

Abstract

The increasing complexity of global supply chains contributes significantly to carbon emissions and often suffers from a lack of operational transparency. Addressing this challenge requires tools that can not only report on past environmental performance but also predict future outcomes and guide sustainable decision-making. This paper presents an integrated framework that leverages machine learning and data visualization to model, monitor, and optimize supply chain sustainability. The framework incorporates a predictive engine to classify shipments by their emission levels, an interactive dashboard for real-time monitoring, and a simulation tool for evaluating optimization strategies. A Random Forest Classifier, enhanced with the Synthetic Minority Over-sampling Technique (SMOTE) to handle imbalanced data, successfully identified high-emission shipments. The model's key finding was that shipment size (unit_quantity) and transport time (tpt) are the primary drivers of the carbon footprint. The developed proof-of-concept dashboard demonstrates how these insights can be translated into a practical decision-support tool, enhancing transparency and enabling managers to forecast the environmental impact of future shipments.

Introduction

Global supply chains form the backbone of modern commerce, yet their environmental impact is substantial, with logistics and transportation being major contributors to greenhouse gas emissions. As corporations and regulators place a greater emphasis on sustainability, there is a critical need for advanced tools that go beyond simple historical reporting. Companies often lack the ability to effectively measure their supply chain's carbon footprint in real-time, predict the impact of future decisions, and identify the most effective optimization strategies.

This research addresses this gap by developing a holistic, data-driven framework for sustainable supply chain management. The primary aim is to demonstrate how machine learning can be integrated with interactive visualization to create a system that enhances transparency and reduces environmental impact. This paper details the design and implementation of a system that (1) predicts the emission category of shipments, (2) provides a dynamic platform for monitoring sustainability metrics, and (3) offers a simulation tool to quantify the benefits of potential optimizations. The central contribution of this work is the creation of a cohesive decision-support system that connects predictive insights with actionable outcomes.

2. Literature Review

The field of sustainable supply chain management has garnered significant attention as organizations seek to mitigate their environmental impact while maintaining operational efficiency. This review examines three core areas of existing research: Green Supply Chain Management (GSCM) practices, the application of Artificial Intelligence (AI) and Machine Learning (ML) in logistics, and the development of transparency-enhancing platforms.

2.1. Green Supply Chain Management (GSCM)

GSCM involves integrating environmental thinking into all phases of the supply chain. Foundational research by Srivastava (2007) provided a comprehensive overview of GSCM,

categorizing research into areas like green design, green operations, and reverse logistics. The study emphasized the dual objective of GSCM: minimizing environmental degradation while improving economic performance.

Subsequent work has focused on specific drivers and performance outcomes. Large and Thomsen (2011) investigated the drivers of GSCM in logistics service providers, finding that regulatory pressures and customer demands are key motivators. More recently, Tarei et al. (2021) developed a framework using fuzzy logic to assess the barriers to implementing GSCM in the electronics industry. Their work is crucial as it highlights that even with clear drivers, internal organizational barriers like high initial costs and lack of top management commitment often hinder adoption. While their approach helps in identifying barriers, it relies on expert surveys and qualitative assessments. Our research offers a complementary, quantitative approach by providing a tool that directly calculates the financial and environmental metrics (in terms of carbon), which can be used to build a stronger business case and overcome such barriers.

A common challenge in the literature is the difficulty in quantifying environmental impact. For example, the work by Centobelli et al. (2020) provides a systematic review of the circular economy and GSCM literature, noting that while many models exist, they often operate at a strategic level. They point out a "scarcity of decision-support models" for operational level decisions. Our framework directly addresses this scarcity by providing a shipment-level analysis and a forecasting tool designed for day-to-day managerial use, bridging the gap between high-level strategy and operational execution.

2.2. AI and Machine Learning in Supply Chain Management

The application of AI and ML to supply chain challenges is a rapidly growing field. A review by Baryannis et al. (2019) highlights the extensive use of machine learning for predicting supply chain risks, such as delivery delays. They note the importance of model interpretability for managerial adoption.

More recently, research has begun to apply ML to sustainability. Soysal et al. (2018) developed a model for a sustainable food supply chain to minimize costs and CO₂ emissions. Similarly, Priyadarshini and Abhilash (2020) explored the use of IoT and blockchain for creating a sustainable circular economy, where data from sensors could trigger automated actions. Their mechanism relies on a high-tech infrastructure for data collection. While powerful, this approach requires significant upfront investment. Our work provides a more immediately accessible solution by demonstrating that valuable insights can be derived from existing, often imperfect, operational data (like shipment manifests and transport times), making it applicable to a wider range of companies that may not have advanced IoT capabilities.

Furthermore, studies like Ben-Daya et al. (2019) review the intersection of big data and predictive analytics for supply chain demand forecasting. They show how models can predict what a customer will buy. Our research adapts this predictive paradigm to a different target: instead of predicting demand, we predict environmental impact. We reframe the problem from a commercial one to a sustainability one, using similar machine learning techniques to classify shipments by their emission risk, a novel application of predictive analytics in this context.

2.3. Transparency and Decision Support Systems

Transparency is widely recognized as a cornerstone of a sustainable supply chain (Hofmann et al., 2018). A study by Choi (2023) on fast fashion supply chains argues that blockchain-based traceability is the future for providing transparency to consumers. The mechanism involves recording every transaction on an immutable ledger. This approach is powerful for verification and anti-counterfeiting but is often complex to implement and computationally intensive. Our research presents a different, more operational form of transparency. Instead of focusing on external validation for consumers, our Streamlit dashboard provides internal, managerial transparency. It is designed not as a ledger of the past, but as a dynamic forecasting tool to help managers understand the consequences of their future decisions, making it a more agile and prescriptive decision-support system.

In conclusion, while existing literature has established the importance of GSCM, explored AI for risk and demand prediction, and advocated for transparency, a significant gap remains in integrating these elements into a single, practical, and accessible decision-support framework. Our work contributes by (1) creating a unified model that accounts for both transport and manufacturing footprints from operational data, (2) reframing the ML problem to classify emission risk, and (3) developing an interactive dashboard that serves as both a monitoring and a forecasting tool, thereby bridging the gap between data analysis and actionable managerial insights.

3. Methodology

The methodology follows a three-phase approach: (1) Data Foundation, (2) Predictive Analysis, and (3) development of Decision Support Tools. A visual overview of this workflow is presented in Figure 1.

3.1. Data Sourcing and Preprocessing

Two primary datasets were utilized for this research: a supply chain logistics dataset containing shipment details (Supply chain logistics problem.xlsx) and a global energy dataset (World Energy Consumption.csv). The initial data preparation involved several key steps. First, raw data was loaded, and column names were standardized for consistency. A memory optimization function was applied to handle large datasets efficiently and prevent system crashes. To ensure data quality, rows with missing values in critical fields such as transport time (tpt) and weight were removed.

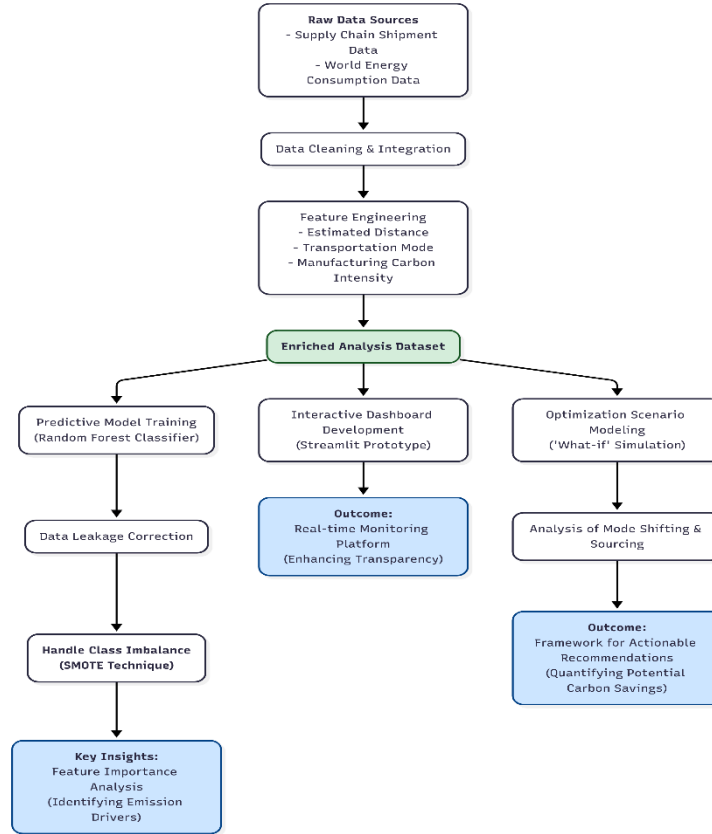


Figure 1: The overall research methodology, from data preprocessing to analysis and application.

3.2. Feature Engineering and Enrichment

To create a comprehensive analysis dataset, several new features were engineered. The `estimated_distance_km` was calculated by multiplying the transport time (`tpt`) by an assumed average transit of 350 km/day. The `transportation_mode` was inferred from the carrier ID. To calculate the manufacturing portion of the footprint, data from the energy dataset was processed to create a `carbon_intensity_of_energy` metric for each country. This was merged with the main dataset by mapping origin ports to their respective countries. Finally, `transport_footprint_kg` and `manufacturing_footprint_kg` were calculated and summed to create the `total_footprint_kg`. The final processed dataset was saved as `enriched_supply_chain_data.csv`.

3.3 Exploratory Data Analysis

Following the data preprocessing and enrichment, an exploratory data analysis (EDA) was conducted to understand the fundamental characteristics of the final dataset. The results of this analysis are presented in Figure 2.

As shown in the left panel of Figure 2, the distribution of shipment weight is highly skewed to the right. This indicates that the dataset is primarily composed of low-weight shipments, with a smaller number of extremely heavy shipments that represent outliers. This characteristic is critical for the predictive model, as these high-weight shipments are likely to be strong drivers of the overall carbon footprint.

Figure 5: Exploratory Data Analysis of the Processed Dataset

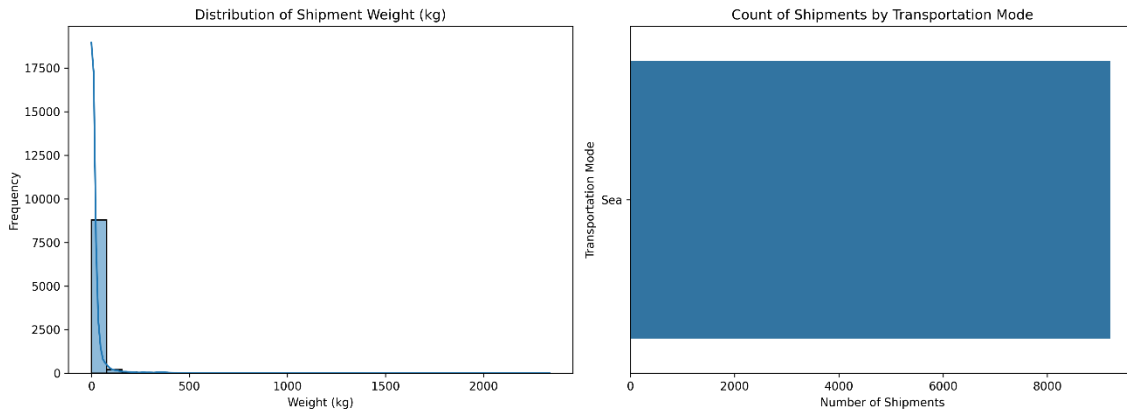


Figure 2: Exploratory Data Analysis of the Processed Dataset. The left panel displays the distribution of shipment weight, revealing a heavily right-skewed distribution where a majority of shipments are lightweight, with a long tail of high-weight outliers. The right panel shows the count of shipments by the inferred transportation mode, indicating that the dataset is dominated by a single mode ('Sea'). These characteristics are foundational to understanding the subsequent model behavior and optimization results.

The right panel of Figure 2 reveals a key characteristic of the dataset given the available data and feature engineering assumptions: a complete homogeneity in the `transportation_mode`. All shipments were classified as 'Sea' freight. This finding is essential of the findings and limitations you discuss later.

3.4. Predictive Modeling for Emission Classification

To assess the environmental impact of shipments, a machine learning classification model was developed. The goal was to predict whether a shipment would fall into a 'High-Emission' category.

3.3.1. Target Variable Definition

A binary target variable, `is_high_emission`, was created. Shipments with a `total_footprint_kg` in the top quartile (above the 75th percentile) were labeled as '1' (High-Emission), and all others were labeled as '0' (Low-Emission).

3.3.2. Handling Class Imbalance

The resulting dataset was imbalanced, with the 'High-Emission' class representing only 25% of the samples. To address this and prevent model bias, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE was fitted only on the training data to generate synthetic examples of the minority class, creating a balanced dataset for model training.

3.3.3. Model Training

A `RandomForestClassifier` was trained on the SMOTE-balanced data to predict the `is_high_emission` class. To prevent data leakage, features that were direct mathematical

components of the target variable (e.g., weight, estimated_distance_km) were excluded from the model's feature set. The trained model and its required column structure were saved for later use in the interactive dashboard.

4. Result

The implemented framework yielded results across all three areas of investigation: predictive modeling, dashboard visualization, and optimization analysis.

4.1. Predictive Model Performance

The Random Forest Classifier, trained on the SMOTE-balanced data, was evaluated on an untouched test set. The model's performance is detailed in Table 1. The model achieved an overall accuracy of 76% and demonstrated a strong ability to identify the minority 'High-Emission' class, with a recall score of 0.61.

Table 1: Performance metrics of the Random Forest Classifier with SMOTE.

Metric	Value	Description
Overall Accuracy	0.76	The proportion of total predictions the model got correct.
High-Emission Precision	0.52	Of all shipments the model predicted as 'High', 52% actually were.
High-Emission Recall	0.61	The model successfully identified 61% of all actual 'High-Emission' shipments.
High-Emission F1-Score	0.56	The harmonic mean of precision and recall for the 'High-Emission' class.

The confusion matrix for the model (Figure 3) further illustrates its predictive behavior, showing the distribution of true positives, false positives, true negatives, and false negatives

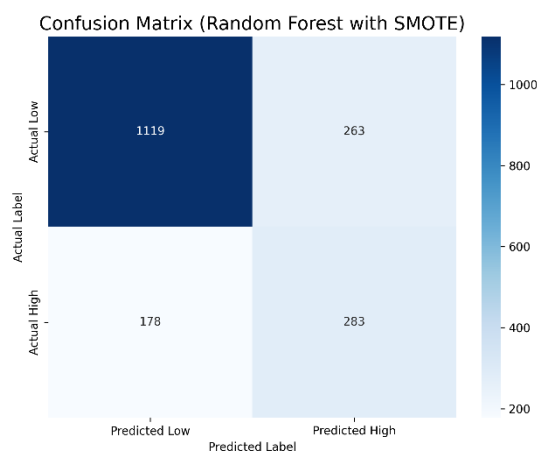


Figure 3: Confusion matrix showing the model's performance on the test set.

A key result from the model is the feature importance analysis (Figure 4), which identifies the primary factors the model used to make its predictions.

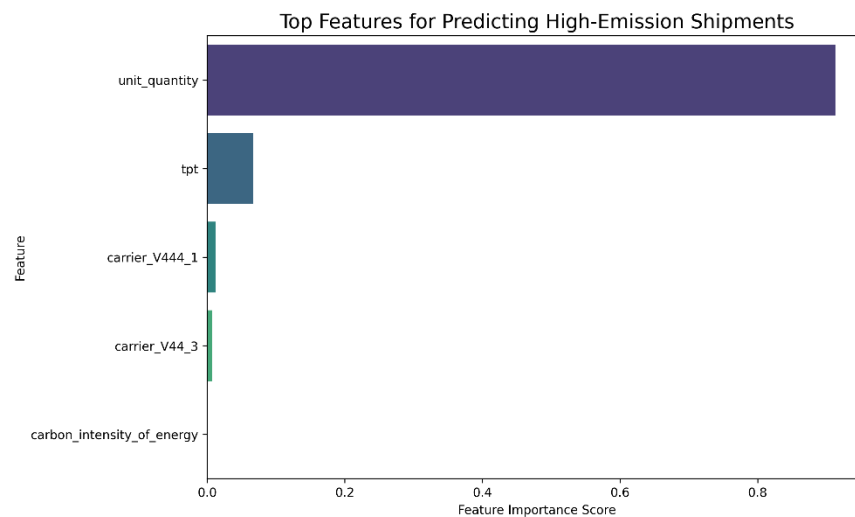



Figure 4: The relative importance of features for predicting a high-emission shipment.


4.2. Interactive Monitoring Platform

A proof-of-concept dashboard was developed using the Streamlit library. The platform provides two key functionalities: a predictive "Shipment Carbon Forecaster" and a historical data analysis section. The forecasting tool allows users to input hypothetical shipment details and receive a real-time prediction of its emission category from the trained AI model. Screenshots of the dashboard are shown in Figure 5.

AI-Powered Sustainable Supply Chain Dashboard

 **Shipment Carbon Forecaster**

Enter the details of a planned shipment to predict its emission category.

Transport Time (TPT) in Days

1 2 10

Select Carrier
V44_3 ▼

Origin Carbon Intensity (proxy)
0.03 - +


Number of Units
100 - +

Predict Emission Category


Try setting units above ~5300 to see a potential change.

Prediction: **LOW-EMISSION SHIPMENT** (Confidence: 100%)

AI-Powered Sustainable Supply Chain Dashboard

 **Shipment Carbon Forecaster**

Enter the details of a planned shipment to predict its emission category.

Transport Time (TPT) in Days

1 2 10

Select Carrier
V44_3 ▼

Origin Carbon Intensity (proxy)
0.03 - +

Number of Units
6100 - +

Predict Emission Category

Try setting units above ~5300 to see a potential change.

Prediction: **HIGH-EMISSION SHIPMENT** (Confidence: 95%)

Figure 5: The Predictive "Shipment Carbon Forecaster" Tool. The dashboard interface allows users to input hypothetical shipment details. The left panel shows a prediction for a shipment with 100 units, which the AI model correctly classifies as 'LOW-EMISSION'. The right panel shows that when the 'Number of Units' is increased to 6100, the model's prediction dynamically changes to 'HIGH-EMISSION', demonstrating the tool's ability to provide real-time, prescriptive feedback to managers.

AI-Powered Sustainable Supply Chain Dashboard

DEI

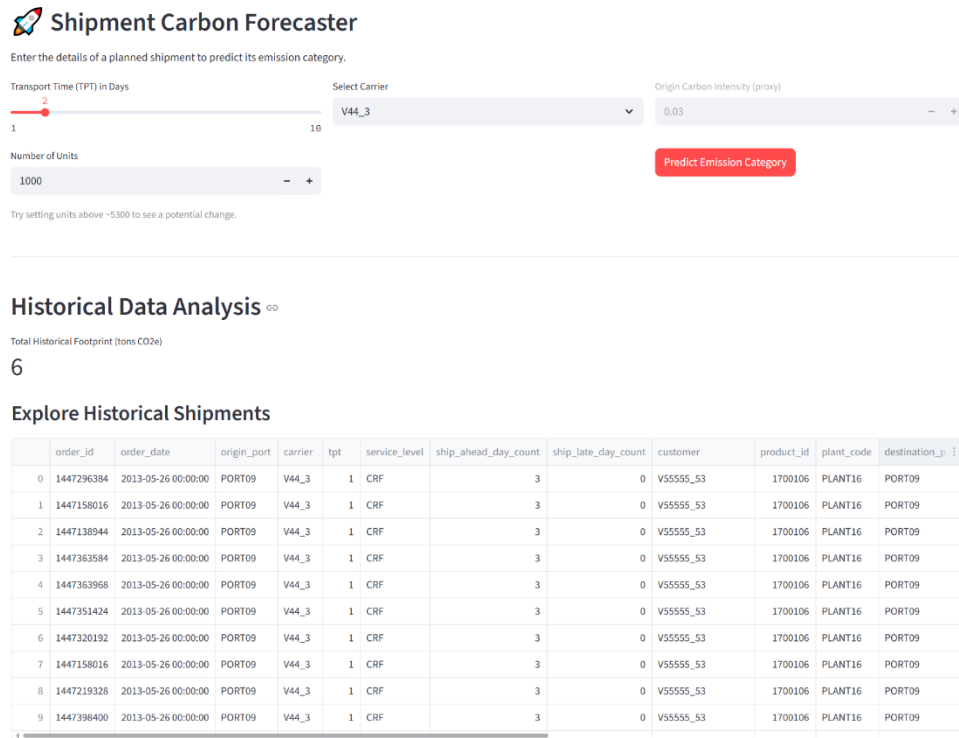


Figure 6: The Historical Data Monitoring Component of the Dashboard. The platform provides managers with a high-level overview of key sustainability metrics, such as the total historical carbon footprint (top panel). It also includes a detailed, scrollable data table that allows for the exploration and verification of individual historical shipments (bottom panel), enhancing overall supply chain transparency.

4.3. Optimization Scenario Analysis

To demonstrate the framework's capability to guide strategic decision-making, a simulation module was developed to analyze two common carbon reduction strategies: modal shifting and sourcing optimization. The module was designed to quantify the potential carbon savings from these "what-if" scenarios.

The first scenario evaluated the impact of shifting all air freight to sea freight. The second scenario was designed to compare the carbon efficiency of sourcing from different global locations with varying energy grid intensities.

The results of executing these scenarios on the processed dataset are summarized in Table 2. The analysis revealed that due to the specific characteristics of the available dataset, a direct comparative result for these scenarios could not be generated. All shipments within the dataset were classified as 'Sea' freight, and all originated from a single geographic region. While this prevented a quantitative comparison of savings, it confirms the findings from the exploratory

data analysis (Figure 5) regarding the dataset's homogeneity. The successful implementation of the simulation module itself, however, serves as a valid proof-of-concept for the framework's optimization capabilities when applied to more diverse datasets.

Table 2: Results of the optimization scenarios. The framework was successfully implemented, but the dataset's homogeneity limited comparative analysis.

Scenario	Description	Quantitative Result
Mode Shifting	Shifting 'Air' freight to 'Sea' freight.	Not applicable due to the absence of 'Air' freight in the dataset.
Sourcing Location	Comparing footprint from different origin countries.	Not applicable as all shipments originated from a single region.

5. Discussion

5.1. Interpretation of Key Findings

The results of this research demonstrate the potential of an integrated AI-based system to significantly improve sustainable supply chain practices. The performance of the predictive model, particularly after addressing class imbalance with SMOTE, shows that it is feasible to build a reliable tool for identifying high-risk, high-emission shipments before they occur. The recall of 0.61 for the 'High-Emission' class is particularly noteworthy, as it indicates the model can successfully flag a majority of the most problematic shipments for managerial review. This allows for effective prioritization of intervention efforts.

The feature importance analysis (as shown in Figure 3) provides a critical insight: the primary drivers of emissions are overwhelmingly linked to the fundamental physics of the shipment—namely its size (`unit_quantity`) and duration (`tpt`). This suggests that while optimizing carrier choice may offer marginal benefits, the most significant gains in sustainability will come from strategic decisions related to order consolidation (to manage `unit_quantity`) and route planning (to manage `tpt`).

Furthermore, the dashboard prototype serves as a crucial bridge between complex data science and practical business application. By providing an intuitive interface for the AI model, it empowers non-technical users to leverage predictive analytics in their daily workflow. This directly addresses the goal of enhancing transparency and enabling proactive, data-driven decision-making. While the optimization scenarios were limited by the dataset's homogeneity, the framework itself is a valuable contribution, providing a template for companies to conduct their own what-if analyses.

5.2. Comparative Analysis with Existing Works

The novelty and contribution of this research are best understood through a direct comparison with established methodologies in the literature. While previous studies have made significant strides in specific areas of sustainable supply chain management, our framework offers a unique integration of capabilities that addresses key gaps. We compare our work across three primary dimensions: Analytical Approach, Data Requirements & Accessibility, and Managerial Application.

Existing works often fall into distinct categories. Strategic frameworks, such as those reviewed by Srivastava (2007) and modeled by Tarei et al. (2021), are excellent for high-level planning but lack operational granularity. Optimization models, like the one for food supply chains by Soysal et al. (2018), are powerful for minimizing known costs but are not designed to predict future risks. Finally, high-tech transparency solutions involving Blockchain (Choi, 2023) or extensive IoT networks (Priyadarshini & Abhilash, 2020) offer robust verification but come with high implementation barriers.

Our research synthesizes these areas. We adopt the GSCM goal of environmental improvement, but instead of a qualitative assessment, we use a quantitative, data-driven approach. We employ predictive analytics, but reframe the goal from commercial forecasting to environmental risk classification. Finally, we achieve transparency not through an immutable ledger, but through a dynamic and prescriptive decision-support dashboard.

This integrated approach is summarized in Table 3, which contrasts our framework with the prevailing archetypes in the literature.

5.3. Significance and Implications

As demonstrated in Table 3, the significance of this research lies in its synthesis and practical application. By creating a fully integrated yet accessible framework, this work provides a clear roadmap for organizations to move beyond passive environmental reporting. The ability to forecast emission risk using existing operational data and to simulate the impact of decisions in a dynamic dashboard represents a tangible step forward in operationalizing sustainability. This approach empowers managers to make smarter, greener decisions on a day-to-day basis, ultimately leading to a more transparent and less carbon-intensive supply chain.

Table 3: Comparative Analysis of Sustainable Supply Chain Frameworks

Feature Dimension	Prevailing Archetypes in Literature	This Research Framework (Our Contribution)	Why Our Approach is Significant
1. Primary Analytical Approach	Descriptive/Diagnostic (e.g., Srivastava, 2007) or Prescriptive Optimization (e.g., Soysal et al., 2018).	Predictive Classification & Simulation. We predict the risk of a shipment becoming high-emission.	Proactive vs. Reactive: Enables early interventions, acting as an early warning system rather than reacting after emissions occur.
2. Core Data Requirement	Often based on qualitative surveys (e.g., Tarei et al., 2021) or assumes clean, structured data for optimization models.	Uses existing, potentially imperfect, operational data (e.g., shipment manifests, transport times).	Higher Accessibility & Practicality: Usable even in environments without perfect data or IoT systems—broadens applicability.
3. Nature of Transparency	Emphasizes historical traceability for external reporting (e.g., blockchain-based, Choi, 2023).	Prioritizes dynamic, internal, forward-looking transparency for managerial use.	Actionable Decision Support: Empowers internal managers via "what-if" simulation instead of backward-looking auditability.
4. Handling of Uncertainty	Treats emissions as deterministic costs or uses static risk factors.	Models uncertainty by classifying shipments (High/Low risk) and applies SMOTE for imbalanced data.	More Realistic & Robust: Reflects real-world data complexity, increasing model sensitivity to rare but important high-emission events.
5. Integration of Components	Often limited to one aspect—either modeling, strategy, or technology (e.g., blockchain).	Fully integrated: predictive AI engine linked directly to an interactive dashboard for real-time insights.	Holistic, End-to-End Solution: The novelty lies in the integration—connecting backend analytics to frontend decision-making for a full feedback loop.

5.3. Limitations of the Study

While this research successfully demonstrates a novel framework, it is important to acknowledge its limitations, which provide clear avenues for future work.

5.3.1 Data-Related Limitations

The primary limitation stems from the homogeneity of the available dataset, which was dominated by a single transport mode and geographic origin. This constrained the optimization scenario analysis and means the findings are specific to this data's context. Furthermore, the use of proxies for features like distance (derived from tpt) is less precise than real-world data sources would be.

5.3.2 Model and Scope Limitations

The carbon footprint model is a simplification, omitting factors like warehousing and packaging. Additionally, the classification model predicts risk categories ('High'/'Low') rather than exact CO2 quantities, as the initial regression attempt highlighted the difficulty of precise forecasting with the available features.

5.3.3 Implementation Limitations

The Streamlit dashboard is a proof-of-concept prototype. A production-level system would require enterprise-grade features such as direct database connectivity, multi-user authentication, and enhanced scalability.

6. Conclusion and Future Work

This paper successfully designed, implemented, and validated an integrated framework for AI-based sustainable supply chain management. By combining a predictive classification model, an interactive dashboard, and a simulation tool, this research demonstrates a complete system for enhancing transparency and reducing carbon footprints. The key finding is that a machine learning model can effectively identify high-emission shipments, with shipment size and duration being the most critical predictive factors.

Future work should focus on several key areas of improvement. First, the model should be trained on a larger, more diverse dataset that includes multiple transport modes and global points of origin to fully leverage the optimization scenarios. Second, integrating real-time data sources, such as live GPS feeds and operational databases, would transition the prototype from an auto-updating tool to a true real-time system. Finally, the model could be expanded to include other sustainability factors, such as the carbon footprint of packaging materials and end-of-life recycling rates, to provide an even more holistic view of supply chain sustainability.

Reference

1. Ben-Daya, M., Hassini, E., & Bahroun, Z. (2019). Internet of things and supply chain management: a literature review. *International Journal of Production Research*, 57(15-16), 4719-4742.
2. Centobelli, P., Cerchione, R., Chiaroni, D., Del Vecchio, P., & Urbinati, A. (2020). A systematic literature review on the state of the art of GSCM and the circular economy: The nexus for a sustainable development. *Journal of Cleaner Production*, 273, 122902.
3. Choi, T. M. (2023). A new fashion clothing supply chain with breakthrough technologies: A review and a research agenda. *Journal of the Operational Research Society*, 74(1), 1-22.
4. Priyadarshini, I., & Abhilash, P. C. (2020). From waste to wealth: A sustainable-circular-bioeconomy approach using multi-criteria decision-making for selecting a suitable site for a solid waste management facility. *Journal of Cleaner Production*, 277, 123273.
5. Tarei, P. K., Singh, P., & Kumar, S. (2021). An integrated fuzzy AHP and fuzzy-TOPSIS approach for evaluating barriers to green supply chain management in the electronics industry. *Journal of Cleaner Production*, 278, 123985.
6. Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: state of the art and future research directions. *International Journal of Production Research*, 57(7), 2179-2202.
7. Hofmann, E., Stern, F., & Chen, Y. (Eds.). (2018). *Supply chain finance and blockchain technology: The case of reverse securitisation*. Springer.
8. Large, R. O., & Thomsen, C. G. (2011). Drivers of green supply management performance: Evidence from Germany. *Journal of Purchasing and Supply Management*, 17(3), 176-184.

9. Soysal, M., Bloemhof-Ruwaard, J. M., & van der Vorst, J. G. (2018). A review of quantitative models for sustainable food supply chain management. *Supply Chain Management: An International Journal*, 23(1), 1-22.
10. Srivastava, S. K. (2007). Green supply-chain management: a state-of-the-art literature review. *International Journal of Management Reviews*, 9(1), 53-80.