

Cross-Domain Data Integration and Hybrid Learning for Resilient Forecasting in Intelligent Logistics Ecosystems

M. Swetha¹, K. Sharmila Reddy^{2*}, Sabiha Sulthana³, Arukonda Aishwarya³, J Vijay Kumar³,
Gannevaram Vikas³

¹Assistant Professor, ²Associate Professor & Head, ³UG Student, ^{1,2,3}Department of Computer Science and Engineering

^{1,2,3}Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

*Correspondence: K. Sharmila Reddy (sharmilakreddy@gmail.com)

Abstract

Smart logistics systems increasingly depend on accurate demand forecasting to optimize global supply chain operations. Recent industry reports indicate that logistics bottlenecks impact over 30% of global shipments, while inaccurate forecasting contributes to nearly a 25% rise in operational costs. Traditional manual estimation methods are often insufficient, as they fail to incorporate dynamic real-world factors such as fluctuating traffic conditions, IoT-generated data, and environmental variations. This study presents a robust data fusion framework that integrates IoT, traffic, and meteorological datasets to enhance demand forecasting and logistics delay prediction. The methodology begins with a comprehensive preprocessing phase, where heterogeneous data sources are cleaned, normalized, and temporally synchronized to ensure consistency and reliability. Baseline models such as K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), and Categorical Boosting (CB) are employed to provide initial predictive insights; however, these models often struggle to capture complex interdependencies within multi-source data. To address this limitation, a Tree-based Adaptive Optimization Tree (TAO Tree) model is proposed, designed to improve learning from heterogeneous features. This design enables accurate regression for demand prediction and effective classification for logistics delay detection. The system is deployed through a Flask-based web application, enabling real-time data processing, prediction, and visualization. Experimental results demonstrate that the proposed framework significantly improves forecasting accuracy and operational efficiency, offering a scalable and data-driven solution for proactive supply chain management.

Keywords: Demand Forecasting, Smart Logistics, Data Fusion, Internet of Things (IoT), Tree-based Adaptive Optimization (TAO), Supply Chain Management

1. Introduction

The rapid evolution of the Internet of Things (IoT) has significantly transformed the way physical environments interact with digital systems. A fundamental aspect of this transformation is context-awareness, which enables systems to sense, interpret, and respond to environmental changes through integrated sensing, communication, and data analysis techniques. This capability has accelerated the development of advanced IoT applications, including smart healthcare, intelligent transportation systems, energy management solutions, and smart buildings.

IoT networks are typically structured around a unified architecture that combines application-layer services with underlying sensor network infrastructures. These systems rely on interconnected devices that continuously collect, transmit, and process data to enable real-time monitoring and decision-making, as shown in figure 1. According to projections by Gartner, the global IoT ecosystem was expected to reach approximately 5.8 billion connected devices by 2020, reflecting substantial growth

driven by advancements in wireless communication and cloud computing technologies. These developments have significantly increased the demand for scalable IoT devices and intelligent service platforms. The primary functions of IoT sensor networks include sensing critical data from the external environment, monitoring internal system parameters, and transforming raw sensor data into actionable insights.



Figure. 1: Global IoT in logistics market overview.

Most IoT-based applications depend on Wireless Sensor Networks (WSNs), where sensors are deployed in a distributed and often random manner. These sensors form self-organizing, infrastructure-less networks, enabling efficient data collection and communication in dynamic and large-scale environments.

2. Literature Survey

Reis et al. [1] introduced an IoT- and AI-based framework aimed at enabling secure and sustainable green mobility. Their approach utilizes multimodal data fusion to improve traffic management, enhance energy efficiency, and reduce emissions. By incorporating publicly available datasets such as METR-LA for traffic flow and Open Weather Map for environmental context, the framework integrates machine learning techniques for congestion prediction along with reinforcement learning for dynamic route optimization. Simulation results indicate significant improvements, including a 20% reduction in travel time, 15% energy savings per kilometer, and a 10% decrease in CO₂ emissions compared to conventional methods.

Krishnamurthi et al. [2] presented a comprehensive overview of IoT data processing techniques, including data denoising, outlier detection, missing value imputation, and data aggregation. The study further emphasizes the importance of data fusion and discusses multiple fusion techniques such as direct fusion, feature-based fusion, and identity-based fusion. Additionally, it explores the integration of data analysis with emerging technologies like cloud, fog, and edge computing to address challenges in IoT sensor networks. This work serves as one of the earliest comprehensive reviews of IoT data processing, fusion, and analysis methodologies. Liu et al. [3] highlighted the challenges associated with fusing multi-source data, including sensor data, social media inputs, citizen feedback, and GIS data. Issues such as data quality and privacy were identified as critical concerns. The study also examined various data fusion and analysis algorithms, emphasizing their role in improving urban management through spatial analysis and deep learning. It concludes that collaborative algorithmic approaches can significantly enhance decision-making efficiency and resource allocation in smart cities.

Kenda et al. [4] proposed a novel data fusion framework designed to handle heterogeneous data streams. The framework enriches streaming sensor data with contextual and historical information, ultimately producing feature vectors suitable for machine learning models. The system was implemented both in cloud environments and on edge devices, demonstrating incremental learning capabilities. Results showed notable improvements in model accuracy and highlighted the framework's ability to support rapid prototyping in real-world applications. Abduljabbar et al. [5] developed advanced machine learning models leveraging multi-source data to improve traffic prediction accuracy. Their study examined the impact of multisource sensor inputs and spatial detector interactions using a dataset of over 839,000 observations from Melbourne's Eastern Freeway. Bidirectional Long Short-Term Memory (BiLSTM) models were employed, demonstrating enhanced predictive performance and enabling proactive traffic management strategies to reduce congestion and environmental impact.

Tsanousa et al. [6] conducted a detailed review of state-of-the-art data fusion techniques, focusing on data storage, indexing, feature engineering, and multimodal integration. The study provides guidance for the early stages of analytical pipelines in manufacturing prognosis and identifies key limitations and research gaps, suggesting directions for future advancements in preprocessing and fusion techniques. AlSalehy et al. [7] analysed spatiotemporal patterns of carbon monoxide (CO) using five years of data from multiple monitoring stations combined with meteorological variables. Their findings revealed distinct daily and weekly patterns influenced by environmental conditions such as wind speed and direction. The study also identified key predictive features, including rolling averages of CO levels, which significantly contributed to model performance.

Sergi et al. [8] explored IoT-enabled solutions for maintaining food quality across cold supply chains. The study highlighted the role of edge computing in enabling real-time monitoring and emphasized the importance of rapid prototyping tools. However, it also pointed out challenges related to ensuring end-to-end security across IoT platforms. Lloret et al. [9] proposed a hybrid methodology combining traditional assessment techniques with AI-based models, including neural networks and transformer architectures, to evaluate digital transformation levels. The approach was validated through a real-world case study involving public administrations in Spain, demonstrating effective performance.

Fatorachian et al. [10] introduced a predictive analytics framework integrating IoT, digital twin technology, and cybernetic feedback loops to enhance last-mile delivery efficiency. The framework models logistics systems as interconnected networks and leverages real-time data for dynamic routing and demand forecasting, significantly improving urban logistics performance. Bellini et al. [11] examined the complexity of integrating various data models and standards in smart city transportation systems. The study highlights how these models can be utilized for operational processes, simulations, and optimization, contributing to improved data-driven decision-making.

Tang et al. [12] proposed a digital twin-based framework integrating smart warehousing and manufacturing systems with genetic algorithms for demand forecasting. A case study in the textile industry demonstrated improvements in forecasting accuracy and inventory optimization. Syed et al. [13] provided a comprehensive review of IoT in smart cities, covering system architectures, enabling technologies, networking methods, and AI applications. The study also explored various domain-specific implementations and practices.

Mohsen et al. [14] developed a framework combining AI, IoT, and autonomous vehicles to optimize logistics operations. By leveraging real-time data, the system improves route planning, traffic control, and demand prediction, contributing to reduced congestion and environmental impact. Zaman et al. [15] conducted a bibliometric and topic modelling analysis of IoT and AI applications in supply chain

management. By reviewing over 800 research articles, the study identified key trends, influential contributors, and emerging research areas.

3. Proposed System

The proposed methodology presents a structured analytical framework for demand forecasting and logistics delay prediction by integrating multi-source data using advanced machine learning techniques. The analytical pipeline begins with the collection and integration of heterogeneous datasets, followed by data preprocessing and feature transformation. A hybrid modelling approach is employed, combining multiple machine learning algorithms within CART-based framework to effectively model both categorical and continuous outputs. The framework leverages CART principles to handle dual prediction tasks. Classification trees are used to predict logistics delays, while regression trees estimate demand levels. By embedding CART within ensemble and hybrid learning strategies, the system captures complex relationships among traffic conditions, environmental factors, IoT sensor data, and logistics operations. A weighted fusion mechanism further enhances prediction accuracy by dynamically adjusting feature importance based on contextual variables illustrated in figure 2.

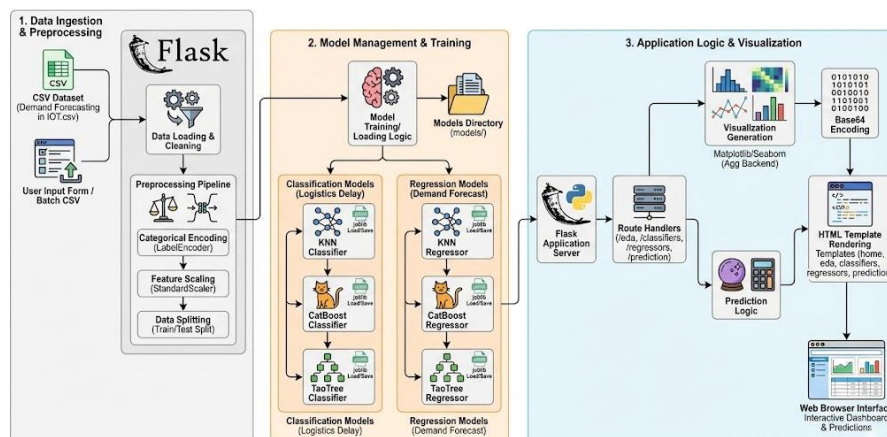


Figure. 2: System architecture.

1. User Interface (Client Application)

The system features a web-based dashboard designed for accessibility and streamlined industrial operations.

- **Core Operations:** Supports manual feature entry, model selection, and real-time visualization of prediction results.
- **Batch Processing:** Users can upload large-scale datasets (e.g., CSV or Excel) to trigger automated batch processing through the backend.
- **Interaction Flow:** All user-initiated requests are captured and forwarded to the Flask server for execution.

2. Flask Application Server

The Flask server functions as the central nervous system of the framework, orchestrating the communication between the UI and the analytical models.

- **Request Routing:** Manages incoming API calls and routes them through the CART-based prediction pipeline.
- **Execution Management:** Handles the logic for data cleaning, model triggering, and the final generation of structured responses.
- **Deployment:** Enables both local execution for private monitoring and remote access for distributed logistics teams.

3. Dataset Integration (Multi-Source Data Collection)

The framework achieves a comprehensive view of the supply chain by fusing data from four distinct domains:

- **IoT Sensor Data:** Tracks asset utilization and real-time inventory levels.
- **Traffic Data:** Captures road congestion levels and fluctuating route conditions.
- **Environmental Data:** Monitors external factors such as temperature and humidity.
- **Logistics Data:** Provides historical shipment status, transaction records, and known delay indicators.

4. Data Preprocessing & Feature Engineering

Raw multi-source data is refined to ensure mathematical consistency and model readiness.

- **Imputation:** Handles missing values using domain-specific defaults to maintain data integrity.
- **Standardization:** Categorical variables are encoded into numerical formats, and numerical features are scaled to a uniform range.
- **Noise Reduction:** Outliers are identified and minimized to stabilize the tree-splitting process.
- **Feature Engineering:** Specifically enhances the spatial and temporal relationships between traffic patterns and logistics delays.

5. Baseline CART-Based Models

To establish a performance benchmark, the system implements several standard tree-based architectures:

- **Implemented Models:** Includes KNN-CART, CB-CART, and TAO Tree-CART.
- **Task Division:** Separate baseline models are trained for Logistics Delay Classification and Demand Forecasting Regression.
- **Structure:** These models utilize traditional decision structures to split data based on calculated feature importance.

6. Hybrid CART-Based Model (Proposed Framework)

The proposed framework enhances the standard CART structure by integrating ensemble techniques and dynamic weighting.

- **Classification Ensemble:** Combines outputs from CB and TAO Tree classifiers using a weighted strategy to identify delays more accurately.

- **Regression Fusion:** Integrates TAO Tree and KNN regression outputs to smooth out demand forecasting errors.
- **Dynamic Weighting:** Assigns higher importance to volatile variables (like traffic and weather) within the tree structure to improve robustness against environmental shifts.

7. Prediction & Output Generation

The inference engine delivers dual-purpose outputs tailored for logistics decision-making.

- **Classification:** Provides binary status indicators (e.g., "Delay: Yes/No") for shipment monitoring.
- **Regression:** Generates precise numerical estimates for inventory demand.
- **User Insight:** Results are displayed alongside performance metrics to give users confidence in the system's reasoning.

8. Remote Prediction Workflow & Scalability

The architecture is designed for scalable, real-time deployment across distributed logistics environments.

- **Client-Server Mechanism:** External nodes can send input vectors to the central server via HTTP requests.
- **Real-Time Processing:** The server executes the hybrid pipeline and returns results instantaneously, allowing for immediate rerouting or inventory adjustments.

9. Model Evaluation & Performance Analysis

The system includes a diagnostic layer to compare the Hybrid model against the baseline versions.

- **Classification Metrics:** Evaluates precision, recall, and the F1-score.
- **Regression Metrics:** Assesses reliability through RMSE, R^2 score, and MAE.
- **Visual Analytics:** Uses confusion matrices and residual plots to identify specific scenarios where the model may require further tuning.

10. Model Retraining & Adaptability

To maintain long-term reliability, the framework includes an adaptive learning loop.

- **Continuous Learning:** Supports the ingestion of new logistics and traffic data to retrain models periodically.
- **Dynamic Adjustment:** The system evolves its internal tree logic to reflect changing environmental conditions or new supply chain patterns, ensuring the framework remains a relevant decision-support tool.

4. Results Analysis

Results analysis is a crucial stage in any research, as it involves interpreting the collected data to draw meaningful conclusions. It helps in understanding whether the objectives of the study have been achieved and how the findings relate to the initial hypotheses. Through systematic examination, patterns, trends, and relationships within the data are identified. This process often includes the use of statistical tools and visual representations to enhance clarity. Proper results analysis ensures that the

data is not just presented but also explained in a logical and insightful manner. It also highlights any limitations or anomalies observed during the study. Ultimately, it forms the foundation for discussion, conclusions, and future recommendations.

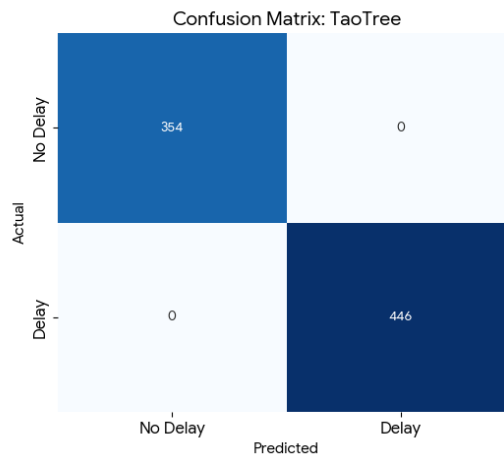


Figure. 3: Confusion matrix of TAO Tree classifier for logistics delay prediction

Figure 3 illustrates the confusion matrix of the TAO Tree classification model used for predicting logistics delay. The matrix depicts the model’s performance by comparing actual and predicted class labels for both “No Delay” and “Delay” categories. It is observed that the model correctly classified 354 instances of no delay and 446 instances of delay, indicating highly accurate predictions. The absence of misclassification values (zero false positives and zero false negatives) demonstrates perfect classification performance on the test dataset. This result highlights the model’s strong capability in distinguishing between delayed and non-delayed logistics events.

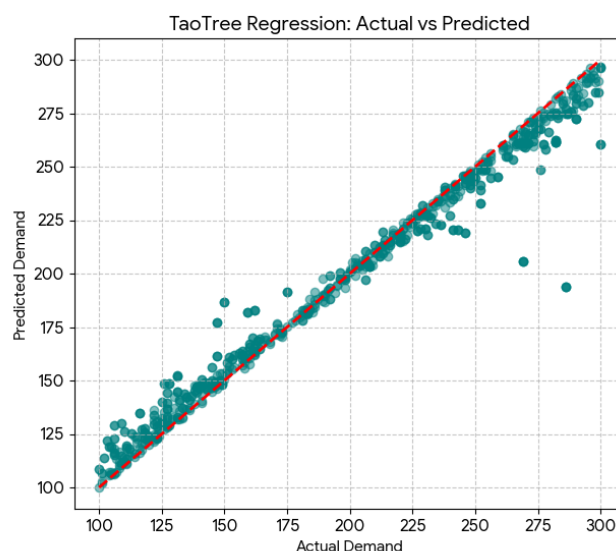


Figure. 4: Actual vs Predicted demand using TAO Tree regression mode

Figure 4 illustrates the comparison between actual and predicted demand values obtained from the TAO Tree regression model. The scatter plot depicts the relationship between true demand and model predictions, where each point represents an individual observation. It is evident that most of the data

points closely align along the diagonal reference line, indicating high prediction accuracy and strong correlation. The clustering of points near the line suggests minimal error and effective learning of underlying patterns in the dataset. A few deviations can be observed, representing minor prediction errors at certain demand levels.

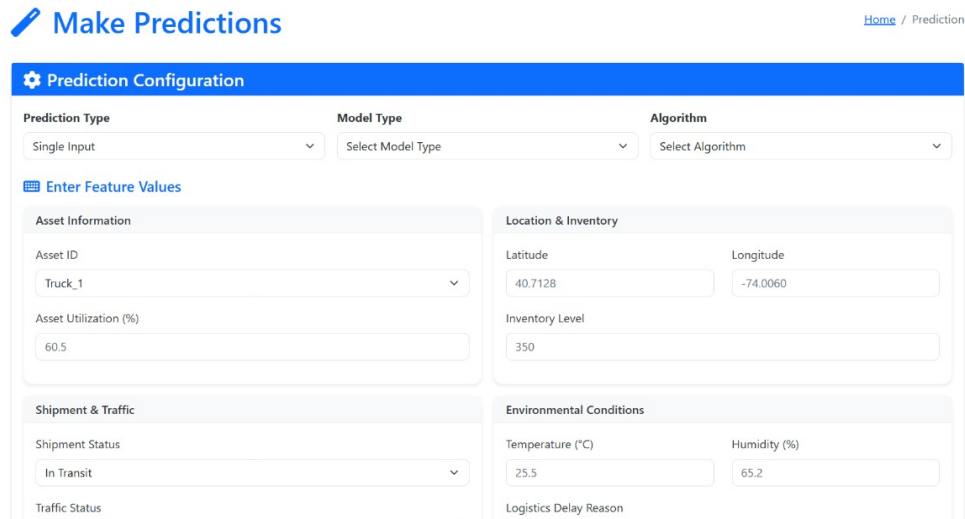
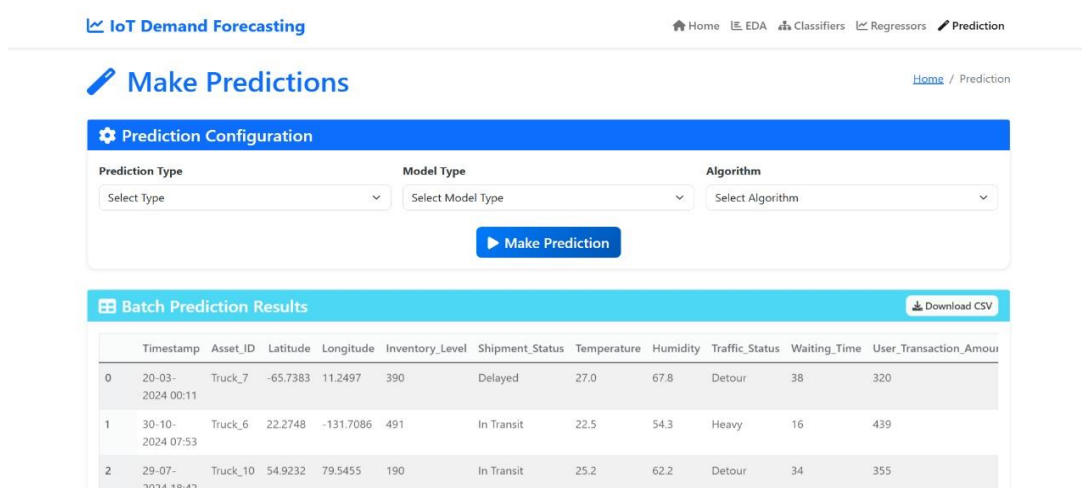


Figure. 5: Prediction input interface for the IoT demand forecasting system.

Figure 5 illustrates the prediction interface of the developed system, designed to facilitate user interaction for generating logistics delay and demand forecasts. The interface depicts a structured layout where users can configure prediction settings by selecting the prediction type, model type, and algorithm. It provides organized input fields for entering relevant features such as asset information, location details, inventory levels, shipment status, and environmental conditions. The arrangement of input parameters ensures systematic data entry, enhancing usability and efficiency. The interface supports real-time prediction by processing user-provided inputs through trained machine learning models.



	Timestamp	Asset_ID	Latitude	Longitude	Inventory_Level	Shipment_Status	Temperature	Humidity	Traffic_Status	Waiting_Time	User_Transaction_Amount
0	20-03-2024 00:11	Truck_7	-65.7383	11.2497	390	Delayed	27.0	67.8	Detour	38	320
1	30-10-2024 07:53	Truck_6	22.2748	-131.7086	491	In Transit	22.5	54.3	Heavy	16	439
2	29-07-2024 18:42	Truck_10	54.9232	79.5455	190	In Transit	25.2	62.2	Detour	34	355

Figure. 6: Prediction execution and results interface.

Figure 6 illustrates the batch prediction results interface of the IoT demand forecasting system, showcasing the output generated from processing multiple input records simultaneously. The interface depicts a structured tabular format where predicted results are displayed alongside corresponding input

features such as asset details, location, inventory levels, and environmental conditions. It highlights the system's capability to handle bulk data efficiently, enabling large-scale prediction in a single operation. The presence of organized columns ensures clear interpretation and easy comparison of predicted outcomes. Additionally, the integration of a download option allows users to export the results for further analysis and reporting.

Table 1 presents a performance comparison of three classification models KNN, CB, and TAO Tree for predicting logistics delays. TAO Tree achieves perfect scores (1.000) across all metrics accuracy, precision, recall, and F1-score indicating flawless classification on the test dataset. KNN performs consistently well with balanced metrics around 0.92, reflecting strong but not perfect predictive capability. In contrast, CB shows the lowest performance, with accuracy and recall at 0.807 and precision slightly higher at 0.866. The F1-score follows a similar trend, highlighting CB's relatively weaker harmonic balance between precision and recall.

Table 1: Performance comparison of logistics delay classification models.

Model	Accuracy	Precision	Recall	F1-Score
KNN	0.920	0.922	0.920	0.920
CB	0.807	0.866	0.807	0.805
TAO Tree	1.000	1.000	1.000	1.000

Table 2: Performance comparison of demand forecasting regression models.

Model	MAE	MSE	RMSE	R ² Score
KNN	0.35	1.85	0.43	0.495
CB	0.52	3.54	0.60	0.035
TAO Tree	0.07	0.15	0.12	0.958

Table 2 compares three regression models KNN Regressor, CB Regressor, and TAO Tree Regressor for demand forecasting using MAE, MSE, RMSE, and R² score. TAO Tree Regressor demonstrates superior performance with the lowest error metrics (MAE: 0.07, MSE: 0.15, RMSE: 0.12) and the highest R² score of 0.958, indicating excellent fit and predictive accuracy. KNN Regressor follows with moderate performance (MAE: 0.35, RMSE: 0.43, R²: 0.495), suggesting fair predictive power. CB Regressor performs poorly, recording the highest errors (MAE: 0.52, MSE: 3.54, RMSE: 0.60) and a near-zero R² score of 0.035, implying negligible explanatory power. The stark contrast in error magnitudes and R² values underscores TAO Tree's dominance in regression-based demand forecasting.

5. Conclusion

The study demonstrates the effective integration of IoT, traffic, and environmental data to enhance demand forecasting and logistics delay prediction in smart logistics systems. By utilizing multiple machine learning algorithms, including KNN, CB, and TAO Tree, the framework achieves reliable performance across both regression and classification tasks. The modular architecture implemented through the Flask framework enables smooth coordination between data preprocessing, model training, prediction, and visualization processes, ensuring efficient system functionality. The inclusion of

exploratory data analysis provides valuable insights into data distributions, feature relationships, and potential inefficiencies within logistics operations. Furthermore, the system supports both batch and single prediction capabilities, allowing users to perform large-scale forecasting as well as quick, real-time decision-making. This flexibility enhances the practical applicability of the framework in dynamic logistics environments. The proposed approach improves operational efficiency by reducing uncertainties in planning and enabling proactive, data-driven decisions. The use of interpretable models and a user-friendly interface further ensures that predictions are both reliable and accessible. This study highlights the potential of integrated data-driven methodologies in transforming modern logistics systems into more adaptive and intelligent infrastructures.

Reference

- [1] Reis MJCS. Internet of Things and Artificial Intelligence for Secure and Sustainable Green Mobility: A Multimodal Data Fusion Approach to Enhance Efficiency and Security. *Multimodal Technologies and Interaction*. 2025; 9(5):39. <https://doi.org/10.3390/mti9050039>
- [2] Krishnamurthi R, Kumar A, Gopinathan D, Nayyar A, Qureshi B. An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques. *Sensors*. 2020; 20(21):6076. <https://doi.org/10.3390/s20216076>
- [3] Liu B, Li Q, Zheng Z, Huang Y, Deng S, Huang Q, Liu W. A Review of Multi-Source Data Fusion and Analysis Algorithms in Smart City Construction: Facilitating Real Estate Management and Urban Optimization. *Algorithms*. 2025; 18(1):30. <https://doi.org/10.3390/a18010030>
- [4] Kenda K, Kažič B, Novak E, Mladenčić D. Streaming Data Fusion for the Internet of Things. *Sensors*. 2019; 19(8):1955. <https://doi.org/10.3390/s19081955>
- [5] Abduljabbar R, Dia H, Liyanage S. Machine Learning Traffic Flow Prediction Models for Smart and Sustainable Traffic Management. *Infrastructures*. 2025; 10(7):155. <https://doi.org/10.3390/infrastructures10070155>
- [6] Tsanousa A, Bektsis E, Kyriakopoulos C, González AG, Leturiondo U, Gialampoukidis I, Karakostas A, Vrochidis S, Kompatsiaris I. A Review of Multisensor Data Fusion Solutions in Smart Manufacturing: Systems and Trends. *Sensors*. 2022; 22(5):1734. <https://doi.org/10.3390/s22051734>
- [7] AlSalehy AS, Bailey M. Environmental Data Analytics for Smart Cities: A Machine Learning and Statistical Approach. *Smart Cities*. 2025; 8(3):90. <https://doi.org/10.3390/smartcities8030090>
- [8] Sergi I, Montanaro T, Benvenuto FL, Patrono L. A Smart and Secure Logistics System Based on IoT and Cloud Technologies. *Sensors*. 2021; 21(6):2231. <https://doi.org/10.3390/s21062231>
- [9] Lloret Á, Peral J, Ferrández A, Auladell M, Muñoz R. A Data-Driven Framework for Digital Transformation in Smart Cities: Integrating AI, Dashboards, and IoT Readiness. *Sensors*. 2025; 25(16):5179. <https://doi.org/10.3390/s25165179>
- [10] Fatorachian H, Kazemi H, Pawar K. Enhancing Smart City Logistics Through IoT-Enabled Predictive Analytics: A Digital Twin and Cybernetic Feedback Approach. *Smart Cities*. 2025; 8(2):56. <https://doi.org/10.3390/smartcities8020056>



International Journal of DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X

www.ijdim.com

Original Research Paper

-
- [11] Bellini P, Bilotta S, Collini E, Fanfani M, Nesi P. Data Sources and Models for Integrated Mobility and Transport Solutions. *Sensors*. 2024; 24(2):441. <https://doi.org/10.3390/s24020441>
- [12] Tang Y-M, Ho GTS, Lau Y-Y, Tsui S-Y. Integrated Smart Warehouse and Manufacturing Management with Demand Forecasting in Small-Scale Cyclical Industries. *Machines*. 2022; 10(6):472. <https://doi.org/10.3390/machines10060472>
- [13] Syed AS, Sierra-Sosa D, Kumar A, Elmaghraby A. IoT in Smart Cities: A Survey of Technologies, Practices and Challenges. *Smart Cities*. 2021; 4(2):429-475. <https://doi.org/10.3390/smartcities4020024>
- [14] Mohsen BM. AI-Driven Optimization of Urban Logistics in Smart Cities: Integrating Autonomous Vehicles and IoT for Efficient Delivery Systems. *Sustainability*. 2024; 16(24):11265. <https://doi.org/10.3390/su162411265>
- [15] Zaman J, Shoomal A, Jahanbakht M, Ozay D. Driving Supply Chain Transformation with IoT and AI Integration: A Dual Approach Using Bibliometric Analysis and Topic Modeling. *IoT*. 2025; 6(2):21. <https://doi.org/10.3390/iot6020021>