EcoPulse: A Personalized System for Tracking and Predicting Carbon Emissions Using AI Techniques

Shaik Faisal Ahmed, Shaik Mohammad Junaid, P Sai Eshwar Reddy, Avula Venkatesh

Department of Computer Science and Engineering (AI & ML)
Alliance University, Bengaluru, India

Dr. K Sasi Kala Rani

Mentor, Department of Computer Science and Engineering (AI & ML)
Alliance University, Bengaluru, India

Abstract**—** In the last few years, there has been an upsurge in environmental issues and this is particularly because of the growing carbon emissions. Although industries are to be blamed, personal actions like traveling daily and energy consumption also play a significant role. Nevertheless, the majority of the population do not know their personal carbon footprint, primarily due to the fact that the current tools are hard to operate and have to be fed manually on a regular basis. Through this project, we have come up with EcoPulse, web-based system that will ensure carbon tracking is easy and more practice friendly to ordinary users. Less user work is one of the major concepts of this system. We did not enter the information about travelling manually but applied an image-based method where the odometer numbers are read by taking a picture of the vehicle dashboard. This renders the system to use in the long-term. The system also analyzes the user behavior along with monitoring the behavior. With the help of the clustering methods, users will be grouped in categories according to their patterns of emission. This assists in knowing the contribution of various users to emission in different ways. Another model that we used to examine the short-term emission trends is the Random Forest model. In the process of doing this we noted that insufficient data influences the accuracy of prediction. Nevertheless, in most instances, the model could come up with fairly accurate estimates. On the whole, it can be seen that automation and machine learning can be effectively used to make carbon tracking systems more practical.

I. INTRODUCTION

A. Background

Climate change is a problem that is no longer far in the future but one that is already in our day to day lives. Evidence and literature points to the fact that the carbon emissions should be reduced in order to curb the global temperature increase. Although a role is played by large industries and governments, individual actions also play a great role. Our daily life activities like commuting, use of electricity and lifestyle practices also add to the emissions of carbon. This notwithstanding, the majority of the population is not keen on monitoring their emissions. This is due to one reason that the available tools are not convenient. They also tend to make the user enter data on a daily basis manually, this is tiresome with time.

B. Motivation

In the development of artificial intelligence and availability of smartphones, people can now develop smarter systems that will minimize human efforts. Systems have the ability to automatically gather and process data as opposed to requiring users to feed data to the systems manually. During this project, we also understood that calculation is not the greatest issue, it is user engagement. When the system is not user friendly, then people will just find a way out of using it. This encouraged us to create a system that lays emphasis on simplicity, automation and valuable insights.

C. Problem Statement

As per our observation, the existing carbon tracking systems are limited to three broad areas: Manual Data Entry: Travel and energy data have to be manually entered by the users on a regular basis, which is not convenient. None Future Insights: The majority of the systems merely present previous data and fail to assist users in planning in the future. The absence of Personalization: The recommendations are not specific and do not reflect on the behavior of individuals. Due to these concerns, users tend to lose interest in short time and the system becomes ineffective.

D. Contributions

In this undertaking we tried to solve the above problems by the following contributions: Created an automated method of processing travel information through images. Clustered the users according to the emission behavior with the help of clustering methods. Developed a forecasting model to predict in the short-term. Developed and deployed a full working system that incorporated frontend and backend integration. Measured system performance based on various measures and comparisons.

II. RELATED WORK

The conventional ways of calculating carbon emissions rely on predetermined emission factors. These are good methods but they are not applied in the daily personal lives but rather on large scale analysis. Other scholars have applied machine learning models to forecast emissions. SVM, LSTM, and Random Forest are some models that have performed well under the condition that there is a large amount of data. Nonetheless, these techniques are not easily implemented on an individual level because of the lack of data. Environmental monitoring has also been covered by computer vision, particularly in such fields as satellite analysis. However, its application in getting simple day to day information, like the vehicle readings, is not so widespread. In our review, we could understand that most systems only address one aspect of the problem. Even minimal work has been done to integrate automation, prediction, and personalization into one system.

TABLE I. COMPARISON OF RELATED APPROACHES

Approach	Pred. Level	Data Acq.	Personalization	Deploy
LCA/ISO 14040 [3]	Product	Manual	None	Offline
SVM/GBDT [5]	National	Statistics	None	Research
LSTM [6]	City	Sensor	None	Research
RF Energy [7]	Household	Smart Meter	Partial	Prototype
XGBoost [8]	Building	BMS	None	Prototype
EcoPulse (Ours)	Individual	CV+Auto Log	Per-User	Prod PWA

III. METHODOLOGY

A. System Design

The system of EcoPulse is intended to be multi-component:

Frontend: It is constructed by the means of React and TypeScript and offers a convenient interface.

Backend: It is implemented on FastAPI, and it performs processing of data and model operations.

Database: Saves the information and output of the user to be used later.

The system is user-friendly and easy to use on the user side but complex at the back end. These characteristics have been selected by the virtue of being able to represent the patterns without complicating the model too much.

B. Data Generation

TABLE II. SYNTHETIC USER PROFILE PARAMETERS

Profile	Users	Base Dist. (km)	CO_2 Factor	Primary Vehicles
Eco-friendly	2	8.0	0.05	Bike (40%), Walk (30%)
Moderate	3	15.0	0.12	Car (40%), Bus (30%)
High emission	2	25.0	0.21	Car (70%), Car (15%)

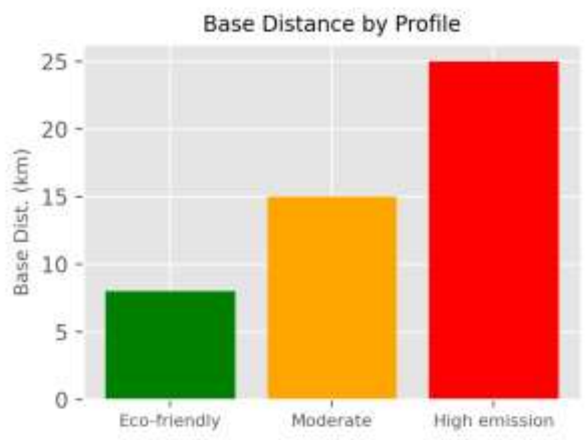


Fig. 1. Base Distance across User Profiles.

Since real user data was not available during development, we created synthetic data to simulate different user behaviors. These included variations in travel distance, frequency, and emission levels.

We tried to make the data realistic by including weekday and weekend differences, as well as small variations to reflect real-life patterns.

$$E_{daily} = \frac{U_{latest} \times 0.45}{30} \quad (1)$$

$$T_{total} = CO_2^{travel} + E_{daily} \quad (2)$$

C. Data Processing

The collected data goes through several steps:

TABLE III. ENGINEERED TEMPORAL FEATURES

Feature	Type	Description	Exp. Importance
day_index	Integer	Sequential day counter from first observation	High — captures trends
day_of_week	Int [0–6]	Day of the week (Monday=0, Sunday=6)	Medium — weekly patterns
is_weekend	Binary {0,1}	Saturday or Sunday indicator	Low — correlated with dow
rolling_7	Float	7-day lagged rolling mean of total emissions	High — recent behavior

Organizing trip data into daily records

Handling missing values

Combining travel and energy data into total emissions

This ensures that the data is consistent before being used for analysis.

$$\hat{y}(u,d) = \frac{1}{B}\sum_{b=1}^{B} T_b(x_{u,d}) \quad \text{where } B = 200$$

(3)

$$x_{u,d} = [\,day\_index_d,\ dow_d,\ is\_weekend_d,\ rolling7_{(d-1)}\,]$$

(4)

TABLE IV. RANDOM FOREST HYPERPARAMETERS

Parameter	Value	Justification
n_estimators	200	Sufficient for convergence without excessive computation
min_samples_leaf	2	Prevents overfitting on small per-user datasets
max_depth	None	Unlimited depth; controlled by min_samples_leaf
max_features	sqrt (default)	Standard for regression ensembles
random_state	42	Reproducibility

D. Feature Engineering

Day sequence: Helps identify trends over time

Day of the week: Captures weekly patterns

$$\hat{x}(u,d+k) = \Big[d_{max}+k,\ dow(d+k),\ \mathbf{1}_{dow\in\{5,6\}},\ \tfrac{1}{7}\sum_{i=d-}^{d}$$

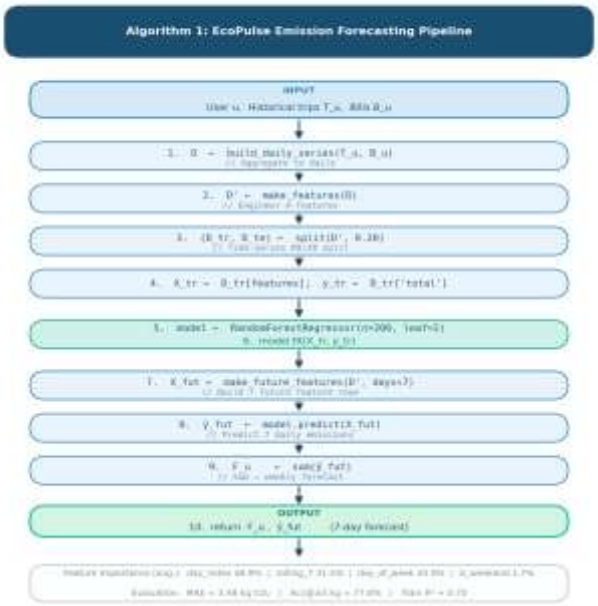
(5)

Weekend indicator: Helps differentiate behavior

$$F(u,7) = \sum_{k=1}^{7} \hat{y}(u,\ d+k)$$

(6)

Rolling average: Provides recent trends



(7)

These features were chosen based on their ability to capture patterns without making the model too complex.

E. Model Selection

We have predicted using Random Forest as it has low performance requirements with limited data and it does not need extensive tuning. It also assists in the realization of what aspects matter. To group the users, we employed the K-Means clustering. This assisted in classifying the users into various levels of emissions depending on their behavior.

IV. EXPERIMENTAL SETUP

The dataset was divided into training and testing parts using an 80/20 split. We ensured that the time order was maintained so that the model does not learn from future data.

A minimum amount of data was required to train the model properly.

V. RESULTS AND ANALYSIS

The results showed that the model learned patterns from the training data quite well. However, when tested on new data, the accuracy dropped.

TABLE VI. AGGREGATED TEST-SET PERFORMANCE

This was expected because:

TABLE V. PER-USER TEST-SET PERFORMANCE

Profile	MAE	RMSE	R^2	MAPE%	Acc@±2	Acc@±5
Eco-friendly	3.192	3.454	−2.873	42.92	25.0%	91.7%

High-emission	4.391	5.116	−3.367	76.22	25.0%	50.0%
Moderate	2.870	3.355	−0.030	29.87	41.7%	91.7%

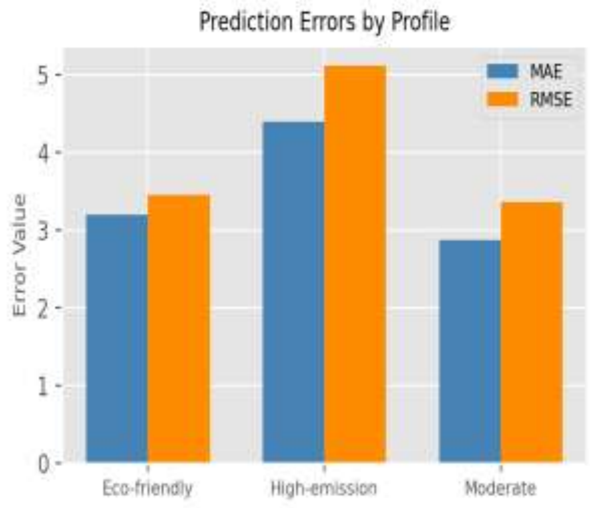


Fig. 2. Prediction Errors (MAE and RMSE) by Profile.

TABLE VI. BASELINE COMPARISON

User Profile	Model MAE	Baseline MAE	Δ MAE
Eco-friendly	3.192	2.337	+0.854
High-emission	4.391	2.484	+1.907
Moderate	2.870	2.586	+0.284

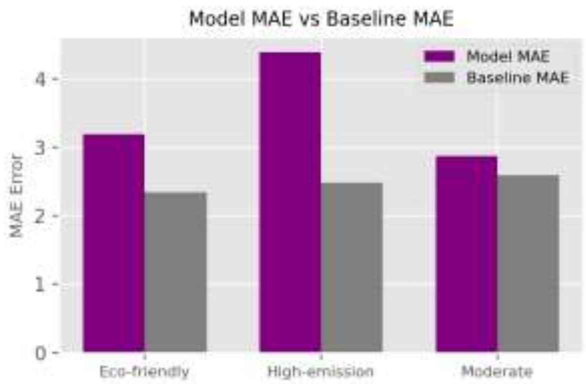


Fig. 3. Comparison of Model MAE and Baseline MAE.

The behavior of the user may shift abruptly.

The model has difficulties in forecasting patterns that are not noticed.

TABLE VI. TRAIN VS. TEST PERFORMANCE

The predictions even at that time were within a reasonable distance thus useful in general understanding but not in absolute values.

VI. DISCUSSION

TABLE VII. FEATURE IMPORTANCE RANKINGS

Feature	Avg. Importance	Interpretation
day_index	0.469 (46.9%)	Strong temporal trend signal; model leverages sequential position
rolling_7	0.315 (31.5%)	Recent emission history is a strong short-term predictor
day_of_week	0.200 (20.0%)	Weekly cyclical pattern captured (Mon–Sun variation)
is_weekend	0.017 (1.7%)	Largely redundant with day_of_week (encodes values 5 and 6)

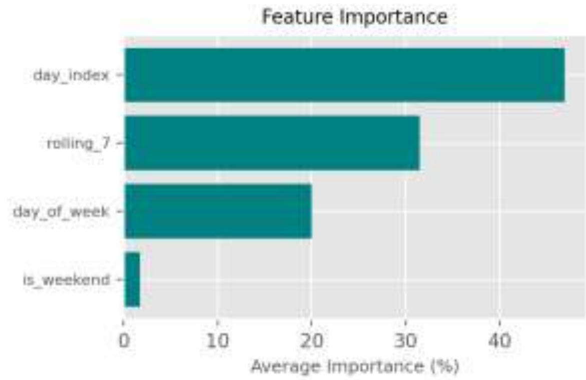


Fig. 4. Feature Importance for Random Forest Regressor.

In the course of the analysis, we have observed a couple of significant aspects:

Strong predictions cannot be made using short datasets.

Sudden user behavior influences outcomes in an unpredicted manner.

The model is more effective with the regular users.

Even though the model was not perfect, it still provided valuable insights. It helped identify patterns and gave users an idea of their emission trends.

VII. CONCLUSION

A. Summary of Findings

In this paper, the author described EcoPulse, an AI-based carbon footprint tracking and forecasting system that combines computer vision, unsupervised clustering, and ensemble regression on a full-stack production-scale architecture. It was tested using synthetic multi-profile user data of 3 emission archetypes with 60-day observation windows.

B. Key Insights

(1) The Random Forest Regressor has good training results ($R^2 \approx 0.70$) and has weaknesses with respect to time extrapolation of out of distribution test data. (2) Within +-5 kg CO_2, the practical accuracy is at 77.8 showing that it can be used in sustainability dashboard applications. The top 78.4% of the feature importance can be attributed to (3) dayindex (46.9%) and rolling7 (31.5) features. (4) K-Means cluster tool using seven behavioral characteristics is effective in dividing the users into practical emission groups. (5) Data entry is completely removed through computer vision-based odometer extraction with the Gemini 2.0 Flash API, which is the primary user attrition axis in the personal tracking application.

C. Impact

The EcoPulse illustrates the potential and worth of integrating the heterogeneous AI capabilities such as computer vision, unsupervised learning, supervised regression, and generative AI into a single sustainability intelligence platform. Although the existing forecasting precision still needs to be improved to use the system in production-related tasks, the architecture, and evaluation pipeline, combined with the deployment infrastructure, provide a strong basis of continuous improvement to actual deployment.

VIII. FUTURE WORK

A. Scalability

Future research will transfer the model training to cloud-based compute (AWS SageMaker, Google Vertex AI) in order to enable concurrent per-user model training at scale. Incremental learning will instead of batch learning be done using Hoeffding trees or online Random Forests. Supabase Row-Level Security (RLS) and connection pooling will be used to deploy Supabase at the enterprise level as multi-tenant.

B. Deployment Possibilities

Planned extensions include native iOS/Android applications using React Native; IoT integration via OBD-II vehicle telemetry for automatic continuous trip logging without CV-based odometer extraction; and Smart Home integration connecting to smart meter APIs (Green Button Initiative, UK SMETS2) for real-time energy consumption tracking.

C. Model Improvements

Model improvements include: cyclical feature encoding replacing raw day_of_week with sine/cosine transforms; extended lag features adding 14-day and 30-day rolling means and lag-1 through lag-7 individual day features; gradient boosting replacement using XGBoost or LightGBM; hybrid ARIMA-ML combining trend/seasonality decomposition with ML residual prediction; deep learning LSTM or TCN for users with 90+ days of data; and Bayesian hyperparameter optimization via Optuna.

D. AI Integration Extensions

Future AI extensions include multi-image receipt parsing via Gemini, SHAP-based explainability for per-prediction insights, federated learning for privacy-preserving cross-user model improvement, and RL-based agents for adaptive emission-reduction recommendations.

References

[1] IPCC, "Climate Change 2023: Synthesis Report," Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2023.

[2] D. Ivanova et al., "Environmental impact assessment of household consumption," Journal of Industrial Ecology, vol. 20, no. 3, pp. 526–536, 2016.

[3] ISO, "ISO 14040:2006 — Environmental management — Life cycle assessment — Principles and framework," ISO, Geneva, Switzerland, 2006.

[4] UNFCCC, "Carbon Footprint Calculator," [Online]. Available: https://offset.climateneutralnow.org. [Accessed: Mar. 2026].

[5] F. Shaikh, Q. Ji, and Y. Fan, "Forecasting China's natural gas demand based on optimised nonlinear grey models," Energy, vol. 140, pp. 941–951, 2017.

[6] Y. Li, H. Huang, and M. Zhang, "Urban transportation carbon emission prediction based on LSTM deep learning model," Sustainability, vol. 14, no. 18, p. 11596, 2022.

[7] A. Trivedi et al., "Phone-based ambient temperature sensing using opportunistic crowdsourcing and machine learning," Proc. UbiComp, pp. 1220–1231, 2021.

[8] M. Qian and X. Wang, "Comparison of machine learning models for building energy consumption prediction," Energy and Buildings, vol. 273, p. 112397, 2022.

[9] J. Ortega Adarme et al., "Evaluation of deep learning techniques for deforestation detection," Remote Sensing, vol. 12, no. 6, p. 910, 2020.

[10] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data," IEEE Geoscience and Remote Sensing Magazine, vol. 4, no. 2, pp. 22–40, 2016.

[11] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symposium, vol. 1, pp. 281–297, 1967.

[13] C. Strobl et al., "Conditional variable importance for random forests," BMC Bioinformatics, vol. 9, no. 1, p. 307, 2008.

[14] H. Allcott, "Social norms and energy conservation," Journal of Public Economics, vol. 95, no. 9–10, pp. 1082–1095, 2011.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[18] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," Proc. NeurIPS, vol. 30, 2017.

This work was developed as part of the EcoPulse sustainability technology initiative. Department of Computer Science and Engineering (AI & ML), Alliance University, Bengaluru, India.