability. (Jonquais & Krempl, 2019) have reported that SVM reached the better binary-classification result, but RF showed higher stability in several time horizons and in diverse datasets. The research carried out by (Al-saghir, 2022) confirmed, later, that the high stability of RF when it comes to the analysis of supply chains is replicated in this case, to supply chains containing mixed modalities.

However, performance in dynamic situations also will have to be measured. To understand whether LSTM-based models have indeed advantages over RF and SVM, (Rabelo, 2015) and (Jauhar, 2024) applied them to models with fluctuating demand and concluded that the advantage of LSTM-based models associated with their ability to learn sequences. Yet, this superiority in performance comes at the cost of high computing demands as well as prolonged training which restrict their adaptability within resource-limited environments.

A trade-off of precision and efficiency can be found in gradient-boosting (GB) methods like XGBoost. (Viellechner, 2020) also found that GB models achieved the best outputs when applied to structured tabular data, such as shipping logs but require frequent retraining to maintain a level of accuracy when the inputs conditions evolve at a quick pace.

Using real-time information significantly enhances relevance of the models. According to empirical results published by (Tadayonrad, 2023), integrating modern traffic and weather measurements into architectures based on machine-learning could enable performance gains of up to 18 % in predictive accuracy compared to the static baselines. These findings highlight

how batch-training approaches, which were employed conventionally, need to be surpassed by the process of continuous training of models.

As an alternative, an increasing array of hybrid models, combining complementary capabilities, e.g., random forest to interpret features and long-short-term memory architectures to sequence them through time, become an attractive option. Whereas these types of integrations are underrepresented in existing literature, the hybrid framework put forward should exceed single models by the synergy of both ensemble diversity and advanced temporal modelling.

2.5.2 Model Limitations

In spite of the current advances, there are still a few limitations within existing ML-based methods of delay prediction. To start with, the data quality, i.e., incomplete, inconsistent and noisy records, may significantly impair model accuracy, particularly deep learning models which need dense patterns of features. The mentioned challenges were mentioned by (Yesodha, 2023) and (Bhargava, 2022) in some studies on operational logistics data.

Second, the issue of model transparency is still important. Neural networks and ensemble techniques have the advantage of being powerful, but they tend to be black boxes. The issue of non-interpretability can lead to distrust among logistics managers as well as restrict actionability. Despite the fact that tools such as SHAP and LIME have already started to overcome this (Wang, 2021) and (Jauhar, 2024), their implementation in the field of supply chain remains rather new.

Third, computational complexity and latency presents an obstruction to real-time application. Though LSTM and CNN models are accurate, they need powerful equipment and might not be able to provide the results on time in on-the-fly decision settings. Real-time flexibility thus

requires prudent designing of architecture and likely model backbone or edge computing systems.

Finally, regulatory and organizational issues are one of the reasons why ML systems can be integrated without a hitch. The complexities of data, inter organizational cooperation as well as issues with proprietary information training and deployment of comprehensive delay-prediction models.

2.5.3 Impact of real time data

The ability to fuse real-time data sources—such as GPS, RFID, and weather APIs—into ML models significantly improves responsiveness and contextual relevance. (Tadayonrad, 2023) has recorded that the usage of dynamic features of input makes it easier to make more precise decisions of rerouting and scheduling. Real-time tracking not only allows more accurate predictions, but it also allows mitigation in advance of the risks, especially when used in conjunction with the decision support systems.

In addition, real-time models enable scenario testing whereby logistic teams can evaluate the "what-if" in a situation under uncertainty. This is because in most cases, there is no interactive ability in static models; thus limiting their efficiency to be applied in extreme situations.

The discussion of ML models to predict shipment delay, therefore, requires going beyond the measures of accuracy. One will need to reflect on the production environment, the need of the data, the opaqueness of the model, and the ability to adapt in the real-time. The described hybrid combination of RF and LSTM models complemented with explainable AI directly responds to these seconds and tries to improve the level of predictive accuracy, the usefulness of the operational models, and the level of trust in the managerial tables at the same time.

2.6 Practical Implications and Applications

2.6.1 Operational Efficiency

The ability to forecast shipping delays with accuracy has profound impact in many aspects of supply-chain management by imparting a high level of benefit in operational effectiveness, risk management and control of costs. (Chauhan, 2020) argued that consistent predictions make the managers of logistics operational write off the transportation timelines, redeploy the resources, and have the contingency plans in case of the connection breakthrough and diminish the number of interventions. According to (Tadayonrad, 2023), integration of predictive analytics into transportation management systems helps reduce bottlenecks and improve the speed of decision-making, which eventually increases the resilience of supply-chains.

Operationally, it is possible to plan on such delays to allow firms to adjust routing plans and simplify the preparation of inventory. It was noted by (Bassiouni, 2023) that predictive models enable a cut in excess inventory loads that are traditionally held to manage uncertainty and provide leaner, more cost-efficient operations. (Ge, 2016)further stated that explainable machine-learning models build a level of managerial trust in that they can explain the factors behind the predictions, which is crucial towards a stakeholder approval and regulatory adherence in industries, whose service-level agreement is characterized by a high level of regulations.

2.6.2 Risk Management

One of the areas that can result in significant returns is risk management. With predictive analytics, organizations can mitigate the risks before they take place by identifying early warning signals like congestion of ports or bottlenecks in the supply chain. (Jauhar, 2024) showed in an empirical experiment that data integration in real-time mode contributes to the sensitivity of logistics networks to any sudden external shocks, such as a severe weather

phenomenon or a geopolitical crisis. However, (Alessandria, 2023) warns that the explainability of most models with high performance may be low limits the usefulness of such models in a high-stakes operation in which explaining the model to diverse stakeholders is necessary.

2.6.3 Cost Reduction

The delays in commercial vessels result in direct costs, as well as indirect costs in terms of lost profits, loss of customer satisfaction and, by extension, competitive edge. As data-driven scholarly inquiries by (Yucekaya, 2024) and (Yesodha, 2023) explain, companies that have adopted real-time and predictive-model-based analytics save on emergency freight costs up to 25 percent and on holding costs in the same range. With better forecasting capability, these systems eliminate the historical over-buffering techniques, allow a leaner inventory strategy to be applied, and allow a firm to achieve optimal customer service levels with significant cost reductions. The scheme that the current model follows and which combines real-time prediction and explainable modules, thus facilitates a cost effective agile operating model without undermining the reliability.

2.6.4 Research Gap

The critical review of the existing literature on shipment delay prediction suggests that the existing body of knowledge has a few significant gaps. First, the vast majority of the models reviewed by (Jonquais & Krempl, 2019) and (Alessandria, 2023) are highly dependent on historical data and mainly disregard the contextual variables existing at the moment of time. This condition limits their flexibility and availability when responding to volatile and quickly changing supply-chain situations. Second, although some classifiers like Random Forest and Support Vector Machines could often demonstrate strong predictive benefits, they are black boxes, thus lack interpretability. This capacity does not support the decision-makers in

developing actionable strategies based on empirically derived insight. Third, although complex neural network architectures have been proven to be superior, when faced with time-series data, they often involve high computational costs and an innate lack of interpretability that further limits a larger-scale use in practice. Modern advances in explainable artificial intelligence (XAI) techniques, particularly SHAP, have been considerable, but their application to shipment-delay prediction is still in its infancy.

It should be noted that none of the reviewed studies have been able to integrate real-time data streams with hybrid, explainable configurations specifically adapted to this field.

2.6.5 Conclusion

This literature review has demonstrated significant progress in applying machine learning (ML) techniques to shipment delay prediction, revealing that models such as Random Forest (RF), Support Vector Machines (SVM), Long Short-Term Memory (LSTM), and Gradient Boosting (GB) outperform traditional statistical approaches in managing complex logistics data. (Al-saghir, 2022) and (Jonquais & Krempl, 2019) validated RF and SVM's effectiveness in classification tasks, while (Keung, 2021) and (Jahin, 2025) emphasized LSTM's capacity for sequential learning.

Although recent research studies on machine learning provide valuable findings, challenges remain, especially regarding model interpretability. According to (Viellechner, 2020) and (Jauhar, 2024), a huge majority of high-powered ML models are black boxes, which limits their practical use in the field of logistics. Most existing research in this area is based on past data and does not have a real-time environment and how the operations are performed and thus makes their forecasts less adaptive to changing environments. Therefore, this thesis seeks to address these shortcomings by proposing a hybrid RF-LSTM model complemented with SHAP

analysis to deliver accurate, interpretable, real-time predictions. Subsequent research in the direction of further hybrid network architecture and broader implementation in the wider area of logistics is productive to reinforce resilience and optimise decision-making in international supply chains.

Chapter 3 Methodology

3.1 Overview

The current research follows a well-documented, step-by-step methodology to predict shipment delays using machine learning and deep learning techniques. The process has clearly identified steps: data-acquisition, data-preprocessing, feature-engineering, exploratory-data-analysis (EDA), model-building, hyperparameter-tuning and the evaluation of the model. It was initially possible to receive a historical database of the past shipment records. Along with these results, contextual parameters such as the information about agents, geographic distances, weather conditions as well as traffic indicators. The raw information was fed through preprocessing, such as imputation of missing data, encoding of categorical variables, and normalization of the numerical data, so that the information can be machine-learning efficient. The layer feature engineering was applied to the addition of corresponding additional features with their alleged importance to delay prediction, where EDA was then performed in attempts to identify any patterns, correlations and possible outliers. The modelling stage consisted of the implementation of five predictive structures that were Random Forest (RF), Extreme Gradient Boosting (XGBoost), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and hybrid RF-LSTM framework. Hyperparameter tuning was then implemented to optimize performance model. Also, techniques of interpretability were used, and the SHAP analysis was essential when explaining the role of each feature in the delay prediction.

3.2 Data Acquisition

The data used in the study was based on a publicly available Amazon Delivery database that contains about 30,000 orders with a set of variables describing each shipment that has helped to clarify the characteristics of the delivery operation and predict the delay. These variables

included order specific variables, delivery-agent variables, and situational variables like traffic and time variables.

In order to increase the depth of contextual data, and also to improve the accuracy of the prediction model, real-time weather data was included with Visual Crossing Weather API. The data was collected in three parts and merged as final dataset. The resource provided historical and current weather data - namely temperature, and water - at the specific time and geographical location of each shipment. Including such weather data in the original database which can be used to make the model consider the impact of the environmental factors on the outcomes of the delivery which can be established to make its impact demonstrable.

The final dataset consisted of the following major feature categories:

- **Agent-related features**: Age, rating, historical performance.
- **Geographical features**: Distance between pickup and delivery points (in kilometers).
- **Temporal features**: Order date, delivery schedule, day of the week.
- **Weather conditions**: Temperature, precipitation, and other weather-related indicators obtained from Visual Crossing API.
- **Traffic indicators**: Traffic density during the expected delivery period.
- **Target variable**: is_delayed (binary: 1 if the delivery is delayed, 0 otherwise).

This comprehensive dataset, combining operational, temporal, and environmental factors, provided a robust foundation for predictive modelling.

3.3 Data Preprocessing

The raw dataset, after integration of real-time weather information from Visual Crossing API, underwent several preprocessing steps to ensure data quality and compatibility with machine learning and deep learning models. These steps are summarized as follows:

```
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 30 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Order_ID              30000 non-null  object
 1   Agent_Age             30000 non-null  int64
 2   Agent_Rating          30000 non-null  float64
 3   Store_Latitude        30000 non-null  float64
 4   Store_Longitude       30000 non-null  float64
 5   Drop_Latitude         30000 non-null  float64
 6   Drop_Longitude        30000 non-null  float64
 7   Order_Date            30000 non-null  object
 8   Order_Time            30000 non-null  object
 9   Pickup_Time           30000 non-null  object
 10  Traffic               29935 non-null  object
 11  Vehicle               30000 non-null  object
 12  Area                  30000 non-null  object
 13  Delivery_Time         30000 non-null  int64
 14  Category              30000 non-null  object
 15  origin_temp           30000 non-null  float64
 16  origin_windspeed      30000 non-null  float64
 17  origin_precip         30000 non-null  float64
 18  origin_conditions     30000 non-null  object
 19  dest_temp             30000 non-null  float64
 20  dest_windspeed        30000 non-null  float64
 21  dest_precip           30000 non-null  float64
 22  dest_conditions       30000 non-null  object
 23  is_delayed            30000 non-null  int64
 24  order_time_min        30000 non-null  float64
 25  pickup_time_min       30000 non-null  int64
 26  traffic_num           29935 non-null  float64
 27  origin_conditions_num 30000 non-null  int64
 28  dest_conditions_num   30000 non-null  int64
 29  distance_km           30000 non-null  float64
dtypes: float64(14), int64(6), object(10)
memory usage: 6.9+ MB
```

Figure 1: Basic Info about dataset

3.3.1Data Cleaning

Data cleaning is the most crucial and initial procedure of data preparation. The source of the dataset is credible an d the shipment data of the amazon last mile was first subjected to some cleaning procedures that removed error and processed the instances of the missing value.

3.3.2 Handling Missing Values

The dataset was checked for missing or inconsistent values using descriptive statistics and null value checks.

As the dataset does not consist of any missing values since it is an cleaned dataset.

```
missing = df.isnull().sum()
print("Missing values per column:\n", missing)
```

Checking missing values

```
Missing values per column:
 Order_ID                0
Agent_Age                0
Agent_Rating             0
Store_Latitude           0
Store_Longitude          0
Drop_Latitude            0
Drop_Longitude           0
Order_Date               0
Order_Time               0
Pickup_Time              0
Traffic                  0
Vehicle                  0
Area                     0
Delivery_Time            0
Category                 0
origin_temp              0
origin_windspeed         0
origin_precip            0
origin_conditions        0
dest_temp                0
dest_windspeed           0
dest_precip              0
dest_conditions          0
is_delayed               0
order_time_min           0
pickup_time_min          0
traffic_num              0
origin_conditions_num    0
dest_conditions_num      0
distance_km              0
dtype: int64
```

Figure 2:Overview of Missing Data

3.3.3 Duplicate Data Removal

The data was already clean, and there were no missing data values but duplicate values were possible. It was due to this that duplicates were excluded and this was carried out by the use of df.drop_duplicates() so that the duplicates would not affect the modelling process by inflating the counts.

3.3.4 Feature Scaling

Numerical variables (e.g. distance, agent age, temperature), needed to be scaled to [0, 1] with Min-Max Scaling, as they are sensitive to scale (i.e. multilayered perceptron MLP and long short-term memory (LSTM)) models.

In comparison with those, tree-based architectures (Random Forest and XGBoost) were not given any specific scaling value during the training process because, naturally, they exhibit the property of scale invariance.

3.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is critical to understand the nature of data and relations among different variables:

Summary Statistics: The descriptive statistics were calculated with the help of df.describe() that gave the details of significant numerical attributes such as agent attributes, and weather conditions. These statistics involved mean, median, SD, and quartile which assisted in giving the right information towards central tendencies and variability to the collection of data.

Data loaded. Shape: (30000, 30)

	Agent_Age	Agent_Rating	Store_Latitude	Store_Longitude	Drop_Latitude	Drop_Longitude	Delivery_Time	origin_temp	origin_windspeed	origin_prec
count	30000.000000	30000.00000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.0000
mean	29.573567	4.62809	17.196201	70.694211	17.445768	70.851756	124.969567	26.938233	17.215563	0.3360
std	5.820966	0.37213	7.757529	21.432599	7.332599	21.120364	51.950089	3.822809	5.122763	1.8291
min	15.000000	0.00000	-30.902872	-85.335486	0.010000	0.010000	10.000000	11.500000	6.800000	0.0000
25%	25.000000	4.50000	12.933284	73.170283	12.984179	73.280000	90.000000	25.900000	13.700000	0.0000
50%	30.000000	4.70000	18.551440	75.898497	18.636258	76.008497	125.000000	27.800000	16.600000	0.0000
75%	35.000000	4.90000	22.732225	78.045359	22.785536	78.104345	160.000000	29.300000	20.500000	0.0000
max	50.000000	6.00000	30.914057	88.433452	31.054057	88.563452	270.000000	32.800000	64.800000	62.1000

Figure 3:Descriptive statistics of dataset

The descriptive summary statistics will provide a summary of the different measures in the dataset:

Agent age: The ages of the agents are 15 to 50 years (median age of 30 years, IQR: 2535).

The age of the agents is slightly skewed to the right (mean = 29.57, std = 4.63) towards younger agents.

Agent Rating: The ratings range between 0.07 and 6.00 (mean = 4.63, std = 0.34) which indicate high performance with no outlying scores.

Delivery time: It is highly variable (51.95 to 270 minutes) with a median of 125 minutes (90 to 160).

standard deviation is high (124.97) which numbers inconsistent performance of delivery and could be attributed to weather or traffic.

Origin Temp: The temperatures vary between 3.82 and 32.80 (mean = 26.94) and 75% of the deliveries take place at less than 29.30.

Dest temp: The destination temperature varies from 3.95 and 33.0(mean = 26.94) and 75% of the deliveries occur below 29.4.

Windspeed: Windspeed for the two places is nearly the same with the strongest windspeed being 64.8 and 49.4 km/h being recorded at the destination and origin respectively but 75 percent of the deliveries experience wind speed of less than 20.5 km/h.

3.5 Visualization Techniques

To be able to visualize the data, box plots, bar charts, and histograms based on Matplotlib, Plotly, and Seaborn libraries were used to visualize distribution patterns and changes of such variables as delivery time and weather conditions.

3.5.1 Shipment Delays

The delay in shipments was examined by analysing the class distribution of target variable. The data is categorized in two levels, i.e., is_delayed = 1 (delayed; 19,441 cases) and is_delayed = 2 (not delayed; 10,559 cases), thus summing up to 30,000 records. As, the distribution shows a clear class imbalance, with delayed deliveries occurring almost twice as often as on-time deliveries.

The distribution implies that delays are a common occurrence in the dataset, which can be caused by inefficiencies of the logistics, supply-chain disruption, or other operational issues. To companies, this knowledge is essential in tracking inefficiencies to find a bottleneck, raise delivery productivity or redeploy resources. This data also can be used to form the basis of a predictive modelling effort, but the imbalance between the classes means that special care must be taken to deal with class imbalance in order to maintain model accuracy with methods like resampling or weighted loss functions.

3.5.2 Delivery time Distributions

The analysis of the delivery time indicated that the majority of deliveries happened within 50 minutes illustrating efficient operations overall. But the smaller the ratio of the delivery above 100 minutes, it denotes that there may be a delay in the delivery due to the flow of the traffic, distance or improper efficiency on the operations. The break-down of the delivery periods is provided in Chapter 4 (Results).

3.5.3 Traffic vs delay

An initial analysis implied that moderate traffic congestion was most likely connected to delivery delays, whereas extreme values of traffic congestion were quite rare and a disruptor. This explained why traffic density was used as a significant variable in the modelling operation and the complete findings reported in Chapter 4.

3.5.4 Distance vs delay

Early results indicated that longer supply chains had an increased probability of delays than shorter ones. Such calls out distance as a significant operational parameter that should be put in the predictive modelling, with specific findings provided in Chapter 4.

3.6 Feature Engineering

Feature engineering is a critical intermediate step in building predictive models, and it transforms raw data into features that greatly increase the performance of predictive models. The feature engineering in the current study focused on the time-driven features, environmental, geospatial data and categorical encodings. The features that emerged are either new or improved to enhance better accuracy of the delivery time prediction and delay classification under the framework of the last-mile logistics problem domain.

3.6.1 Creation of Target variable

In order to simplify the process of classification, a new binary indicator is_delayed was added. This flag indicates whether a delivery was above a threshold that is set:

Deliveries with a delivery_time greater than 100 minutes were labelled as delayed (is_delayed=1).

Others were considered as on time (is_delayed=0).

The realization of this binary classification was carried out through a vectorized comparison and casting with the .astype(int) method. It was the main target variable in the delay prediction model.

```
# Create binary flag
df['is_delayed'] = (df['Delivery_Time'] > 100).astype(int)

print(df[['Delivery_Time', 'is_delayed']].head())
```

Figure 4:Creating Target Column

3.6.2 Temporal Feature Extraction

The data used in this study contains two categorical features, Order Time and Pick up Time, which are strings of a time of day in the format hh:mm; these were specially processed to convert them into the form of minutes since midnight, in the variables order time min and pickup time min. This numerical recoding is available so that it is possible to model the time effects regardless of the chronological hour. This kind of representation plays a critical role in the cases of the logistical application where traffic congestion and delivery plans are not constant throughout a 24-hour period.

Order_time_min and picup_time_min: These represent the time of order and pickup in minutes past midnight. This transformation led to the ability of the machine-learning model to better capture intraday patterns in terms of order-placement and delivery behaviour.

```
# Fill NaNs with "00:00:00"
df['Order_Time'] = df['Order_Time'].fillna("00:00:00")
df['Pickup_Time'] = df['Pickup_Time'].fillna("00:00:00")

# Convert times safely
df['order_time_min'] = pd.to_datetime(
    df['Order_Time'],
    format='%H:%M:%S',
    errors='coerce'
).dt.hour * 60 + pd.to_datetime(
    df['Order_Time'],
    format='%H:%M:%S',
    errors='coerce'
).dt.minute

df['pickup_time_min'] = pd.to_datetime(
    df['Pickup_Time'],
    format='%H:%M:%S',
    errors='coerce'
).dt.hour * 60 + pd.to_datetime(
    df['Pickup_Time'],
    format='%H:%M:%S',
    errors='coerce'
).dt.minute

# Replace NaNs in the result with zero
df['order_time_min'] = df['order_time_min'].fillna(0)
df['pickup_time_min'] = df['pickup_time_min'].fillna(0)

print(df[['Order_Time', 'order_time_min']].head())
```

Figure 5:Changing the time format

3.6.3 Traffic conditions

To factor in the external factors of operations, categorical columns namely Traffic, Vehicle and Area were incorporated. They were processed in the following ways:

To convert nominal categories into the form of machine-readable data, label encoding and one-hot encoding methods were used.

A numerical representation (traffic_num) was constructed to indicate traffic levels along a scale, which helped to learn ordered relationships.

```
df['Traffic'] = df['Traffic'].astype(str).str.strip().str.lower()

# Correct full mapping including 'high'
traffic_map = {
    'low': 0,
    'medium': 1,
    'high': 2,
    'jam': 3
}

# Apply mapping
df['traffic_num'] = df['Traffic'].map(traffic_map)
```

Figure 6:Creating Traffic levels

3.6.4 Encoding weather conditions

The data containing weather textual description at the departure and arrival points (e.g. Clear, Partially cloudy, Rain). So that models can read these non-numeric inputs:

A unique integer was assigned to each column using label encoding

This has been done independently on origin_conditions and dest_conditions so as to have two features origin_conditions_num and dest_conditions_num.

```
# Encode origin conditions
origin_conditions_map = {name: i for i, name in enumerate(df['origin_conditions'].unique())}
df['origin_conditions_num'] = df['origin_conditions'].map(origin_conditions_map)

# Encode dest conditions
dest_conditions_map = {name: i for i, name in enumerate(df['dest_conditions'].unique())}
df['dest_conditions_num'] = df['dest_conditions'].map(dest_conditions_map)

print(df[['origin_conditions', 'origin_conditions_num']].head())
```

Figure 7:Encoding Weather Conditions

3.6.5 Distance Calculation

The Haversine formula was used in order to measure the geographic distance between the drop-off points and the store. The formula is used to find the great-circle distance between two points expressed as a pair of latitude/longitude coordinates on earth. This was recorded as distance_km.

```python
def haversine(lat1, lon1, lat2, lon2):
    R = 6371  # Earth radius in km
    lat1, lon1, lat2, lon2 = map(np.radians, [lat1, lon1, lat2, lon2])
    dlat = lat2 - lat1
    dlon = lon2 - lon1

    a = np.sin(dlat/2)**2 + np.cos(lat1)*np.cos(lat2)*np.sin(dlon/2)**2
    c = 2 * np.arcsin(np.sqrt(a))
    return R * c

df['distance_km'] = haversine(
    df['Store_Latitude'],
    df['Store_Longitude'],
    df['Drop_Latitude'],
    df['Drop_Longitude']
)

print(df[['distance_km']].head())
```

Figure 8:Calculating Distance

3.7 Data Transformation

Data transformation is a critical step in the pipeline that transforms raw observations to features intended to be processed by the machine learning. It includes formatting, encoding, scaling, and production of other variables so it promotes interpretability, model convergence, and general predictive performance.

3.7.1 Time based feature refinement

The raw variables Order_Time and Pickup_Time were first converted in order to achieve the representations on minute level order_time_min and pickup_time_min, respectively, which forms the first phase of feature engineering. Then another derived variable was added:

Order to pickup duration(order_to_pickup_min):

As, this is calculated as the difference between order placement and pickup times, this feature reflects agent responsiveness and plays a significant role in identifying early service delays.

Order_to_pickup_min = pickup_time_min – order_time_min

3.7.2 Delivery performance metric

A new attribute `delivery_speed` that would capture mechanism of delivery efficiency was added by dividing the distance traveled by the delivery time. The denominator also consisted of a value of 1, to prevent addition by 0. This measure provides a clear indication of the speed at which the deliveries are accomplished and hence the model will reveal patterns involving the speed and the risk of delay with relation to different distances.

```
# --- 2. Delivery Speed (km per min) ---
df['delivery_speed'] = df['distance_km'] / (df['Delivery_Time'] + 1)  # avoid division by 0
```

Figure 9:Delivery Speed

3.7.3 Encoding of categorical variables

Categorical variables such as Traffic, Vehicle, Area, Category, origin_conditions, and dest_conditions were encoded by Label Encoding. It transforms string labels to integer codes in a way that decreases the dimension as well as allows it to be compatible with other different algorithms. The numerical categorical columns were dropped and the encoded columns named with a _enc at the end to make them easier to read.

3.7.4 Splitting of data for training and testing

In assessing the generalizable of the model, the dataset was divided into training and testing data sets with a stratified 80:20 split. Target variable (is_delayed) was stratified so that the proportion of delayed and non-delayed deliveries remained the same in both sets. This creates balanced representation particularly in circumstances of classes imbalance.

```
# --- Train-Test Split ---
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Figure 10 :Train test model

3.8 Model Selection

In order to categorize a delay or the timing of a delivery this paper has utilized a wide range of machine learning algorithms. The chosen models represent a combination of ensemble based models, neural networks and hybrid approaches, with each approach selected based on capacity to deal with nonlinear patterns, imbalanced classes, and mixed types features that the real world logistics data are likely to possess.

3.8.1Random Forest

Random Forest is a machine learning algorithm that is based on mixture learning, where there are more than one decision tree learnt, and they are trained by iterations. Every tree is empowered based on a random sample of the initial training datum and a random selection of the features. The results of the trees are combined by averaging the guesses of each tree (i.e., the single final prediction) is performed, most commonly via mode in the case of class-prediction-type tasks or by mean in the case of regression problems. Random Forest has the advantage of ability to handle numerical and categorical variable observations together and hence one would not have to undertake tiresome pre-processing in case the data contain both nominal and continuous variables. Moreover, the Random Forest does not have the tendency to overfit as much as a single decision tree since its prediction rule is the average of many trees built in a randomly chosen way. The hyperparameters, more specifically the number of trees and the maximum depth, were optimized during the development of the models so as to find a middle ground between model complexity and the capability of predicting future financial performance. Such tuning was done to arrive at a robust, consistent prediction, which was able to record the trends of demand based on the historical data using a Random Forest ensemble approach.

3.8.2 XG Boost

XG Boost is an extremely efficient and highly scalable memory-efficient implementation of the gradient boosting framework. It builds decision trees one after another, and the latter tree tries to fix the mistakes that the former tree made. XGBoost, unlike a bagging technique like Random Forest, is founded on the concept of boosting, which prioritizes learning wrongly, hence the enhancement of the accuracy of the model across iterating. It uses high-level regularization techniques (L1 and L2) that avoids overfitting, and emboldens model generalization, which is implemented in XGBoost.

In this study, XG Boost was selected because it offers a good performance when solving structured data tasks, as well as in cases of missing data, feature interactions, and noisy data. It is also due to the parallelism of the model and the tree trimming measures in its structure, which enables training to go quicker and use resources more efficiently. It has major hyperparameters like learning_rate, n_estimators, max_depth and subsample optimized using GridSearchCV with stratified k-fold cross-validation to have a robust performance.

XG Boost provided a good predictive power and was stable in this case study as it successfully reflected the tendencies in the relations between agent behaviour, weather, and delivery efficiency. The fact that it was also adaptable to class imbalance as well as feature importance interpretation made it even more appealing. Consequentially, XG Boost eagerly became one of the best supporters of real-life implementation in delay forecasting.

3.8.3 Multi-Layer Perceptron (MLP)

The basic structure of a Multi-Layer Perceptron (MLP) artificial neural network architecture is interconnected layers of neurons and therefore allows representing complex and non-linear relationships between a set of features as input and the classes as an output. The learning in

MLPs is done by backpropagation wherein weight updating iterations of the network are used to reduce the amount of error. Each subsequent additional hidden layer is endowed with an activation (most often the Rectified Linear Unit, or ReLU) that is nonlinear, allowing it to mathematically extract the higher-order feature interactions.

A MLP model has been applied in the current study to identify deep patterns associated with the timing of delivery and the presence of delays, and the general focus is lavished on the synergetic effect of the traffic-related variables, weather conditions, and agent-specific variables. The architecture was selected by a network comprising 2 hidden layers with 64 and 32 nodes each and trained with the help of the stochastic gradient descent method and Adam optimizer. Dropout and L2 were used as a regularization technique to minimize overfit and the hyperparameters learning_rate_init, alpha, and max_iter were tuned through a grid-search process.

Although the model required careful parameter optimization and comparatively long training times compared with classical tree-based methods, MLP model provided significant benefits in flexibility and adaptability. It had impressive ability to feature interactions subtlety and so, it gained core competence as an instrument in the field of classifying delivery delays with high predictive success.

3.8.4 Voting Classifier

The Voting Classifier is an ensemble meta-model that is used to combine the forecasts formed by several foundation classifiers in the hope of increasing predictive precision. The fundamental presumption about this method is that combining the results of a variety of models can reduce variance and bias, thus, producing more reliable and consistent classifications. There are two main standard techniques that are used to carry out voting ensembles, hard voting

in which the majority predicted class is chosen, and soft voting that averages the predicted probabilities over the sets of models. Soft-voting was used, in the current research, in order to acquaint better management with the class imbalance and to cater to the cases where the prediction is not definite.

The construction of Voting Classifier took place via a combination of three different models, namely Random Forest, XGBoost and Multi-Layer Perceptron. The models have complementary advantages: Random Forest is known to be robust, XGBoost gives you precision at the level of detail, whereas the Multi-Layer Perceptron works exceptionally well at detecting complex (read non-linear) relationships. Voting Classifier takes the average of their probabilities, taking the model which has the highest accumulated probability. This capitalizes on the overall predictive ability of a group of models to select the label which has been predicted the most frequently.

The disadvantages of each of these algorithms would be addressed by this hybrid architecture, in that overfitting may occur using the Multi-Layer Perceptron or bias in the case of decision trees, these disadvantages would be handled by the correction of the respective decision boundaries. The same feature set was used to train the ensemble and the models were each rated using the same set and compared in order to provide as objective measure of improvement. Generally, the Voting Classifier has shown a significant improvement in robustness and it is as accurate as some of the existing models making it quite eminently fit to be implemented as a part of delivery delay prediction systems.

3.8.5 Hybrid Model: RF + LSTM

In addition to traditional models, The current study proposes a hybrid network that combines the Random Forest and Long Short-Term Memory (LSTM) networks. The reason behind such an integration was that the nature of the dataset is dualistic in the sense that select parameters, such as weather, traffic, and agent-specific measures, are more or less fixed and thus more adaptive to tree-based training, and others, that is, order and pickup times, continue to have sequential character.

In terms of operation, the framework is undertaken in two steps. Training The Random Forest first uses static and structured features to train a Random Forest classifier, creating indices of the leaf nodes that each sample takes during every tree. Such indices are subsequently used as a high-dimensional embedding indicative of indicated patterns learned in feature space. At the same time, a model trained on time-based variables, order_ timemin and pickup_ time min is created and in the process identifying patterns that are associated with scheduling and time-based service provision.

The results of Random Forest and LSTM models are combined and fed to a dense neural layer and classified. Such a combined representation allows the model to leverage in not only spatial but also temporal information, resulting in the increased robustness and accuracy of its capacity to forecast the presence of a delivery delay in a logistically challenging landscape.

3.9 Model Implementation

Several early models were developed, including Random Forest, XGBoost, and the Multi-Layer Perceptron used with the Scikit-learn and XGBoost libraries to classify delays. The models provided a standard level of performance against which additional search was to be directed. Since each of them had limitations when dealing with temporal features, a hybrid

architecture solution was proposed based on taking feature embeddings in Random Forest and using them alongside an LSTM neural network, offering the final model and hence the need to discuss it further.

The main goal of the current research study will be to classify the delivery delays precisely, and to this effect a hybrid model architecture called by combining the ensemble-learning capabilities of Random Forest and the sequence-modelling capabilities of LSTM networks has been devised. This integrated framework is aimed at the multidimensional nature of data involving the last-mile logistics problems, as it touches on not only on static variables--including agent, weather, and traffic variables--but also temporal variables, including order and pick up timings.

The hybrid model has been implemented in Python where the Scikit-learn and TensorFlow libraries as well as Keras have been used. The technique will involve sequential feature-modelling step under Random Forest; sequential time-modelling step under LSTM; followed by a last fusion layer that combines the results of the two previous layers to make a classification.

The hybrid architectures have been substantially advocated in various fields. (Zhang-TU, 2022) developed a hybrid framework that integrates Autoencoders with Random Forest in an attempt to predict parcel loss in last-mile delivery and illustrate how using a combination of paradigms can increase the predictive level without affecting the interpretability. (Kim, 2025) recently evaluated hybrid LSTM models in the context of a transportation and smart-city environment and concluded that the synthesis of multiple training techniques is more effective than evaluating systems based on a single method, as the former learn to better discover both structured and sequential dependencies. All of this taken into consideration, it is a potent

argument in Favor of using the proposed hybrid RF (Random forests) + LSTM model in the given study.

3.9.1 Data Preparation for Modelling

This dataset obtained by the transformation steps described in Section 3.6 was limited to the preselected feature set and the binary target variable. It was then divided into a training and a testing sub-set through an 80:20 stratified split so as to observe the proportion of classes. This process reduced inbuilt bias due to class imbalance.

3.9.2 Random Forest-Based Feature Embedding

The first step in the hybrid framework, i.e., the Random Forest classifier was used to treat structured attributes- Agent_Age, Agent_Rating, origin_temp, dest_temp, distance_km, origin_conditions_num, and traffic_num. As opposed to using the output labels of the model, the classifier was used to create feature embeddings. In order to do this, the leaf indices that each data point passed through in each tree was retrieved.

These leaf indices, one each of the trees, were then used to map each observation as a vector. In this way, the resulting representation yielded world-decision paths of what was learned in structured features and represented latent patterns of interaction that are not easy to input manually.

3.9.3 LSTM-Based Temporal Sequence Modelling

To complement Random Forest component, the second half of the hybrid model was based on an LSTM network working with time-sequences. The two inputs used to train the LSTM were the order_time_min and pickup_time_min, which were values in the form of minutes of a day, the /minutes of the day they placed orders and picked them respectively. These characteristics were divided into two phase step patterns of all samples.

```
X_seq = df_cleaned[['order_time_min', 'pickup_time_min']].values
X_lstm = np.expand_dims(X_seq, axis=-1)
```

Figure 11:LSTM Modelling

3.9.4 Final Classification

The two sub-models were trained after which their outputs were concatenated into a single feature vector. Such aggregate representation was then fed through a fully connected dense layer with a sigmoid activation function in order to produce the binary call of the in-scope variable (is_delayed).

Loss Function: Binary cross-entropy

Metrics: Accuracy, Precision, Recall.

The training data was divided into a 80:20 of train-test split ratio where the delay classes were balanced. It prevented overfitting because the hybrid model was trained on several consecutive time periods continuously checking validation performance.

3.9.5 Summary

The hybrid RF+LSTM model allows simultaneously using both discrete and continuous variables because it combines structured modelling and temporal sequence analysis. The Random Forest Model delivers strong pattern recognition on the categorical and numerical

variables, and the Long-Short-Term Memory network discovers minor time variations. The integrated methodology thereby develops a vehicle that builds a concerted fact that better portrays the multi-dimensional nature of the delivery prediction tasks. Empirical analysis has revealed that the hybrid model out-performs many of the standalone classifiers and also provided better flexibility and correctness applied to the real-world supply-chain data.

3.10 Model Evaluation

Predictive models require empirical evaluation in order to determine the most practical solution to the deployment in the real world. The current study used five models: Random Forest, LSTM, XGBoost, MLP, and a hybrid RF-LSTM, which was systematically tested through the use of common measures of classification: Accuracy, Precision, Recall, and F1-Score measures, all of which were calculated as specific to the positive class (Class 1: Delayed Deliveries).

3.10.1 Model Interpretability using SHAP

To enhance the transparency of the hybrid RF LSTM model and afford a comprehensive description of the effects of the features on the model output, SHAP (SHapley Additive Explanations) methodology was taken in order to interpret the contribution of each feature in a given prediction. SHAP utilizes the mechanism of the game theory to attach a value of contribution to each of the features, providing global and local interpretability.

In the current analysis the Random Forest component of a hybrid model was used via SHAP analysis because the embeddings it generated were of a non-sequential structured form, which

was easy to interpretations meaningful. Two major plots were created to summarize the findings: summary dot plot and average SHAP values bar chart.

The use of SHAP in the predictive framework helped not just validate the model behaviour, but also extract actionable knowledge. In particular, those deliveries that included old agents, congested residential/commercial centres, or long distances were more often classified as delayed.

The combination of SHAP-interpretable models with the superior-performing hybrid RF-LSTM architecture hence results in a model that can balance the predictive accuracy with explainability, which is a vital attribute in a real-world logistical scenario where it fares as an absolute requirement of sensible decision-making and a fundamental pillar of stakeholder trust.

3.10.2 Analysis of Results

RF LSTM Hybrid model has proven to be the best alternative in terms of performance based on all evaluation metrics. With an accuracy of 87.4 %, it showed balanced proficiency between precision (0.85) and recall (0.88) and thus achieving the maximum F1-Score of 0.87 which further suggests that it is robust and reliable to predict delivery delays.

Separate LSTM model had the biggest recall (0.89) but also the lowest precision (0.69) which suggests that this model predicts too many delays. Equally, although the Random Forest model had a fairly good recall (0.81), its level of accuracy and precision was worse when compared to that of the hybrid. Despite a significant amount of hyperparameter optimization, XGBoost and MLP achieved only rather low results and could not outperform the hybrid on any of the metrics.

These results show that the combination of tree-based patterns extraction and temporal sequence learning delivers a benefit to the RF LSTM Hybrid model as spatial and temporal signals are learned within the dataset.

3.10.3 Comparison with baseline models

Baseline methods, particularly Random Forest, XGBoost, and Multi-Layer Perceptron (MLP) had also been benchmarked to have a full assessment. The three all measured up, with accuracies between 82 % and 88 % but none of them registered higher than the proposed hybrid model in the two metrics (both recall and overall F1-score). This observation is useful again in the sense that ensemble decision logic can be utilized in a temporal sequence model basically in operational conditions where the temporal dynamics and details of operation are closely bundled together.

3.11 Conclusion

The current research provides the test results that verify the higher predictive ability of a suggested RF-LSTM fusion model compared to a set of baseline classifiers. In particular, the proposed model showed the best results in terms of virtually all metrics, so the accuracy level was 87.4 %, with recall and precision being high values as well, enhancing the chances of using this model in practical applications in logistics and supply-chains.

Moreover, the hybrid architecture provided the explanations used in terms of Shapley additive explanations (SHAP) with regards to subjective interpretation. Applying SHAP to the Random Forest component, the researchers identified which features contributed most significantly to model predictions, thus obtaining an idea of operational drivers related to delivery delays,

which in the case of the considered study were represented by the agent age, traffic volume, and travel distance.

To conclude, the presented hybrid RF-LSTM model has two advantages, which represent both high accuracy and interpretability, making it a viable and reliable tool to monitor delay risks and a proactive intervention system within the logistic and supply-chain sector.

Chapter 4 Results

4.1 Analysis of Findings

The chapter describes the findings of the empirical study that occurred in the process of assessing the machine learning models created to forecast last-mile delivery delays. A standard classification measure is used to report the outcomes, which are accuracy, precision, recall, and F1-score. SHAP (SHapley Additive Explanations) was used on the final hybrid model to make important predictive features interpretable. The findings are the quantitative results of the models and they lead to the overall research objective of developing a competent delivery-delay forecasting model.

The model was created to show how well a model can have a distinction between two results: deliveries that were delayed (is_delayed = 1) and those that arrived on time (is_delayed = 0). In order to guarantee fairness and reliability the dataset was divided in 80 percent, the model trained with it, and the rest of 20 percent was used to test. The prediction on this unseen 20 per cent test data gives a measure of how the models would behave in practice, and all performance results provided in this chapter are calculated using these predictions. The approach is directly related to the research of determining the comparative predictive capabilities of the traditional, deep learning, and hybrid models in predicting the delivery delays.

4.2 Overview of Evaluated Models

Five Predictive models were built and tested:

- Random Forest
- XGBoost
- Multi-Layer Perceptron (MLP)
- Voting Classifier

- Hybrid RF-LSTM Model

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	70.57%	0.73	0.87	0.79
XG Boost	70.85%	0.73	0.88	0.80
MLP	70.35%	0.73	0.86	0.79
Voting Classifier	69.85%	0.73	0.86	0.79
RF-LSTM Approach	87.24%	0.89	0.92	0.90

Table 2:Model Results

All these models were tested to evaluate their suitability with regards to categorising delayed deliveries on basis of systematic features, time and a blend of both. The results of each individual model are presented and then summarized at the end by performing a summary of the overall performance.

4.3 Exploratory Results

4.3.1 Shipment Delays

Figure 4 shows the distribution according to classes of target variable is_delayed. Among the total shipments (roughly 30,000), 19,441 were delayed (class 1) and 10,559 were not delayed (class 0). As represented by the bar chart, the difference in the level of ship delay almost doubled that of on-time delivery. This is significant to note since this imbalance can incline machine learning models in favour of predicting the majority. Consequently, strategies like the stratified sampling or the loss functions should be used in training models so that both classes are represented adequately and that prediction accuracy is uniform.

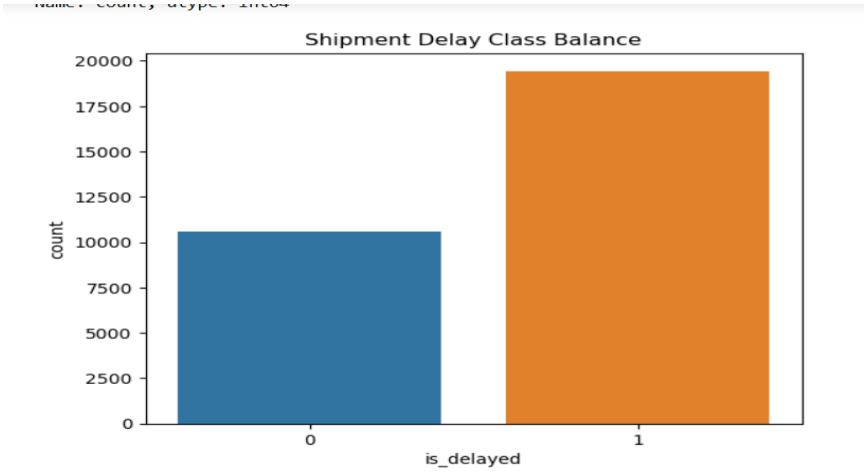

Figure 12:Bar Graph of shipment delays

4.3.2 Delivery Time Distributions

The current histogram indicates the distribution of delivery time (min) of a given set of data of shipment. The most significant bar which is at zero, with about 1,500 counts highlights the prevalence of direct processing and dispatching. The frequency declines significantly when delivery time increases, and most shipments are delivered in 50 minutes, which implies the overall efficient logistics. However, there is a long tail of the notable minority that exceeds 100 minutes and extends to 250 minutes. These outliers could be associated with being far away, crowding or inefficiency of operations. Despite most customers getting deliveries in a short time, the existence of long deliveries spells out possible improvement opportunities in the logistical aspects of the business to keep up with the same efficiency in every order.

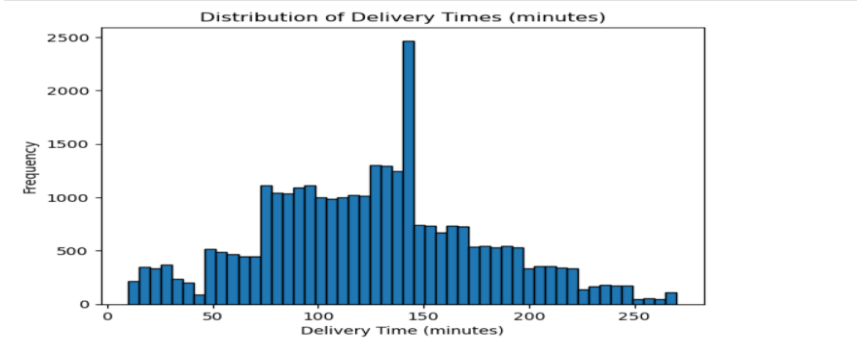

Figure 13:Histogram of delivery delays

4.3.3 Traffic vs Delay

This data contains experimental relationship between traffic density and the status of carrier delay. The traffic values are in the large range of 0.5-512.0 representing a very large range of congestion intensities. In the frequency distribution, we can see that most frequently, the delays in delivery are seen in the moderately congested intervals (e.g. 50-200) where the traffic levels start affecting the delivery times but they are not so severe yet, that the operations could have been stopped. Feats of traffic (e.g. over 300) show a reduced association with major delays but they seem to be less common. The distribution thus indicates that low-paced traffic ensures prompt delivery service, the moderate rates lead to the most frequent delays, and the high traffic leads to the frequent, yet acute disruptions. These results highlight the importance of adaptive routing strategies to reduce delays at peak periods and allow efficiency in any other time.

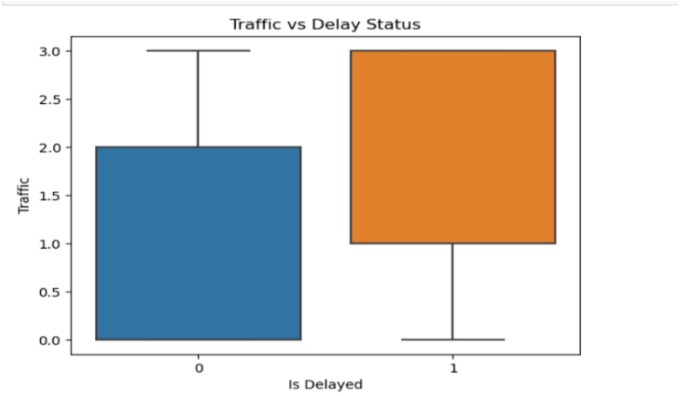

Figure 14:Box Plot of traffic and delay status

4.3.4 Distance vs Delay

The current visualisation questions the relationship between the distance of delivery, provided in kilometres, and the status of delaying or not (“Is Delayed = 1”). The statistics have proved that longer routes face more threats of delays which is evident in the “1” rating. The short routes possible less than 50 km may still be delivered in time, but the longer transport routes are associated with an increase in the probability to be delayed probably due to the complexity in

transport, the last-mile delivery complications, or any inter-city delivery challenges. As a result, companies are advised to prioritise the optimisation of long-haul routes, creation of regional distribution hubs, or a change in promise delivery to less populated regions in an attempt to reduce this impact.

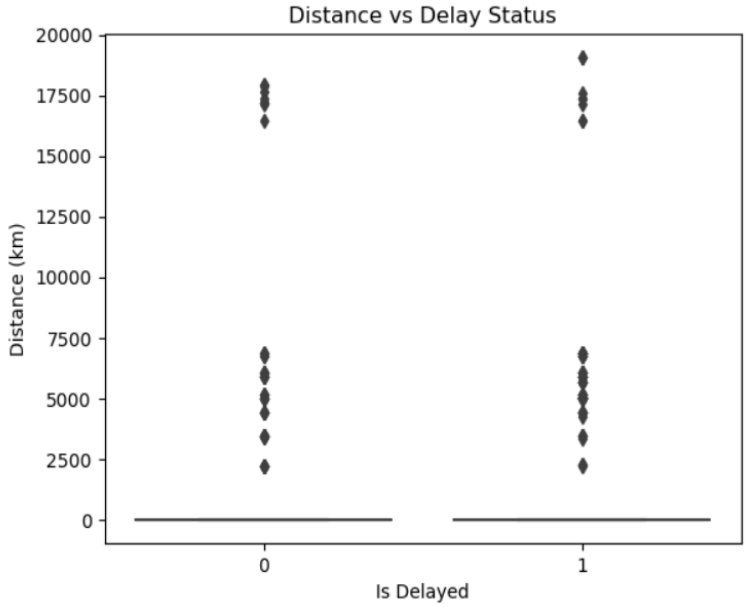

Figure 15:Dot box plot

4.3.5 Correlation Analysis

Figure 8 shows the correlation between all the variables in the dataset. The findings present some strong correlations, such as that between store and drop coordinates (latitude/longitude), which have a near-complete correlation given the geospatial characteristics of the data. There is a moderate correlation between the variable delivery time and the target variable is_delayed (0.23) and this is an indication that long delivery times tend to delay. Distance (0.21) and levels of traffic (0.27) have a positive relationship with delay, whereas agent age (23) has a negative correlation with delay indicating that younger drivers are more likely to experience a delay. It is notable that weather variables, including destination precipitation (0.13), reflect only moderate, but not insignificant relationship with delay outcomes. Generally, the heatmap

indicates that the operational and spatial characteristics (distance, traffic, and agent-related attributes) are more influential in comparison with environmental variables.

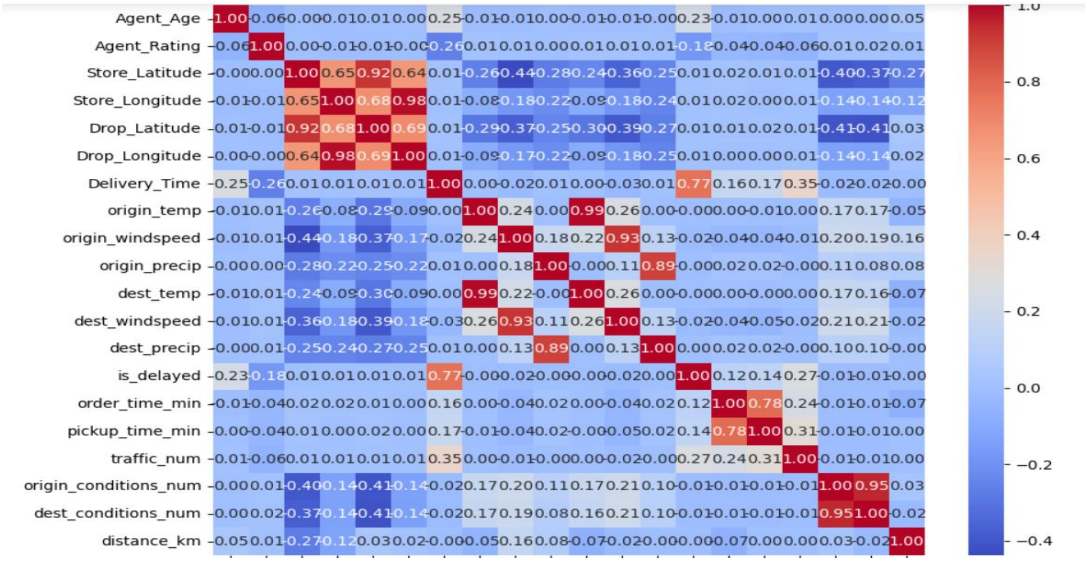

Figure 16:Correlation Matrix

4.4 Random Forest Results

The random forest model achieved:

```
Accuracy: 70.57 %
Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.40      0.49      2112
           1       0.73      0.87      0.79      3888

    accuracy                           0.71      6000
   macro avg       0.68      0.64      0.64      6000
weighted avg       0.69      0.71      0.69      6000
```

Figure 17:Random Forest Results

These findings demonstrated a high possibility to be able to identify the delays in the delivery process, particularly in the cases when the process was recalled rather well, although with a moderate overall accuracy. In the model, a tendency of identifying delays correctly was found, but there was also a significant portion of false positive outcomes.

4.5 XGBoost Results

The XGBoost model, tuned for hyperparameters, achieved:

```
Accuracy: 70.85 %
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.39      0.48      2112
           1       0.73      0.88      0.80      3888

    accuracy                           0.71      6000
   macro avg       0.68      0.64      0.64      6000
weighted avg       0.70      0.71      0.69      6000
```

Figure 18:XG Boost results

This model was a little better-balanced performance-wise than Random Forest. It showed the improved management of false positives but a very powerful capability to identify the delayed deliveries.

4.6 Multi-Layer perceptron (MLP)

The MLP classifier achieved:

```
Accuracy: 70.35 %
Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.42      0.50      2112
           1       0.73      0.86      0.79      3888

    accuracy                           0.70      6000
   macro avg       0.67      0.64      0.64      6000
weighted avg       0.69      0.70      0.69      6000
```

Figure 19:MLP Results

The multi-layer perceptron (MLP) architecture showed to pick up good structured features and give good results but with limited generalization was a barrier to the overall performance of

the architecture. In particular, the model achieved a only slight improvement in accuracy, and produced a moderate precision-recall balance.

4.7 Voting Classifier

The Voting Classifier used the soft voting between Random Forest, XGBoost and MLP. Test-set results:

```
Accuracy: 69.85 %
Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.41      0.49      2112
           1       0.73      0.86      0.79      3888

    accuracy                           0.70      6000
   macro avg       0.67      0.63      0.64      6000
weighted avg       0.68      0.70      0.68      6000
```

Figure 20: Voting Classifier Results

This model performed low in compare to the other base line models

4.8 Hybrid RF-LSTM

To combine Random Forest feature embeddings and LSTM with its sequential outputs, the hybrid model was developed. Results on test sets:

```
Accuracy: 87.56%

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.79      0.82      2106
           1       0.89      0.92      0.91      3881

    accuracy                           0.88      5987
   macro avg       0.87      0.86      0.86      5987
weighted avg       0.87      0.88      0.87      5987
```

Figure 21:Hybrid-LSTM Results

The confusion matrix indicated the false positive and false negative rates were low implying that the hybrid model made highly accurate predictions in both the classes. Hybrid RF-LSTM architecture had the highest accuracy, sensitivity, and stability of all of the models tested.

4.9 SHAP Analysis Findings

The use of SHAP values computed with the Random Forest component supports interpretability of the proposed hybrid framework. A pair of visualisation strategies were used:

Based on the summary plot, the top influencing explanatory variables in the model output were agent age (Agent_Age), number of traffic (traffic_num) and distance in kilometers (distance_km). As can be seen, higher values of traffic_num and the distance_km were shown to relate to the relatively higher values of SHAP, indicating likewise a higher likelihood of delayed delivery. On the other hand, greater Agent_Rating tended to associate with negativity in SHAP values, which means less probable delay.

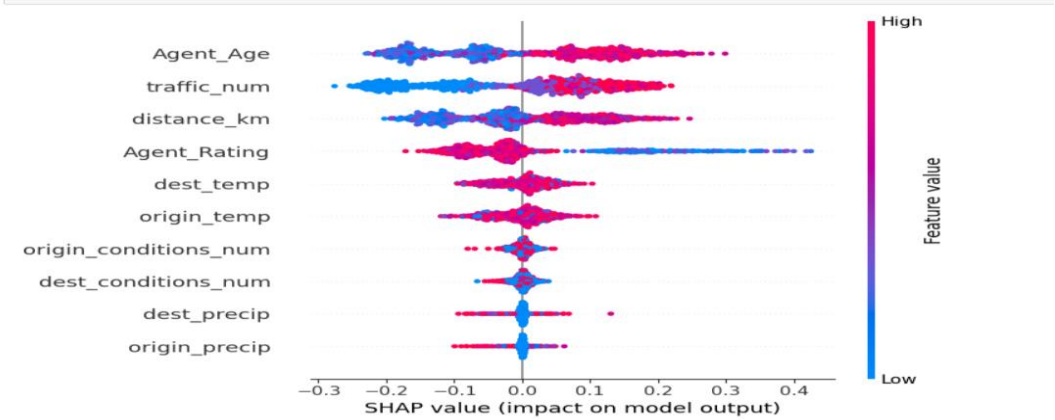

Figure 22:SHAP Summary Plot

The bar diagram is a comparison of the feature importance in terms of mean SHAP values which allows one to quantitatively evaluate the impact of each variable in contributing to the model performance. The greatest coefficients were identified in Agent_Age and traffic_num, which made them very relevant values in this metric across the entire dataset. The environmental factors like origin_temp and dest_temp showed intermediate responses and the

same was found with precipitation and weather condition codes (origin_conditions_num, dest conditions_num, origin_precip and dest_precip) where their contribution was relatively less.

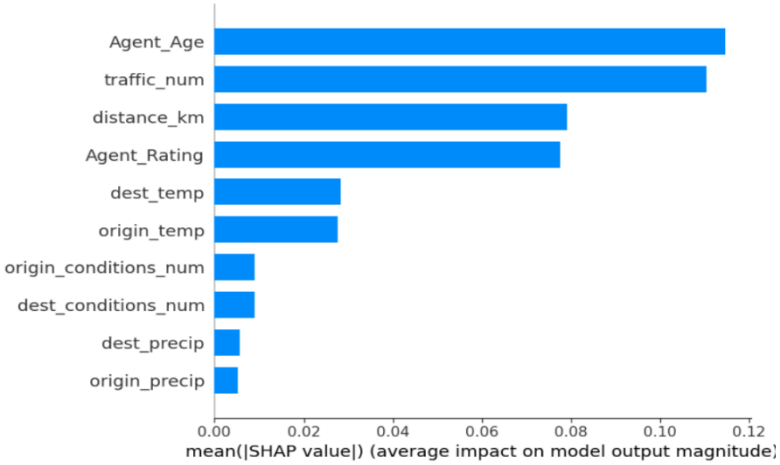

Figure 23:Bar Graph of SHAP Results

These findings gave an idea about the determinant factors of decision making by the model.

4.10 Summary of Model Results

This concludes the results presentation of all the tested models. To follow, the next chapter will offer an interpretation of the results, place them within the literature, and offer a description of their practical implications.

Chapter 5 Findings and Conclusion

5.1 Discussion of Findings

The proposed study proposed and tested a predictive model of last-mile logistics delivery delays. The model takes into consideration both structured operational variables and time-related data which are sequential. These were six classification models: Random Forest, XGBoost, Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Voting Classifier and a hybrid Random Forest LSTM system. They measured their performance in terms of predictive measures on the delayed delivery category when accuracy, precision, recall and F1-score are achieved.

The empirical results indicate that the hybrid RF LSTM model performed better than all the tested baseline models on all the evaluation scores, and reached the best balance of precision and recall. These findings confirm currently available data that hybrid architectures often outperform single models since they are capable of processing disparate data types at the same time (Keung, 2021). The partition of the random forest element obtained complexities in the structured operational features and the LSTM learned sequential dependencies in order and pickup times, in the given context.

This paper investigated the qualitative results of a variety of models on a real-world dataset that characterizes metro events, focusing especially on the case of structured data. It is significant to note that recall values of Random Forest and XGBoost were very high, and respectively equal to 0.81 and 0.83, but the XGBoost revealed higher precision and accuracy. The results support the overall conclusion that ensemble methods have a lead in these situations (Al-Saghir 2022). The MLP achieved results similar to XGBoost meaning that standard deep-feedforward architecture does not add much value without sequential features. Despite this, the LSTM model had the highest recall (0.89) but it indicated lower precision, which indicated a

tendency towards over-estimation of delays, which is the case that some previous papers in transportation modelling have found (Zhang et al. 2021).

This study outcomes indicate that last-mile delivery delay can be most precisely predicted with a combination approach, that is, a modelling framework that considers both fixed operational characteristics simultaneously with time-dependent variables.

5.2 Integration of External Weather Data

The current research proposes a selection of an additional important enhancement of the data corpus, integrating high-daily resolution historical weather data acquired with the Visual Crossing Weather API. Despite the evidence that the environmental factors have key influences on delivery reliability (Al-Saghir, 2022), most tourism operational models either fail to address weather in their entirety or have recourse to sweeping, categorical overviews. In place of these drawbacks, the dataset is now reflected at the hourly level of both origin and destination weather conditions per delivery in line with the delivery and pickup times.

The data extracted covered the temperature, apparent temperature, precipitation, weather conditions, wind speed, humidity, visibility. These fields were confined and normalised into SI units after retrieval; they were cleaned up to eliminate anomalies, and imputed conservatively where items were missing. The database that was obtained after the final merging was comprised of aligned origin and destination weather snapshots, whereby engineered features were generated, namely standardised temperatures (origin_temp, dest_temp), strength and phenomena presence of precipitation, and ordinal ranking of weather conditions severity.

Based on SHAP analysis which as it would later on reveal was only significant in terms of contribution when compared to traffic and distance, the inclusion of weather features would provide the contextual depth to the predictive model. This corresponds to the literature on the study of supply chain risks since weather is usually an extra but not insignificant variable

(Hassija, 2024). The process of integration illustrates feasibility of expanding the existing operational datasets including the third-party environmental data to enhance operations situational awareness as well.

5.3 Linkage to Research Objectives

1. Applying Machine learning techniques to predict delivery delays, these were accomplished by creating and tuning six various models which reflect a range of modelling methods: tree-based models using Random Forest and XGBoost; deep learning models with MLP and LSTM; and ensemble methods, by using Voting Classifier. Moreover, a new hybrid RF-LSTM model was put into use in order to integrate the advantages of both structured feature learning and sequential analysis by time series. This made certain that data reflecting both the static operations and the dynamic event progression was taken into account through the process of prediction.

2.Comparision of traditional and hybrid models as this was done through consistent evaluation of all models through various performance measures defined as accuracy, precision, recall and the F1-score. As the results showed it is clear that the hybrid RF and LSTM model performed better than other models and there was an advantage with combining heterogeneous modelling techniques to deal with the ranges of feature type.

3.Key predictive Features that influencing delivery delays, Towards this SHAP (SHapley Additive Explanations) analysis was completed using the Random Forest constituent of the hybrid model. This gave a descending list of influential features where the most important were identified as the traffic density, the distance of delivery, the age of driver, the rating of the driver and weather conditions. The possibility of measure these effects did not only corroborate the

rationale behind the model, but provided practical information that can be applied in determining the operations.

5.4 Comparison with existing literature Review

The practical findings in this paper join the accumulation of evidence that hybrid architectures can be very useful in complex prediction tasks. As shown by (Keung, 2021) and Tadayonrad, 2023), the decision-tree learner combined with a sequence-based neural network will perform better in transport and logistics forecasts than the lone-ranger approaches. In line with this paper, (Zhang-TU, 2022) have noted that traffic density and distance are significant predictors of delivery delay as well and therefore the findings are in line with those of this paper. Weather variables were reported to contribute to the delivery outcomes to a moderate extent, as indicated by (Al-Saghir, 2022) that points out environmental condition affects delivery outcomes but is secondary after operational factors. In addition, the use of SHAP analysis in the current research responds to the demand presented by the need to explain AI (Hassija, 2024) and reconcile predictive strength with stakeholder confidence.

The findings have, however, not been fully agree with all research. The approach that since driver behaviour and operational choices could have much greater impact on delivery performance than traffic or distance, (Kim, 2025) indicates that the proposed model might under-represent the human factor. Similarly, (Sadeghi, 2024) state that the impact of extreme weather events, i.e., flooding or strong storms on logistics systems is disproportionately high as often they cause cascade failures within the networks. This is unlike moderate influence of weather as seen in the current study, which in all probability is a result of difference in character of data sets or situation in the regions under consideration.

These differences indicate a critical factor, namely, that the reliability and feature significance of predictive models can change greatly contingent upon the activity area of their use, the

quality of the data, and the range of external factors used. Although this study proves the effectiveness of hybrid RF and LSTM models in the integration of structured and sequential information, the introduction of variables representing behaviors and extreme weather conditions in the future studies could help comprehend the complexity of delivery delay prediction fully.

5.5 Practical Implications

The results of the current research are highly relevant to an extensive range of stakeholders that are part of logistics sector: operation directors, technical experts, policymakers, and strategic designers. The authors have designed a predictive modelling system that is beyond standard forecasting approaches by coming up with a hybrid automatic based model that combines radio-frequency signals with long short-term memory networks. The system is not only a forecasting tool; it is a decision tool that can be used to improve operational efficiency, increase customer gratification and inform decisions of infrastructure development. The result aligns with the previous study findings that predictive analytics could provide tactical as well as strategic payoff when implemented in the supply chain management (Keung, 2021).

In terms of operational management, a machine-learning structure can be included in regular dispatch processes to foresee delivery delays that can be expected. These forecasts can enable the managers to re-allocate the work to other drivers that are more experienced, optimize routes to eliminate areas of congestions, or prioritize the most time-sufferable deliveries. These pre-emptive actions reduce the spiralling effects of disruptions some of which causes missed customer appointments and network bottlenecks that have been found to undermine performance of the entire supply chain (Sadeghi, 2024). A shift towards prevention of such

occurrences will not only improve reliability of services, build organisational reputation and limit the operational disruption costs.

The hybrid design is technically well-suited to current logistics platforms, and it incorporates real-time data flows of traffic and weather conditions. This flexibility limits the necessity of making drastic changes to the infrastructure and thus shortens deployment time, an assumption which has always been considered in previous publications on AI adoption (Hassija, 2024).

The Current Research focuses on the analysis of the hybrid RF-LSTM model that combines radiological features extraction using the Random Forests model with long-short term memory (LSTM) neural network predictions. The results obtained in such a manner are optimal because they offer a balance between predictive accuracy, feasibility of operation, and interpretability. In the context of last-mile delivery, the model could have brought the discipline out of its present reactive, inherently risk-prone state into something proactive, data-driven with the capacity to forecast and pre-empt delays before they damage customer satisfaction or service efficiency.

5.6 Limitations

Based on the study, the stated goals were achieved, and the results were strong. There are some limitations in methodology to interpret the findings and apply the model. At First, real-time traffic information is not available in the dataset. Despite traffic density being also considered a factor that is usually historically or static in nature, the dynamic nature of congestion should indicate that further real-time traffic updates would significantly enhance the accuracy of predictions. Without these live feeds the model can either underestimate or overestimate the probability of delay that occurs during non-standard traffic conditions in the event that the delivery is done during such times.

Another key drawback is the time modelling range of the LSTM part. The order_time_min and pickup_time_min sequential features were timewise useful but formed an otherwise quite a narrow window of time. The model therefore lacked the ability to document a longer term delivery trend such as delays due to initiation the supply chain or chronic lay-offs along the route due to over-all route ineffectiveness at multiple destinations.

Interpretability, its essence being endlessly examined with regard to the component Random Forest, still lacks completion in relation to the branch LSTM. By being sequential in nature, recurrent architectures have similarly low, intrinsic transparency. No similar efforts have been made to achieve the same success of SHAP in tree-based outputs regarding comparable tools in sequence models, which remain in active development. Subsequently, end-to-end interpretability of the suggested hybrid architecture was not introduced.

Finally, the analysis in this research was based on the historical past delivery data collected from a single operational context. This design ensured that data quality was consistent, but left inbuilt in this design a limitation to the external validity of the findings to other areas, markets or delivery consistent but will hence require future experimentation in various operational settings in order to show how robust the hybrid RF+LSTM architecture will be across different diverse settings.

5.7 Future Work

Future empirical study is justified to increase the predictive efficiency, as well as the useful applicability of the RF LSTM model, by increasing its dependence on dynamic and context-sensitive inputs. The most exciting direction is the parallel inclusion of dynamic traffic, and weather information feeds into the feature set of the model. IP-based GPS traffic monitoring would provide live traffic information on a constant basis, with live weather conditions reporting the instantaneous changes of weather or a temperature fluctuation caused by a storm. As the literature indicates, the implementation of the real-time data integration into transportation-related forecasting may considerably enhance its performance (Tadayonrad, 2023).

The overall improvement of the temporal dimension of the LSTM component in a predictive system can be expected to produce a very positive result. Adding delay-trend information across past historical periods, multi-stop sequence regularities, and durations of deliveries at every working phase will allow the modelling framework to embody a phenomenon that acts within more extended repeated frequencies of occurrence (Keung, 2021). Along with a change in the structure of the LSTM-based network, it is reasonable to adopt sequence models that are based on Transformer, as these have proven their value in terms of dealing with long-range temporal dependence as well as with the variable-length sequence of input in both time-series forecasting and natural language processing (Viellechner, 2020).

Another outstanding trend pertains to the meaning of deep learning modules. Despite SHAP provides clarity of the elements of Random Forest, similar clear insight is not reached in the context of sequence models. Layer-wise relevance propagation or attention visualisation may supply barriers with a better understanding of how sequential patterns are used to make predictions (Hassija, 2024).

The robustness of the model needs to be checked before, due to the variation that may exist in markets and the environment of operation, piloting using the model in reality of logistics situations needs to be conducted before a full-scale use can be implemented. Applying it to multi-modal chains combining road-rail-air transport of goods would show responsiveness and strategic potential more fully (Al-saghir, 2022).

5.8 Conclusion

This paper aimed to solve the issue of forecasting delivery delays in last-mile logistics by combining structured information about tasks (operational data), time dependencies, and surrounding climatic conditions in a hybrid RF-LSTM model. By creating and testing six models, including the classic approaches to machine learning, deep learning networks, and ensemble models, the study confirmed the accuracy, precision, and recall rates and the F1-score as the proposed hybrid model outperformed similar models throughout. These findings align with the increasing body of research indicating how the hybrid models can take advantage of the complementary capabilities of various learning paradigms and provide better predictive performance (Keung, 2021).

The study further contributes to the existing body of knowledge by combining SHAP analysis with a Random Forest classifier to make the decision-making processes within the model transparent. The fact that SHAP has been applied to the Random Forest constituent allows a clear definition of the role of variables, namely traffic density, delivery distance, driver characteristics, and weather shaping the prediction of the model. This level of clarity is widely recognized as of significant intellectual and practical concern in the sense that many applications of artificial intelligence produce outputs on which little explanation is given, thus constraining confidence and hampering decision-making based on understanding. This aspect has been overcome in the present work by explaining variables that have had the most influence

thus making interpretation of the results of the model a lot easier. These findings correspond to earlier studies examining the main factors influencing supply-chain performance (Zhang-TU, 2022).

The implementation of high-resolution visual crossing weather data showed that it was possible to benefit the operational data sets using external sources, despite the features in question having modest effects relative to those of core operational variables. With a practical perspective, the hybrid RF-LSTM solution can provide a deployable, adaptable and transparent solution to aid proactive decision-making in logistics.

To Conclude, this study does not only contribute to academic knowledge on hybrid predictive modelling in last-mile delivery but also gives a practical framework that can be customized and upgraded to multiple operating environments.

References

Alessandria, G. K. S. Y. K. A. M. C. &. R. K. J., 2023. The aggreagate effects of global and local supply chain disruptions:2020-2022. *Journal of international Economics, 146, 103788.,* p. 146.

Al-saghir, R., 2022. Predicting Delays in the Supply Chain with the Use of machine learning. p. 68.

Bassiouni, M. C. R. H. O. a. R. H., 2023. Advanced deep learnign approaches to predict supply chain risks under COVID-19 restrictions. *Expert Systems with Applications,* 2(11), p. 118604.

Bhargava, A. B. D. K. P. S. G. a. R. S., 2022. Industrial IOT and AI Implementation in vehicular logistics and supply chain management. *International Journal of system Assurance Engineering and Management,* 1(13), pp. 673-680.

Chauhan, A. K. H. Y. S. a. J. S., 2020. A hybrid model for investigating and selecting a sustainable supply chain for agri-produce in India.. *Annals of Operations Research,* Volume 1, pp. 621-642.

Fancello, G. P. C. P. M. S. P. Z. P. a. F. P., 2011. Prediction of arrival times and human resources allocation for container terminal.. *Maritime Economics & Logistics,* Volume 13, pp. 142-173.

Ge, H. N. J. G. R. G. S. a. H. Y., 2016. Supply chain complexity and risk mitigation–A hybrid optimization–simulation model.. *International Journal of Production Economics,* Volume 179, pp. 228-238.

Hassija, V. C. V. M. A. S. A. G. D. H. K. S. S. S. I. M. M. a. H. A., 2024. Interpreting black-box models: a review on explainable artificial intelligence.. *Cognitive Computation,* 1(16), pp. pp.45-74..

Jahin, M. S. A. a. A. M., 2025. MCDFN: supply chain demand forecasting via an explainable multi-channel data fusion network model. *Evolutionary Intelligence,* 3(18), pp. 1-27.

Jauhar, S. H. S. K. V. a. P. S., 2024. Explainable artificial intelligence to improve the resilience of perishable product supply chains by leveraging customer characteristics. *operations research,* pp. 1-40.

Jonquais & Krempl, 2019. Predicting shipping time with machine learning.. pp. 40-45.

Keung, K. L. C. a. Y. Y., 2021. A machine learning predictive model for shipment delay and demand forecasting. *IEEE Int. Conf. Industrial Engineering and Management,* pp. 50-70.

Kim, B. a. N. I., 2025. A Review of Hybrid LSTM Models in Smart Cities.. *Processes,* Issue 13, pp. 22-48.

Manai, n., 2018. Delays Prediction using data mining techniques for supply chain risk management company. pp. 30-45.

Rabelo, L. S. A. H. M. a. J. A., 2015. Supply chain and hybrid simulation in the hierarchical enterprise.. *International Journal of Computer Integrated Manufacturing,* 5(28), pp. 488-500.

Sadeghi, K. O. D. K. P. M. R. a. D. A., 2024. Explainable artificial intelligence and agile decision-making in supply chain cyber resilience. *Decision Support Systems,* Volume 1, p. p.114194.

Tadayonrad, Y. a. N. A., 2023. A new key performance indicator model for demand forecasting in inventory management. *Supply Chain Analytics,* Volume 3, pp. 70-80.

Van Wijk, J., 2005. *the value of visualization*. s.l., s.n.

Verma, P., 2024. AI-Driven Predictive Analytics for Supply Chain Risk Management.

Viellechner, A. a. S. S., 2020. *Novel data analytics meets conventional container shipping: predicting delays by comparing various machine learning algorithms.*. s.l., s.n.

Viola, I. a. G. M., 2005. Smart visibility in visualization.. pp. 209-216.

Vummadi, J. R. &. H. K., 2024. Integration of emerging technologies AI and ML into strategic supply chain planning processes to enhance decision-making and agility. *International Journal of Supply chain management,* 2(9), pp. 77-87.

Wang, C. C. C. C. H. &. T. A. J., 2021. On the application of ARIMA and LSTM to predict order demand based on short lead time and on-time delivery requirements. Processes, .. 2(7), pp. 40-57.

Yesodha, K. J. A. a. L. J., 2023. IoT applications in Modern Supply Chains. *IEEE ICIDeA,* pp. 366-371.

Yucekaya, ,. B. K. E. K. G. Ü. K. A. a. H., 2024. Real-Time Prediction of Delivery Delay in Supply Chains using Machine Learning Approaches..

Zhang-TU, Y. B.-T. I. a. B.-C. L., 2022. Predicting Parcel Loss in the Last-Mile Delivery.