Vol. 06, No. 02, pp. 116 –123 (2025)
ISSN: 2708-0757

JASTT

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

OptiShip: Predictive Freight Cost Modeling Incorporating Regional Fuel Variability

Aye Thiri Nyunt, Brij Kotak, Ravi Chauhan, Rituraj Jain*, Vedant Keshariya

Department of Information Technology, Marwadi University, Rajkot, Gujarat, India,
ayethirinyunt3011@gmail.com, brijkotak20@gmail.com, ravirajchauhan2112@gmail.com, jainrituraj@yahoo.com, keshariyavedant0@gmail.com

**Correspondence: jainrituraj@yahoo.com*

Abstract

Properly estimating the cost of freight transportation has always been a challenge to the logistics sector, especially when considered in a region like India, where some factors are dynamic, such as the price of diesel, which keeps on fluctuating and the variable nature of delivery costs to be accounted for. The traditional estimation methods tend to be based on either contractual or stationary models, which do not capture spatial and time eras of fuel prices, and as such, they cannot cope with reality in the real world. The proposed study presents OptiShip, machine learning-based infrastructure that could provide realistic and situationally aware freight cost forecasts. The major goal is to improve the accuracy of cost estimation, which involves incorporating regional avenues of fuel price variability alongside the fundamental logistical components, namely distance and delivery time. The framework is based on a five-phase approach, which includes the data gathering, preprocessing, and pre-alignment of the diesel prices based on the spatial locations in terms of a KDTree model, training and analyzing of the models. To model the non-linear, multidimensional components of the relationship between the input features, three algorithms of ensemble learning, Random Forest, Gradient Boosting, and XGBoost are used. The hyperparameter tuning with the help of GridSearchCV and the evaluation of the performance of the models are performed through R2, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) scores. The Random Forest Regressor is found to be the best according to experimental results, having an R2 value of 0.97, an RMSE of 12.69 and an MAE of 4.94 which proves that it can model the actual representations of the highly accurate real-world logistic situation. The results indicate that regional economic indicators can be very useful to integrate into cost forecasting models and indicate that OptiShip could also be employed in real-time logistics platforms.

Keywords: *Ensemble models, freight cost, fuel pricing, logistics, machine learning*

Received: June 01st, 2025 / Revised: July 18th, 2025 / Accepted: July 20th, 2025 / Online: July 22nd, 2025

I. INTRODUCTION

Logistics industry is an essential industry that exists in any country and plays an essential role in sustaining the economic infrastructure of the country. Freight transport on the other hand, is the very spine of the economic activity in developing nations India, connecting manufacturers and suppliers across vast and disparate geographies with their consumers. On the other hand, optimizing the freight costs is still an ongoing challenge because there are various dynamic factors including the change of the fuel prices, change of delivery duration, or change of transportation routes. Currently, freight cost estimation is based on manual calculation or fixed rate contract which does not take into effect real time economic variable and operational variables. Logistics providers and businesses seldom enjoy suboptimal budgeting, improper resource allocation and financial inefficiencies from the use of these old approaches.

Data driven method to the future success and game changing for accurate freight cost prediction is now finding itself with the ever-growing rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML). These challenges are exactly what ML models can alleviate, as they are particularly good at detecting complicated, nonlinear relationship between the factors influencing shipping cost which include distance, delivery time, regional fuel price, and geographic location.

doi: 10.38094/jastt62244

Rapidly moving industry of logistics and supply chain industry is Artificial Intelligence and freight cost optimization is one of the innovations in the area. AI enables the organizations to enhance planning, optimize route selection, optimize resources, and better delivery effectiveness [1]. Businesses use AI algorithms to process and analyze huge amounts of historical and real time data, in order to understand customer behaviors, the demand trends and also other operational factors to know which customers are actually buying and for which regions. Therefore, more accurate forecasting and optimal inventory management [1,2] is enabled.

Although AI's impacts on transportation and logistics do not always pertain to automation or road safety, they affect many areas of the supply chain. This includes planning and scheduling, execution and monitoring, of the whole logistics chain [3]. Both machine learning and neural network techniques are being more and more used to optimize energy consumption in the transportation sectors and turn to significant cost saving and environmental benefits [4].

In addition, AI based logistics management can significantly improve the operational effectiveness, resource allocation, AI and data driven decision making, and the entire end to end supply chain management [5]. In this case, the integration of AI into the freight cost estimation can be said to be a huge step forward moving away from the traditional method of using human intelligence to manage freight cost and supply chain control. As such, the use of AI in freight cost optimization is turning transportation logistics into a smarter, faster and lower cost decision making science. AI technologies are showing to be crucial tools in the current supply chain strategy as they have the ability to minimize the impact on environment, minimize the operational expenses, and increase competitive positioning [4,6,7]. With AI becoming more and more established in logistics, it is crucial for logistics enterprises to spend money into these technologies to exploit their true potential and compete effectively in the ever more perplexed and dragged global market.

Although freight cost modeling has developed considerably in recent years, most of the available methods fail to suitably accommodate the reality of logistics in a country like India. Traditional models usually assume constant fuel prices or operate with average nationwide values without considering the large regions and time variations which have a direct effect on the accuracy of estimating costs. Additionally, these models have been found to be simplistic in their illumination of linear assumptions that fail to capture complex interdependencies of such factors as distance, delivery time and behavior of fuel consumption.

The suggested OptiShip framework answers these deficiencies by presenting a spatially conscious and machine learning-based pipeline that is responsive to the current fuel price conditions and logistical changes. It uses geographic alignment based on KDTree to match the contextual prices of diesel and applies effective ensembles of regression models to discover complicated, non-linear trends of cost-driving factors. As far as to illustrate the ability of OptiShip framework to address some crucial gaps in the literature, Table I provides the qualitative comparison between the existing model limitations and the areas of innovations developed in our framework.

The main contributions of this study are the following:

- A predictive framework that estimates the cost of freight, OptiShip, is introduced where the variability in the diesel price in a particular region, combined with core logistical factors can be embedded to provide better contextual booking.
- A dynamic trip origin association method on the basis of the KDTree algorithm is used to associate trip origin with region-specific fuel pricing data on the basis of geographic and temporal proximity.
- The Random Forest, Gradient Boosting model and XGBoost ensemble machine learning models are used to capture non-linear relationships between distance, delivery time, and fuel prices.
- Full hyperparameter tuning has been done on the model selection process by the usage of the GridSearchCV, and the evaluation shows that Random Forest Regressor is the best performing model with R^2 of 0.97, RMSE of 12.69, and MAE of 4.94.
- It has a framework that can be deployed in the real-time by using containerized APIs and lightweight model inference pipelines and is able to integrate with logistics dashboards and ERP systems.

In this paper, we introduce OptiShip, an AI based predictive system implemented for freight cost prediction and powered by the ensemble-based machine learning models, Random Forest, Gradient Boosting, and XGBoost that are able to accurate predict freight cost. What is notable about this system is the integration of real-world diesel price data via a KDTree based spatial mapping algorithm, allowing fuel costs to be really tightly tied to trip origins coordinates. Real trip data of Tripura, Gujarat and Maharashtra are sourced to train and validate the model.

The remainder of this paper is structured as follows. Section 2 presents a review of related literature and identifies key research gaps in current freight cost estimation methods. Section 3 describes the methodology adopted, including data preprocessing, feature engineering, and model configuration. Section 4 outlines the experimental results and model performance evaluation. Section 5 highlights computational feasibility, deployment considerations, and a roadmap for future enhancements and section 6 specifies roadmap for the future enhancements. Finally, Section 7 concludes the paper with key findings.

II. Literature Review

Freight transport is a primary input for logistics industry that, in its turn, is one of the significant components of the economic infrastructure and a uniting force of supply chains. Considering all the factors influencing the industry, it is a big issue in developing nations to optimize freight costs [8]. The logistics industry is adopting green freight practices in order to reduce environmental concerns and increase their image in the eyes of buyers. The issue with this, however, is that implementation presents barriers in developing countries. Particularly, those barriers are important societal and managerial barriers of the third world nations that need more attention from the policymakers and industry managers to develop efficient strategies for green freight implementation [9]. The new logic of the logistics is evolving on the basis of innovations achieved by integrating the advanced technologies and data-driven approaches. Thus, the transportation and logistics has applied machine learning algorithms that enabled optimizing routes, forecasting demand as well as developing autonomous vehicles, which result in less operational expenses, moreover improve safety [10]. Moreover, the electric adoption of freight transport can help in minimizing environment impacts and greenhouse gas emissions. For efficient transition towards electric freight vehicles, optimal placement of charging infrastructure from travel and parking patterns is important [11].

AI is disrupting the shipping cost estimation and freight cost prediction in logistics and supply chain management marketplace. There are advanced technologies that improve the accuracy and efficiency by several orders of magnitude compared to the traditional techniques. In fuel consumption cost prediction for shipping companies, machine learning algorithm, CatBoost, has proved to be very effective. CatBoost algorithm achieved 0.976 R2 value in the study of a large PCTC shipping company in South Korea, which scores 0.976 on top of 18 other [12]. Also examined in the model were ship size, route, distance, speed, sea day, port call day and duration. This approach can assist shipping companies in optimally estimating the fuel consumption costs and meeting environmental requirements, so as to improve the operation efficiency. However, interestingly, while AI based methods tend to be a better solution for freight rate forecasting, the econometric model Auto-Regressive Integrated Moving Average (ARIMA) still outperforms for demand prognosis [13].

In shipping cost estimation and cost of freight managing in the field of logistics and supply chain operations, AI has made a complete revolution. With the help of advanced algorithms and machine learning techniques, businesses are now able to predict and optimize the shipping costs better [1,6]. These AI powered systems take in data from vast amounts of historical data, real time market conditions, and dozens of other factors affecting shipping costs to do that. These help companies figure out what is the best decision, reduce the overall transportation cost and reduce time and distance between the vendors and the company. By deducting machine learning algorithms to analyze patterns in ship data, shipping data can help make proactive decisions and risk management in the freight cost estimate [1,14]. In particular, although shipping autonomic does exist in developed countries such as the US, less advanced countries tend to only adopt it more slowly because of their infrastructure limitations. Nevertheless, there have been attempts to address these challenges using innovative approaches based upon mobile technologies, as well as based on cloud-based solutions [15]. AI integration to shipping cost estimation and freight cost management has made shipping costing and freight costing operationally much more efficient and cost effective. AI is changing the way businesses ship the costs by predicting accurately, optimizing the routes and data-driven making decision, overall makes businesses more profitable and competitive in the global shipping industry [14,16].

Furthermore, to strengthen the technical foundation of ensemble learning in logistics, recent work by [17] provides an in-depth overview of modern ML techniques, applications, and trends which align with our ensemble-based model strategy. Additionally, Asaad et al. [18] demonstrated a hybrid deployment model (Wi-Lo), showing how location-aware solutions can improve contextual predictions paralleling our use of KDTree-based spatial alignment in OptiShip.

Recent advances in machine learning, particularly ensemble-based techniques, have demonstrated strong potential for solving non-linear prediction problems across domains. Abdullah et al. [19] present a comprehensive analysis of modern ML approaches, emphasizing the suitability of models like Random Forest and XGBoost for applications requiring interpretability and robust performance. Their findings support the adoption of ensemble learning in real-world systems, reinforcing the design choices made in the OptiShip framework, which leverages these models for accurate and scalable freight cost estimation.

To summarize the key challenges identified from existing literature and to position our proposed solution effectively, Table I outlines the major research gaps and how the OptiShip framework addresses them through its design and implementation.

TABLE I. RESEARCH GAPS IN FREIGHT COST ESTIMATION AND CORRESPONDING SOLUTIONS IN OPTISHIP FRAMEWORK

Research Gap Identified	Resolution by OptiShip Framework
Conventional models do not account for regional and time-based variations in diesel fuel prices, leading to poor cost accuracy.	OptiShip incorporates a KDTree-based spatial mapping technique that aligns diesel prices with trip origins and dates to enhance prediction accuracy.
Freight cost estimation often ignores the complex and nonlinear relationships among key factors such as distance, delivery time, and fuel prices.	The framework uses ensemble machine learning models Random Forest, Gradient Boosting, and XGBoost to capture and learn intricate dependencies in the data.
Manual or fixed-rate cost estimation lacks flexibility and fails to adapt to real-time logistical and economic conditions.	OptiShip provides a fully automated, AI-driven prediction model that dynamically integrates current fuel prices and trip parameters.
Existing models rarely incorporate actual region-specific diesel price data, especially for diverse geographies like India.	The system uses real-world logistics and diesel pricing datasets from multiple Indian states, making the model geographically and economically contextual.
Freight prediction tools are often not optimized for the variability and uncertainty typical in developing nations' logistics sectors.	OptiShip is specifically tailored for high-variability scenarios, with extensive preprocessing and hyperparameter tuning for improved performance.

III. Methodology

This study employs a structured five-phase pipeline methodology, meticulously following the established machine learning (ML) modeling pipeline to ensure the accurate prediction of freight costs using ML models. As depicted in Figure 1, the process initiates with Phase 1: Data Preparation, involving crucial steps such as loading and preprocessing logistics and diesel price datasets, converting timestamps to standardized datetime formats, and executing KDTree-based nearest neighbor searches to accurately attribute fuel prices to specific trip data. Subsequently, Phase 2 focuses on Model Setup, where the dataset is carefully separated into an 80% training set and a 20% test set, features are explicitly defined, and hyperparameter optimization is conducted using GridSearchCV. During Phase 3 which is Model Development, the setup of a strong ensemble of models, such as Random Forest, Gradient Boosting and XGBoost, occurs and their performance is extensively measured using regression methodologies involving MSE, RMSE and R² score. It also includes how to select the most appropriate resources (e.g hyperparameters) and continuing to develop other models if necessary. In Phase 4: Evaluation and Refinement, overfitting is addressed, and mitigated with adjustments made on a model by model basis, for instance, via decreasing tree depth or increasing training samples. Then, Phase 5: Finalization will deploy the best model onto real world use. This is the stepwise approach of making a robust, scalable, and accurate freight cost predicting system.

A. Dataset and Preprocessing

As a basis of this predictive modeling, there are two distinctive yet closely related datasets, the logistics trip dataset [20] and the diesel price dataset [21]. Also, it has two crucial columns in case of India showing regions and collection dates from different states, respectively. It is in the synergistic combination of spatial and economic variables in these datasets that the datasets can be seamlessly incorporated in a coherent machine learning framework for freight cost prediction.

The logistics dataset consists of the historical freight trip records, each record has a variety of important data such as total distance travelled, hours spent in transportation, shipper and the consignee, and finally the cost related to the freight. The main pillar on which these predictive models are built upon are these comprehensive features. At the same time, the diesel dataset contains daily diesel prices observed in many cities and towns in India. This dataset consists of different record of every city name, its exact geographic coordinates (latitude and longitude), exact diesel price (in INR per liter) and date on the diesel price was recorded.

Aligning diesel price data together with the logistics dataset as a key preprocessing step was important. I was able to achieve this using a fancy spatial nearest neighbor search technique which employed KDTree. As a perfect chronological analysis about the geographical coordinates of origin of every logistics trip. Thereafter the KDTree algorithm was used to find the closest city in its diesel dataset with both geographical proximity and temporal correspondence in the date of travel to the trip. In addition to this, an accurately assigned corresponding diesel price from the identified location and date, which was then another vital feature in this logistics trip, then was assumed to be given. It is the spatial integration within this model that enables the model to integrate region specific diesel costs hence making the freight cost predictions very contextual. Derivatives of diesel prices have the significant impact on the freight transportation costs and therefore directly affect the firm profitability. Through incorporation of historical diesel price data into the cost prediction model through this KDTree based spatial mapping, the model becomes more proficient in estimating expenses associated with diesel trips at a more fine grain than before. It take into account local and temporal variations in diesel purchase price, permitting higher precision of overall freight cost forecasting and the capacity to use such information for more strategic planning and budgeting of logistics operations in areas of volatile diesel prices.

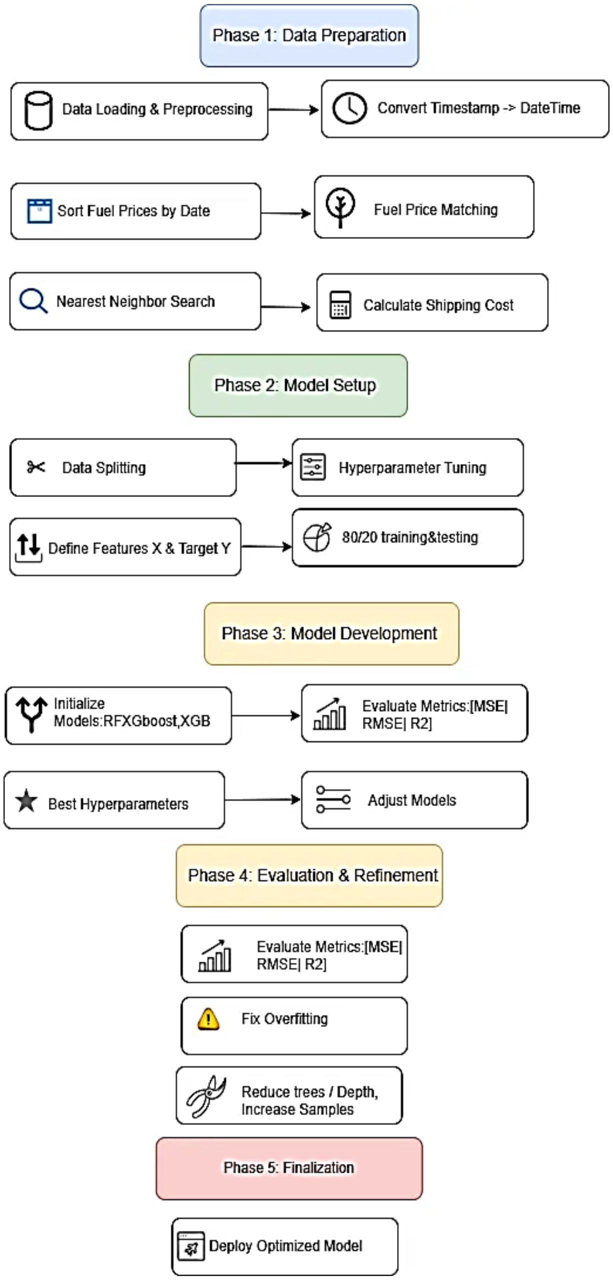
Phase 1: Data Preparation
Data Loading & Preprocessing
Convert Timestamp -> DateTime
Sort Fuel Prices by Date
Fuel Price Matching
Nearest Neighbor Search
Calculate Shipping Cost
Phase 2: Model Setup
Data Splitting
Hyperparameter Tuning
Define Features X & Target Y
80/20 training&testing
Phase 3: Model Development
Initialize Models:RFXGboost,XGB
Evaluate Metrics:[MSE| RMSE| R2]
Best Hyperparameters
Adjust Models
Phase 4: Evaluation & Refinement
Evaluate Metrics:[MSE| RMSE| R2]
Fix Overfitting
Reduce trees / Depth, Increase Samples
Phase 5: Finalization
Deploy Optimized Model

Fig. 1. Block diagram of proposed system.

Data preprocessing was done so that features of both the datasets were obtained in the desired format to feed the machine learning algorithms. The overall preprocessing dealt with in this work included different ways of imputing missing variables by removal if possible or suitable imputation version where applicable. Besides, location names were encoded in accordance to appropriate encoding schemes to transform categorical variables into numeric representations. For the last, feature values were normalized so that all the variables would influence the model's learning process in a similar manner and thus none of the variables with a large scale will dominate the predictions. Upon completion of this preprocessing phase, the final set of features fed into the model comprised the travel distance, delivery time, and the diesel price meticulously mapped using the KDTree algorithm. Consequently, the total shipping cost was designated as the target variable for our prediction task. To prevent the model from being trained on data it would later be tested against, thereby ensuring an unbiased evaluation of its generalization capabilities, the dataset was rigorously split into an 80:20 ratio for training and testing, respectively. Table II provides a summary of the key features utilized in the dataset, along with a brief description of each.

B. Model Selection and Configuration

This study strategically employs three ensemble-based supervised machine learning models to accurately estimate freight costs: Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor. These models are particularly well-suited for problems involving high-dimensional, non-linear data and excel in regression tasks that necessitate inferring complex relationships among variables, which is crucial for predicting logistics costs effectively.

TABLE II. DATASET FEATURE DESCRIPTION

Feature Name	Type	Description
Distance(km)	Numerical	Total distance in the freight trip
Delivery Time (days)	Numerical	Number of days taken for the shipment to reach its destination
From Location	Categorical	Origin city/state of the shipment
To Location	Categorical	Destination city/state of the shipment
Diesel Price (INR)	Numerical	Mapped diesel price (via KDTree) for the origin location and trip date
Trip Cost (INR)	Numerical	Actual cost incurred for the freight trip (target variable)

The Random Forest Regressor is an ensemble learning technique that constructs a multitude of decision trees based on random sub-samples of the training data. These trees are built independently, and their individual predictions are aggregated (e.g., averaged for regression tasks) to produce a final, more robust prediction. This aggregation process significantly reduces the risk of overfitting and substantially improves the model's generalization capabilities, essentially achieving a balance between bias and variance reduction.

On the contrary, Gradient Boosting models usually construct a series of decision trees. The new tree built is a so as to minimize the errors (residuals) made by the tree (one before) in the ensemble. With every step, the model is able to learn from its mistakes, and capture more subtle and more complex patterns of the data. Gradient Boosting brings out its strength by forming an additive model which learns about intricate patterns sewn into the data, and makes highly accurate predictions.

Extreme Gradient Boosting (XGBoost) ... this is an optimized and highly efficient implementation of gradient boosting. Besides that, it embraces diverse optimization and regularization methods (L1 and L2 regularization) to avoid overfitting and improve its training speed and model performance. Moreover, as XGBoost is capable of handling sparse data, it is much more effective on big and noisy datasets. Being a powerful architecture, it is a good choice for solving complex regression problems such as freight cost prediction.

Hyperparameter tuning was done systematically complete to find the optimum setting for it using GridSearchCV so that we are able to achieve the best possible model performance. It exhaustively tries each set of hyperparameters for the given model using a predefined set of hyperparameters, trying each option and checking which gives the highest performance with respect to the chosen metric, e.g. across validated R² score. Using the models, key parameters (number of estimators (number of trees), learning rate (step size shrinkage to prevent overfitting in boosting models), maximum tree depth (to reduce variance and help control complexity) and subsample ratios (fractions of samples used to train each tree), were systematically evaluated. The procedure to meticulously tune this model was fruitful in selecting the best configuration for each model and in reducing prediction error by also serving as a guard against overfitting.

Table III gives the model-specific hyperparameter choices of each model, including Random Forest, and Gradient Boosting, and XGBoost. Such setups were central to the improved predictive capacity of the OptiShip framework as they strengthened its real, practical robustness, as well as increased its flexibility in interacting with real freight data.

TABLE III. HYPERPARAMETER SEARCH SPACE AND FINAL SELECTED VALUES

Model	Parameter	Search Range	Selected Value
Random Forest	n_estimators	[100, 200, 300]	200
	max_depth	[4, 6, 8, None]	8
	min_samples_split	[2, 5, 10]	5
	min_samples_leaf	[1, 2, 4]	2
Gradient Boosting	n_estimators	[100, 200, 300]	200
	learning_rate	[0.01, 0.05, 0.1]	0.05
	max_depth	[3, 5, 7]	5
	subsample	[0.6, 0.8, 1.0]	0.8
XGBoost	n_estimators	[100, 200, 300]	200
	learning_rate	[0.01, 0.05, 0.1]	0.05
	max_depth	[3, 5, 7]	5
	subsample	[0.6, 0.8, 1.0]	0.8
	colsample_bytree	[0.6, 0.8, 1.0]	0.8

The selected models were fed with three main inputs namely "Distance (km)", "Delivery Time (days)" and the "Mapped Diesel Price (INR)". The "Trip Cost (INR)" was our target variable which our models were trained to predict. As it was stated before, the entire dataset was divided by us meticulously into an 80:20 ratio, in order to have training and testing sets of appropriate size. This segregation shields the model from being handed over seen data, so that the performance of the model can hence be tested on unseen data for an unbiased assessment on the generalization capabilities. For effective evaluation of the performance of each model on the test set, standard regression performance metrics were used. They included Coefficient of Determination (R^2), which represents the proportion of the variance in the dependent variable that is explained by the independent variables, MAE that measures the average magnitude of the errors without regard of their direction and RMSE, which gives the measure of the error in the same units of the target variable, thus, it is more interpretable than MSE, and is more sensitive to large errors. Taken together, these metrics give an overall view of how accurate the model is and how the distribution of the prediction errors.

IV. Result And Discussion

We evaluate the predictive performance of the ensemble learning models by comparison of their R^2 scores, RMSE, MAE among others using the cross validation. Here, Table IV displays, that the Random Forest Regressor had the highest R^2 score of 0.97 and this showed the relationship between actual and predicted freight costs was strong. The ensembling property provides it a better capability to mitigate the overfitting and effectively model complex, nonlinear relationships, resulting in a superior performance.

TABLE IV. Model Performance Comparison

Model	R^2 Score	RMSE	MAE
Random Forest	0.97	12.69	4.94
Gradient Boosting	0.96	66.57	44.92
XGBoost	0.96	96.02	66.02

Gradient Boosting and XGBoost also did very well, achieving an R^2 score of 0.96 as well. However, their error values were way higher, especially on XGBoost, which had RMSE of 96.02 and MAE of 66.02, whereas Gradient Boosting yielded RMSE of 66.57 and MAE of 44.92. Although R^2 values are similar, the higher error values show that these models provide less consistent prediction accuracy than Random Forest. Random Forest model outperforms others for all the metrics as it is shown in Figure 2. This helps further reinforce its robustness in capturing the spatial and economic process embedded in multi-faceted cost dynamics, including regional characteristics, namely, diesel price. RMSE and MAE do not suffice for extended analysis of error distribution and model reliability other than correlation.

A polar comparison of RMSE and MAE values of three ensemble learning models: Random Forest, Gradient Boosting, and XG Boost, is given in Figure 3. For example, radial plots of two different bars in dark blue for RMSE and dark orange for MAE for each model for easy visualization. As can be seen from the chart, Random Forest is best among the rest, attaining the lowest error values in both metrics. XGBoost has the highest RMSE and MAE which means it has lower predictive precision but still very high R^2 score while Gradient Boosting ranks second. As the polar layout efficiently represents the magnitude of and comparative performance among each of the models in a visually compact and straightforward style, it is also highly appropriate for this use case.

The experimental findings demonstrate the importance of integrating real-world spatial and economic features, such as region-specific diesel prices mapped via KDTree, to improve freight cost estimation accuracy. The Random Forest model's capacity to adapt to the heterogeneity of the dataset comprising multiple geographic zones, diesel price fluctuations, and varying travel distances makes it an ideal choice for such a contextualized prediction task.

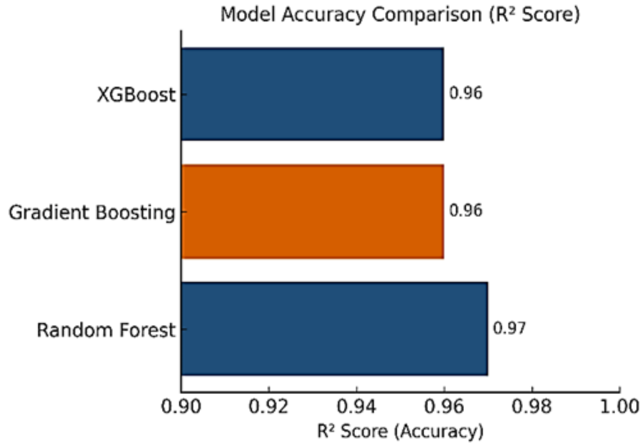

Fig. 1. Comparison of the used model's accuracy

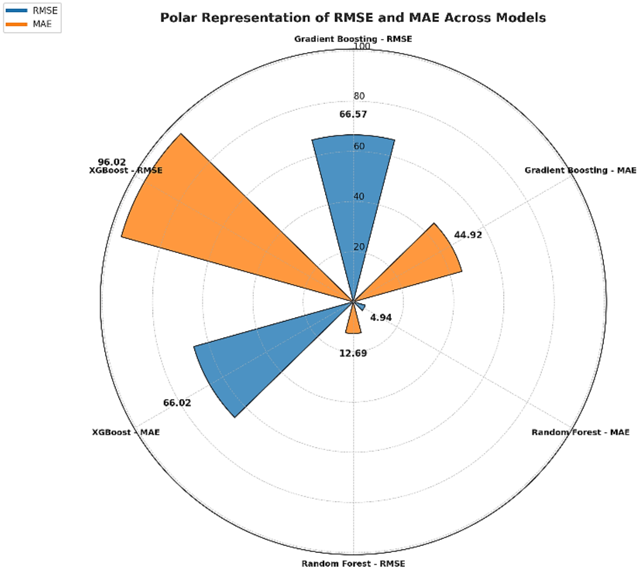

Fig. 2. Comparison of model performance metrics (RMSE and MAE) for Random Forest, Gradient Boosting, and XGBoost. Each axis represents the magnitude of error; smaller radii indicate better performance. Random Forest shows superior accuracy across all metrics.

To bridge the existing research gaps, Table V systematically links previously identified issues in freight cost estimation with the contributions of the OptiShip framework. For example, the integration of regionally mapped diesel prices through KDTree addresses the lack of geographical contextualization in traditional models. Furthermore, the use of ensemble learning methods accounts for complex interactions between logistical variables, which were often oversimplified in earlier rule-based or linear models.

In conclusion, these results substantiate the robustness, contextual adaptability, and high predictive performance of the OptiShip framework, particularly with the Random Forest Regressor at its core. Future work may involve further interpretability techniques such as SHAP analysis and broader validation across diverse logistics scenarios, including multimodal transport networks.

TABLE V. FULFILLMENT OF IDENTIFIED RESEARCH GAPS THROUGH IMPLEMENTATION AND OUTCOMES IN THE OPTISHIP FRAMEWORK

Gap Addressed	Implemented Feature	Outcome/Validation
Lack of regional fuel price integration	KDTree-based diesel price mapping	Enhanced contextual accuracy; improved prediction metrics ($R^2 = 0.97$, MAE = 4.94)
Failure to capture nonlinear dependencies	Ensemble ML models (Random Forest, XGBoost, Gradient Boosting)	Accurately modeled complex patterns in freight data; superior performance compared to other models
Use of outdated/manual estimation methods	Automated ML pipeline using real-world logistics data	Replaced static rate methods with dynamic, data-driven predictions
Absence of region-specific dataset usage	Real logistics data from Tripura, Gujarat, Maharashtra	Regional diversity included in training; improved generalization and relevance
Inadequate tuning and adaptation for high-variability environments	Hyperparameter tuning via GridSearchCV; overfitting mitigation	Optimal model configuration achieved; reliable performance in variable logistics contexts

V. COMPUTATIONAL FEASIBILITY AND DEPLOYMENT CONSIDERATIONS

OptiShip model is characterized with deployment efficiency in mind. Ensemble models like Random Forest and LGBF/XGBoost have an advantage of predictive accuracy and computational time. Benefiting from its low inference latency due to the initial training stage taking less than five minutes (on a mid-range workstation with Intel i7 processor and 16 GB RAM and without GPU acceleration), the model can be used to not only drive real-time prediction pipelines but also implement an OOD detection approach.

To be integrated with operational systems, OptiShip may be containerized (e.g., by using Docker) and deployed in any cloud environment (e.g., AWS, Azure, GCP) or traditionally on infrastructure. The trained models can be made available through lightweight REST APIs through Python frameworks such as Flask or FastAPI and thereby they could be connected directly to logistics dashboards and enterprise resource planning (ERP) systems.

Moreover, due to the small size of the feature space (three fundamental predictors: distance, delivery time, and mapped fuel price), preprocessing and handing real time input is kept at minimal overhead. This lends credence to the fact that the model can be integrated into current supply chain management systems as a microservice that can provide an on-demand freight cost estimate. This type of modular deployment would similarly support future data feed integrations (e.g. toll APIs, weather services) with a small reconfiguration of the system. To conclude, the OptiShip framework is not only time-efficient but also architecturally adaptive, which makes it an appropriate candidate to be deployed in a scalable manner in the real logistics setting.

VI. LIMITATIONS AND ROADMAP AND FUTURE ENHANCEMENTS

While the OptiShip framework demonstrates promising results in freight cost prediction using real-world logistics and diesel pricing data, several limitations should be acknowledged. First, the current study is geographically limited to road freight operations in India, which may affect the generalizability of the model to other countries with different logistics patterns, fuel pricing mechanisms, and regulatory environments. Second, although the model incorporates region-specific fuel prices using KDTree alignment, it does not yet account for other operational variables such as vehicle type, toll costs, or weather conditions factors that can influence freight costs significantly. Third, the study focuses exclusively on structured, tabular data and does not explore advanced deep learning or time-series forecasting methods, which could offer improved performance in highly dynamic environments. Finally, model interpretability tools such as SHAP were not included in this version, limiting transparency into feature-level decision influences. These limitations provide meaningful directions for future work, where model enhancements and broader deployments can be explored.

Looking ahead, several enhancements are envisioned to further advance the capabilities and applicability of the OptiShip framework. One key direction involves integrating vehicle-type information into the prediction model, recognizing that different categories of freight vehicles exhibit distinct fuel consumption patterns and cost dynamics. By accounting for truck-specific attributes, the model's estimations can be made more granular and operationally realistic.

Another important enhancement is the inclusion of toll data and weather conditions through real-time API integration. These variables, which significantly impact transit duration and cost, will enable OptiShip to respond dynamically to real-world disruptions or changes in operating conditions. This addition will improve the robustness and adaptability of the framework under varying logistical scenarios. Furthermore, to ensure sustained model accuracy over time, especially in dynamic economic or seasonal environments, an adaptive retraining mechanism will be implemented. This will allow the system to learn from new data continuously and address concept drift, thereby maintaining its predictive reliability across evolving contexts.

Finally, a web-based deployment dashboard is proposed to bring OptiShip closer to end-users. This interface will serve as a decision-support system, offering intuitive visualizations, key performance indicators, and actionable insights for logistics managers and planners. In parallel, future extensions will also focus on enriching the dataset to better represent varied logistics contexts such as rural last-mile deliveries, cross-border long-haul freight, and multimodal transport. This will improve the generalizability of the model and ensure its applicability across a broader spectrum of real-world freight scenarios. Collectively, these enhancements aim to transform OptiShip into a comprehensive and intelligent freight cost management solution for modern logistics ecosystems.

VII. Conclusion

In this paper, the OptiShip system, a machine learning framework that enables forecasting shipping prices based on spatial and economic characteristics, such as distance, delivery period, and prices of diesel in regions was introduced. Data contextualization was also considerably increased by employing a new method of KDTree-based pricing to align fuel prices to trip origins and date. The ensemble models analyzed were Random Forest, Gradient Boosting, and XGBoost, with Random Forest Regressor performing the best by having a score of 0.97 R^2, which is a sign of high predictive potential in a lot of technical and non-linear logistics. Such findings validate the powerfulness of ensemble learning in the freight cost estimation present in actual world variables. The research only looks at the Indian road freight data that can affect its applicability to other regions or transport modes. Such other factors as the type of vehicle, toll prices, and the weather were not examined, and the approaches to model interpretability, such as SHAP, were not provided. Neural networks in deep learning and temporal forecasting work was also beyond the scope of the work. The future extensions will investigate vehicle-specific attributes, real-time tolls and weather APIs, and adaptive retraining approaches to cover concept drifts. Also, there are plans to give real-time and explainable predictions and aid logistics decision-making in various operational conditions through a web-based dashboard.

References

[1] I. A. Shah, N. Z. Jhanjhi, and S. K. Ray, “Artificial Intelligence Applications in the Context of the Security Framework for the Logistics Industry,” igi global, 2024, pp. 297–316. doi: 10.4018/978-1-6684-6361-1.ch011.

[2] Vandana. M. Dr. Nagaraju Ellaturu Naveena M, and Dr. T. L. K. M. Rajalakshmi Dr. Shweta Bambuwala, “Ai-Driven Solutions for Supply Chain Management,” Journal of Informatics Education and Research, vol. 4, no. 2, May 2024, doi: 10.52783/jier.v4i2.849.

[3] X. Mou, Artificial Intelligence: Investment Trends and Selected Industry Uses. finance corporation washington dc, 2019. doi: 10.1596/32652.

[4] T. Olatunde, Z. Sikhakhane, D. Akande, and A. Okwandu, “Reviewing The Role Of Artificial Intelligence In Energy Efficiency Optimization,” Engineering Science & Technology Journal, vol. 5, no. 4, pp. 1243–1256, Apr. 2024, doi: 10.51594/estj.v5i4.1015.

[5] A. S. Shatat and A. S. Shatat, “Artificial Intelligence Competencies in Logistics Management: An Empirical Insight from Bahrain,” Journal of Information & Knowledge Management, vol. 23, no. 01, Nov. 2023, doi: 10.1142/s0219649223500594.

[6] T. V and D. Krisknakumari, “Artificial Intelligence in Enhancing Operational Efficiency in Logistics and SCM,” International Journal of Scientific Research in Science and Technology, vol. 11, no. 5, pp. 316–323, Oct. 2024, doi: 10.32628/ijsrst24115107.

[7] W. Guo, “Exploring the Value of AI Technology in Optimizing and Implementing Supply Chain Data for Pharmaceutical Companies,” Innovation in Science and Technology, vol. 2, no. 3, pp. 1–6, May 2023, doi: 10.56397/ist.2023.05.01.

[8] R. Siddiqui, “Quantifying the Impact of Development of the Transport Sector in Pakistan,” The Pakistan Development Review, vol. 46, no. 4II, pp. 779–802, Apr. 2024, doi: 10.30541/v46i4iipp.779-802.

[9] S. Singh, S. Luthra, A. Kumar, A. Barve, and K. Muduli, “Evaluating Roadblocks to Implementing a Green Freight Transportation System: An Interval Valued Intuitionistic Fuzzy Digraph Matrix Approach,” IEEE Transactions on Engineering Management, vol. 71, pp. 2758–2771, Jan. 2024, doi: 10.1109/tem.2022.3188643.

[10] N. L. Rane, J. Rane, S. K. Mallick, and Ö. Kaya, “Applications of machine learning in healthcare, finance, agriculture, retail, manufacturing, energy, and transportation: A review,” deep science, 2024. doi: 10.70593/978-81-981271-4-3_6.

[11] J. Fu, H. J. Bhatti, and A. Nåbo, “Locating charging infrastructure for freight transport using multiday travel data,” Transport Policy, vol. 152, pp. 21–28, Apr. 2024, doi: 10.1016/j.tranpol.2024.04.007.

[12] M. Su, H. J. Lee, X. Wang, and S.-H. Bae, “Fuel consumption cost prediction model for ro-ro carriers: a machine learning-based application,” Maritime Policy & Management, vol. 52, no. 2, pp. 229–249, Jan. 2024, doi: 10.1080/03088839.2024.2303120.

[13] E. Liachovičius, E. Šabanovič, and V. Skrickij, “Freight Rate And Demand Forecasting In Road Freight Transportation Using Econometric And Artificial Intelligence Methods,” Transport, vol. 38, no. 4, pp. 231–242, Dec. 2023, doi: 10.3846/transport.2023.20932.

[14] N. L. Rane, M. Paramesha, J. Rane, and P. Desai, “Artificial intelligence, machine learning, and deep learning for sustainable and resilient supply chain and logistics management,” deep science, 2024. doi: 10.70593/978-81-981367-4-9_5.

[15] A. Atadoga, O. Obi, A. Daraojimba, F. Osasona, S. Dawodu, and S. Onwusinkwue, “AI in supply chain optimization: A comparative review of USA and African Trends,” International Journal of Science and Research Archive, vol. 11, no. 1, pp. 896–903, Jan. 2024, doi: 10.30574/ijsra.2024.11.1.0156.

[16] S. Krishnamurthy, K. Tirupati, E. Shrivastav, S. Jain, S. Ganipaneni, and P. Vashishtha, “Leveraging AI and Machine Learning to Optimize Retail Operations and Enhance,” Darpan International Research Analysis, vol. 12, no. 3, pp. 1037–1069, Sep. 2024, doi: 10.36676/dira.v12.i3.140.

[17] A. A. Abdullah, N. S. Mohammed, M. Khanzadi, S. M. Asaad, Z. Kh. Abdul, and H. S. Maghdid, “In-depth analysis on machine learning approaches,” *ARO-The Scientific Journal of Koya University*, vol. 13, no. 1, pp. 190–202, May 2025, doi: 10.14500/aro.12038.

[18] S. M. Asaad, H. S. Maghdid, and Z. Kh. Abdul, “Hybrid positioning technique using the existing Wi-Fi and LoRa technologies (Wi-Lo),” *Expert Systems With Applications*, p. 127127, Mar. 2025, doi: 10.1016/j.eswa.2025.127127.

[19] R. A. Saleh and S. R. M. Zeebaree, “Transforming Enterprise Systems with Cloud, AI, and Digital Marketing,” *International Journal of Mathematics Statistics and Computer Science*, vol. 3, pp. 324–337, Mar. 2025, doi: 10.59543/ijmscs.v3i.13883.

[20] K. Sudhir, “Fuel price in India,” Kaggle, Jan. 27, 2021. https://www.kaggle.com/datasets/sudhirnl7/fuel-price-in-india/data

[21] Devaraj V, “Delhivery Logistics Dataset,” Kaggle, Jun. 03, 2024. https://www.kaggle.com/datasets/devarajv88/delhivery-logistics-dataset