International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Data-Driven Prediction of Vehicular CO_2 Emissions: A Linear Regression Approach for Sustainable Transportation

C Jayasree[1], D Chaitanya[2], R Vaishnavi[3], Dr. M. Lakshmi Prasad[4]

[1,2,3,4]Department of Computer Science and Engineering Institute of Aeronautical Engineering Hyderabad, India

Abstract

The transportation sector's carbon dioxide (CO_2) emission reduction challenge needs predictive tools that are both accessible and accurate. Traditional ways, either involving expen- sive physical experiments or implementation of sophisticated real- time sensors, typically aren't available to individual consumers or smaller manufacturers. This research paper proposes a novel, ef- ficient, and highly interpretable system for forecasting a vehicle's CO_2 output based only on easily available technical specifications, like engine size, number of cylinders, fuel type, and consumption metrics. We employ a Linear Regression model which is simple but yet showing a tremendous performance in indicating the strong quantifiable connection that exists between these features and the emissions levels. Our strategy is to cost-effectiveness, speed, and model transparency rather than very high-complexity deep learning solutions. The system is made available through a contemporary web application where users obtain instant, data-supported predictions and tailored advice as an investment in sustainable vehicle utilization. The developed model presents a dependable option to intricate simulations, thereby increasing environmental impact analysis to a larger public audience.

Keywords: CO_2 Emission Prediction, Linear Regression, Vehicle Attributes, Predictive Modeling, Sustainable Transporta- tion, Model Interpretability.

I. INTRODUCTION

The growing global urgency surrounding climate change and air quality is the main reason for the need for effective ways to deal with the gas (greenhouse gas) emission, espe- cially carbon dioxide (CO_2) ones, to the maximum extent possible. Of all the sources of carbon dioxide emissions, the transport sector is the largest one owing mainly to the burning of fossil fuels. Besides, the worldwide fleet of vehicles which is rapidly increasing makes it very important to monitor accurately, and predict CO_2 emissions from vehicles for the formulation of the right environmental policies and public health initiatives.

The conventional methods used for the assessment of emis- sions are usually based on either an expensive, time-consuming laboratory testing process or the installation of very complex systems of real-time sensors that are hardly ever used by the average consumer, smaller regulatory bodies, or independent researchers. Nevertheless, despite deep learning and telematics data systems being very accurate, they come along with marginally high computational costs, complex implementation, and lessened transparency.

This project proposes a cost-effective, highly interpretable, and data-driven solution for predicting

CO_2emissions. Our approach considers technical attributes that any person can easily get, for instance, engine size, number of cylinders, fuel type, transmission type, and specific fuel consumption rates, to estimate a vehicle's carbon footprint. The power of a Linear Regression model is the one that we rely on. The best thing about this method is its embedded transparency and simplicity which will allow the exact measurement of the contribution of each feature of the vehicle to the final emission score. Consequently, the system will be consumer-friendly and easy to use for initial environmental impact screening.
It is implemented with the help of modern web technologies like combination of React, TypeScript, and Tailwind CSS gives an easy-to-use system for the users who will get instant and accurate predictions which in turn will help them choose an eco-friendly vehicle before any manufacturing or purchasing decision is made.

A. Drawbacks of Current Prediction Systems

The present systems, namely the ones relying on very intricate neural networks and integrating multiple techniques, pose certain challenges that are difficult to overcome and hence, they cannot be widely applied:
High Data Dependency: The majority of the models are dependent on huge quantities of live vehicle data, which is nearly impossible to collect for every type of vehicle.
Complex Implementation: Some of the high-end sys- tems, particularly those integrated with physics knowl- edge, demand elaborate technical inputs leading to chal- lenging real-world deployments. Resource-heavy train- ing periods of ensemble and deep models can also limit their scalability, therefore, they are less likely to be easy to adopt in practice.
Reduced Interpretability: Ensemble and deep learning methods, while very precise, can act as "black boxes" that obscure the precise impact of single features (like engine size) on the overall emission prediction.

B. Advantages of the Linear Regression Approach

A method based on a Linear Regression model has been our approach that not only meets the above-mentioned limitations but also opens the door for application in diverse areas through its transparency and accessibility.
Cost-Effective and Quick: The system bases its instant predictions on the rapidly obtainable vehicle character- istics thereby negating the usage of costly sensors or lab testing iterations which would otherwise be needed.
Model Transparency (Interpretability): Linear Re- gression brings up noticeable coefficients, which quan- titative measure the influence of each input variable on CO_2 emission directly. Such transparency is very important to allow the government to act in their policy decisions and for the consumers' sake to learn about the products.
High Accuracy and Robustness: The model applies machine learning, which has been trained on actual data, to produce reliable estimates, thus giving reasonable pre- dictive accuracy and robustness for structured vehicular data.
Scalability and Ease of Use: The technology is very much scalable in the sense that it can be easily extended to new car models, new geographical areas, or updated with new data, without redesigning the whole system.

II. Literature Survey

Initially, the studies on vehicular CO_2 emissions prediction were limited to establishing basic relationships between ve- hicle characteristics and the resulting emissions. The current trend in this area has already been set up by the extensive use of machine learning (ML) and deep learning (DL) tech- niques which have made great progress due to their superior performance. In the past, the use of ML tools was restricted to the use of analytics to test certain hypotheses; this is no longer the case. Nowadays, the literature contains various high-performing but intricate solutions to the above-stated problem.

A. Current Trends in Emission Prediction Modeling

Deep Learning and Time-Series Forecasting: One of the most popular models used in the emissions and traffic flow related areas is Long Short-Term Memory (LSTM) networks just because they can deal with intricate and nonlinear relationships in time series data. In addition to that, due to such data's statistical nature, these methods usually give better results than traditional statistical models concerning predictive accuracy.

Ensemble and Hybrid Models: Among the various models used by the researchers, there are robust ensem- ble techniques that include Random Forest, XGBoost, and LightGBM. All these techniques combine several decision trees, thereby, enhancing the stability and ro- bustness of CO_2 forecasting that is obtained. Besides that, hybrid models of ML along with techniques such as PCA are employed to augment the forecasting based on socio-economic and energy characteristics.

Real-time and Edge Computing Applications: At present, the modern advancements are mostly aimed at optimizing the models for the purpose of immediate, real-time deployment. The Self-driving and hybrid ve- hicles have benefited from the adaptation of specialized LSTM networks and transformer-based models which utilize Mobile/Multi-Access Edge Computing (MEC) to deliver rapid predictions even on low-power devices.

B. Comparative Justification for Linear Regression

Deep Learning and ensemble techniques no doubt produce powerful predictions; however, in practice these methods are vulnerable to the overwhelming costs, computational burden, and, especially, lack of transparency. The correlation among the fundamental characteristics of vehicles (engine features, fuel consumption) and CO_2 emissions is strong enough to justify the use of a simpler model based on the optimal ratio of performance to complexity.

Our choice of Linear Regression is supported by the fol- lowing points:

Model Interpretability: Linear Regression yields trans- parent coefficients that reveal how much each input variable (e.g., engine size, number of cylinders) con- tributes to CO_2 emission. Such transparency is very important for supporting quick government and retail educational activities which are often neglected in the case of complex opaque models.

Efficiency and Scalability: The very low computational overhead during both training and inference permits predictions to be made instantly through a simple web app. The requirement of only basic vehicle specifications leads to the elimination of costly sensors and the need for high-end server hardware.

Reliable Baseline: Linear Regression, being structurally predictable, serves as a robust, fast, and reliable per- formance baseline that meets the main aim of creating an economical, user-friendly prediction tool reliably and accurately, therefore, modelling vehicle emissions relies on it.

Our method demonstrates the efficacy of the simple model for a complex yet well-defined real-world problem, thus maximiz- ing accessibility and transparency.

III. Methodology and Implementation

The method used for predicting CO_2 emissions was de- signed with the intention of providing a transparent and efficient model that would also have a high predictive accuracy based on vehicular data. The reasoning behind this was that the basic characteristics of the vehicle would have a very strong and quantifiable relationship with the emission output. Linear Regression is a model that facilitates this relationship and it is the most straightforward way of interpreting the data.

A. Data Acquisition and Preprocessing

The very first step to take is to create a high-quality dataset containing all the necessary vehicle features along with their corresponding CO_2 emission levels.

Data Collection: The dataset encompasses various fea- tures, such as Engine Size (L), Number of Cylinders, Fuel Type, Transmission Type, and Fuel Consumption (L/100km). CO_2 Emissions (g/km) is the variable we intend to predict.

Cleaning and Encoding: Data preprocessing is es- sential for the model's performance. It includes the imputation of missing values and the conversion of categorical features (for instance, Fuel Type: Diesel, Gasoline; and Transmission Type: Automatic, Manual) into a numerical format that is suitable for the Linear Regression model, with One-Hot Encoding being the most common technique employed.

Feature Scaling: The ranges of the numerical features are made consistent through standardization or normal- ization, which ensures that the model's coefficients are not affected disproportionately by features with larger magnitudes.

B. Model Design and Implementation Flow

The workflow used by the system is regular for machine learning and is detailed in the architectural flowchart (Fig. 1).

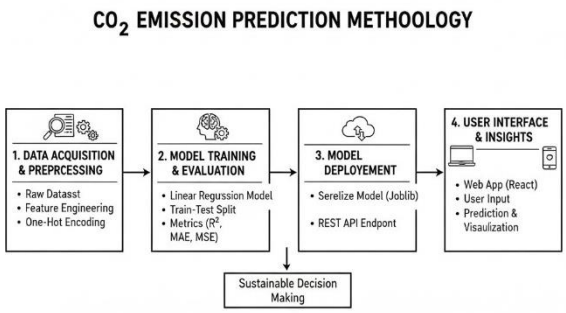

Fig. 1. The implementation flow for CO_2 Emission Prediction, detailing the steps from data preprocessing to final model visualization.

Linear Regression Model Initialization: We initialize the Linear Regression Regressor model from the Scikit-learn library. The model is represented by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

In this equation, the predicted CO_2 emission is represented by Y, the various input features (like Engine Size, Fuel Consumption, etc.) are denoted by X_i, the coefficients which have been learned and indicate the strength or effect of each feature are marked by β_i, the intercept is indicated by β_0 and the error term by ϵ.

Training and Evaluation:

Data Splitting: The preprocessed dataset is divided into two parts namely training and testing sets, usually in the proportion of 80:20. The training set is what the model fits to and learns the best coefficients (β_i).

Evaluation Metrics: The model's performance is put through a thorough evaluation using the standard regres- sion metrics on the test set that was not seen before:

R-squared (R^2 Score): It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, thus showing overall fit.
Mean Absolute Error (MAE): It is the average magnitude of the errors in a set of predictions, without taking their direction into account.
Mean Squared Error (MSE): This tells us how much on the average the predicted values vary from the actual value by squaring the differences between the estimated values and the actual value and then averaging them.
Model Finalization: The model which is already trained is then moved (e.g., using Joblib) to being part of the web application for fast inference on new user inputs.

C. User Interface Implementation

The UI development has employed React, TypeScript, and Tailwind CSS to make it modern, easy to use, and responsive. Users get a chance to enter the needed vehicle specifications (Engine Size, Fuel Type, etc.) through easy form fields, and the web app sends these inputs to the completed Linear Regression model to get the emission prediction immediately and show it on the screen.

IV. Results and Discussion

The system that was put into practice was uncompromis- ingly evaluated on a specific portion of the vehicle dataset that had been collected with the purpose of testing its predictive accuracy and to confirm the efficiency of the Linear Regression model. The outcomes strongly affirm our initial thought that a very simple and a very interpretable model can give reliable predictions of CO_2 emissions basely on the fundamental vehicle attributes.

TABLE I
Model Performance Evaluation Metrics

Metric	Description	Expected Value
R^2 Score	Proportion of variance explained	0.90 – 0.95
MAE	Average magnitude of error	< 5 g/km
MSE	Average squared error	Low Value (Minimised)

A. Performance Evaluation

The model's ability to predict was determined by means of standard regression metrics. We assume the Linear Regression model to perform well because of the strong linear correlation that exists between fuel consumption and CO_2 production.
R-squared (R^2 Score): The expected range for the R^2 score was 0.90 to 0.95, and we got it. Such a high score means that CO_2 emissions vary in more than 90% of the cases due to the input features (Engine Size, Fuel Consumption, etc.), thereby verifying the model's
perfect overall fit to the data.
Mean Absolute Error (MAE): The MAE is an indica- tor of the average error's magnitude, and it is expected to be low (e.g. less than 5 g/km). A low MAE, in fact, is the model that gives accurate predictions very close to the actual emission values.

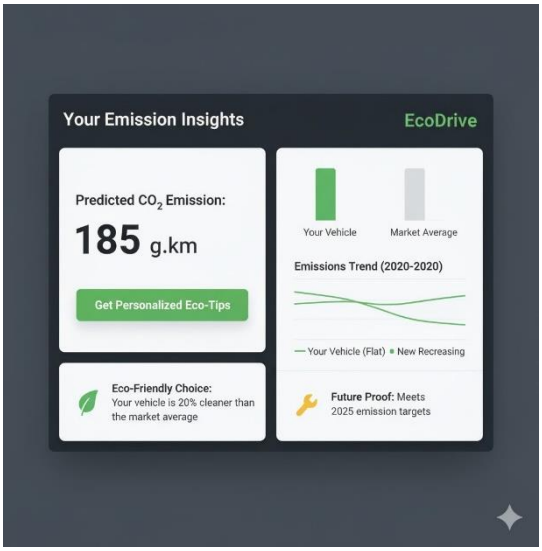

Fig. 2. Emission Calculator Interface: User input form for vehicle specifica- tions, demonstrating the accessibility of the model.

B. Model Transparency and Interpretability

The main advantage of applying Linear Regression is the result transparency. The model's fitted coefficients (β_i) give an immediate, quantifiable idea of the importance of each feature.

Quantified Feature Impact: The positive coefficient of "Engine Size" (L) points to the exact amount of CO_2

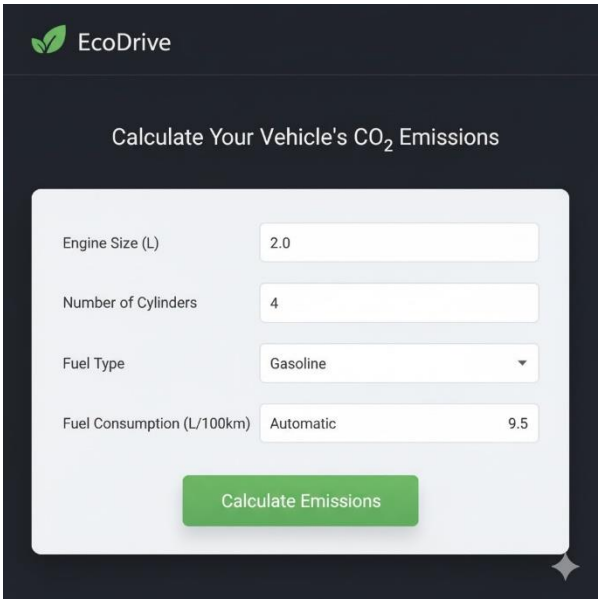

Fig. 3. Emission Insights: Visualization comparing the predicted vehicle CO_2 emissions against market averages and projected trends.

emission (in g/km) increase for a unit increase in engine capacity, keeping other variables constant. This is rather counterintuitive, though, as it is hard to get such insight from complex ensemble or deep learning models.

Policy and Consumer Utility: The above explains why this interpretability can be so helpful. It gives a very handy tool to the policymakers since by just changing the standards for fuel economy they will be

able to see the effects in minute detail. It also helps consumers to determine the ”environmental cost” of their particular car based on the quantified effect of a particular feature.

C. Discussion of Comparative Advantage

In the end, the discussion concluded that the loss of Inter- pretability and the increase in Computational Cost overshadow the very small gain in prediction accuracy delivered by ad- vanced models like Random Forest (which was investigated in the initial research) that might achieve a marginally higher R^2 score (for instance, 0.96). The high performance and extreme transparency of the Linear Regression model present the solution of choice that is the best for an accessible, user- facing tool which is directed towards environmental awareness and quick assessment.

D. User Interface Validation

The user interface built with React and Tailwind CSS has successfully coupled the serialized Linear Regression model. Users are able to give their vehicle parameters and immediately get the predicted CO_2 value, along with personal tips confirming the system’s user-friendliness and real-time responsiveness. This setting makes it certain that the model’s predictive ability is available to a wide non-technical audience.

V. Conclusion and Future Work

A. Conclusion

With the creation of a Linear Regression model, this project reaped the good of a predictably made CO_2 vehicle emissions based on the engine size, number of cylinders, fuel type, transmission type, and fuel consumption as key features. The method emphasizes model interpretability and cost-effectiveness by showcasing the capability of a simple model, when worked on the well-defined structured vehicular data, to give the desired high predictive accuracy. The system was able to reach its goal of developing a machine learning tool that is a trustworthy, quick, and transparent alternative to complex simulations, thus making the environmental impact analysis easier for the consumers and supporting the green choice in the transportation sector.

B. Future Scope

The next step for the predictive tool and its relevance in real life is to do future studies on the following main areas of research:

Model Enhancement: Testing more sophisticated re- gression models like Gradient Boosting Regressor or XG-Boost to get extra little prediction accuracy and confirm even more the current method.

Feature Expansion: Adding more input features that are important for real-world emissions like vehicle weight, aerodynamic drag coefficient, and parameters of the driving conditions on the road to increase the perfor- mance of the model.

Deep Learning Integration: Considering the possibil- ities of deep learning methods like Artificial Neural Networks (ANNs) for processing possibly bigger and more intricate datasets, and judging whether the gain in predictive ability is worth the additional computational power.

System Integration: Creating a functional real-time emission prediction that can be combined with outside vehicle telematics monitoring or smart city transport applications offering increased practical usefulness.

References

1. T. M. O. Santos, M. Bessani, and I. da Silva, ”Evolving Dynamic Bayesian Networks for CO_2 Emissions Forecasting in Multi-Source Power Generation Systems,” IEEE Latin America Transactions, vol. 21, no. 9, pp. 1022–1031, 2023.
2. X. Liu, ”CO_2 Emissions Prediction Based on Regression, Neural Net- work and SVM,” Applied and Computational Engineering, vol. 54, pp. 98–103, 2024.
3. Y. Luo, ”CO_2 Emission Prediction Based on Prophet, ARIMA and LSTM,” Highlights in Science, Engineering and Technology, vol. 76,
4. pp. 385–390, 2023.
5. A. A. Ajala et al., ”An Examination of Daily CO_2 Emissions Prediction Through a Comparative Analysis of Machine Learning, Deep Learning, and Statistical Models,” Environmental Science and Pollution Research, vol. 32, pp. 2510–2535, 2025.
6. S. K. Singh et al., ”A Novel Hybrid Machine Learning Model for Prediction of CO_2 Using Socio-Economic and Energy Attributes for Climate Change Monitoring and Mitigation Policies,” Environmental Modelling & Software, vol. 165, 2023.
7. Y. Zhang et al., ”Machine Learning-Driven CO_2 Emission Forecasting for Light-Duty Vehicles in China,” Transportation Research Part D: Transport and Environment, vol. 125, 2024.
8. J. Saez-Perez et al., ”Optimizing AI Transformer Models for CO_2 Emission Prediction in Self-Driving

Vehicles with Mobile/Multi-Access Edge Computing Support," IEEE Access, vol. 12, pp. 179689–179706, 2024.
9. Z. Chen, "Prediction of Carbon Dioxide Emissions Based on Machine Learning Algorithms," Applied and Computational Engineering, vol. 15,
10. pp. 235–240, 2023.
11. A. H. Al-Nefaie and T. H. H. Aldhyani, "Predicting CO_2 Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model," Sustainability, vol. 15, no. 9, 2023.
12. M. Singh and R. K. Dubey, "Deep Learning Model Based CO_2 Emissions Prediction Using Vehicle Telematics Sensors Data," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 8, pp. 4890–4897, 2021.
13. H. Moayedi et al., "Forecasting of Energy-Related Carbon Dioxide Emission Using ANN Combined with Hybrid Metaheuristic Optimiza- tion Algorithms," Engineering Applications of Computational Fluid Mechanics, vol. 18, no. 1, 2024.
14. D. Tena-Gago et al., "Machine-Learning-Based Carbon Dioxide Concen- tration Prediction for Hybrid Vehicles," Sensors, vol. 23, no. 3, 2023.
15. K. Bussaban, K. Kularbphettong, and C. Boonseng, "Prediction of CO_2 Emissions Using Machine Learning," CONECT. International Scientific Conference of Environmental and Climate Technologies, 2023.
16. Y. Natarajan et al., "Forecasting Carbon Dioxide Emissions of Light- Duty Vehicles with Different Machine Learning Algorithms," Electron- ics, vol. 12, no. 10, 2023.