World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

OPTIMIZING SUSTAINABLE SUPPLY CHAIN EFFICIENCY THROUGH DATA VISUALIZATION AND AI-ENHANCED PREDICTIVE ANALYTICS

Kayode L. Ogunsusi *

Department of Mathematics and Statistics, Austin Peay State University, Clarksville, TN 37044, USA.
* Corresponding Author

ORCID Details
Kayode L. Ogunsusi: https://orcid.org/0009-0002-7392-7141

World Journal of Advanced Research and Reviews, 2026, 31(03), 641–656

Publication history: Received on 02 August 2026; revised on 09 September 2026; accepted on 11 September 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.3.2341

ABSTRACT

Sustainable supply-chain management increasingly requires analytical approaches that can translate complex freight data into timely insights for reducing operational inefficiencies and transportation-related greenhouse-gas emissions. This study develops an automated data-analytics framework that integrates freight-data processing, carbon-emission estimation, data visualization, machine learning, and anomaly detection to support sustainability-oriented supply-chain decision-making. A randomly ordered working subset of 250,000 shipment records from the 2022 U.S. Commodity Flow Survey was examined, with 121,801 shipments remaining for the emissions analysis after transportation-mode and data-validity restrictions. Shipment-level greenhouse-gas emissions were estimated using mode-specific factors from the U.S. Environmental Protection Agency's 2025 Greenhouse Gas Emission Factors Hub. A Random Forest classifier was subsequently developed to identify shipments in the upper quartile of estimated carbon emissions using operational characteristics available for decision support. The model achieved an accuracy of 89.7%, recall of 91.5%, F1-score of 81.6%, and ROC-AUC of 0.964. A robustness analysis excluding shipment weight retained strong discriminatory capability, achieving a ROC-AUC of 0.932. Permutation importance identified shipment weight and transportation mode as the most influential predictors of high-carbon risk. Isolation Forest analysis was further applied within transportation modes to identify unusually carbon-inefficient shipment patterns while controlling for inherent differences in modal emission intensity. The results demonstrate how automated visualization and AI-enhanced predictive analytics can support earlier identification of carbon-intensive freight activities and enable stakeholders to concentrate on targeted, data-informed sustainability decisions.

Keywords: Sustainable supply chain, Carbon emissions, Freight transportation, Data visualization, Predictive analytics, Machine learning, Random Forest, Isolation Forest, Supply-chain automation.

1 INTRODUCTION

Supply-chain efficiency has traditionally been evaluated through operational measures such as transportation cost, delivery time, inventory availability, shipment reliability, and resource utilization. Growing emphasis on environmental sustainability, however, has expanded the concept of supply-chain efficiency to include the greenhouse-gas consequences of freight transportation and logistics decisions. Transportation mode, shipment weight, travel distance, commodity characteristics, and operational requirements can substantially alter the environmental burden associated with moving goods. Consequently, improving supply-chain performance increasingly requires analytical systems capable of evaluating operational efficiency and environmental impact together rather than treating carbon emissions as a separate reporting consideration [1], [2].

The increasing availability of high-volume freight and logistics data creates opportunities for more systematic sustainability analysis, but it also introduces challenges for decision-makers. Large datasets may contain millions of shipment observations and numerous operational attributes that are difficult to evaluate effectively through manual processes alone. Data visualization can reduce this complexity by revealing differences in shipment activity, transportation patterns, carbon intensity, and operational behavior. Automation can further reduce repetitive data-processing requirements, allowing supply-chain stakeholders to devote greater attention to interpretation, risk identification, and decision-making. However, descriptive visualization alone is primarily retrospective and may not provide sufficient capability for identifying carbon-intensive shipments before operational decisions are completed [3], [4].

Machine-learning techniques provide an opportunity to extend visualization from descriptive analysis toward predictive decision support. Rather than relying exclusively on historical summaries, predictive models can evaluate combinations of shipment characteristics and estimate whether particular freight movements exhibit elevated carbon-emission risk. Such capability is particularly relevant when transportation planners must evaluate large numbers of shipments and prioritize those requiring intervention. Machine learning can also support anomaly detection by identifying shipment patterns that differ substantially from typical carbon-efficiency behavior. When integrated with automated processing and visualization, these methods can create a more proactive analytical framework for sustainable supply-chain management [3], [5], [6].

This study develops such a framework using publicly available U.S. freight and environmental data. Shipment-level information from the 2022 Commodity Flow Survey Public Use Microdata Sample [7] is combined with mode-specific greenhouse-gas emission factors from the U.S. Environmental Protection Agency's 2025 Greenhouse Gas Emission Factors Hub [8]. Shipment weight and great-circle distance are used to derive ton-mile activity, which is subsequently combined with transportation-mode emission factors to estimate shipment-level carbon dioxide equivalent emissions. Because the Commodity Flow Survey provides great-circle rather than route-specific transportation distance, the resulting estimates are interpreted as consistent distance-based indicators of freight-related emissions rather than reconstructed measurements of the precise emissions produced along individual transportation routes. Similarly, the EPA transportation factors used in the analysis represent transportation combustion emissions rather than comprehensive life-cycle emissions.

The analytical framework extends this emissions-estimation process through data visualization and AI-enhanced predictive analytics. A Random Forest classifier is used to identify shipments falling within the upper quartile of estimated carbon emissions using operational characteristics that could support earlier decision-making. Shipment distance and variables directly constructed from the carbon-emission calculation, including ton-miles and calculated CO_2-equivalent emissions, are excluded from the predictor set to reduce target leakage and avoid an overly deterministic relationship with the outcome. A robustness model additionally removes shipment weight because weight directly contributes to the emissions-estimation equation. The primary model achieves a ROC-AUC of 0.964, while the weight-excluded robustness model retains a ROC-AUC of 0.932, indicating that meaningful predictive information also exists in transportation mode, shipment value, industrial sector, commodity category, geographic characteristics, and other operational attributes. Isolation Forest is subsequently applied separately within major transportation modes to identify unusually carbon-inefficient shipment patterns without allowing structural differences in modal emission intensity to dominate the anomaly-detection process.

This study originated from a supply-chain data-visualization project developed and presented in 2023. The present research substantially expands that earlier work through the integration of updated freight and emissions data, shipment-level carbon assessment, and AI-enhanced predictive analytics for carbon-risk prediction and anomaly detection, transforming the original visualization-focused project into a broader sustainable supply-chain decision-support framework.

Accordingly, the study addresses five research questions: (1) How do freight transportation modes differ in estimated greenhouse-gas emission intensity and shipment-level emissions? (2) Which operational shipment characteristics are most influential in identifying high-carbon-risk freight movements? (3) To what extent can machine learning predict elevated carbon-emission risk from operational shipment information? (4) Does predictive performance remain strong when shipment weight is excluded from the model? and (5) Can within-mode anomaly detection identify unusually carbon-inefficient freight patterns that may warrant additional stakeholder review?

The principal contribution of the study is an integrated analytical workflow linking automated freight-data processing, EPA-based emissions estimation, sustainability visualization, predictive risk identification, and carbon-efficiency anomaly detection. Rather than introducing artificial intelligence as an independent or purely technological objective, the framework uses machine learning to extend an existing data-visualization and automation approach toward earlier and more targeted sustainability decision support. The resulting methodology provides a reproducible foundation for examining how large-scale freight information can be transformed into actionable insights for reducing carbon-intensive supply-chain activity while maintaining operational awareness.

2 LITERATURE REVIEW

Recent research on sustainable supply chains increasingly emphasizes the integration of environmental performance, digital analytics, and intelligent decision-support methods. Studies have examined freight decarbonization, data visualization, machine learning, predictive analytics, and anomaly detection as complementary tools for improving logistics efficiency and reducing environmental impact. This review synthesizes those research streams to establish the basis for combining shipment-level emissions estimation, visualization, and AI-enhanced analytics within a unified sustainable supply-chain framework.

2.1 Sustainable Supply Chains and Freight-Related Carbon Emissions

Sustainable supply-chain management extends traditional objectives of cost, service quality, and operational efficiency by incorporating environmental considerations into sourcing, production, distribution, and transportation decisions. Transportation is particularly important because freight movement directly connects supply-chain activities while generating greenhouse-gas emissions that vary substantially according to transportation mode, distance, shipment characteristics, and operational configuration. Recent reviews of sustainable logistics emphasize freight decarbonization, transportation optimization, modal choices, and emissions measurement as important components of greener supply-chain systems [1], [2].

The increasing availability of operational data has also shifted sustainability analysis from aggregate environmental reporting toward more detailed assessment of individual activities and decisions. Data-science and big-data methodologies are increasingly used across supply-chain planning levels to improve efficiency, resilience, sustainability, transportation management, forecasting, and decision support [3]. This development is relevant to freight emissions because shipment-level characteristics can provide considerably more decision-relevant information than aggregate carbon inventories alone.

Carbon-aware transportation research increasingly demonstrates that economic efficiency and environmental performance can be evaluated jointly. Sánchez-Pravos et al. [9], for example, combined Random Forest and XGBoost emission prediction with evolutionary route optimization and showed that carbon-aware routing could reduce emissions while explicitly considering associated cost trade-offs. Such research illustrates the transition from retrospective emissions measurement toward analytical systems that can inform operational decisions before transportation choices are finalized.

2.2 Data Visualization, Automation, and Supply-Chain Decision Support

Large supply-chain datasets create an information-management challenge because operational insights may remain difficult to extract when analysts depend on repetitive manual processing. Data visualization provides an interface between large-scale quantitative data and managerial interpretation by transforming complex patterns into forms that can be more readily explored and compared. A systematic review of 364 supply-chain and logistics studies identified data visualization, predictive analytics, machine learning, and advanced data processing as important components of contemporary data-science capabilities across strategic, tactical, and operational decision levels [3].

Visual analytics is particularly relevant to transportation and fleet management. Ferreira et al. [4] developed a visual-analytics approach for industrial vehicle fleet assessment after observing that logistics analysts faced difficulties extracting useful information from large operational databases through manual procedures. Their results indicated that visual representations could support fleet-utilization assessment and identification of optimization opportunities. This demonstrates the value of integrating automated analytical processing with visualization rather than requiring decision-makers to manually derive operational insights from raw logistics records.

For sustainable supply chains, this combination is especially useful because environmental decisions frequently involve multiple dimensions simultaneously, including shipment volume, distance, transportation mode, economic value, and carbon intensity. Visualization can expose patterns within these variables, while automation can perform repetitive transformation and emissions-estimation tasks consistently. The role of visualization in the present study is therefore not merely graphical presentation; it forms part of a decision-support workflow that converts shipment-level freight records into sustainability indicators that stakeholders can interpret efficiently.

2.3 Machine Learning and Predictive Analytics in Freight Transportation

Machine learning has become increasingly prominent in freight analytics because transportation outcomes frequently depend on nonlinear interactions among shipment, geographic, commodity, and industry characteristics. Uddin et al. [10] compared multiple machine-learning classifiers using U.S. Commodity Flow Survey data and found that tree-based ensemble approaches performed particularly well for freight-mode prediction, with Random Forest producing the highest predictive accuracy among the evaluated methods. Their analysis also identified shipment characteristics including distance, industry classification, and shipment size as influential variables in freight-mode decisions.

More recent research has extended machine learning toward environmental objectives. Sizan et al. [5] evaluated machine-learning approaches specifically for shipment-level emission-risk prediction in sustainable supply-chain logistics, emphasizing the value of moving from retrospective emissions monitoring toward proactive classification of potentially high-emission shipments. Similarly, Random Forest has been applied to transportation-emission prediction in maritime operations, where operational characteristics were used to estimate fuel consumption and associated CO_2 emissions [11].

These findings support the use of Random Forest in the present study. The method can capture nonlinear relationships and interactions without requiring the restrictive functional assumptions associated with conventional linear models. It also supports variable-importance analysis, allowing predictive accuracy to be complemented by interpretation of the shipment characteristics that contribute most strongly to carbon-risk classification.

An important methodological consideration, however, is target leakage. When an emissions outcome is mathematically derived from variables such as weight, distance, and mode-specific emission factors, using the same constructed variables indiscriminately as predictors can produce artificially strong model performance. The present study therefore excludes shipment distance, calculated CO_2e, ton-miles, and emission-factor variables from predictive inputs and additionally evaluates a robustness specification that removes shipment weight. This allows the analysis to determine whether other operational characteristics retain meaningful predictive information rather than merely reproducing the emissions equation.

2.4 Anomaly Detection in Logistics Operations

Predictive classification addresses whether a shipment is likely to belong to a predefined carbon-risk category, whereas anomaly detection provides a complementary capability by identifying observations that differ substantially from typical operational behavior. This is valuable in supply chains because anomalous observations may represent operational inefficiencies, unusual transportation requirements, data-quality problems, or activities requiring further managerial investigation.

Isolation Forest is particularly suitable for this purpose because it is an unsupervised method and therefore does not require a pre-existing labeled set of anomalous shipments. Recent logistics research has demonstrated the applicability of Isolation Forest and related approaches to identifying abnormal operational patterns. Xie et al. [6], for example, developed an improved Isolation Forest methodology for multidimensional cold-chain logistics data and reported strong precision, recall, F1, and AUC performance relative to alternative anomaly-detection approaches.

A key consideration for carbon analysis is that transportation modes possess structurally different emission intensities. An overall anomaly model can therefore incorrectly interpret normal observations from inherently carbon-intensive modes as unusual simply because their baseline emissions differ from other modes. For this reason, the present study performs anomaly detection within transportation-mode groups, allowing air, rail, and truck shipments to be evaluated against their respective peer patterns. This produces a more meaningful screening mechanism for carbon inefficiency than applying one unrestricted anomaly threshold across fundamentally different freight technologies.

2.5 Research Gap and Study Positioning

Existing studies show that data analytics and visualization can improve supply-chain decision support [3], [4], machine learning can model freight characteristics using Commodity Flow Survey data [10], predictive models can support carbon-aware transportation analysis [5], [9], [11], and unsupervised methods can detect abnormal logistics behavior [6]. However, these capabilities remain largely separate, creating an opportunity for a simple, reproducible shipment-level framework using large-scale U.S. freight microdata and authoritative government emission factors.

Accordingly, this study integrates four stages within one workflow: automated freight-data processing, EPA-based shipment-level CO_2e estimation, sustainability visualization, and AI-enhanced predictive and anomaly analytics. It combines the 2022 Commodity Flow Survey Public Use Microdata Sample with EPA transportation emission factors [7], [8]. The CFS, jointly produced by the U.S. Census Bureau and Bureau of Transportation Statistics, is a primary national source of domestic freight-movement data.

The study further tests the predictive model without shipment weight and applies Isolation Forest separately within transportation modes. These choices address dependence on a variable embedded in the emissions equation and cross-mode confounding in anomaly detection. The resulting framework moves beyond retrospective emissions estimation by converting freight data into interpretable sustainability indicators for identifying carbon-intensive activities requiring closer attention.

3 DATA AND METHODOLOGY

This study uses a quantitative, data-driven methodology combining public freight-shipment data, standardized greenhouse-gas emission factors, automated processing, visualization, and machine learning. The workflow estimates shipment-level carbon emissions, examines freight sustainability patterns, and applies predictive and anomaly-detection models to support earlier identification of carbon-intensive activities. It integrates descriptive, predictive, and unsupervised analytics with emphasis on reproducibility, interpretability, and practical supply-chain decision support.

3.1 Data Sources and Study Sample

This study combines freight-shipment data from the 2022 Commodity Flow Survey Public Use Microdata Sample (CFS PUMS) with transportation-related greenhouse-gas emission factors from the U.S. EPA 2025 GHG Emission Factors Hub [7], [8]. Jointly produced by the U.S. Census Bureau and Bureau of Transportation Statistics, the CFS provides shipment-level data on transportation mode, weight, value, origin and destination, industry sector, commodity classification, distance, temperature control, hazardous-material status, and exports.

Table 1: Key Variables Used in the Sustainable Freight Analysis

Variable Group	Variables	Purpose
Shipment identification	SHIPMT_ID	Unique shipment record
Geography	ORIG_STATE, DEST_STATE	Origin and destination characteristics
Freight characteristics	MODE, SCTG, SECTOR	Transportation, commodity, and industry classification
Shipment attributes	SHIPMT_VALUE, SHIPMT_WGHT, SHIPMT_DIST_GC	Shipment attributes used across emissions estimation and predictive analysis
Operational indicators	TEMP_CNTL_YN, EXPORT_YN, HAZMAT	Additional shipment characteristics

To maintain computational efficiency while preserving a sufficiently large analytical sample, 250,000 shipment records were selected from the CFS PUMS. The PUMS shipment identifier (SHIPMT_ID) was assigned after the shipment records were randomly sorted and then numbered sequentially, supporting the use of the selected records as a randomly ordered working subset for the present analysis [7]. Initial data-quality assessment identified no missing observations or duplicate shipment identifiers within the selected sample. Following transportation-mode restrictions and the exclusion of records with nonpositive shipment weight or distance, 121,801 shipments remained for the emissions analysis. For emissions estimation, the analysis was restricted to transportation modes that could be directly matched to EPA freight emission factors: truck, rail, waterborne transportation, and air freight. Modes without directly comparable EPA ton-mile factors, including multimodal shipments, parcel services, pipeline transportation, customer pickup, and unknown modes, were excluded from the primary emissions analysis.

3.2 Automated Data Processing and Transportation-Mode Classification

The analytical workflow automated the preparation, transformation, and classification of freight-shipment records to support consistent emissions estimation and subsequent predictive analysis. Transportation-mode codes in the Commodity Flow Survey were mapped to corresponding freight categories used in the EPA emissions framework. For-hire and company-owned truck shipments were consolidated into a single truck category, while inland water, Great Lakes, and deep-sea shipments were grouped as waterborne transportation. Rail shipments were retained as a separate category, while CFS mode 14, defined as “Air (includes truck and air),” was mapped to the air-freight category for the emissions analysis [7].

This automated classification process reduced repetitive manual handling of shipment records and ensured that transportation modes were consistently aligned with the appropriate emission factors. The resulting structured dataset provided the basis for calculating shipment-level carbon emissions, comparing sustainability performance across freight modes, and developing predictive models for identifying high-carbon-risk shipment patterns.

3.3 Shipment-Level Greenhouse-Gas Emissions Estimation

Shipment-level greenhouse-gas emissions were estimated using shipment weight, shipment distance, transportation mode, and EPA mode-specific emission factors. Shipment weight reported in pounds was first converted to short tons as:

$$W_i = \frac{W_{lb,i}}{2000}$$

where W_i represents shipment weight in short tons and $W_{lb,i}$ represents shipment weight in pounds. Ton-mile activity was subsequently calculated as:

$$TM_i = W_i \times D_i$$

where TM_i is the short ton-miles associated with shipment i, and D_i represents the great-circle shipment distance in miles.

The EPA 2025 freight transportation factors applied in the study were approximately 0.186 kg CO_2 per short ton-mile for truck transportation, 0.021 for rail, 0.077 for waterborne transportation, and 1.086 for air freight [8]. Methane and nitrous-oxide factors were also incorporated using 100-year global warming potentials of 28 for CH_4 and 265 for N_2O [8]. Because the CFS air category includes truck-and-air movements whereas the applied EPA factor represents aircraft freight, emissions estimated for this category should be interpreted as approximations of air-dominant freight activity.

Table 2: EPA Freight Transportation Emission Factors Used in the Study

Transportation Mode	CO_2 (kg/short ton-mile)	CH_4 (g/short ton-mile)	N_2O (g/short ton-mile)	CO_2e (kg/short ton-mile)
Truck	0.186	0.0016	0.0054	0.187476
Rail	0.021	0.0016	0.0005	0.021177
Waterborne	0.077	0.0310	0.0020	0.078398
Air	1.086	0.0000	0.0334	1.094851

The combined CO_2-equivalent emission factor was calculated as:

$$EF_{CO_2e,m} = EF_{CO_2,m} + \frac{28EF_{CH_4,m}}{1000} + \frac{265EF_{N_2O,m}}{1000}$$

where m represents transportation mode. Shipment-level greenhouse-gas emissions were then estimated as:

$$CO_2e_i = TM_i \times EF_{CO_2e,m}$$

with the resulting quantity expressed in kilograms of CO_2-equivalent.

Because the CFS distance variable represents great-circle distance between shipment origin and destination, rather than the exact physical transportation route, the resulting emissions estimates are interpreted as standardized distance-based indicators of freight-related emissions rather than direct measurements of actual route-specific fuel consumption. Similarly, the EPA factors represent transportation-related emissions and are not intended to constitute a complete life-cycle carbon assessment.

3.4 Carbon-Efficiency Indicators and Data Visualization

In addition to absolute shipment-level emissions, two carbon-efficiency indicators were derived to support comparative analysis. Carbon emissions relative to shipment economic value were calculated as:

$$CEV_i = \frac{CO_2e_i}{V_i} \times 1000$$

where V_i is shipment value in U.S. dollars and CEV_i represents kilograms of CO_2e per $1,000 of shipment value.

Carbon emissions relative to shipment weight were calculated as:

$$CET_i = \frac{CO_2e_i}{W_i}$$

where CET_i represents kilograms of CO_2e per short ton transported.

Visualization was used to examine differences in freight-mode emission intensity, median shipment emissions, relationships between shipment distance and emissions, predictive-model performance, feature importance, and anomalous carbon-efficiency patterns.

3.5 High-Carbon-Risk Classification

To extend descriptive analysis toward predictive decision support, shipments were classified into high- and lower-carbon-risk groups. A shipment was defined as **high carbon risk** when its estimated CO_2-equivalent emissions fell within the upper quartile of the analyzed shipment distribution:

$$Y_i = \begin{cases} 1, & CO_2e,i \geq Q_{0.75} \\ 0, & CO_2e,i < Q_{0.75} \end{cases}$$

where $Q_{0.75}$ represents the 75th percentile of shipment-level estimated emissions. This threshold defines a relative screening category within the study sample and should not be interpreted as a regulatory or externally established carbon-emission limit.

A Random Forest classifier [12] was trained using operational shipment characteristics including shipment weight, shipment value, transportation mode, commodity classification, industrial sector, temperature-control status, export status, hazardous-material status, origin state, and destination state.

Table 3: Predictive Modeling Design

Component	Specification
Model	Random Forest Classifier
Target	High-carbon-risk shipment
High-risk definition	Upper quartile of estimated shipment CO_2e
Training/testing split	80% / 20%
Primary predictors	Weight, value, mode, commodity, sector, origin, destination, temperature control, export, hazmat
Excluded target-construction variables	Shipment distance, CO_2e, ton-miles, EPA emission factors
Robustness test	Shipment weight excluded
Evaluation metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC

Shipment distance and variables directly generated from the emissions calculation—including estimated CO_2e, ton-miles, and EPA emission factors—were excluded from the predictor set to reduce target leakage and avoid an overly deterministic relationship with the outcome. Shipment weight was retained as an operational predictor in the primary model and subsequently removed in the robustness analysis. The dataset was divided into 80% training and 20% testing subsets using stratified sampling to preserve the high-carbon-risk class distribution.

3.6 Model Evaluation and Robustness Analysis

Predictive performance was assessed using accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC). These measures were selected because high-carbon-risk identification requires consideration of both overall classification performance and the model's ability to correctly identify carbon-intensive shipments.

The principal metrics were calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$ and

$$F1 = 2\left(\frac{Precision \times Recall}{Precision + Recall}\right)$$

A robustness analysis was additionally performed by removing shipment weight from the predictor set. Because shipment weight directly contributes to the mathematical calculation of ton-miles and therefore estimated emissions, this sensitivity test was used to determine whether the model retained meaningful predictive capability from other operational characteristics rather than primarily reproducing the underlying emissions equation.

Permutation feature importance was subsequently used to evaluate the relative contribution of the original shipment variables to predictive performance.

3.7 Within-Mode Carbon-Efficiency Anomaly Detection

An unsupervised Isolation Forest model [13] was applied as a complementary analytical technique for detecting unusual carbon-efficiency patterns. The anomaly analysis used CO_2e per $1,000 of shipment value and CO_2e per short ton as its principal variables. Log transformations were applied to reduce the influence of strong right-skewness in these measures.

Preliminary analysis demonstrated that applying a single anomaly model across all transportation modes disproportionately identified air-freight observations because of the inherently higher EPA emission intensity associated with aviation. The final methodology therefore applied Isolation Forest separately within transportation-mode groups.

Air, rail, and truck shipments were analyzed independently, while waterborne transportation was excluded from this component because its working-sample size was insufficient for reliable within-group anomaly modeling. An anomaly contamination level of 2% was specified to identify the most unusual carbon-efficiency patterns within each sufficiently represented transportation mode.

Accordingly, the identified observations are interpreted as screening candidates for further stakeholder investigation, rather than evidence that 2% of freight shipments are inherently anomalous. This within-mode approach controls for structural differences in transportation technologies and provides a more meaningful comparison of carbon efficiency among otherwise comparable freight movements.

4 RESULTS AND DISCUSSION

The analysis produced complementary findings on freight-related greenhouse-gas emissions, shipment-level carbon risk, and carbon-efficiency anomalies. The results demonstrate substantial differences in emission intensity across transportation modes and show that operational shipment characteristics can be used to identify freight movements associated with elevated carbon-emission risk. The predictive analysis further indicates that useful discriminatory performance can be retained even when shipment weight, the variable most directly related to the emissions calculation, is excluded. These findings support the integration of visualization, predictive analytics, and anomaly detection as complementary components of sustainable supply-chain decision support.

4.1 Freight Transportation and Carbon-Emission Patterns

The estimated emission intensities differed substantially across the four transportation categories evaluated in the study. Based on the EPA-derived factors, air freight had the highest emission intensity at approximately 1.095 kg CO_2e per short ton-mile, followed by truck transportation at approximately 0.187 kg CO_2e per short ton-mile. Waterborne transportation had an estimated intensity of approximately 0.078 kg CO_2e per short ton-mile, while rail exhibited the lowest intensity at approximately 0.021 kg CO_2e per short ton-mile**.**

These differences demonstrate that transportation-mode selection can materially influence the environmental impact associated with moving an equivalent amount of freight over a comparable distance. The considerably higher intensity associated with air transportation illustrates the potential carbon implications of time-sensitive freight movement, while the lower emission intensity of rail highlights the sustainability advantages that may be available when operational requirements permit modal substitution.

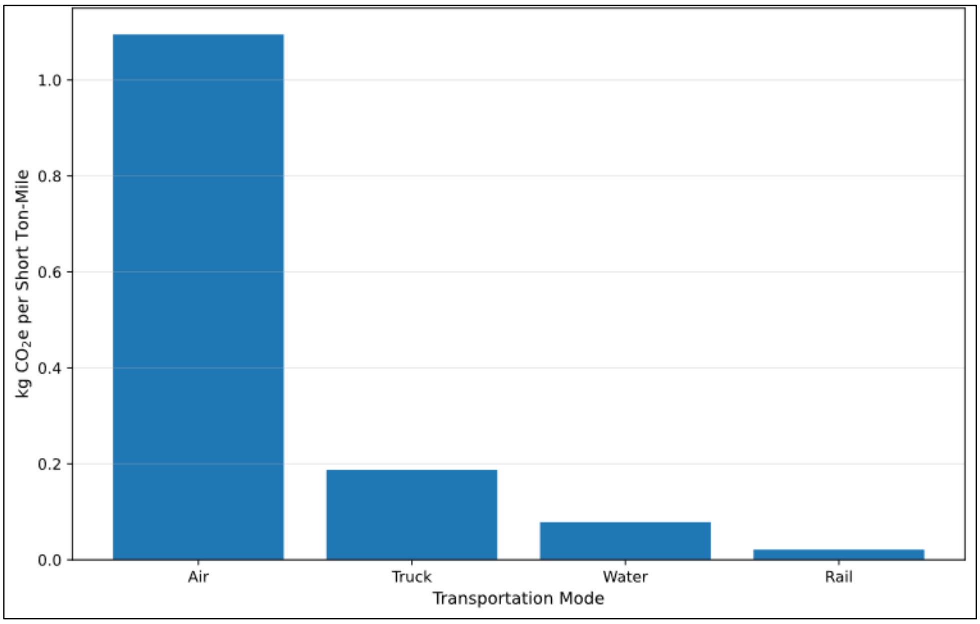

Figure 1: Estimated GHG Emission Intensity by Freight Transportation Mode

The shipment-level analysis further showed that transportation-mode intensity alone does not determine total emissions for an individual shipment. Shipment weight and transportation distance interact with the mode-specific emission factor to determine estimated CO_2e. Consequently, a shipment transported using a comparatively lower-intensity mode may still generate considerable emissions when the shipment is particularly heavy or moves over a long distance.

The median shipment-level comparison provides an additional perspective by incorporating the characteristics of the actual freight records rather than comparing emission factors alone. Notably, although rail exhibited the lowest mode-specific emission intensity in Figure 1, it showed the highest median shipment-level emissions in Figure 2. This illustrates that lower emissions per ton-mile do not necessarily translate into lower absolute emissions per shipment because shipment weight and distance also influence the resulting carbon footprint. The waterborne result should be interpreted cautiously because only 30 usable waterborne observations were available in the working sample.

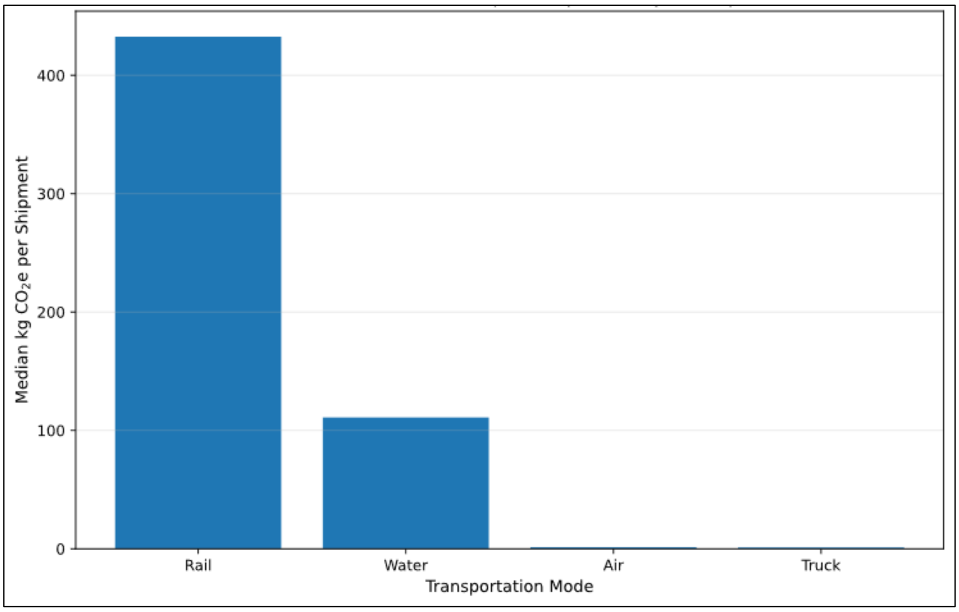

Figure 2: Median Estimated GHG Emissions per Shipment by Transportation Mode

The relationship between shipment distance and estimated greenhouse-gas emissions also revealed substantial variation across freight movements. Although longer transportation distances generally increase the opportunity for higher emissions, shipment weight and transportation mode produce considerable dispersion among observations. Similar travel distances can therefore correspond to substantially different carbon outcomes.

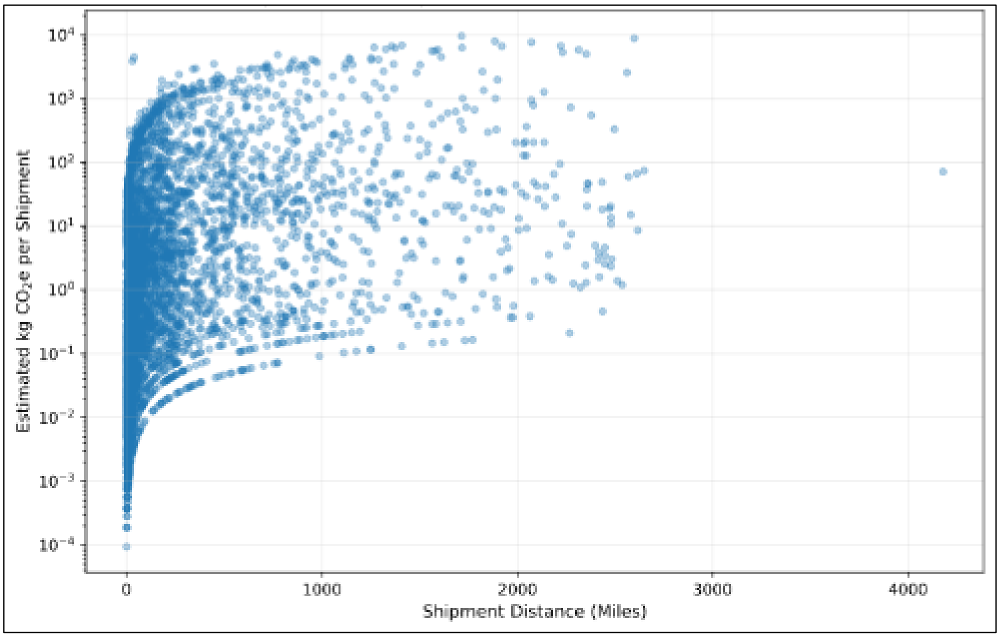

Figure 3: Relationship Between Shipment Distance and Estimated GHG Emissions (log-scaled y-axis)

These findings indicate that sustainable freight evaluation should consider transportation mode, distance, and shipment characteristics jointly. Evaluating only one of these variables may obscure potentially important carbon-efficiency differences among freight movements.

4.2 Predictive Identification of High-Carbon-Risk Shipments

The Random Forest classifier demonstrated strong ability to distinguish shipments categorized as high carbon risk from those in the lower-risk group. The primary model achieved an accuracy of 89.7%, precision of 73.7%, recall of 91.5%, and an F1-score of 81.6%. The ROC-AUC value of 0.964 indicates a high level of discriminatory performance across classification thresholds.

The comparatively high recall is particularly relevant for sustainability-oriented screening. A recall of 91.5% indicates that the model successfully identified a large majority of shipments belonging to the high-carbon-risk category. In a decision-support context, this reduces the likelihood that potentially carbon-intensive freight movements would remain unidentified during initial screening.

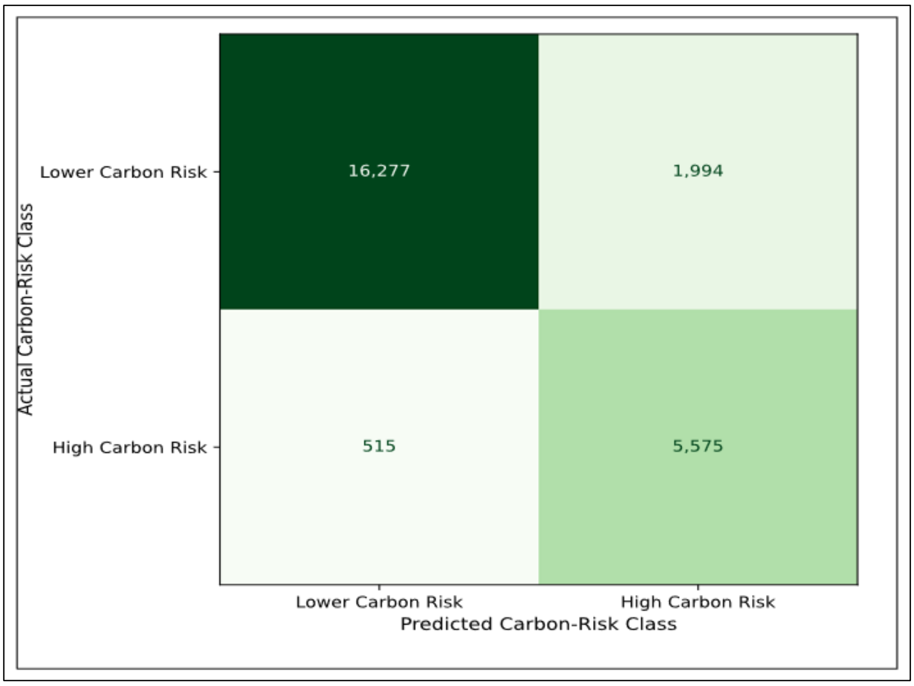

Figure 4: Confusion Matrix for Random Forest High-Carbon-Risk Classification

The confusion matrix provides a visual representation of the model's classification performance and demonstrates that most observations were assigned to their correct carbon-risk category. Although some false-positive classifications occurred, the strong recall suggests that the model favored identifying high-risk shipments rather than overlooking them. Such a trade-off can be appropriate where the analytical objective is to prioritize potentially carbon-intensive shipments for closer review rather than to make fully automated operational decisions.

4.3 Operational Factors Influencing Carbon Risk

Permutation feature-importance analysis provided further insight into the operational characteristics contributing to high-carbon-risk classification. Shipment weight was the dominant predictor, with a permutation importance value of 0.336581. Transportation mode was the second most important feature at 0.045269, followed by shipment value at 0.019152 and industrial sector at 0.017485.

Table 4: Permutation Importance of Operational Shipment Characteristics

Rank	Feature	Permutation Importance
1	Shipment Weight	0.336581
2	Transportation Mode	0.045269
3	Shipment Value	0.019152
4	Industry Sector	0.017485
5	Commodity Classification	0.007062
6	Origin State	0.005722
7	Destination State	0.005719
8	Temperature-Control Status	0.002197
9	Hazardous-Material Status	0.000058
10	Export Status	0.000007

Commodity classification, origin state, and destination state also contributed predictive information, although their individual importance values were substantially smaller. Temperature-control status showed a modest contribution, while hazardous-material and export indicators provided very little incremental predictive value within the fitted model.

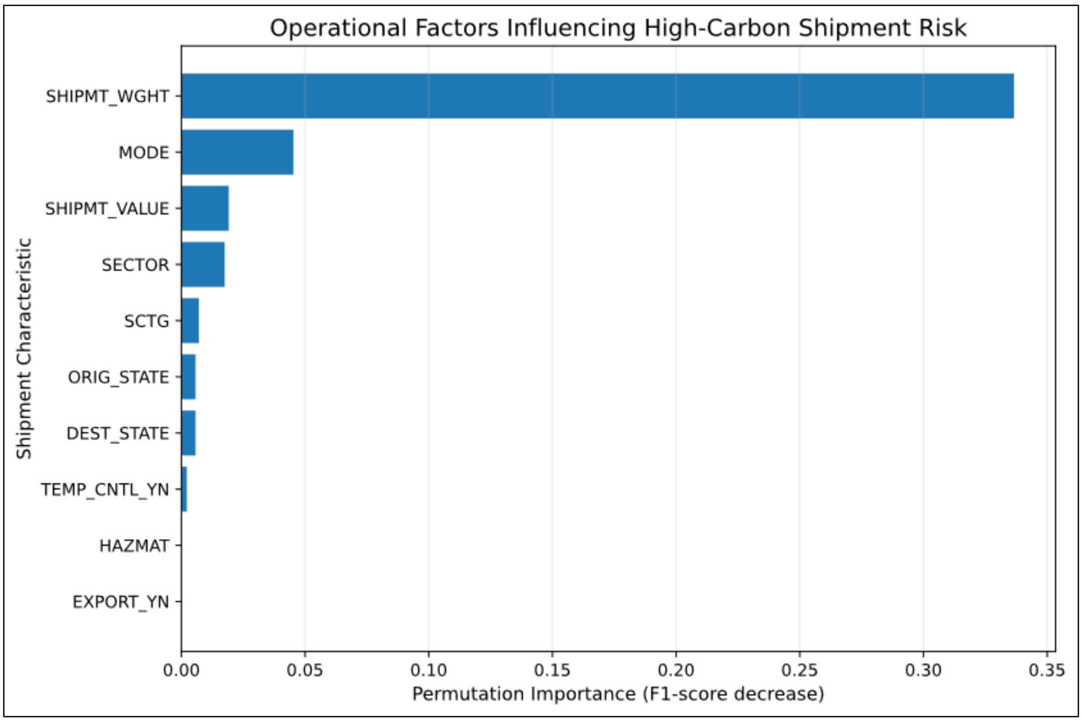

Figure 5: Permutation Importance of Operational Variables in High-Carbon-Risk Prediction

The dominance of shipment weight is consistent with the underlying emissions framework because freight activity is partly determined by the amount of material transported. However, the additional contribution of transportation mode, shipment value, industrial sector, commodity type, and geographic characteristics indicates that carbon-risk classification is not explained by shipment weight alone. The feature-importance results therefore provide a basis for identifying operational dimensions that may deserve greater attention during sustainability-focused shipment planning.

4.4 Robustness of the Predictive Model

A robustness analysis was conducted to determine whether the predictive performance of the primary model depended excessively on shipment weight. This test was particularly important because shipment weight directly contributes to the ton-mile calculation used to estimate shipment-level emissions.

Table 5: Comparison of Primary and Shipment-Weight-Excluded Models

Metric	Primary Model	Without Shipment Weight
Accuracy	0.897	0.865
Precision	0.737	0.683
Recall	0.915	0.858
F1-score	0.816	0.760
ROC-AUC	0.964	0.932

After excluding shipment weight, model performance declined moderately but remained strong. Accuracy decreased from 89.7% to 86.5%, precision from 73.7% to 68.3%, recall from 91.5% to 85.8%, and F1-score from 81.6% to 76.0%. ROC-AUC declined from 0.964 to 0.932.

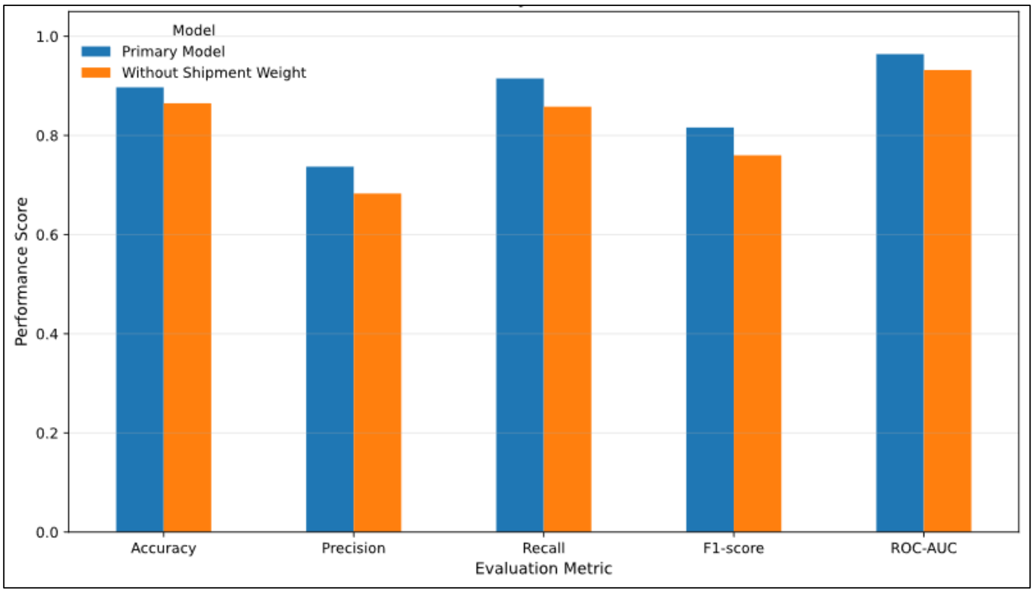

Figure 6: Predictive Performance of Primary and Shipment-Weight-Excluded Carbon-Risk Models

The robustness results strengthen the predictive findings. Although shipment weight contains substantial information about absolute emissions, the weight-excluded model still achieved a ROC-AUC above 0.93 and recall above 85%. This demonstrates that transportation mode, shipment value, industrial sector, commodity characteristics, geography, and other operational variables collectively contain meaningful information about high-carbon shipment risk.

The result also reduces concern that the primary model merely reproduced the emissions-estimation equation. Instead, the predictive framework captured broader operational patterns associated with carbon-intensive freight activity.

4.5 Within-Mode Carbon-Efficiency Anomaly Detection

The Isolation Forest analysis complemented the supervised classification model by identifying shipments exhibiting unusual carbon-efficiency behavior. Anomaly detection was performed separately within transportation modes to avoid treating structural differences in baseline mode-specific emission intensity as anomalous behavior.

Using the predefined 2% screening level, the analysis identified 33 of 1,669 air shipments, 14 of 681 rail shipments, and 2,385 of 119,421 truck shipments as unusual within-mode carbon-efficiency patterns. These corresponded to anomaly-screening rates of approximately 1.98%, 2.06%, and 2.00%, respectively.

Table 6: Within-Mode Carbon-Efficiency Anomaly Screening Results

Transportation Mode	Shipments	Carbon-Efficiency Anomalies	Screening Rate (%)
Air	1,669	33	1.98
Rail	681	14	2.06
Truck	119,421	2,385	2.00

Waterborne transportation was not included in the within-mode anomaly model because only 30 usable observations were available in the working sample, which was insufficient for meaningful group-specific anomaly screening.

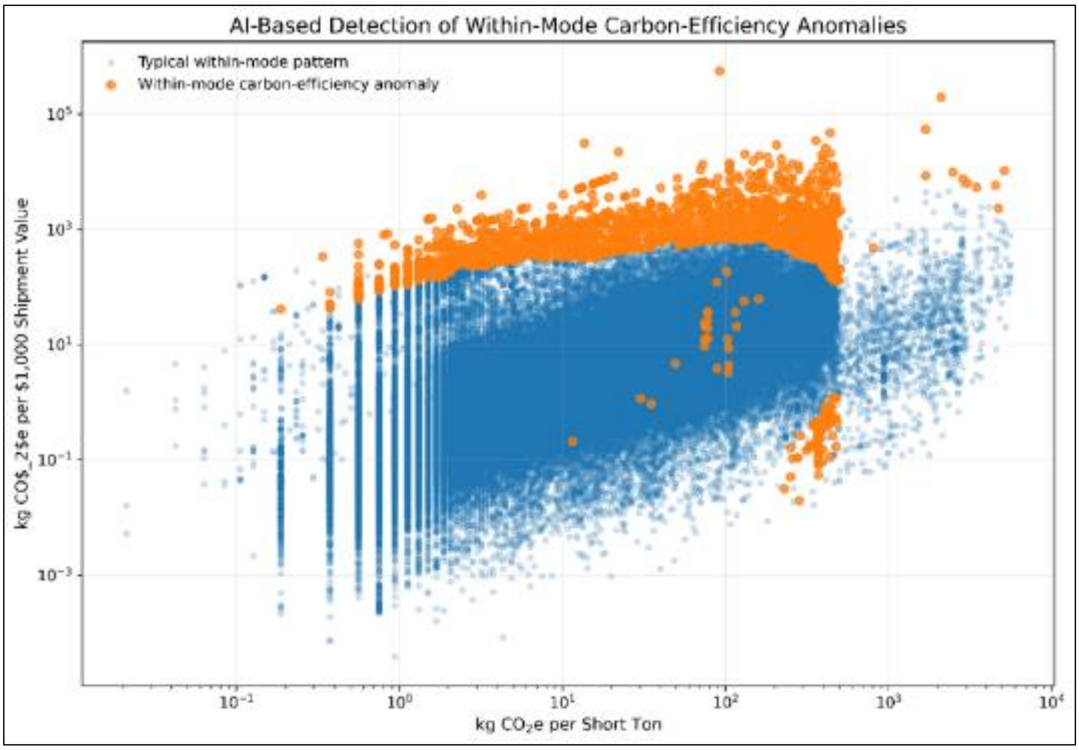

Figure 7: AI-Based Detection of Within-Mode Carbon-Efficiency Anomalies (log-scaled axes)

The approximately 2% screening rates should not be interpreted as estimates of the true prevalence of carbon inefficiency in U.S. freight transportation. Rather, they reflect the predefined Isolation Forest contamination threshold used to identify the most unusual observations within each sufficiently represented transportation mode. The analytical value lies in prioritizing shipments for further examination based on their relative carbon-efficiency characteristics.

Applying anomaly detection separately within transportation modes also addressed an important methodological issue. When all transportation modes were initially analyzed together, air freight was disproportionately identified as anomalous because of its inherently higher emission intensity. Within-mode modeling instead compares each shipment with operational peers using the same general transportation technology, producing a more meaningful basis for sustainability screening.

4.6 Implications for Sustainable Supply-Chain Decision-Making

The combined results illustrate how descriptive visualization, emissions estimation, predictive classification, and anomaly detection can support different stages of sustainable supply-chain analysis. Visualization reveals differences in carbon intensity and shipment-level emissions, while predictive modeling provides a mechanism for identifying potentially high-carbon freight movements before additional assessment. Feature-importance analysis helps explain which shipment characteristics contribute most strongly to predictive risk, and anomaly detection highlights unusual carbon-efficiency patterns that may require further investigation.

From an operational perspective, the results suggest that sustainability interventions should focus particularly on shipment weight and transportation mode while also considering shipment value, industrial sector, commodity characteristics, and geographic factors. Where operationally feasible, the framework can help prioritize evaluation of shipment consolidation, alternative transportation modes, improved freight planning, or closer review of movements exhibiting unusually high carbon intensity.

The proposed framework is not intended to automatically prescribe transportation decisions. Cost, delivery requirements, reliability, infrastructure availability, commodity characteristics, and service commitments remain important considerations in freight planning. Instead, the framework provides an analytical mechanism for making carbon-related information more visible and actionable within broader supply-chain decision processes.

Overall, the findings demonstrate that integrating automated emissions estimation with visualization and AI-enhanced analytics can move sustainability assessment beyond retrospective reporting toward more proactive identification of carbon-intensive freight activity. This capability provides a practical basis for helping stakeholders concentrate attention on shipments and operational patterns where sustainability interventions may offer the greatest potential benefit.

5 CONCLUSION AND FUTURE RESEARCH

This study developed a sustainable supply-chain analytics framework that combines shipment-level emissions estimation, data visualization, predictive machine learning, and anomaly detection. Using 2022 Commodity Flow Survey data and EPA freight-emission factors, the analysis showed clear differences in carbon intensity across transportation modes and demonstrated that shipments classified as high carbon risk under the study's upper-quartile threshold can be identified with strong predictive performance. The Random Forest model achieved 89.7% accuracy and a ROC-AUC of 0.964, while the robustness model retained a ROC-AUC of 0.932 after excluding shipment weight.

The findings show that shipment weight, transportation mode, shipment value, sector, commodity type, and geography provide useful information for sustainability-focused decision support. Within-mode Isolation Forest analysis also identified unusually carbon-inefficient shipment patterns for further review. Although the analysis relies on standardized emission factors, great-circle distance estimates, and an unweighted working subset, it provides a practical basis for identifying sample-level carbon patterns rather than official population estimates. The CFS air category also includes truck-and-air movements, which introduces additional approximation into air-freight emissions estimates. Future research can extend the approach using route-specific data, additional freight modes, and scenario-based carbon-reduction strategies.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The author gratefully acknowledges Dr. Ramanjit Sahi of Austin Peay State University for her guidance and support during the original 2023 supply-chain data-visualization project that provided the foundation for this expanded research.

Disclosure of conflict of interest

The author declares no conflict of interest regarding this manuscript.

REFERENCES

[1] D. Toktaş, M. A. Ülkü, and M. A. Habib, “Toward Greener Supply Chains by Decarbonizing City Logistics: A Systematic Literature Review and Research Pathways,” Sustainability, vol. 16, no. 17, 7516, 2024. doi: 10.3390/su16177516.

[2] C. Qu and E. Kim, “Reviewing the Roles of AI-Integrated Technologies in Sustainable Supply Chain Management: Research Propositions and a Framework for Future Directions,” Sustainability, vol. 16, no. 14, 6186, 2024. doi: 10.3390/su16146186.

[3] H. Jahani, R. Jain, and D. Ivanov, “Data science and big data analytics: A systematic review of methodologies used in the supply chain and logistics research,” Annals of Operations Research, vol. 359, pp. 1297–1354, 2026. doi: 10.1007/s10479-023-05390-7.

[4] G. X. Ferreira, M. M. V. de Paula, R. P. Pagan, and B. G. Batista, “Fleet Profile: Using visual analytics to prospect logistic solutions in industrial vehicles fleet,” Computers in Industry, vol. 151, 103971, 2023. doi: 10.1016/j.compind.2023.103971.

[5] A. A. Sizan, M. S. Uddin, M. Hossain, F. Ahmed, A. Hossain, and R. Hasan, “Comparative Evaluation of Machine Learning Models for Emission Risk Prediction in Sustainable Supply Chain Logistics,” in 2025 International Conference on Intelligent Computing, Information and Control Systems, pp. 919–925, 2025. doi: 10.1109/ICOIICS67115.2025.11390445.

[6] Z. Xie, H. Long, C. Ling, Y. Zhou, and Y. Luo, “An anomaly detection scheme for data stream in cold chain logistics,” PLOS ONE, vol. 20, no. 3, e0315322, 2025. doi: 10.1371/journal.pone.0315322.

[7] U.S. Department of Transportation, Bureau of Transportation Statistics, and U.S. Department of Commerce, U.S. Census Bureau, 2022 Commodity Flow Survey Datasets: 2022 CFS Public Use Microdata Sample (PUMS) File, Jan. 2026.

[8] U.S. Environmental Protection Agency, 2025 GHG Emission Factors Hub, 2025.

[9] L. Sánchez-Pravos, J. Parra-Domínguez, S. Rodríguez González, and P. Chamoso, "A machine learning and evolutionary optimization framework for carbon-aware supply chain routing," Supply Chain Analytics, vol. 13, 100182, 2026. doi: 10.1016/j.sca.2025.100182.

[10] M. Uddin, S. Anowar, and N. Eluru, "Modeling freight mode choice using machine learning classifiers: A comparative study using Commodity Flow Survey (CFS) data," Transportation Planning and Technology, vol. 44, no. 5, pp. 543–559, 2021. doi: 10.1080/03081060.2021.1927306.

[11] C. A. Hargreaves and B. W. N. Toh, "Machine Learning for Sustainable Shipping: Predicting Vessel CO_2 Emissions Using Random Forest Models," in Federated Learning – A Systematic Review, IntechOpen, 2025. doi: 10.5772/intechopen.1008820.

[12] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.