International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php

Leveraging IoT and Machine Learning to Optimize Public Transportation and Reduce Carbon Footprint

Jimson A. Olaybar[1], Jose C. Agoylo Jr.[2]
[1]Faculty of Computer Studies and Information Technology, Southern Leyte State University – Main Campus, Sogod, Southern Leyte, Philippines
[2]BSIT Department, Southern Leyte State University – Tomas Oppus Campus, Tomas Oppus, Southern Leyte, Philippines

ABSTRACT
Public transportation is a key strategy for lowering greenhouse gas emissions, yet inefficiencies in demand prediction and resource allocation reduce its effectiveness. This study presents a data-driven framework that integrates Internet of Things (IoT) behavioral data with machine learning (ML) models to improve transit efficiency and mitigate carbon emissions. The IoT-Carbon Footprint Dataset, containing 10,000 daily activity records on energy use, travel distance, and transport mode, was analyzed using regression, classification, and clustering techniques. Linear Regression showed strong predictive accuracy (R^2 = 0.854, MAE = 2.96), while Logistic Regression achieved 87.1% accuracy in classifying high- and low-emission groups. K-means clustering, despite low cohesion (Silhouette Score = 0.103), identified general mobility profiles such as habitual car users and public transport commuters. The results demonstrate that simple, interpretable ML models can provide robust insights for evidence-based policy, supporting low-emission mobility and efficient transit planning. This framework illustrates the practical application of IoT data and ML in advancing sustainable urban transportation.
Keywords *- Internet of Things (IoT), Machine Learning, Public Transportation, Carbon Footprint Reduction.*

I. INTRODUCTION

Green public transportation is now a global priority as cities face mounting traffic congestion, air pollution, and climate change issues. The transport sector is among the world's largest greenhouse gas emitters, and maximizing its efficiency is crucial to reducing the overall carbon footprint. Public transit systems, when managed efficiently, can significantly reduce emissions by substituting individual vehicles with shared mobility. Yet, scheduling inefficiencies, inadequate demand planning, and a lack of flexibility to respond to real-time conditions commonly hinder their ability to meet sustainability objectives.

The emergence of the Internet of Things (IoT) has created new avenues for gathering and analyzing mobility and energy data. IoT sensors installed in vehicles, smart meters, and GPS-enabled units create unbroken streams of data on passenger flows, energy consumption, and environmental conditions. These sources of data enable researchers and policymakers to track the impact of transport modes on carbon emissions at both individual and system levels.

Machine learning (ML) methods are crucial for interpreting large IoT datasets. Using regression, classification, and clustering models, it is possible to forecast carbon emissions, identify patterns in travel behavior, and inform transportation policy. Decision support systems based on ML can assist public transport planning through demand forecasting, encouragement of low-emission modes, and adaptive policy support.

While past studies have examined either the operational efficiency of transit systems or carbon footprint mitigation in energy and mobility environments, few studies have merged both aspects. This gap exists in recognizing how data mining, driven by IoT and ML predictions, can be integrated into a model that enhances transportation efficiency while also reducing carbon emissions.

Thus, this study proposes to create a data-centric framework that integrates transportation options with ML predictions to reduce the carbon footprint. Through the utilization of IoT datasets, this work provides practical insights for enhancing public transportation systems and promoting sustainable urban mobility.

II. LITERATURE REVIEW

Machine learning (ML) has been widely adopted to improve public transportation systems, particularly in predicting travel times, passenger demand, and service reliability. Early statistical and regression models often struggled to capture nonlinear and dynamic traffic conditions, which led to the development of more advanced ML approaches. Decision trees, support vector machines, and ensemble models have demonstrated higher accuracy in predicting passenger flows and identifying operational bottlenecks [1]; [3]. Deep learning approaches, such as recurrent neural networks (RNNs) and long short-term memory

(LSTM) networks, have further improved travel time and demand prediction by modeling temporal dependencies in transit data [4]. Moreover, smart card and GPS datasets have been analyzed using clustering and pattern mining techniques to uncover passenger behavior trends, peak travel times, and route-level inefficiencies that can inform better scheduling and resource allocation [5], [6].

In parallel, IoT-based data collection has enabled real-time monitoring of mobility and energy consumption, supporting research on sustainability and carbon footprint analysis. Studies have applied ML models to IoT-derived data, such as energy usage, GPS-based travel distance, and smart meter readings, to estimate and predict individual and system-level carbon emissions [7]. IoT sensors embedded in vehicles and infrastructure also provide continuous data streams for monitoring energy efficiency and environmental impact in transportation systems [8]. Researchers have demonstrated that integrating IoT data with ML enables more accurate forecasting of emissions and facilitates policy recommendations for sustainable transport planning [9]. Furthermore, IoT-based carbon monitoring frameworks for smart cities have shown the potential of linking user mobility behaviors with emission reduction strategies [10].

Recent studies highlight the role of ML in emission prediction. Udoh et al. [11] applied ML to predict CO_2 emissions in light-duty vehicles, while [12] extended this to multi-fuel innovative ships using onboard sensor data. Similarly, [13] demonstrated that LSTM-based models can accurately estimate CO_2 emissions from road transport. In parallel, Oladimeji et al. [14] provide a broader overview of IoT and ML applications in intelligent transportation systems, reinforcing the relevance of integrating IoT-driven datasets into sustainable mobility research

Although significant progress has been made in optimizing transit efficiency through ML [1]; [6] and analyzing carbon emissions using IoT data [7]; [10], these research directions have developed mainly in isolation. Most studies either emphasize operational accuracy in public transport optimization or focus on emission monitoring without addressing transit service improvements. Few works have attempted to integrate IoT-driven behavioral data with ML predictions to optimize public transit efficiency while reducing its carbon footprint jointly. This research addresses the gap by developing a framework that leverages IoT data and ML models to predict emissions, analyze transportation mode choices, and propose strategies for optimizing sustainable public transportation.

III. RESEARCH METHODOLOGY

A. Research Framework

This study adopts a data-driven framework that integrates IoT-derived behavioral data with machine learning models to optimize public transportation and reduce carbon emissions. The framework consists of four main stages:

i. Dataset selection and preprocessing – The dataset used in this study is sourced from Kaggle under the title IoT Carbon Footprint Dataset. It contains 10,000 entries, each representing an individual, and includes features related to energy consumption, transportation, and environmental conditions.

ii. Feature engineering and data transformation – Deriving meaningful features, normalizing continuous variables, and encoding categorical attributes.

iii. Machine learning model development – Implementing supervised (regression and classification) and unsupervised (clustering) models to predict and optimize carbon-efficient transport usage.

iv. Performance evaluation and interpretation – Assessing the models using appropriate evaluation metrics and interpreting their implications for sustainable public transportation.

B. Dataset Description

The IoT-Carbon Footprint Dataset is utilized as the primary data source. It contains 10,000 entries representing individuals' daily activities, with attributes such as energy usage, transportation distance, vehicle type, smart appliance usage, renewable energy percentage, environmental conditions (temperature and humidity), and estimated carbon emissions. Transportation-related features (e.g., vehicle type and travel distance) are emphasized to evaluate the impact of different modes of mobility—particularly public transportation—on carbon footprint.

C. Data Preprocessing

The dataset is cleaned to handle missing values, normalize continuous features (e.g., distance, energy usage), and encode categorical variables (e.g., vehicle type, building type). Outliers are identified and removed using interquartile range analysis to improve model robustness. An 80:20 train–test split is applied for model validation.

D. Machine Learning Models
Three categories of ML techniques are employed:

❖ Regression Models – Multiple linear regression and random forest regression are used to predict daily carbon emissions ($kgCO_2$) based on energy usage, travel distance, and transportation choice.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Explain: y is carbon footprint, x_i are predictors (distance, transport mode, energy use), β_i are coefficients, and ϵ is error.

❖ Classification Models – Logistic regression and support vector machines (SVM) are applied to classify individuals into "low-emission" and "high-emission" groups, supporting targeted interventions.

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Explain: This models the probability of a trip being "high-emission" vs. "low-emission."

❖ Clustering Models – K-means clustering is used to identify distinct travel behavior groups (e.g., habitual car users, frequent bus commuters, mixed-mode travelers), providing insights into potential mode-shift strategies.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Explain: k is the number of clusters, C_i is cluster i, and μ_i is the centroid.

E. Data Mining and Pattern Extraction
Beyond predictive modeling, association rule mining is conducted to uncover behavioral patterns, such as correlations between renewable energy usage, transport choices, and emission outcomes. These insights aim to inform public transport policies that promote bus usage and the integration of electric vehicles.

F. Performance Evaluation
The regression models are evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 score, while classification performance is assessed with accuracy, precision, recall, and F1-score. For clustering, silhouette score is applied to measure cluster cohesion and separation. The results are compared across models to identify the most effective ML techniques for emission prediction and transport optimization.

G. Implementation Tools
All experiments are implemented using Python, with libraries such as Scikit-learn for machine learning, Pandas and NumPy for data processing, and Matplotlib for visualization.

IV. RESULTS AND DISCUSSION

A. Regression Analysis

Table 1. Regression Results for Carbon Emission Prediction

Model	RMSE	MAE	R^2

Linear	3.76	2.96	0.854
Random Forest	3.94	3.08	0.840

Table I summarizes the regression results for carbon emission prediction using Linear Regression and Random Forest models. Both models demonstrated strong predictive power, with Linear Regression slightly outperforming Random Forest (R^2 = 0.854 vs. 0.840). The Root Mean Square Error (RMSE) values indicate that both models were able to estimate carbon emissions with minimal deviation from actual values. The lower Mean Absolute Error (MAE) of Linear Regression (2.96) further reinforces its robustness in capturing the linear relationship between travel distance, energy usage, and emission outcomes.

This scatter plot in Figure 1 is for comparing actual vs. predicted carbon emissions (in $kgCO_2$) based on a Random Forest regression model. Each of the blue dots is one data point where the x-coordinate is the actual carbon emission value and the y-coordinate is the prediction produced by the model. Ideally, the perfect predictions should lie on the red dashed diagonal line (line of equality where actual = predicted). The tight clustering of points around this line—particularly within the lower to mid-emission range—means that the model does well, although some deviation can be seen at higher levels of emissions where predictions are slightly lower than actual emissions (points lie below the line). This indicates that the model generalizes well in terms of capturing overall trends but fails to predict extreme or higher values, as is typical of complex nonlinear regressors such as Random Forest.

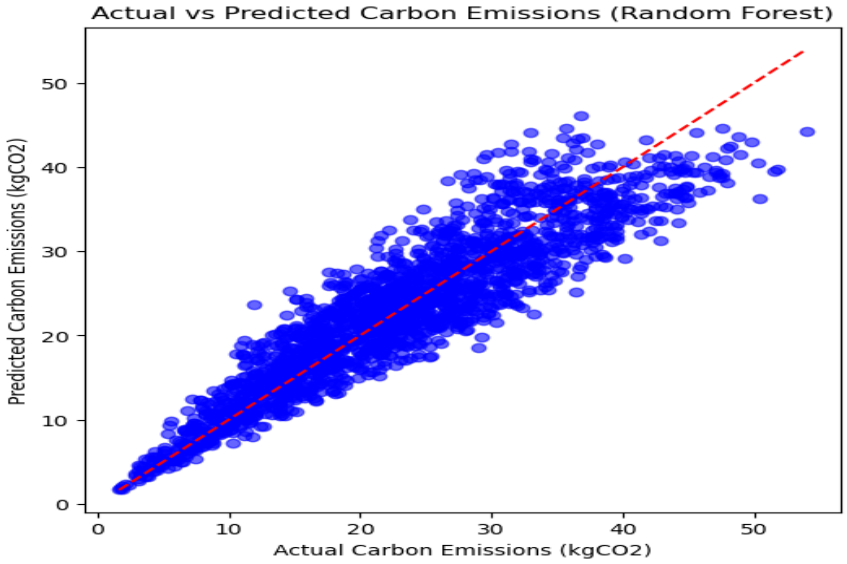


Figure 1. Actual vs Predicted Carbon Emissions

B. Classification Performance

Table 2. Result of the Evaluation Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.871	0.873	0.874	0.873
Random Forest	0.862	0.864	0.863	0.864
SVM (RBF Kernel)	0.860	0.857	0.869	0.863

The classification results in Table 2 demonstrate the effectiveness of Logistic Regression, Random Forest, and Support Vector Machine (SVM) in distinguishing between high-emission and low-emission individuals. Logistic Regression achieved the highest accuracy (87.1%) and F1-score (0.873), outperforming Random Forest and SVM. The results suggest that even a simple linear model is highly effective in this task, highlighting a strong separation between emission categories within the dataset.

This bar chart plots the performance of three machine learning classifiers—Logistic Regression, Random Forest, and Support Vector Machine (SVM)—on four evaluation metrics: Accuracy, Precision, Recall, and F1-score. Each model is shown with grouped bars in different colors representing the metrics, as labeled in the legend. In general, Logistic Regression performs marginally better on all four metrics with scores remaining around 0.87, showing balanced and stable performance. Random Forest has slightly lower values, approximately 0.86, with very close Precision, Recall, and F1-score measures, indicating consistent but slightly below-par predictions than Logistic Regression. SVM has the highest Recall, implying good sensitivity in detecting positive cases, but the trade-off is in terms of lower Precision, resulting in the lowest Accuracy among the three. This compromise is also evident in its F1-score, which is similar but lagging behind Logistic Regression. The plot indicates that although all three models have relatively good performance, Logistic Regression has the overall best compromise between classification metrics in this particular use case.

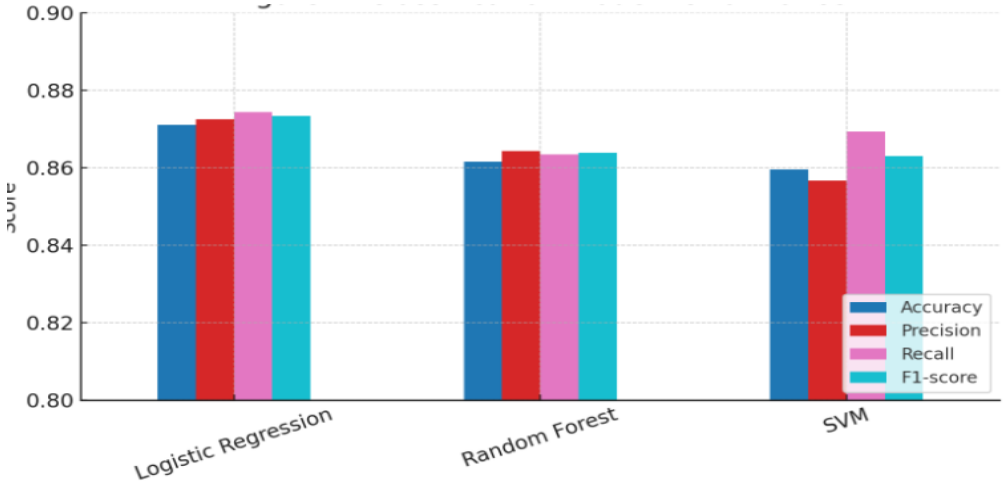


Figure 2. Classification Model Performance

C. Clustering Results

Table 3. Clustering Analysis

Model	**Silhouette Score**
K-Means	0.103

The clustering analysis using K-means yielded a Silhouette Score of 0.103, as shown in Table 3. This relatively low score suggests that the dataset lacks well-defined cluster boundaries, reflecting the complexity and overlap of individual travel behaviors. Despite this, clustering provides valuable exploratory insights, particularly in identifying broad patterns such as habitual car users, frequent public transport commuters, and mixed-mode travelers.

Although the numerical cohesion is limited, the analysis still revealed meaningful behavioral archetypes, such as habitual private car users, frequent public transport commuters, and mixed-mode travelers see Figure 3. These broad groupings are valuable for exploratory analysis, even if the mathematical separation is weak. From a policy perspective, identifying car-dependent segments is particularly important, as they represent high-emission groups that could benefit most from incentives to adopt public transportation or low-emission alternatives.

The low silhouette score can be attributed to two main factors: (i) the overlapping nature of urban mobility patterns, where many individuals cannot be assigned to a single homogeneous category, and (ii) the assumptions of K-means, which works best on spherical, well-separated clusters. In this context, clustering should not be viewed as a precise segmentation method, but rather as a diagnostic tool for uncovering broad mobility tendencies.

For future work, clustering performance could be enhanced by incorporating advanced methods such as DBSCAN, Gaussian Mixture Models (GMM), or spectral clustering, which are better suited for capturing irregular cluster shapes and fuzzy group boundaries. Additionally, the creation of derived features (e.g., emissions per kilometer, proportion of trips by public transport) may improve separability between traveler groups. Overall, while the clustering score is modest, the results demonstrate the potential of unsupervised learning to provide exploratory behavioral insights that complement the stronger regression and classification models.

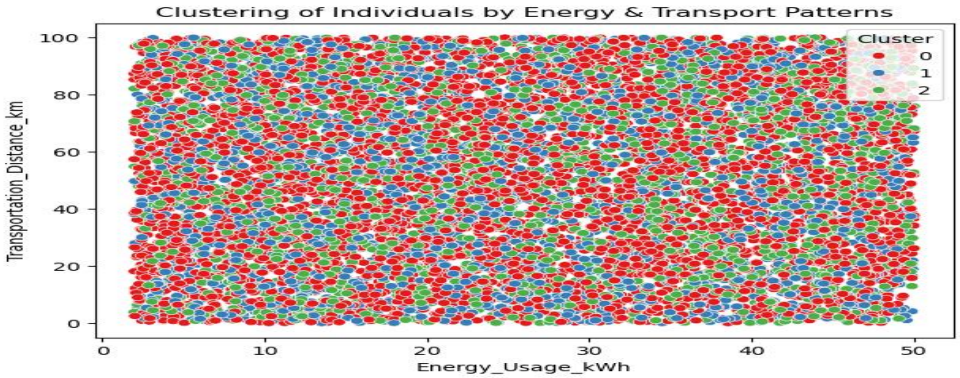


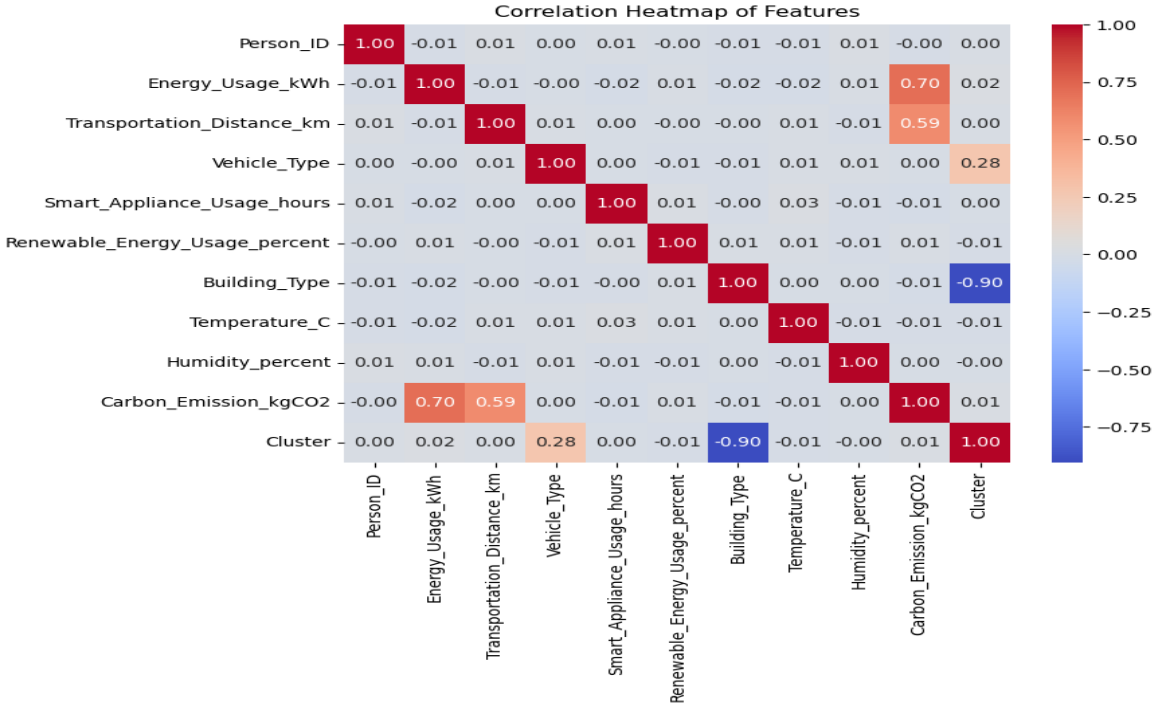
Figure 3. Clustering Performance

D. Heatmap Analysis

The results demonstrate that machine learning models can effectively capture and predict the relationship between transportation choices and carbon emissions see Figure 4. Linear Regression proved sufficient for regression tasks due to the strong linear trends in the dataset, while Logistic Regression was most effective for classification. The clustering results, though limited in cohesion, provided exploratory insights into behavioral segments that could guide targeted interventions.

The findings reinforce the potential of integrating IoT-derived data with machine learning to optimize sustainable transportation. Policymakers can use these insights to promote low-emission behaviors, such as incentivizing public transport and integrating renewable energy use. Moreover, the study emphasizes the importance of simple yet effective models, which provide interpretability, computational efficiency, and predictive accuracy.

Figure 4. Correlation Heatmap



V. CONCLUSION

This study demonstrated the integration of Internet of Things (IoT) data and machine learning (ML) techniques to optimize public transportation and reduce carbon emissions. By analyzing IoT-derived datasets that captured travel behaviors, energy usage, and environmental conditions, the study evaluated regression, classification, and clustering models for their predictive effectiveness.

The regression results showed that linear regression achieved the highest accuracy (R^2 = 0.854), confirming the presence of a strong linear relationship between travel distance, energy usage, and carbon emissions. In classification tasks, Logistic Regression outperformed Random Forest and SVM with an accuracy of 87.1%, suggesting that simple linear classifiers can effectively distinguish between high- and low-emission groups. Clustering analysis with K-means achieved a low Silhouette Score of 0.103, indicating weak cluster separation; however, it still provided exploratory insights into broad travel behavior patterns.

Overall, the findings highlight that ML models, particularly linear approaches, are practical tools for supporting sustainable mobility decisions. The results also highlight the importance of IoT-enabled data collection in providing policymakers with real-time insights. By leveraging these insights, transport authorities can promote low-emission travel behaviors, enhance public transport efficiency, and design data-driven strategies for reducing urban carbon footprints.

Future research can expand on this work by applying advanced clustering methods, incorporating deep learning models for more complex pattern recognition, and testing the framework on larger, real-time IoT datasets. Such improvements will strengthen the predictive capacity of ML-driven systems and further support the transition toward sustainable and intelligent transportation networks.

VI. CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

VII. REFERENCES

[1] Z. H. Jeong and R. Rilett, "Bus arrival time prediction using an artificial neural network model," J. Intell. Transp. Syst., vol. 8, no. 1, pp. 41–50, 2003. https://doi.org/10.1207/s15472450jits0801_4

[2] X. Chen, J. Yu, and L. Chen, "Bus arrival time prediction using GPS data," in Proc. of 2004 IEEE Conf. Intell. Transp. Syst., 2004, pp. 1–6. https://doi.org/10.1109/ITSC.2004.1398961

[3] Z. Tong, Y. Wang, and M. Zhang, "Passenger flow prediction with ensemble learning in urban rail transit," Transp. Res. C, vol. 124, p. 102892, 2021. https://doi.org/10.1016/j.trc.2021.102892

[4] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," Transp. Res. C, vol. 54, pp. 187–197, 2015. https://doi.org/10.1016/j.trc.2015.03.014

[5] J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda, "Origin and destination estimation in New York City with automated fare system data," Transp. Res. Rec., vol. 1817, pp. 183–187, 2002. https://doi.org/10.3141/1817-24

[6] H. Sun, J. Wu, Y. Wu, and S. Yang, "Mining passenger travel patterns from smart card data for transit planning," Transp. Res. C, vol. 36, pp. 1–12, 2013. https://doi.org/10.1016/j.trc.2013.07.010

[7] A. Shinde and R. Dey, "Carbon footprint prediction using IoT data and machine learning techniques," Sustainability, vol. 14, no. 12, pp. 7152, 2022. https://doi.org/10.3390/su14127152

[8] F. Zhang, Y. Qin, D. Zhu, and C. Liu, "Public transit demand prediction with machine learning: A review and outlook," Information, vol. 13, no. 5, pp. 243, 2022. https://doi.org/10.3390/info13050243

[9] M. A. Hannan, M. Faisal, P. J. Ker, and A. Hussain, "Impact of Internet of Things on energy efficiency and sustainability in transport," IEEE Access, vol. 8, pp. 178989–179020, 2020. https://doi.org/10.1109/ACCESS.2020.3026633

[10] P. Saha, A. Karmakar, and S. Saha, "IoT-based carbon emission monitoring for sustainable smart cities," in Proc. of 2021 IEEE Int. Conf. on Smart Cities and Green ICT Systems (SMARTGREENS), 2021, pp. 55–62. https://doi.org/10.1109/SMARTGREENS52560.2021.00015

[11] J. Udoh, J. Lu, and Q. Xu, "Application of machine learning to predict CO_2 emissions in light-duty vehicles," Sensors, vol. 24, no. 24, p. 8219, 2024. https://doi.org/10.3390/s2424821

[12] J. Lee, J. Eom, J. Park, J. Jo, and S. Kim, "The development of a machine learning-based carbon emission prediction method for a multi-fuel-propelled smart ship by using onboard measurement data," Sustainability, vol. 16, no. 6, p. 2381, 2024. https://doi.org/10.3390/su16062381

[13] S. Li, Z. Tong, and M. Haroon, "Estimation of transport CO_2 emissions using machine learning algorithm," Transp. Res. D: Transp. Environ., vol. 133, p. 104276, 2024. https://doi.org/10.1016/j.trd.2024.104276

[14] D. Oladimeji, K. Gupta, N. A. Kose, et al., "Smart transportation: An overview of technologies and applications," Sensors, vol. 23, no. 8, p. 3880, 2023. https://doi.org/10.3390/s23083880