



# OPEN Deep learning model based prediction of vehicle CO<sub>2</sub> emissions with eXplainable AI integration for sustainable environment

Gazi Mohammad Imdadul Alam<sup>1</sup>, Sharia Arfin Tanim<sup>2</sup>, Sumit Kanti Sarker<sup>2</sup>, Yutaka Watanobe<sup>3</sup>, Rashedul Islam<sup>4</sup>✉, M. F. Mridha<sup>2</sup> & Kamruddin Nur<sup>2</sup>✉

The transportation industry contributes significantly to climate change through carbon dioxide (CO<sub>2</sub>) emissions, intensifying global warming and leading to more frequent and severe weather phenomena such as flooding, drought, heat waves, glacier melting, and rising sea levels. This study proposes a comprehensive approach for predicting CO<sub>2</sub> emissions from vehicles using deep learning techniques enhanced by eXplainable Artificial Intelligence (XAI) methods. Utilizing a dataset from the Canadian government's official open data portal, we explored the impact of various vehicle attributes on CO<sub>2</sub> emissions. Our analysis reveals that not only do high-performance engines emit more pollutants, but fuel consumption under both city and highway conditions also contributes significantly to higher emissions. We identified skewed distributions in the number of vehicles produced by different manufacturers and trends in fuel consumption across fuel types. This study used deep learning techniques to construct a CO<sub>2</sub> emission prediction model, specifically a light multilayer perceptron (MLP) architecture called CarbonMLP. The proposed model was optimized by hyperparameter tuning and achieved excellent performance metrics, such as a high R-squared value of 0.9938 and a low Mean Squared Error (MSE) of 0.0002. This study employs XAI approaches, particularly SHapley Additive exPlanations (SHAP), to improve the model interpretation ability and provide information about the importance of features. The findings of this study show that the proposed methodology accurately predicts CO<sub>2</sub> emissions from vehicles. Additionally, the analysis suggests areas for further research, such as increasing the dataset, integrating additional pollutants, improving interpretability, and investigating real-world applications. Overall, this study contributes to the design of effective strategies for reducing vehicle CO<sub>2</sub> emissions and promoting environmental sustainability.

**Keywords** CO<sub>2</sub> emissions, CarbonMLP, EXplainable Artificial Intelligence, Vehicle attributes, Fuel consumption, Environmental sustainability.

In the 21st century, climate change poses a significant challenge to humanity, as it harmfully affects ecosystems, the economy, and human welfare. Carbon dioxide (CO<sub>2</sub>) emission is the primary cause of climate change. Additionally, the reasons for this epidemic include various human activities such as industrial processes, energy production, and transportation. While CO<sub>2</sub> emissions are a global concern, countries such as Canada face unique challenges due to their reliance on fossil fuel-based transportation across vast geographical landscapes. Canada's transportation sector accounts for a significant portion of the country's overall greenhouse gas (GHG) emissions, making it a priority for policy interventions aimed at reducing environmental impacts<sup>1</sup>. Moreover, these emissions threaten the environment and negatively affect human health. Exposure to air pollution from traffic can aggravate respiratory problems such as asthma and chronic obstructive pulmonary disease (COPD) and may increase the risk of developing heart disease or stroke<sup>2</sup>.

Transportation is a crucial foundation of contemporary society, enabling economic transactions, worldwide interconnections, and individual movement. However, reliance on fossil fuels has resulted in a substantial and

<sup>1</sup>School of Science, Engineering & Technology, East Delta University, Chattogram 4209, Bangladesh. <sup>2</sup>Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh.

<sup>3</sup>Department of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu 965-8580, Japan.

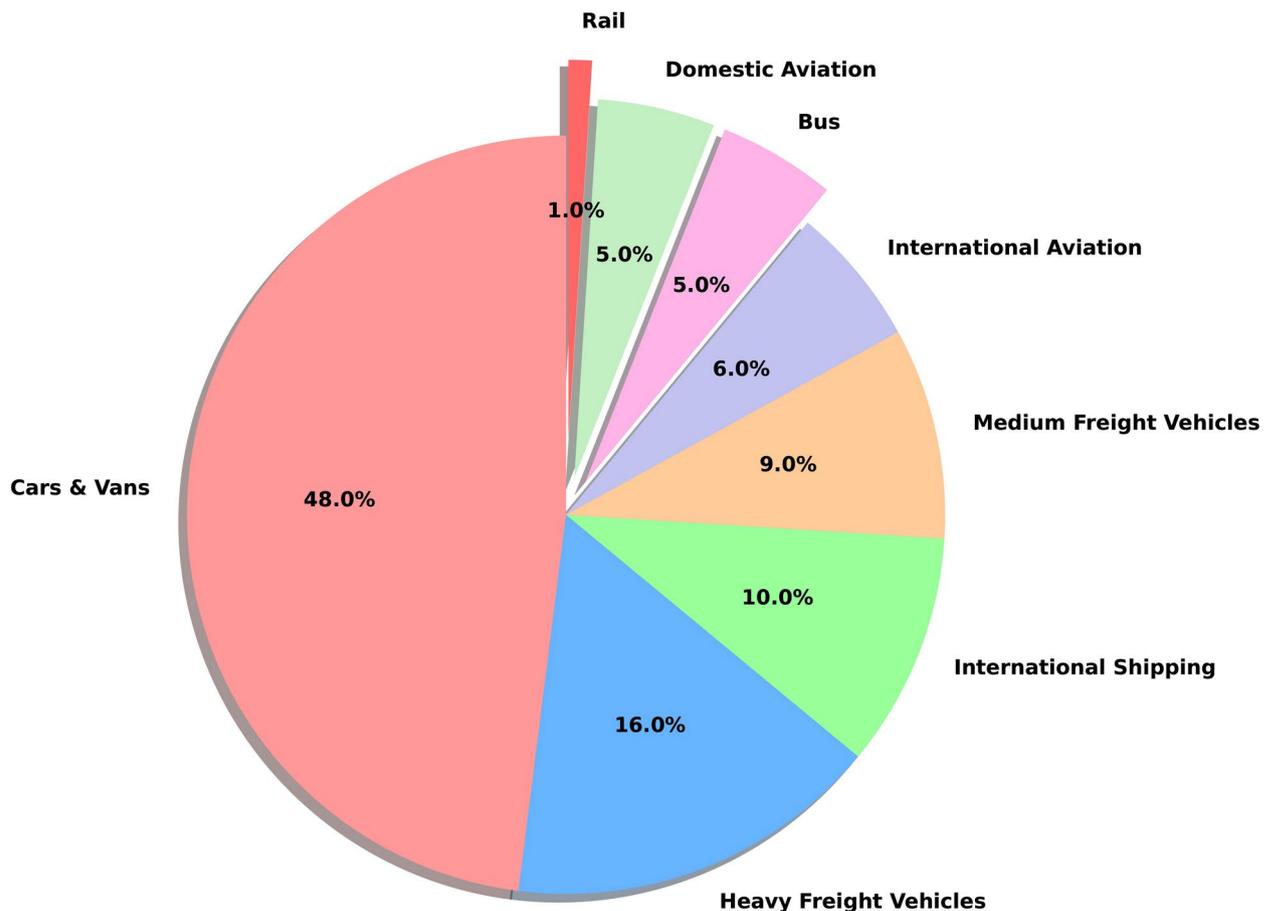
<sup>4</sup>Department of Computer Science and Engineering, University of Asia Pacific, Dhaka 1216, Bangladesh. ✉email: rashed.cse@gmail.com; kamruddin@aiub.edu

increasing environmental predicament, namely, a surge in CO<sub>2</sub> emissions. These emissions are a major cause of the increase in greenhouse gases, which accelerate climate change and lead to environmental problems such as higher sea levels, extreme weather conditions, and disruptions to ecosystems. Motor vehicles have a significant impact on the transportation sector, and are a major contributor to CO<sub>2</sub> emissions. In 2020, cars accounted for approximately 23% of global CO<sub>2</sub> emissions<sup>3</sup>. An average passenger automobile produces 4.6 metric tonnes of carbon dioxide annually<sup>4</sup>. Different transports also emit nitrogen oxides and other pollutants, which also contribute to smog formation and acid rain and are harmful to ecosystems<sup>5</sup>. The World Counts emphasize the significant influence of cars on the condition of the air, highlighting them as a primary contributor to air pollution<sup>6</sup>.

Figure 1 shows the distribution of CO<sub>2</sub> emissions by vehicle type for the year 2022 worldwide<sup>7</sup>. The graph highlights the percentage contribution of each vehicle category to total global CO<sub>2</sub> emissions. Comprehending and forecasting these emissions are essential for formulating efficient measures to combat climate change.

The complex interplay of these factors influences vehicle CO<sub>2</sub> emissions. The most significant contributors are the fuel type, engine characteristics, driving behavior, road conditions, vehicle weight, and aerodynamics. The level of CO<sub>2</sub> emissions produced by vehicles is directly influenced by the type of fuel used.

Biofuels, hydrogen, and electricity produce less CO<sub>2</sub> than fossil fuels such as gasoline and diesel<sup>8</sup>. Biofuels have a lower carbon footprint, while hydrogen fuel vehicles produce only water vapor as a byproduct<sup>9</sup>. Engine displacement, power output, and technology are factors that influence CO<sub>2</sub> emissions. Although advancements in engine technology, such as direct injection and turbocharging, can enhance fuel efficiency and reduce emissions, it is important to note that larger and more powerful engines often generate higher levels of CO<sub>2</sub><sup>10</sup>. Fuel consumption and CO<sub>2</sub> emissions also significantly depend on the gearbox mechanism, whether manual or automatic<sup>11</sup>. CO<sub>2</sub> emissions can be decreased by eco-driving strategies such as maintaining steady speeds, activating traffic lights, and moderately accelerating. Conversely, aggressive acceleration, rapid braking, and frequent idling contribute to higher emissions<sup>12</sup>. Vehicle fuel consumption and CO<sub>2</sub> emissions are influenced by traffic congestion, road surface conditions, and inclines. Stop-and-go traffic results in heightened engine idle and suboptimal fuel use, whereas uneven road surfaces can generate extra friction, thus affecting fuel efficiency<sup>13</sup>. Furthermore, severe temperatures and the level of winds also affect fuel consumption and emissions<sup>14</sup>. A higher vehicle weight requires increased energy consumption, which results in higher levels of CO<sub>2</sub> emissions. Moreover, the presence of aerodynamic drag affects cars because their streamlined structures encounter reduced air resistance, which results in lower emissions<sup>15</sup>. The development of policies and strategies is important for



**Fig. 1.** Transportation Emissions by Vehicle Type Contributing to total CO<sub>2</sub> Emissions.

reducing climate change by estimating vehicle CO<sub>2</sub> emissions. To promote cleaner and more fuel-efficient vehicles, the prediction and monitoring of CO<sub>2</sub> emissions from vehicles remains a difficult task. Traditional methods for estimating emissions, such as emission factors and vehicle testing, have limitations in accuracy, scalability, and adaptability to dynamic driving conditions<sup>16</sup>. In addition, regulatory bodies have established emission factors based on criteria such as vehicle type, fuel type, and other pertinent variables. Although useful, these parameters frequently offer a generic approximation and fail to consider the dynamic interaction of factors that influence actual emissions in the real world<sup>17</sup>.

This research was motivated by the pressing need to address the environmental impact of vehicle emissions and enhance our comprehension of the processes that cause variations in emissions. Policymakers can make well-informed judgements about emission reduction methods, sustainable transportation policies, and infrastructure expenditures by creating accurate forecasting models<sup>18</sup>. Furthermore, precise forecasts of CO<sub>2</sub> outflows can empower consumers to make clear choices when selecting vehicles, emphasize eco-friendly driving practices and support sustainable transportation.

This study aimed to address the limitations of existing methods for estimating vehicle CO<sub>2</sub> emissions by developing a lightweight deep learning (DL) model that leverages real-world data and offers several key advantages:

1. Build and train a lightweight DL model using advanced techniques to predict vehicle CO<sub>2</sub> emissions accurately from comprehensive datasets. Multiple algorithms are compared to select the most accurate and robust predictors.
2. We integrate real-world data to train the deep learning model with eXplainable Artificial Intelligence (XAI) integration, enhancing realism and accuracy in predicting vehicle CO<sub>2</sub> emissions.
3. Ensure computational efficiency and scalability by developing a DL model with minimal data preprocessing, thereby promoting real-world applicability for widespread adoption in on-board emission estimation and regulatory use.

Although the dataset used in this study is specific to Canadian vehicles and emissions data, the developed model and techniques can be generalized and applied to similar datasets from other countries and regions. Vehicle CO<sub>2</sub> emissions are a global issue and are influenced by factors such as engine size, fuel consumption, and vehicle type. The predictive techniques and insights gained from this research apply to various geographic contexts, as they leverage universal vehicle characteristics. Furthermore, the methods demonstrated in this study provide a framework that can be easily adapted to account for regional differences in emissions standards, vehicle types, and driving conditions, making it relevant for national and international efforts to reduce CO<sub>2</sub> emissions.

This study analyzed recent CO<sub>2</sub> emissions trends for a variety of vehicle types and models from multiple manufacturers. This study uses data from the Canadian government's official open data website to show how various car features affect CO<sub>2</sub> emissions. The data preparation process comprises rigorous cleaning, data engineering, and transformation procedures to increase dataset quality and prediction accuracy. This study employs advanced deep learning techniques, specifically a multilayer perceptron (MLP) architecture, to improve the prediction accuracy of CO<sub>2</sub> emissions. Although traditional machine learning models are mentioned for performance comparison, the primary focus of this study is on the deep learning approach, which leverages its strengths to achieve more accurate CO<sub>2</sub> emission projections. Furthermore, XAI methodologies, such as Shapley Additive exPlanations plots, are used to improve model interpretability. This integration not only enhances model transparency but also provides more detailed insights into the impact of numerous vehicle characteristics on CO<sub>2</sub> emissions predictions, allowing for the creation of more effective emission reduction plans.

In this paper, we present an in-depth analysis of the forecasting CO<sub>2</sub> emissions from vehicles using deep learning and XAI techniques. Section '**Related Work**' review important literature on the topic to provide a context for our research. Section '**Method**' describes the approach used, which included data collection, preprocessing, model development, and evaluation techniques. Section '**Result Analysis**' presents the results and discussions that evaluate the performance of the proposed model and discuss valuable achievements. The '**Conclusion and Future Work**' closes by summarizing the main contributions, identifying shortcomings, and suggesting areas for future research.

## Related work

Artificial intelligence (AI) models have shown promise in reducing carbon dioxide (CO<sub>2</sub>) emissions through various mechanisms. Studies have shown that AI can notably lower CO<sub>2</sub> emissions, particularly in regions with advanced industrial structures<sup>19,20</sup>. AI algorithms enhance the Measurement and Verification (MV) protocols for energy-efficient infrastructure, leading to substantial reductions in both energy consumption and emissions<sup>21</sup>. For example, AI techniques, such as multi-gene genetic programming, have been effectively applied to model and optimize CO<sub>2</sub> capture in coal-fired power plants, achieving over 99% accuracy in emission predictions<sup>22</sup>. These AI models provide valuable insights for improving the design and operational strategies of CO<sub>2</sub> capture systems, contributing to long-term decarbonisation efforts<sup>22</sup>. Additionally, AI strategies have been successfully implemented at the city level to further reduce carbon emissions and support carbon neutrality goals<sup>23</sup>.

The application of AI in the chemical industry has also gained recognition for its ability to optimize processes, predict emissions, and support sustainable practices. This contribution is essential for the industry's transition toward net-zero emissions and overall<sup>24</sup>. The impact of AI on carbon reduction varies across countries, with more pronounced effects observed in high-carbon emission and high-income nations. This variation underscores the importance of considering industrial and demographic structures when designing strategies for emission reduction<sup>25</sup>. Collectively, these findings highlight the vital role that AI plays in advancing carbon neutrality and offer insights for policy recommendations and sustainable development strategies.

The transportation sector, which is responsible for approximately 16.2% of global CO<sub>2</sub> emissions, has also benefited from AI-driven advancements<sup>26</sup>. Machine learning (ML) algorithms, such as Random Forest, Decision Tree, and Regression Models, have been employed to accurately predict CO<sub>2</sub> emissions<sup>27</sup>. Among these, Gradient Boosting Regression (GBR) has been proven to be the most effective. These ML algorithms consider both socioeconomic and transportation-related factors when predict emissions<sup>28</sup>. Supervised machine learning regression approaches, validated through metrics such as the Root Mean Square Error (RMSE), have been used to enhance the accuracy of CO<sub>2</sub> emission forecasts<sup>29</sup>. In particular, the real-time forecasting of CO<sub>2</sub> emissions from automobiles in India has demonstrated the effectiveness of constructing ML models and optimizing hyperparameters to achieve precise predictions<sup>30</sup>.

Deep Neural Networks (DNNs) have gained popularity for estimating CO<sub>2</sub> emissions because of their ability to analyze complex patterns and correlations in large datasets. Advanced deep-learning algorithms, such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Recurrent Neural Networks (RNN), have been successfully utilized to forecast vehicle emissions with notable success<sup>31</sup>. Integrated models combining these algorithms have shown promise for enhancing the accuracy of emission predictions. For instance, LSTM and BiLSTM models have been effectively employed to predict CO<sub>2</sub> emissions based on vehicle characteristics, such as engine size, fuel type, and consumption rates<sup>32</sup>. Additionally, ensemble learning techniques that utilize deep neural networks have been suggested to reduce uncertainty in predicting vehicle energy efficiency, further improving the accuracy and robustness<sup>33</sup>.

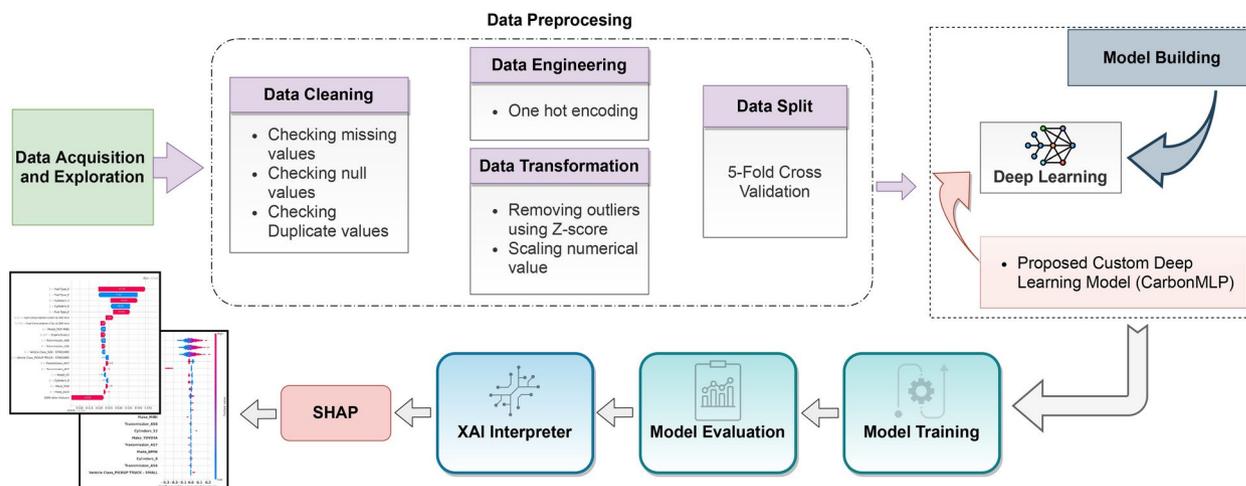
Ensemble approaches, which combine multiple models and algorithms, have proven to be highly interpretable, accurate, and robust in predicting carbon-related parameters. These models are invaluable tools for policymakers and environmental decision-making<sup>34,35</sup>. For example, models trained using real-time sensor data collected through OBD-II ports in vehicles have provided scalable and efficient methods for monitoring emissions at the vehicle level<sup>36</sup>. One particular machine learning model, UWS-LSTM, has demonstrated precise predictions of CO<sub>2</sub> emissions in hybrid vehicles, making it highly effective for smart vehicle applications that require fast and efficient results<sup>37</sup>.

explainable AI models have demonstrated varying levels of precision in estimating CO<sub>2</sub> emissions from vehicles. Studies have indicated that advanced non-linear multivariate models, such as ENGM(1,4), are more effective than conventional statistical models in forecasting transportation sector emissions, achieving superior accuracy<sup>38</sup>. Furthermore, advanced AI approaches, such as feed-forward neural networks (FFNN), adaptive network-based fuzzy inference systems (ANFIS), and LSTM, have been employed to forecast CO<sub>2</sub> emissions, with LSTM exhibiting particularly high accuracy<sup>39</sup>. Additionally, the use of Shapley Additive Explanations (SHAP) for lane-change decisions has provided clear and understandable explanations for the factors influencing AI model decisions, thereby improving the interpretability and trustworthiness of these models<sup>40</sup>.

In summary, AI's potential to reduce CO<sub>2</sub> emissions is widely recognized across various sectors, including energy, chemicals, and transportation. AI not only optimizes processes and predicts emissions but also supports sustainable practices and informs policy-making. The integration of AI in emission reduction efforts represents a critical pathway toward achieving carbon neutrality and addressing climate change on a global scale. An overview of the Literature Review is given in Table 1

Ref.	Dataset	Model	Accuracy	Limitation
32	Canadian government's open data portal	LSTM and BiLSTM	R2= 93.78%	Relies on existing datasets, challenging the capture of real-time or diverse data sources, affecting adaptability to dynamic emission scenarios.
41	Data collected from road tests using PEMS system and OBDII interface	Gradient Boosting	R2= 61.26%	Validation coefficients (R2 = 0.61, MSE = 0.77) suggest improvement needed in predictive performance.
42	Automotive industry dataset, fuel consumption based on NEDC and WLTP tests	MLP	R2 value < 98.00%	Focused on passenger cars with internal combustion engines, limiting scope to this vehicle type.
43	Real-world GPS tracking data from electric vehicles in Japan	LightGBM, XGBoost, ANN, and LR	R2=98.00%	Model overlooks factors such as additional loads, driver behavior, and traffic conditions.
44	Data from literature, covering fluidized bed gasifiers at steady-state conditions	Gradient Boosting	R2=84.68%	Highlights need for Explainable AI (XAI) but doesn't address broader implications of adopting XAI.
37	Data from hybrid vehicle performance and 20+ operational parameters, collected using SprintIR-R 20 CO <sub>2</sub> sensor	UWS-LSTM	R2=97.50%	Dataset standardization not detailed, affecting reproducibility.
45	Government of Canada dataset on light-duty vehicles (2017–2021)	Univariate Polynomial Regression	R2=98.60%	Lacks higher-performing model for predictive fuel consumption and dataset with more features.
46	On-road remote sensing measurements linked to I/M records for 103,831 light-duty gasoline vehicles in Hangzhou, China	NN, RF, XGBoost, and Ensemble model	R2=88.00%	Relies solely on ORRS data, which may misidentify high/low-emitting vehicles.
47	World Bank, Turkish General Directorate of Highways, and Turkish Statistical Institute data	MLP, XGBoost, and SVM	R2=98.86%	Examined only four scenarios based on correlation impact of input values.
48	Data from Portable Emissions Measurement System (PEMS) recording NOx, CO <sub>2</sub> , and other pollutants from diesel vehicles	Gradient Boosting Regression	R2=99.00%	Prediction accuracy limited by minimal input features.
49	Data collected from real-world driving conditions of several braking events	CatBoost	R2=83.00%	Dataset limited to 600 braking events, not fully representing all scenarios.
50	Open Energy Data Initiative (OEDI) dataset	ML Ensemble Model	94.70%	Model is costly and computationally intensive.

**Table 1.** Overview of the literature review.



**Fig. 2.** Methodology Diagram Illustrating the Entire Process Described in the Paper, from Data Collection to Model Evaluation and Interpretation.

Variable	Description	Example Values
Maker	Manufacturer of the vehicle	ACURA, TOYOTA, FORD
Model	Specific model of the vehicle	ILX, MDX 4WD, RDX AWD
Vehicle Class	Classification of the vehicle type	COMPACT, SUV - SMALL, MID-SIZE
Engine Size (dm <sup>3</sup> )	Engine displacement in cubic decimeters	2, 2.4, 3.5
Cylinders	Count of engine cylinders	2, 4, 6
Transmission	Type of vehicle transmission	AS5, M6, AV7
Fuel Type	Type of fuel used	Z (Denotes a specific type, e.g., gasoline, diesel)
Fuel Consumption City (L/100 km)	City fuel consumption rate in liters per 100 km	9.9, 11.2, 6
Fuel Consumption Hwy (L/100 km)	Highway fuel consumption rate in liters per 100 km	8.5, 9.6, 5.9
Fuel Consumption Comb (L/100 km)	Combined fuel consumption rate in liters per 100 km	6.7, 7.7, 5.8
Fuel Consumption Comb (mpg)	Combined fuel consumption rate in miles per gallon	33, 29, 48
CO <sub>2</sub> Emissions (g/km)	Carbon dioxide emissions per kilometer	196, 221, 136

**Table 2.** Description of dataset features.

### Methods and materials

The methodology of this study centers exclusively on deep learning techniques to comprehensively forecast CO<sub>2</sub> emissions from vehicles. The process began with preprocessing and data collection, followed by the creation of the MLP model. Additionally, explainable AI (XAI) methods, particularly Shapley Additive Explanations (SHAP), have been utilized to enhance the interpretability and reliability of predictions. An overview of the methodology used in this study is shown in Figure 2.

### Data collection and exploration

This study’s dataset, which was obtained from Kaggle, provides extensive information on how CO<sub>2</sub> emissions vary depending on various vehicle parameters. The dataset compiles data from the Canadian government’s official open data website, spanning a period of seven years [Dataset](#). The dataset, which includes 12 columns and 7385 rows across a 7-year span is an extensive information source. Information on vehicle models, fuel types, transmissions, city and highway fuel consumption ratings, and CO<sub>2</sub> emission levels is all included. Table 2 presents an overview of the characteristics, explanations, and associated values of the dataset.

The “uel Consumption Comb (mpg)” column in the dataset was originally added to represent fuel consumption in miles per gallon (mpg). However, additional analysis revealed that the reported values did not match the normal conversion from litres per 100 km (L/100 km) to miles per gallon (mpg)<sup>51</sup>. The correct conversion formula is as follows:

$$\text{mpg} = \frac{235.215}{L/100Km} \tag{1}$$

Due to the disparity between the reported “Fuel Consumption Comb (mpg)” values and the expected values estimated using the conversion procedure, this column was removed from the dataset. Instead, the column reflecting fuel usage in litres per 100 km (L/100 km) will be used for additional analysis and modelling.

A distinct pattern emerged from the analysis of CO<sub>2</sub> emissions by vehicle model: vehicles with high-performance engines, such as SRT, Rolls Royce, and Lamborghini, had the highest emissions, whereas fuel-efficient models, such as Smart and Honda, exhibited the lowest emissions. This observation is consistent with the theory that large engines produce more CO<sub>2</sub>.

To quantify these observations, we calculated the mean CO<sub>2</sub> emissions for 41 distinct car models displayed in the visualization. The mean was estimated by grouping the dataset by vehicle maker and applying the following formula<sup>52</sup>:

$$\text{Mean CO}_2 \text{ Emissions} = \frac{1}{n} \sum_{i=1}^n \text{CO}_{2i} \quad (2)$$

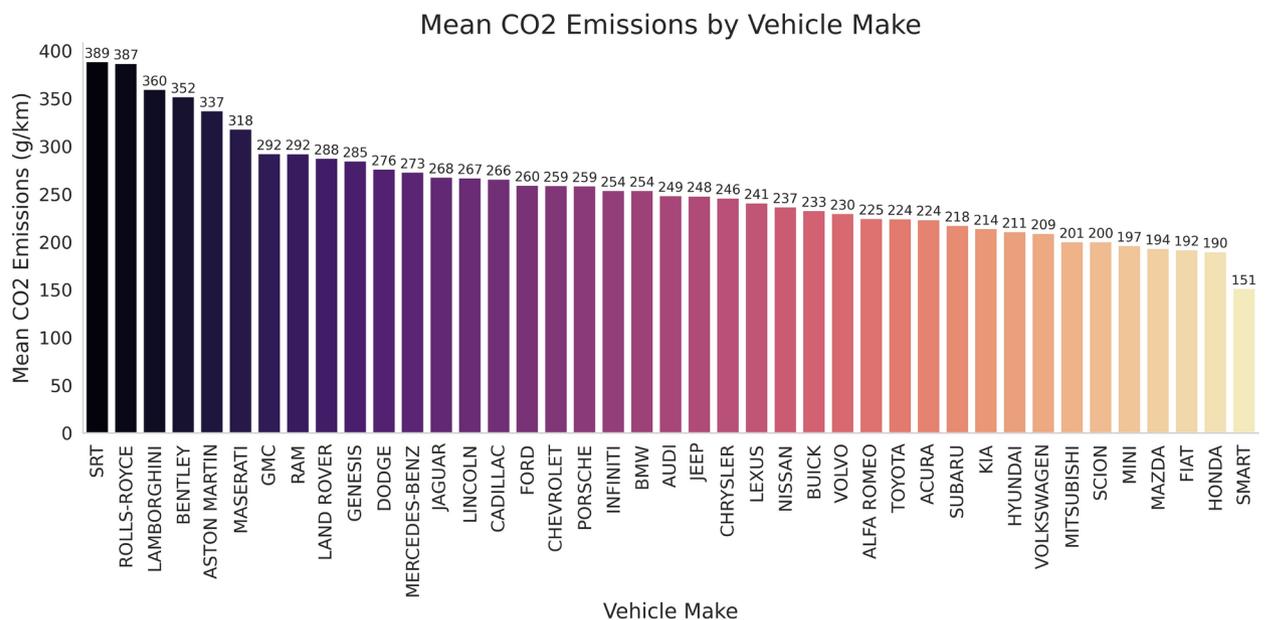
where  $n$  represents the total number of entries for a specific vehicle model, and  $\text{CO}_{2i}$  denotes the CO<sub>2</sub> emissions for each entry.

This analysis, illustrated in Figure 3, provides insightful information on how vehicle design influences the environmental impact and underscores the importance of considering emissions in feature selection for predictive analysis.

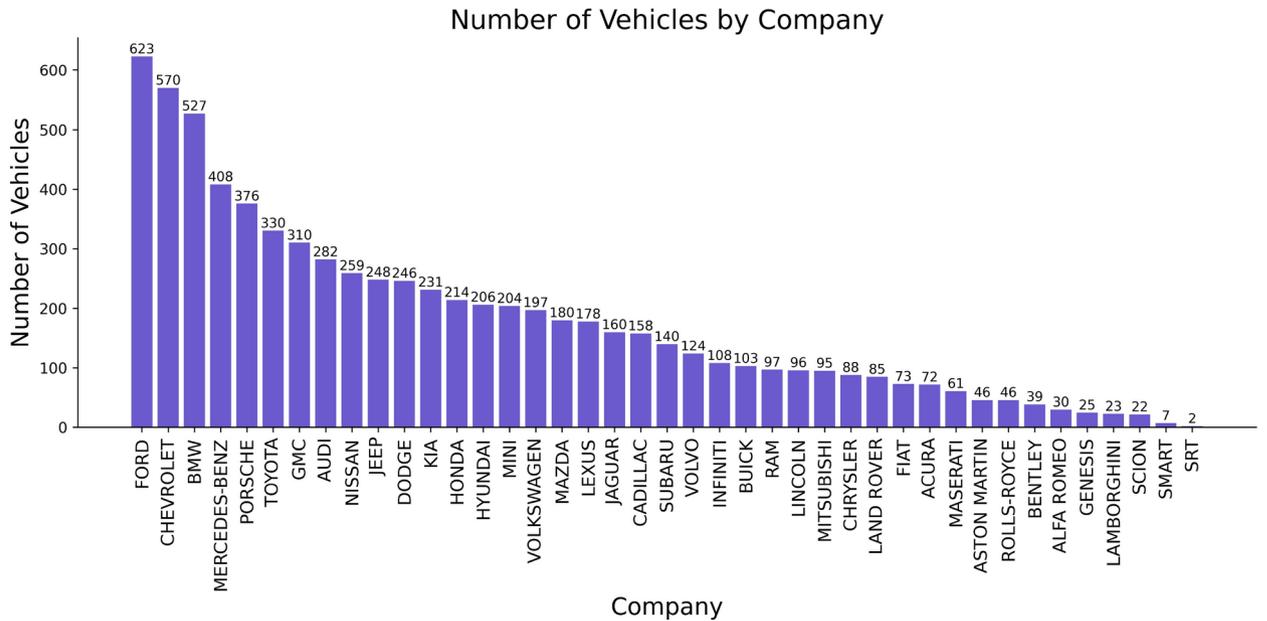
The distribution of vehicles manufactured by each company was skewed when analyzing the dataset collected over a period of seven years in Canada. Ford had the highest number of cars (623); however, its average CO<sub>2</sub> emissions were relatively moderate. This distinction highlights that while Ford’s large vehicle count (623 cars × 270 g/km/car = 168,210 g/km CO<sub>2</sub> emissions) significantly contributes to total emissions, it does not result in the highest average emissions per vehicle. In contrast, SRT had the fewest automobiles in the dataset, whereas Chevrolet was the second-largest manufacturer in terms of vehicle count. Figure 4 provides an overview of the number of vehicles produced by each manufacturer, demonstrating their contributions to overall CO<sub>2</sub> emissions.

The analysis of the fuel usage trends for various fuel types revealed some notable patterns, as shown in Figure 5. Among the fuel types, ethanol (denoted as “E”) exhibited the highest fuel consumption. This increased consumption may be attributed to the lower energy density of ethanol compared to that of gasoline. In contrast, fuels labelled “X” (regular gasoline), “Z” (premium gasoline), and “D” (diesel) demonstrated lower fuel consumption levels. Although natural gas, represented by “N”, is included in the dataset, only one vehicle utilized this fuel, and therefore it is not prominently displayed in Figure 6 due to its limited representation.

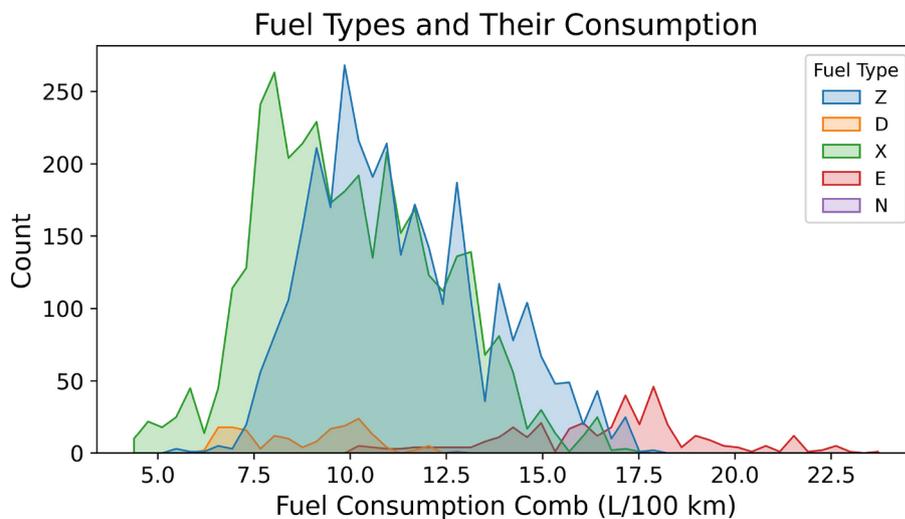
Figure 6 illustrates that despite the higher fuel consumption of ethanol, its CO<sub>2</sub> emissions are comparable to those of other fuels, particularly gasoline and diesel. This suggests that while ethanol may require more fuel per



**Fig. 3.** Mean CO<sub>2</sub> Emissions by Vehicle Make: This figure shows the relationship between vehicle models and CO<sub>2</sub> emissions, helping identify designs that contribute to higher emissions and guiding feature selection in our predictive analysis.



**Fig. 4.** Number of Vehicles Made by Maker (Company): This figure shows the distribution of vehicles by manufacturer, highlighting their impact on overall CO<sub>2</sub> emissions. This insight guides our handling of manufacturer-related categorical variables in the modelling phase.

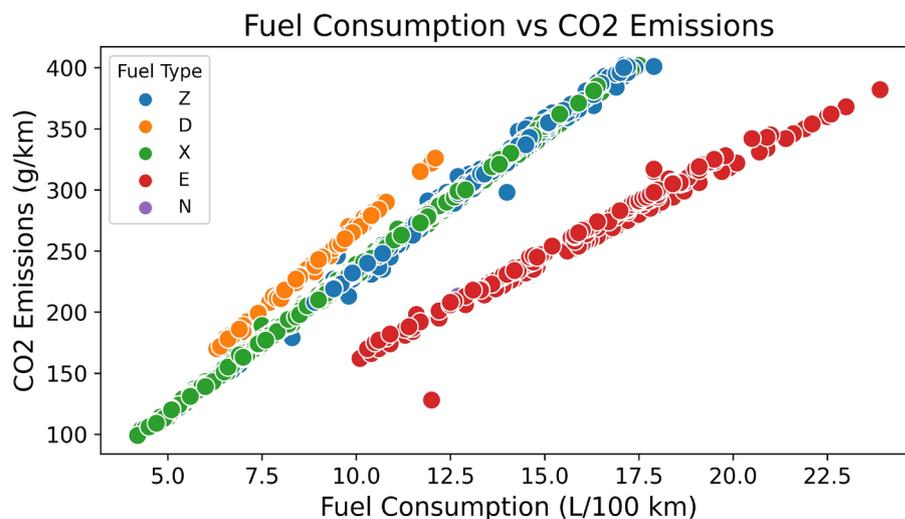


**Fig. 5.** Histogram of Fuel Consumption by Fuel Type: This figure illustrates how fuel types affect fuel consumption, a key predictor of CO<sub>2</sub> emissions. Analyzing these trends enhances our understanding of the impact of fuel type choices on model predictions and emissions.

kilometer, it does not produce proportionally higher CO<sub>2</sub> emissions. Most vehicles in the dataset emit CO<sub>2</sub> in the range of 200 to 300 g/km, regardless of fuel type, with ethanol slightly overlapping this range.

A summary of the CO<sub>2</sub> emissions dataset shown in Table 3 is provided using descriptive statistics. Important details including the engine size (ES) litres, number of cylinders, fuel consumption (FC) in city, highway, combination (litres per 100 km), CO<sub>2</sub> emissions (grams per kilometer), and the total number of observations are listed in the table. As we can see, there is a standard deviation of 1.35 litres and an average engine size of 3.16 litres. A variety of engine sizes, ranging from 0.9 litres to 8.4 litres, are also shown by the statistics. This table also provides information on the differences in CO<sub>2</sub> emissions and fuel usage between the different cars. This synopsis establishes the framework for additional investigation into the variables impacting CO<sub>2</sub> emissions.

Interesting trends were found by analyzing the links between the features (apart from the object features) using the correlation heatmap in Figure 7. Engine size demonstrates a robust positive correlation with cylinder count (0.93), and a moderately strong correlation with fuel consumption measures: city (0.83), highway (0.75), and combined (0.81), as well as CO<sub>2</sub> emissions (0.85). This indicates that larger engines with more cylinders



**Fig. 6.** Scatter Plot of CO<sub>2</sub> Emission by Fuel Consumption: This figure compares CO<sub>2</sub> emissions across fuel types, highlighting their environmental impact.

Statistic	ES (L)	Cylinders	FC City (L/100 km)	FC Highway (L/100 km)	FC Comb. (L/100 km)	CO <sub>2</sub> Emissions (g/km)
Count	7385	7385	7385	7385	7385	7385
Mean	3.160	5.610	12.55	9.040	10.98	250.6
Standard Deviation	1.350	0001	3.500	2.220	2.900	58.51
Minimum	0.900	0003	4.200	0004	4.100	0096
Median	0003	0006	12.10	8.700	10.60	0246
Maximum	8.400	0016	30.60	20.60	26.10	0522

**Table 3.** Descriptive statistics of vehicle parameters.

(0.93) typically lead to higher fuel consumption, with city consumption exhibiting the strongest association. Additionally, vehicles with higher city fuel consumption generally have higher highway and combined fuel consumption, as demonstrated by the high correlations between these measures (all above 0.94). This reinforces the idea that greater fuel consumption during city driving is associated with higher consumption on highways and overall. Correlation analysis was conducted using Python, with Pandas for data handling and Seaborn for generating the heatmap.

### Data Pre-processing

Data pre-processing is a crucial step in any deep learning modelling, as it directly affects the quality of the model's predictions. In this study, we carried out a comprehensive and methodical pre-processing phase to ensure that the dataset was clean, structured, and ready for analysis. Pre-processing involves handling missing values, transforming variables, addressing outliers, and preparing the data for the deep learning pipeline. The goal of this process was to eliminate any inconsistencies or anomalies that might hinder the model's performance. The pre-processing workflow involved several key steps, which are outlined below:

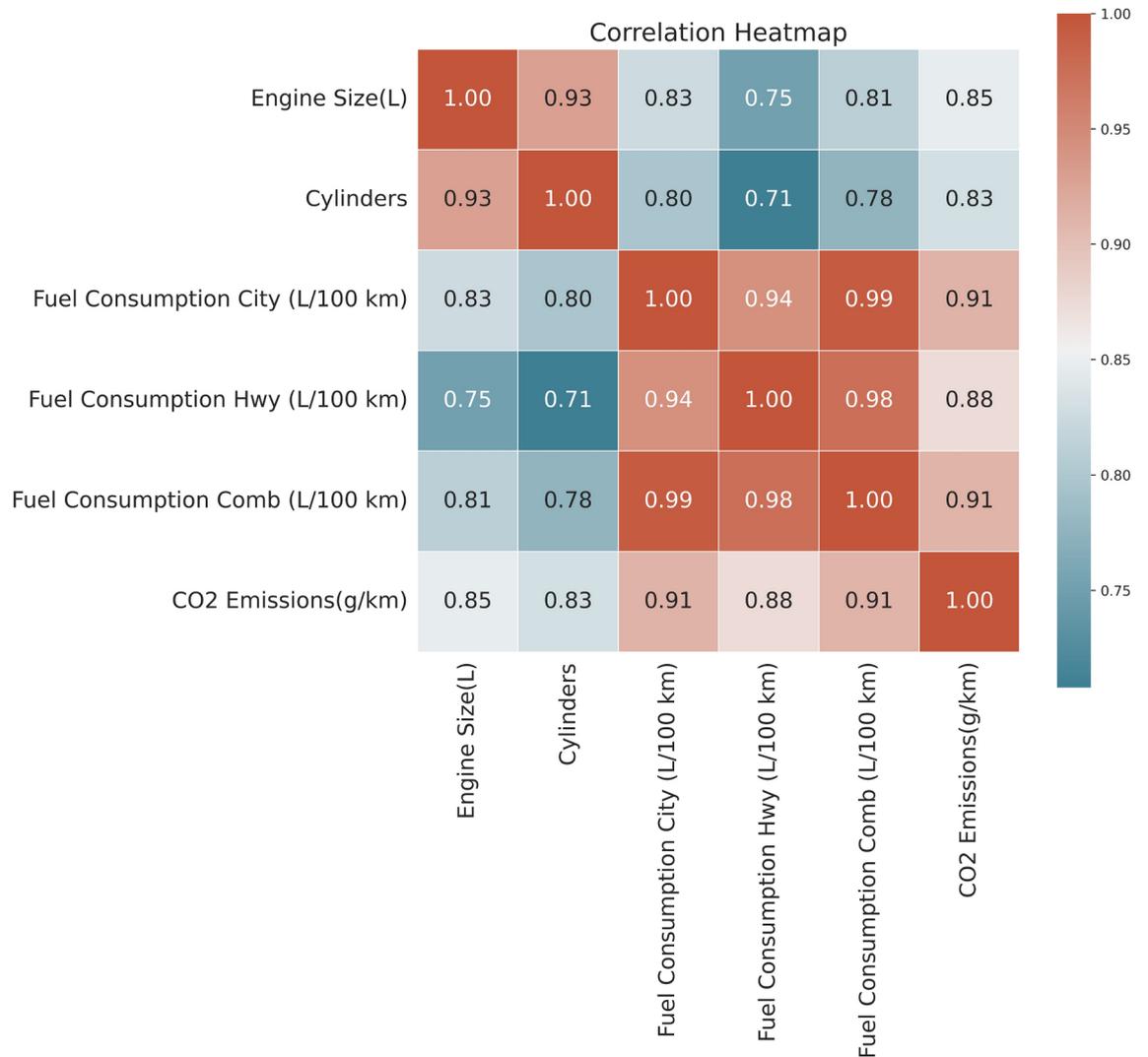
#### Data cleaning

- **Missing and null values:** Missing or null values in the dataset can lead to bias or inaccuracies in model training, especially when important variables are incomplete. The dataset was checked for both the values.
- **Duplicate Values:** Duplicate records in the dataset can skew results, leading to incorrect model outcomes by over-representing certain observations. A thorough check was conducted to identify any duplicate rows. After determining whether any entries in the data were duplicates, values were eliminated. This prevents the analysis from being skewed and guarantees that each data point represents a distinct observation.

#### Data engineering and transformation

Data engineering is crucial for improving the predictive capability of deep learning models. In our study, we used a variety of strategies to prepare and modify the dataset, ensuring that the model could accurately forecast CO<sub>2</sub> emissions based on different vehicle features. The following methods were applied to prepare the dataset:

- **Outlier Detection and Removal:** Outliers were identified using z-score analysis and eliminated from the dataset. This phase is critical for avoiding skewed model predictions induced by extreme values that could



**Fig. 7.** Correlation Heatmap of Vehicle Characteristics: The correlation heatmap shows strong relationships among key numerical features, particularly engine size, cylinder count, and fuel consumption metrics. This suggests multicollinearity that can affect model performance and assists in feature selection for better interpretability.

affect the learning process. Data points with a z-score greater than 2.6 were declared outliers and eliminated, ensuring that the model was trained on clean data.

$$z_i = \frac{x_i - \mu}{\sigma} \quad (3)$$

Where  $x_i$  is an individual data point,  $\mu$  is the mean of the data, and  $\sigma$  is the standard deviation.

- **Categorical Feature Encoding:** To transform categorical variables (such as make, model, vehicle class, transmission, and fuel type) into numerical representations suitable for the model, we first used one-hot encoding. This approach divides each category into binary columns, allowing the model to effectively learn from these variables. However, we also assessed target encoding, which calculates the mean target value for each category and replaces it with mean values. Although target encoding can reduce dimensionality, we discovered that one-hot encoding improved the interpretability of the SHAP values in our models.
- **Normalization and Scaling:** To ensure that every feature contributes equally to the model's performance, the numerical features were min-max scaled. This scaling method converts each feature to a common range (0 to 1), thereby reducing the impact of varying units and magnitudes. This was particularly crucial for features such as fuel consumption and engine size, which had different scales. The equation for min-max scaling is:

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (4)$$

where  $x_{\text{max}}$  and  $x_{\text{min}}$  are the minimum and maximum values, respectively.

In summary, the data engineering process was thorough and aimed at enhancing the predictive capabilities of the model. By detailing the methods used for data preparation and their resultant effects on the dataset, we provide a clearer understanding of how these techniques contribute to overall analysis and modelling efforts.

#### *Dataset splitting and validation*

The performance and robustness of the model were rigorously assessed in this study using a 5-fold cross-validation technique. This approach successfully generalizes the model to unseen data, lowering the possibility of overfitting and enhancing the accuracy of our findings. The dataset was systematically partitioned into five equal sections, denoted as  $D_1, D_2, D_3, D_4$ , and  $D_5$ . In each iteration of the cross-validation process, four of these sections were used for training the model and the remaining section served as the validation set. This process can be mathematically represented as follows:

$$\text{Training Set} = D_i \quad \text{for } i \in \{1, 2, 3, 4, 5\} \quad (5)$$

$$\text{Validation Set} = D_j \quad \text{for } j \in \{1, 2, 3, 4, 5\}, \quad j \neq i \quad (6)$$

This cross-validation procedure was repeated five times to ensure that each subset was used once as the validation set. The overall process is summarized as follows:

#### 1. **Iteration 1:**

- Training on  $D_2, D_3, D_4, D_5$
- Validation on  $D_1$

#### 2. **Iteration 2:**

- Training on  $D_1, D_3, D_4, D_5$
- Validation on  $D_2$

#### 3. **Iteration 3:**

- Training on  $D_1, D_2, D_4, D_5$
- Validation on  $D_3$

#### 4. **Iteration 4:**

- Training on  $D_1, D_2, D_3, D_5$
- Validation on  $D_4$

#### 5. **Iteration 5:**

- Training on  $D_1, D_2, D_3, D_4$
- Validation on  $D_5$

The 5-fold cross-validation method provides a more legitimate estimate of model generalization by testing multiple data partitions, thereby reducing the potential for overfitting. By averaging the performance metrics across all five iterations, we comprehensively evaluated the performance of the model. This approach ensures that the findings are not unduly influenced by a specific train-test split, offering a more reliable assessment of the model's predictive capabilities.

### **Model building**

Deep learning is a subset of machine learning that uses artificial neural networks and is a useful method for discovering intricate patterns and relationships in data<sup>53</sup>. Deep learning models are distinct from typical machine-learning algorithms in that they are composed of numerous layers of interconnected neurons, which allows them to automatically extract meaningful features from raw data<sup>54</sup>. Deep learning models can be especially useful in CO<sub>2</sub> emission prediction because of their capacity to manage non-linear correlations between vehicle features and emissions. Owing to their multi-layered architecture, deep learning models can capture these complicated interactions more efficiently than traditional machine learning models, which makes it difficult to handle such complexities<sup>55</sup>. A variety of deep learning architectures are appropriate for regression problems such as CO<sub>2</sub> emission prediction<sup>56</sup>. This section presents the proposed architecture and its development procedures.

#### *Proposed model development*

This study introduced a novel approach for predicting CO<sub>2</sub> emissions from vehicle attributes. We constructed a light deep learning model using a multilayer perceptron (MLP) architecture. MLPs, which are neural network

foundations, are composed of interconnected neuronal layers. This method harnesses the power of deep learning to identify intricate connections between input features and the target variable (CO<sub>2</sub> emissions).

**Input Format:** The input data for the proposed deep learning model consists of a dataset with multiple features related to vehicle attributes, as described in Table 2. These features include attributes such as Make (vehicle manufacturer), Model, Vehicle Class (VC), Engine Size (ES), Cylinders, and Fuel Type (FT), along with various measures of fuel consumption city (FCCity), fuel consumption highway (FCH), and fuel consumption combined (FCcomb).

In this study multiple densely connected layers were used, each with ReLU activation for non-linearity. The design consists of an input layer, three hidden layers with 128, 64, and 32 neurons each, and a single regression (linear activation) neuron in the final output layer.

The model was built using the Adam optimizer, which is a highly efficient tool known for its effectiveness, especially in large-scale models. The mean squared error (MSE) loss function is also employed. The design was facilitated using TensorFlow’s Keras Application Programming Interface (API). Figure 8 provides a visual representation of the architecture of the proposed deep learning model.

The goal of the proposed model is to forecast CO<sub>2</sub> emissions using the provided dataset. In its formulated form, the model architecture is:

$$\text{Input: } x_{\text{train}} \quad (\text{shape: } (n, m)) \tag{7}$$

$$\text{Hidden 1: } h^{(1)} = \text{ReLU}(x_{\text{train}} \cdot W^{(1)} + b^{(1)}) \tag{8}$$

(shape: (n, 128))

$$\text{Hidden 2: } h^{(2)} = \text{ReLU}(h^{(1)} \cdot W^{(2)} + b^{(2)}) \tag{9}$$

(shape: (n, 64))

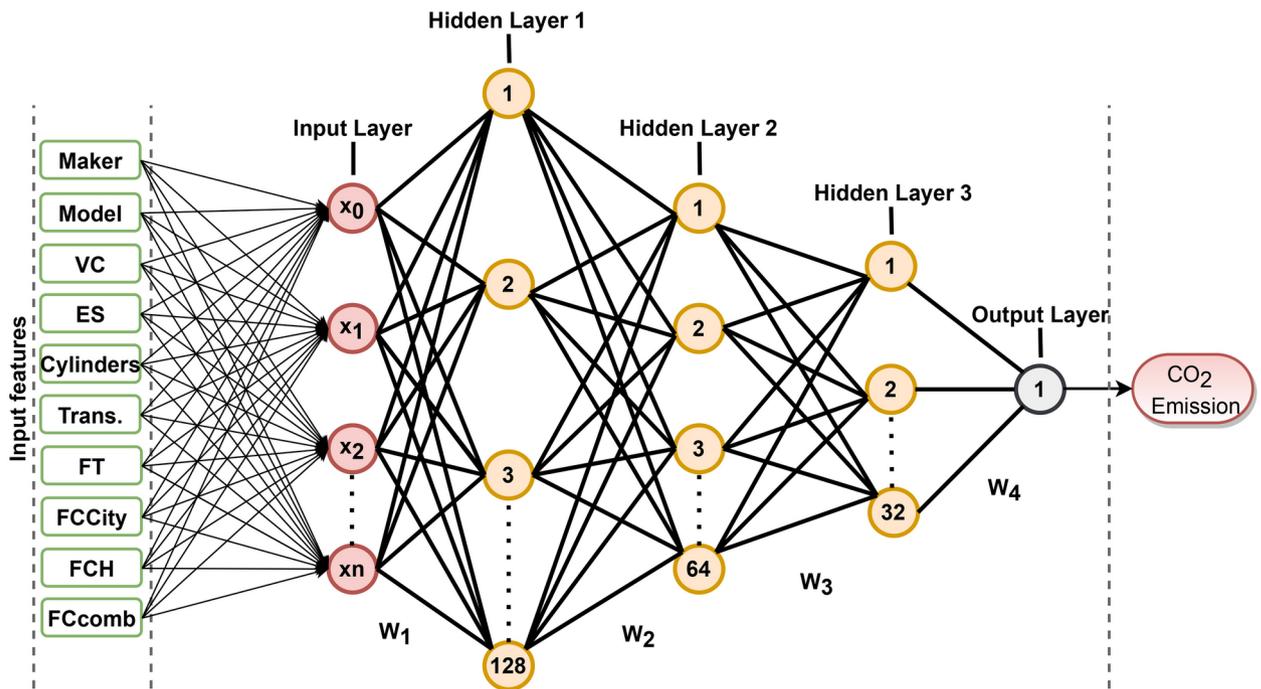
$$\text{Hidden 3: } h^{(3)} = \text{ReLU}(h^{(2)} \cdot W^{(3)} + b^{(3)}) \tag{10}$$

(shape: (n, 32))

$$\text{Output: } Y_{\text{pred}} = h^{(3)} \cdot W^{(4)} + b^{(4)} \tag{11}$$

(shape: (n, 1))

In the given neural network architecture, the entire process from input to output can be traced through several equations, each detailing a specific layer or operation. The input layer defined in Equation (7), uses the training data  $x_{\text{train}}$  which has  $n$  samples and  $m$  features. Equation (8) describes the first hidden layer  $h^{(1)}$  where the ReLU activation function is employed after combining the inputs with weights  $W^{(1)}$  and biases  $b^{(1)}$



**Fig. 8.** Architectural Diagram of the Proposed CarbonMLP Model: Shows the input layer with three hidden layers (128, 64, and 32 neurons, respectively), and the output layer with a single regression neuron. Each layer uses ReLU activation functions, except for the final output layer, which uses linear activation.

. Further transformations occurred in subsequent hidden layers. The second hidden layer  $h^{(2)}$  is represented by Equation (9), again using the ReLU function. However, the outputs from the first hidden layer are processed with a new set of weights  $W^{(2)}$  and biases  $b^{(2)}$ . The third hidden layer followed a similar pattern, as detailed in Equation (10), processing the output of the second hidden layer using the weights  $W^{(3)}$  and biases  $b^{(3)}$ . The network output  $Y_{\text{pred}}$ , which predicts CO<sub>2</sub> emissions, was calculated using Equation (11). This output is the result of processing the activation of the third hidden layer using the final set of weights  $W^{(4)}$  and biases  $b^{(4)}$ . Each step relies on weights  $W^{(i)}$  and biases  $b^{(i)}$  for each layer, driving forward the network's ability to learn and make accurate predictions based on input data.

#### Optimized parameters

Our deep learning model underwent extensive fine-tuning using various hyperparameters to ensure reliable and consistent CO<sub>2</sub> emissions forecasts. After evaluating the multiple configurations, we identified the optimal settings for the proposed architecture. The selected design employs ReLU activation for non-linearity and consists of three hidden layers striking an effective balance between efficiency and complexity. The Mean Squared Error (MSE) loss function aligns well with our regression objectives, whereas the Adam optimizer enhances training efficiency.

In order to mitigate overfitting, the model was trained for 100 epochs with a batch size of 8, using a 5-fold cross-validation approach to ensure effective learning from the data. These carefully selected hyperparameters significantly improved the robustness and accuracy of the model in predicting CO<sub>2</sub> emissions. Table 4 provides a comprehensive overview of the strategies employed to optimize performance, detailing the various hyperparameters involved in the model optimization process.

#### Explainable AI interpretation

explainable AI (XAI) methods are designed to enhance the interpretability of models and provide insights into the elements that influence predictions. In this study, we employed robust XAI techniques, specifically SHapley Additive exPlanations (SHAP), to gain insight into the impact of features on CO<sub>2</sub> emission projections. We used a series of visualizations, including SHAP Summary, waterfall, force, and dependence plots.

#### SHAP (SHapley Additive exPlanations)

The SHAP values offer a robust framework for explaining individual predictions by quantifying the contribution of each of the  $i^{\text{th}}$  feature to the output of the model. The SHAP value for feature  $i$  is calculated using the following equation:

$$SHAP_i = \phi_0(f) + \sum_{j=1}^M \frac{(M-j)! \cdot j!}{M!} (\phi_j(f) - \phi_{j-1}(f)) \quad (12)$$

where;  $SHAP_i$  represents the SHAP value for the  $i^{\text{th}}$  feature,  $\phi_0(f)$  denotes the baseline contribution of the model output,  $f$  is the proposed CarbonMLP model that maps the input features to the predicted CO<sub>2</sub> emissions,  $M$  is the total number of features.

To calculate the contribution of the features, the SHAP method considers all possible combinations of feature values and their respective outputs, providing a fair distribution of the model's prediction among the input features. This means that each feature's contribution is evaluated in the context of all other features, ensuring that the interactions are properly accounted for. The features used in this study and their descriptions are listed in Table 2. These include attributes such as make (vehicle manufacturer), Model, Vehicle Class, Engine Size, Cylinders, and Fuel Type, as well as various measures of fuel consumption (city, highway, combined), which are essential to understanding how vehicle characteristics affect CO<sub>2</sub> emissions. The SHAP value explains the impact of each feature on model output predictions. SHAP values not only quantify the contributions of individual features but also allow for a deeper understanding of how vehicle characteristics influence CO<sub>2</sub> emission

Hyperparameter	Value
Number of Hidden Layers	3
Neurons per Hidden Layer	128, 64, 32
Activation Function	ReLU (Hidden Layer) Linear (Output Layer)
Optimizer	Adam
Loss Function	Mean Squared Error
Epochs	100
Batch Size	8
Validation Split	5-Fold

**Table 4.** Optimized hyperparameters of the proposed CarbonMLP model.

predictions. This methodology enhances model transparency and aids stakeholders in making informed decisions based on analysis.

#### SHAP summary plot

The SHAP summary plot provides a global view of the importance of the mean absolute SHAP value of each feature. This helps with model interpretation and validation by enabling the identification of important predictors and their particular effects on model predictions. In Equation (13), Where  $N$  represents the number of features. The plot was calculated as follows:

$$\text{SHAP Summary Plot} = \sum_{i=1}^N |SHAP_i| \quad (13)$$

#### SHAP waterfall plot

The SHAP waterfall plot shows how each feature contributes to the variation in the base value, providing detailed insights into individual predictions. This makes it easier to interpret certain forecasts by emphasizing the variables that influence model outputs and possible areas for development. In Equation (14), is calculated as follows:

$$\text{SHAP Waterfall plot} = \text{Base Value} + \sum_{i=1}^N SHAP_i \quad (14)$$

#### SHAP force plot

The SHAP force plot illustrates how each characteristic affects a single prediction, and shows how the model determines the output for a given instance. This allows feature impacts to be examined, highlighting how each contributes to the final prediction and improves model transparency. In Equation (15), is calculated for the force plot as:

$$\text{SHAP Force plot} = \text{Base Value} + \sum_{i=1}^N SHAP_i \quad (15)$$

#### SHAP dependence plot

The SHAP dependence plot considers the relationships with other variables and shows the relationship between a feature and the model output forecast. It provides important insights into the feature behavior and model performance by assisting in the discovery of complex patterns and nonlinear relationships. The plot was calculated using Equation (16):

$$\text{Output Prediction} = f(x) + \sum_{i=1}^N SHAP_i \quad (16)$$

### Evaluation metrics

Evaluation metrics are crucial for evaluating the effectiveness and performance of predictive models in practical applications. This component contained the metrics used to assess the performance of the proposed model. The metrics included the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2), and Mean Absolute Percentage Error (MAPE).

#### Mean Squared Error (MSE)

The efficacy of the model was assessed using Mean Squared Error (MSE), which measures the average squared difference between the predicted and observed results. This calculation is expressed by the following equation:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (17)$$

where,  $n$  are the number of samples,  $y_i$  represents the target value, and  $\hat{y}_i$  denotes the target value, respectively.

#### Root Mean Squared Error (RMSE)

The square root of the Mean Squared Error (MSE), also known as the Root Mean Square Error (RMSE), is the mean difference between the observed and predicted outcomes. The RMSE was calculated as follows :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (18)$$

### R-squared ( $R^2$ )

The R-squared ( $R^2$ ) statistic illustrates the extent to which the independent variables account for the variance in the dependent variable. On the scale, which goes from zero to one, higher numbers denote a better model fit. The  $R^2$  formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (19)$$

where,  $\bar{y}$  represents the mean of the observed values.

### Mean Absolute Percentage Error (MAPE)

The average percentage variation between the actual and anticipated values is measured by the Mean Absolute Percentage Error (MAPE), which sheds light on the accuracy of the predictions of the magnitude of the actual value. It is calculated as:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (20)$$

where,  $y_i$  denotes the actual value,  $\hat{y}_i$  the predicted target value of the  $i^{\text{th}}$  sample,  $n$  the number of samples.

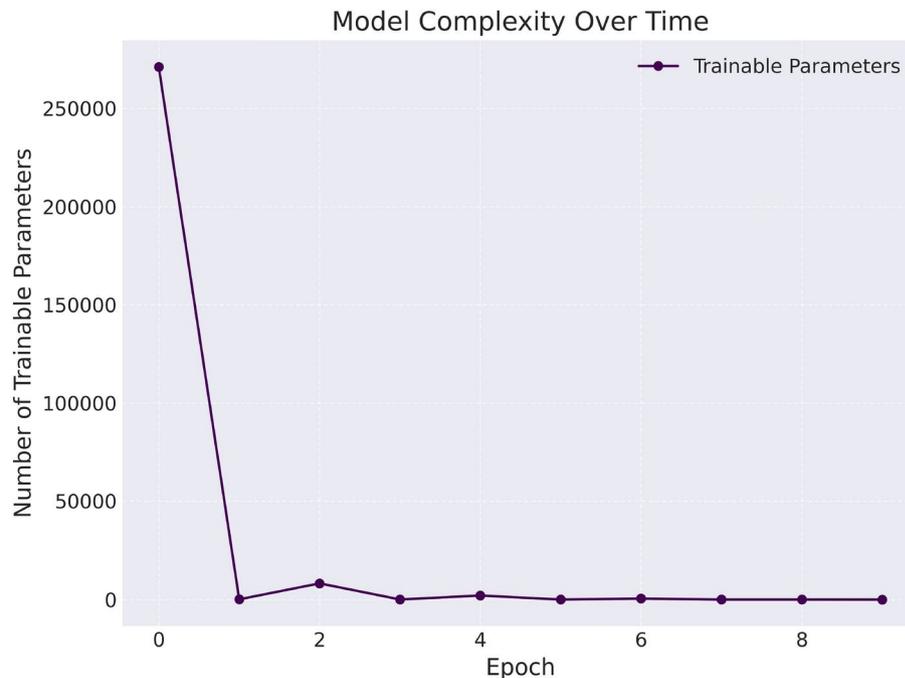
## Experimental results

### Hardware setup and model training

The proposed deep learning model was trained using a powerful GPU P100 accelerator in a Kaggle notebook environment to enhance the processing efficiency. This platform offers a scalable solution for computationally intensive tasks, including deep neural network training. The Keras API was employed for the model training.

Batches of pre-processed data were fed through the model during training, and the Adam optimizer was used to minimize the Mean Squared Error (MSE) loss function. After fine-tuning various hyperparameters, the optimal values were selected: a batch size of 8 and training for 100 epochs with 5-fold cross-validation to ensure robust generalization and avoid overfitting. This carefully designed training procedure allows the model to effectively capture the intricate relationships between vehicle attributes and  $\text{CO}_2$  emissions, resulting in highly reliable prediction capabilities. The complexity of the model over time, as measured by the number of trainable parameters, is shown in Figure 9.

As the training progressed, Figure 9 shows a significant initial drop in the number of trainable parameters within the first two epochs, followed by stabilization across subsequent epochs. This indicates that the model was



**Fig. 9.** Model complexity over time of CarbonMLP model.

quickly adjusted to optimally represent the relationships between vehicle attributes and CO<sub>2</sub> emissions, without over-complicating the structure.

This gradual reduction in model complexity helps to avoid overfitting, demonstrating that the model learns efficiently from the data without increasing unnecessary parameters. The modest increase in trainable parameters in the first few epochs reflects the model's capability to represent intricate patterns, after which it stabilizes, reinforcing the robustness of the model.

### Computational efficiency comparison

To further justify our selection of the proposed CarbonMLP model, we conducted a detailed computational efficiency comparison with more complex LSTM<sup>57</sup> and BiLSTM<sup>58</sup> models. The comparison was based on major performance measures, such as training time, inference time, memory usage, parameter count, and latency. We chose LSTM and BiLSTM models because they are widely used for time-series prediction as well as sequence modeling tasks like CO<sub>2</sub> emissions, making their computational effectiveness benchmark suitable for comparing with the current deep learning models. These measures are calculated under equal conditions to ensure fairness. The results of this comparison are presented in Table 5.

#### *Training and inference time*

Compared to LSTM and BiLSTM, the proposed CarbonMLP model significantly reduces the training and inference times. These reductions make CarbonMLP better suited for real-time and large-scale deployment. CarbonMLP requires 54.7% less time for training than LSTM and 43.5% less time than BiLSTM. Similarly, the inference time for CarbonMLP was 34.8% shorter than LSTM and 25% shorter than BiLSTM, indicating its computational efficiency in production environments.

#### *Memory usage and model size*

Memory usage and model size are critical in contexts with limited computational resources. CarbonMLP had a significantly reduced footprint in both measurements. CarbonMLP consumes 3% less memory than LSTM and BiLSTM. Additionally, the model size of CarbonMLP was 76.6% smaller than that of LSTM and 88.7% smaller than that of BiLSTM, making it highly efficient for deployment in devices with limited storage capacity.

#### *Latency and batch processing time*

Latency and batch processing times are crucial for models to make quick predictions. CarbonMLP outperforms LSTM and BiLSTM in these areas. The proposed CarbonMLP offers a 61.5% reduction in latency compared with LSTM and a 31.8% reduction compared with BiLSTM. Similarly, its batch processing time is 19% faster than that of LSTM and 15% faster than that of BiLSTM, further demonstrating its efficiency in high-throughput environments.

#### *Model complexity and number of parameters*

A model with fewer parameters is easier to train, deploy, and maintain, particularly in cases with limited computational power. CarbonMLP was developed to be lightweight while still providing a competitive performance. CarbonMLP contained 76.8% fewer parameters than the LSTM model and 88.8% fewer parameters than the BiLSTM model, making it highly efficient in terms of model complexity without sacrificing performance.

#### *Performance evaluation*

Finally, the predictive performance of the models was evaluated based on the Mean Squared Error (MSE), with the proposed CarbonMLP outperforming both the LSTM and BiLSTM models in this metric. The proposed CarbonMLP outperformed both the LSTM and BiLSTM models in terms of prediction accuracy, as evidenced by its reduced MSE value. This demonstrates the usefulness of the model, despite its relatively simplistic development.

The computational efficiency comparison establishes that the proposed CarbonMLP model is highly efficient in terms of training and inference time, memory utilization, latency, and model size, while still providing competitive predictive performance. The capabilities of CarbonMLP make it the best choice for real-world CO<sub>2</sub> emission prediction tasks, particularly in contexts with limited computational resources.

Metric	Proposed MLP Model	LSTM Model	BiLSTM Model
Training Time (seconds)	1.96	4.33	3.47
Inference Time (seconds)	0.15	0.23	0.20
Memory Usage (MiB)	137.15	141.39	141.39
Latency (seconds)	0.15	0.39	0.22
Batch Processing Time (seconds)	0.17	0.21	0.20
Mean Squared Error (MSE)	0.0002	0.0003	0.0004
Model Size (MB)	3.25	13.90	28.76
Number of Parameters	281,601	1,212,321	2,506,561

**Table 5.** Computational efficiency and performance comparison.

### Performance evaluation of the proposed model

The proposed deep learning model can fully represent the complex relationships between CO<sub>2</sub> emissions and vehicle attributes. After conducting 5-fold cross-validation, excellent performance metrics were obtained. The average R-squared value was 0.9938, demonstrating a 99.4% explained variation, with an extremely low Mean Squared Error (MSE) of 0.0002 and a Root Mean Squared Error (RMSE) of 0.0142. Additionally, the model's Mean Absolute Percentage Error (MAPE) was 2.5%, indicating that the predictions were, on average, within 2.6% of the actual values.

These metrics are summarized in the performance curve shown in Figure 13, where R-squared, MSE, RMSE, and MAPE are depicted, providing a holistic view of the model's predictive accuracy and its ability to generalize across all cross-validation folds.

The results are further supported by the training and validation loss curves shown in Figure 10, which display the loss trends for each of the five folds. The consistency across these curves highlights the ability of the model to generalize across validation sets without overfitting. Figure 11 presents a summary of these loss curves, further demonstrating that both training and validation losses decreased steadily throughout the training process.

Moreover, the accuracy of the model was confirmed by the actual versus predicted plot in Figure 12, where the data points were closely clustered around the diagonal line, reinforcing the high correlation between the predicted and actual CO<sub>2</sub> emissions. While an isolated outlier with a significantly lower actual value compared to its prediction was observed, this may be attributed to data anomalies, underrepresented patterns, or the model's limitations in capturing edge cases. Importantly, the predicted values were restored to their original scale after normalization, ensuring that the performance metrics accurately reflected the true CO<sub>2</sub> emissions. These compelling performance indicators, combined with insightful visualizations, underscore the robustness of the model and its potential for practical applications in CO<sub>2</sub> emission prediction tasks.

To entirely evaluate the "CarbonMLP" model's effectiveness in estimating CO<sub>2</sub> emissions, we used a Taylor diagram, as shown in Figure 14. This useful picture summarizes three critical metrics: correlation coefficient (R), standard deviation (SD), and centered root mean square difference (RMSD). R denotes the linear relationship between the expected and observed emissions, with a value of one indicating complete agreement. The correlation coefficient (R) was calculated as follows:

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (21)$$

where,  $x_i$  and  $y_i$  represent the individually predicted and observed values, respectively.  $\bar{x}$  and  $\bar{y}$  denote the respective means.

The radial distance from the origin should be close to the observed standard deviation (SD) is calculated as follows:

$$SD = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (22)$$

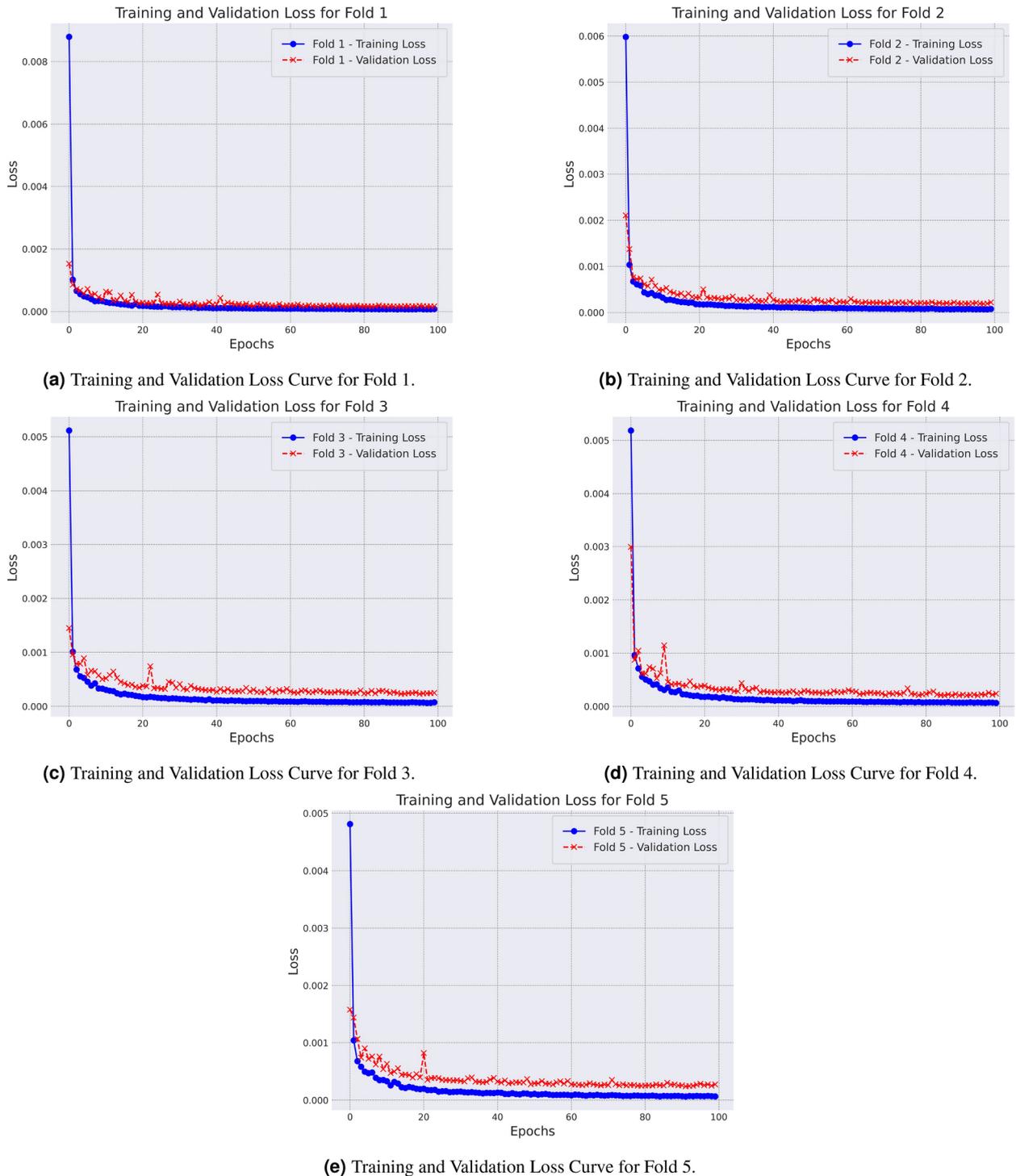
Finally, the smaller angle between the data point and the reference point for the centered root mean square difference (RMSD) reflects a good alignment of cyclical patterns in emissions. The centered RMSD was calculated as follows:

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (23)$$

We can acquire significant information by analyzing the exact location of our "CarbonMLP" model's data point on the Taylor diagram in Figure 14. Ideally, the data point should be close to the circle displaying a correlation coefficient of one, suggesting a high level of agreement between the predictions and observations. Furthermore, the radial distance from the origin should be similar to the reported standard deviation, indicating that the model adequately captured the variability. Finally, a lower angle between the data point and the reference point for the centered RMSD indicates good synchronization of cyclical patterns in emissions. By examining these characteristics of the Taylor diagram, we may accurately assess the "CarbonMLP" model's ability to estimate CO<sub>2</sub> emissions from automobiles.

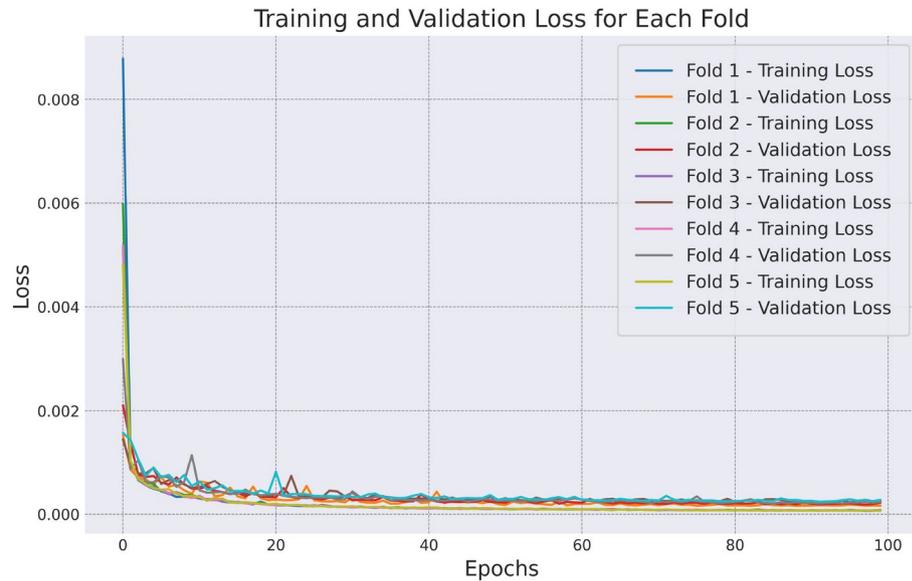
### Comparison of model performance with existing approaches

In predicting CO<sub>2</sub> emissions, our proposed deep learning model outperformed complex deep learning architectures and traditional machine learning algorithms. As the metrics of performance shown in Table 6, the proposed model has the highest R-squared value of 0.9938 and the lowest MSE value of, 0.0002 signifying the high accuracy of the model and the higher correlation of the predicted values with the observed actual CO<sub>2</sub> emissions.

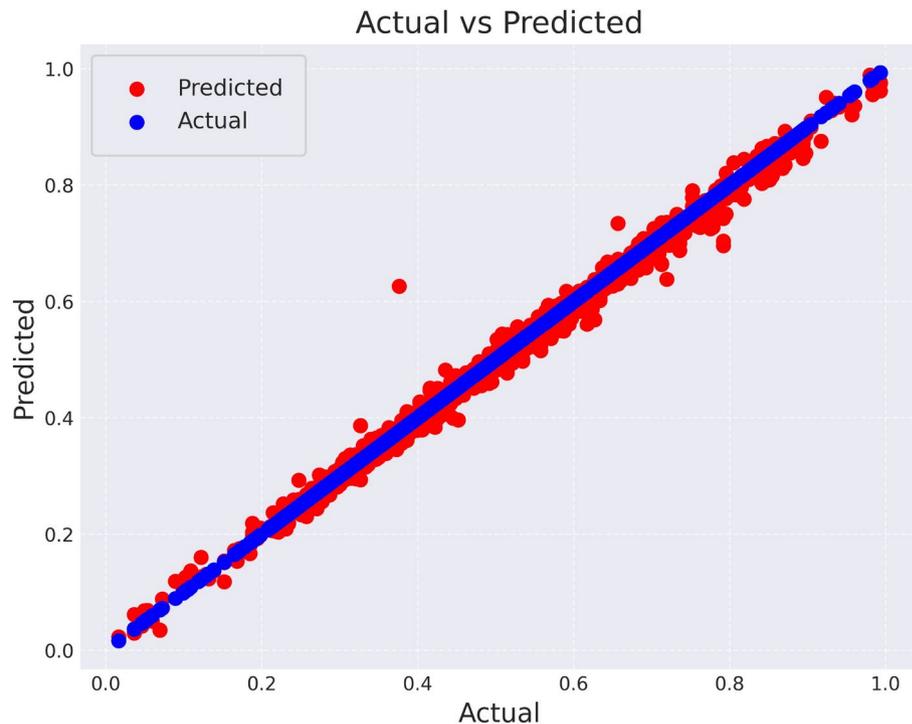


**Fig. 10.** Training and validation loss curves for each of the five folds during 5-fold cross-validation. Each subplot shows the trend of the loss function across epochs for both the training and validation datasets, demonstrating consistent convergence and the model's generalization ability across the different folds. The loss curves show a steady decrease, indicating effective training and no significant overfitting.

In the machine learning module, various algorithms such as Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and XGBoost were applied to compare their performance. These algorithms played different roles in the analysis<sup>59</sup>. Decision Trees are simple, interpretable models that can handle both the linear and non-linear relationships between vehicle attributes and CO<sub>2</sub> emissions. They segment the data into smaller, more manageable chunks, allowing for an intuitive understanding of which features lead to different levels of emissions. However, they may lack the flexibility to capture complex patterns in the data when used alone. K-Nearest Neighbors (KNN) regressor works by measuring the proximity of data points in the feature space.



**Fig. 11.** Average training and validation loss curve of CarbonMLP model.



**Fig. 12.** Actual vs predicted plot of CarbonMLP model.

It is useful in scenarios where the relationship between features is non-linear. However, KNN's performance can degrade when faced with high-dimensional data or large datasets, as seen by its relatively higher MSE and lower R-squared in Table 6. The Support Vector Machine (SVM) regressor aims to find a hyperplane that best fits the data points, minimizing the prediction error. While effective in some cases, SVM struggled to handle the intricacies of the CO<sub>2</sub> emission data in this study, resulting in higher errors compared to the deep learning approaches. XGBoost, a highly efficient gradient boosting algorithm, is well-suited for complex, non-linear problems. It showed relatively strong performance but still fell short of the proposed deep learning model due to its inability to capture deeper patterns in the data without overfitting.

These models were also combined in an ensemble approach, integrating the strengths of Decision Trees, KNN, and XGBoost to improve the generalization and accuracy. By leveraging the diversity of predictions from these models, the ensemble achieved a lower MSE (0.0003) and higher R-squared (0.9889) than individual machine

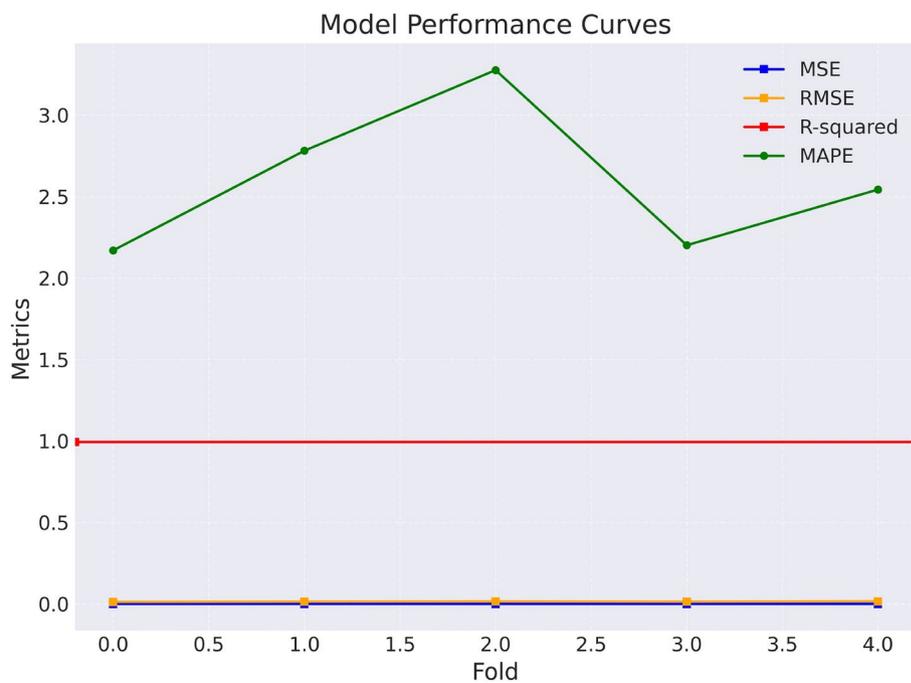


Fig. 13. CarbonMLP model’s performance curve.

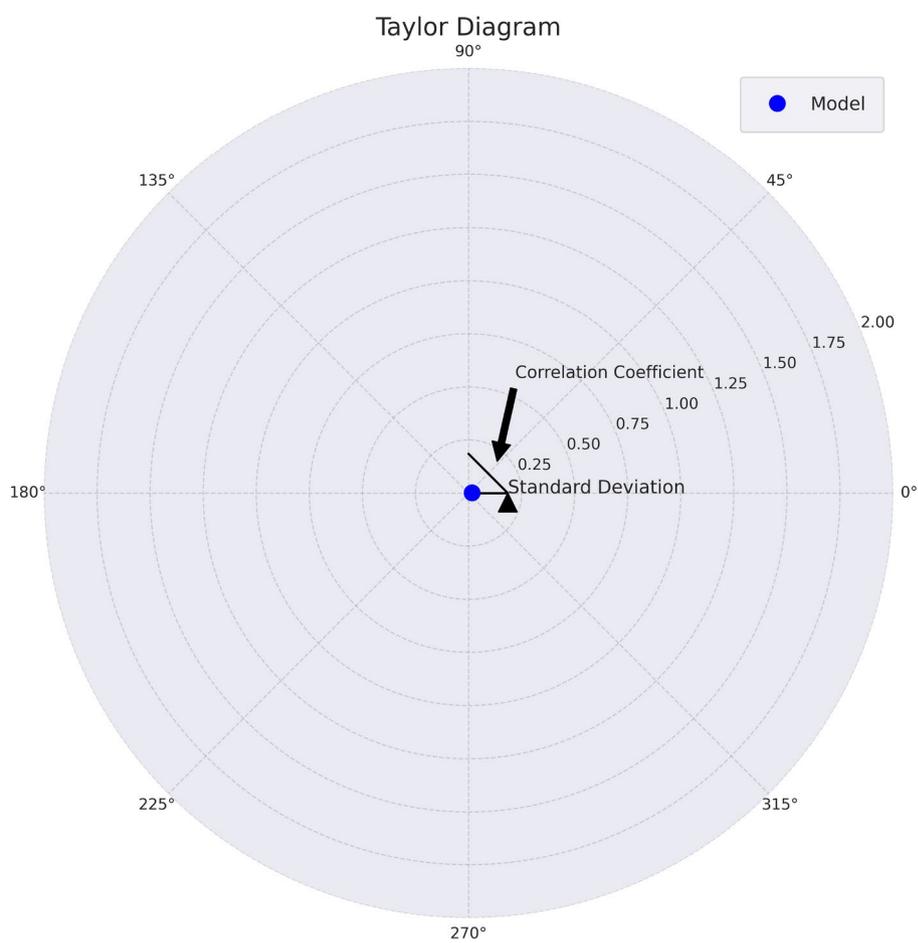


Fig. 14. Taylor diagram for “CarbonMLP” model performance.

Model	MSE	RMSE	R-squared	MAPE
BiLSTM	0.0004	0.0199	0.9878	3.24%
LSTM	0.0003	0.0195	0.9883	3.30%
Decision tree regressor	0.0005	0.0234	0.9827	5.69%
K neighbour regressor	0.0006	0.0256	0.9794	5.18%
SVR	0.0026	0.0514	0.9172	11.46%
XGB regressor	0.0007	0.0280	0.9753	5.07%
Adaboost regressor	0.0015	0.0394	0.9513	8.10%
Catboost Regressor	0.0025	0.0503	0.9207	10.89%
Ensemble (decision tree, knn, xgb)	0.0003	0.0187	0.9889	3.19%
<b>Proposed (CarbonMLP)</b>	<b>0.0002</b>	<b>0.0142</b>	<b>0.9938</b>	<b>2.59%</b>

**Table 6.** Performance comparison among all models.

learning algorithms, although it still did not outperform the proposed deep learning model. The outcomes show that even though such basic techniques, such as Decision Trees or KNN, provide satisfactory predictions, they are not able to embrace the non-linearity of the factors inside the CO<sub>2</sub> emission.

The superiority of these models in sequential data processing was considered when using the LSTM and BiLSTM models in our deep learning module. By their nature, LSTM models cover distances in the sequence and implement memory over long intervals, making them suitable for time-series forecasting. This is important for establishing a relationship between vehicle characteristics and CO<sub>2</sub> emissions over time. BiLSTM improves the above in the sense that word sequences are analyzed in forward and backward ways, capturing contexts in past and future data points. This split view is most helpful in emissions predictions, where the coefficients may be a function of prior and subsequent values. LSTM and BiLSTM are used because they are among the best algorithms that work for sequential data, and sequential data is which are required for prediction tasks. The emission data obtained always have temporal characteristics related to things, such as the specifications of the automobile and changes in legislation. These dynamics can be captured using these models and hence provide better forecasts of CO<sub>2</sub> emissions. It is also important for a predictive model to be built with data pattern changeability so that it can adapt to current data patterns. This is evident in the proposed CarbonMLP model where the deep learning approach efficiently captures these complex interactions and is therefore the best approach towards achieving highly precise CO<sub>2</sub> emission prediction in real environments.

### XAI interpretation using SHAP

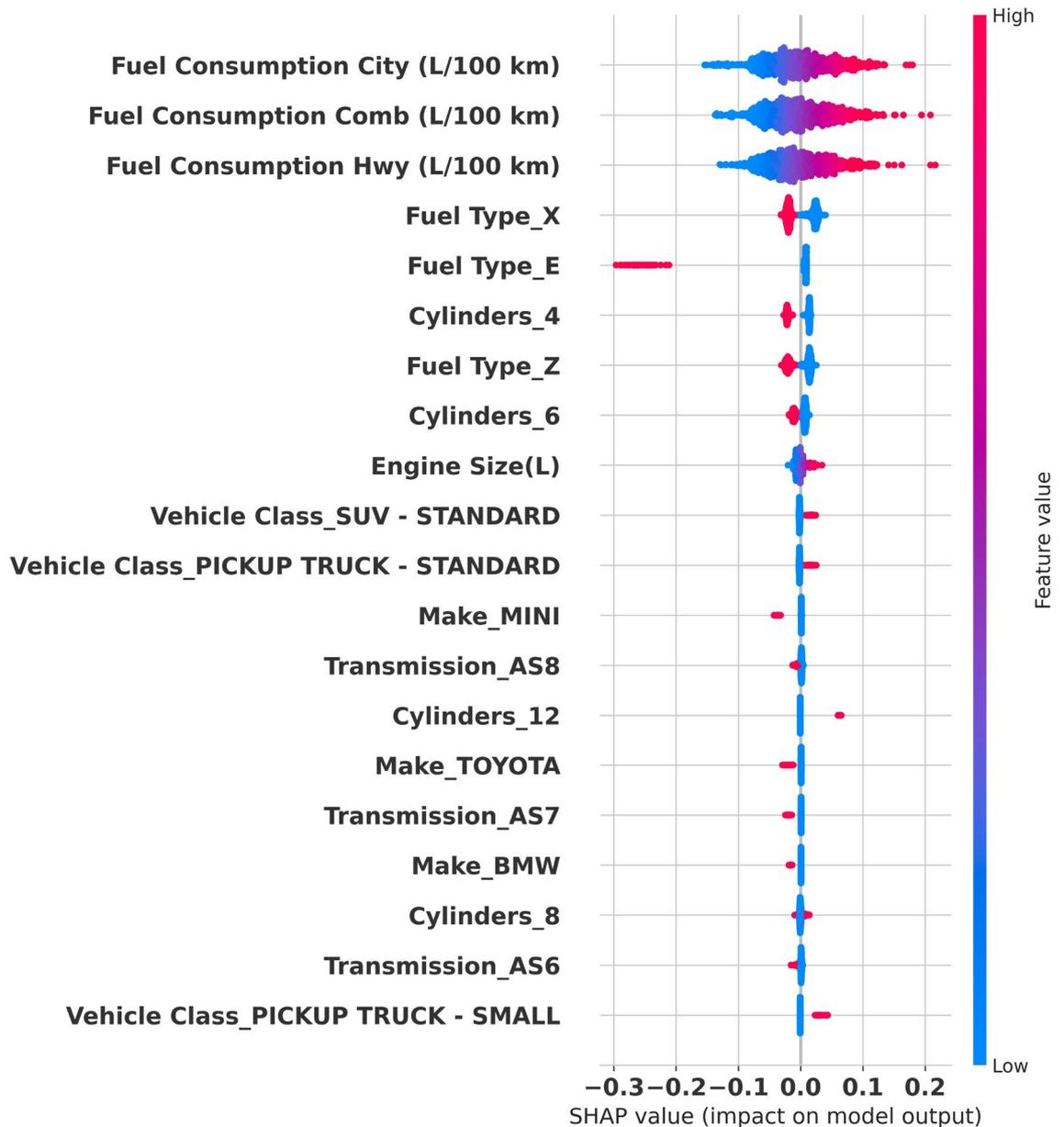
We employed the (SHAP) values to acquire a better understanding of the underlying mechanics of the advanced deep learning model and how different features influence its predictions. This model-agnostic technique allows us to understand how each variable affects the model's CO<sub>2</sub> emission forecasts. With an impressively low Mean Squared Error (MSE) of 0.0002 and a high R-squared value of 0.9938, our proposed model performed well, establishing it as the best option for SHAP explanation. Analyzing the SHAP values associated with each prediction allowed us to determine which features had the greatest positive or negative impact on the model's CO<sub>2</sub> emission predictions.

#### SHAP summary plot: Global feature importance for CO<sub>2</sub> emission prediction

The SHAP summary plot in Figure 15 provides a comprehensive view of how different features affect the CO<sub>2</sub> emissions predictions across the entire dataset. The features were ranked by their importance, and the plot shows the distribution of their impact on the model's output. The key elements of the SHAP summary plot are as follows:

- **Feature Importance (Y-axis):** The features are ordered by importance, with the most impactful features listed at the top. The top three features in this plot are *Fuel Consumption City (L/100 km)*, *Fuel Consumption Comb (L/100 km)*, and *Fuel Consumption Hwy (L/100 km)*, indicating that fuel consumption metrics are critical predictors of CO<sub>2</sub> emissions.
- **SHAP Value (X-axis):** SHAP values represent the contribution of each feature to the prediction. Positive SHAP values increased the predicted CO<sub>2</sub> emissions, whereas negative SHAP values decreased it. A SHAP value of 0 indicates no impact on the prediction for that instance.
- **Color Gradient (Feature Value):** The colors represent feature values, with blue indicating low values and red indicating high values. For example, low fuel consumption values (blue) tend to decrease CO<sub>2</sub> predictions, whereas high fuel consumption values (red) increase the predictions.
- **Distribution of Impact:** The spread of points for each feature indicates how much variation exists in its impact. Features such as *Fuel Type\_X* and *Cylinders\_6* show a wide distribution of SHAP values, indicating that their impact can vary significantly depending on the vehicle configuration.

The summary plot reveals that the fuel consumption metrics, engine size, number of cylinders, fuel type, and vehicle class are among the most important predictors of CO<sub>2</sub> emissions. Features such as *Fuel Consumption City* and *Fuel Consumption Comb* have a consistent positive relationship with CO<sub>2</sub> emissions-vehicles with higher fuel consumption produce more CO<sub>2</sub>. Additionally, the impact of certain features, such as *Fuel Type\_X*



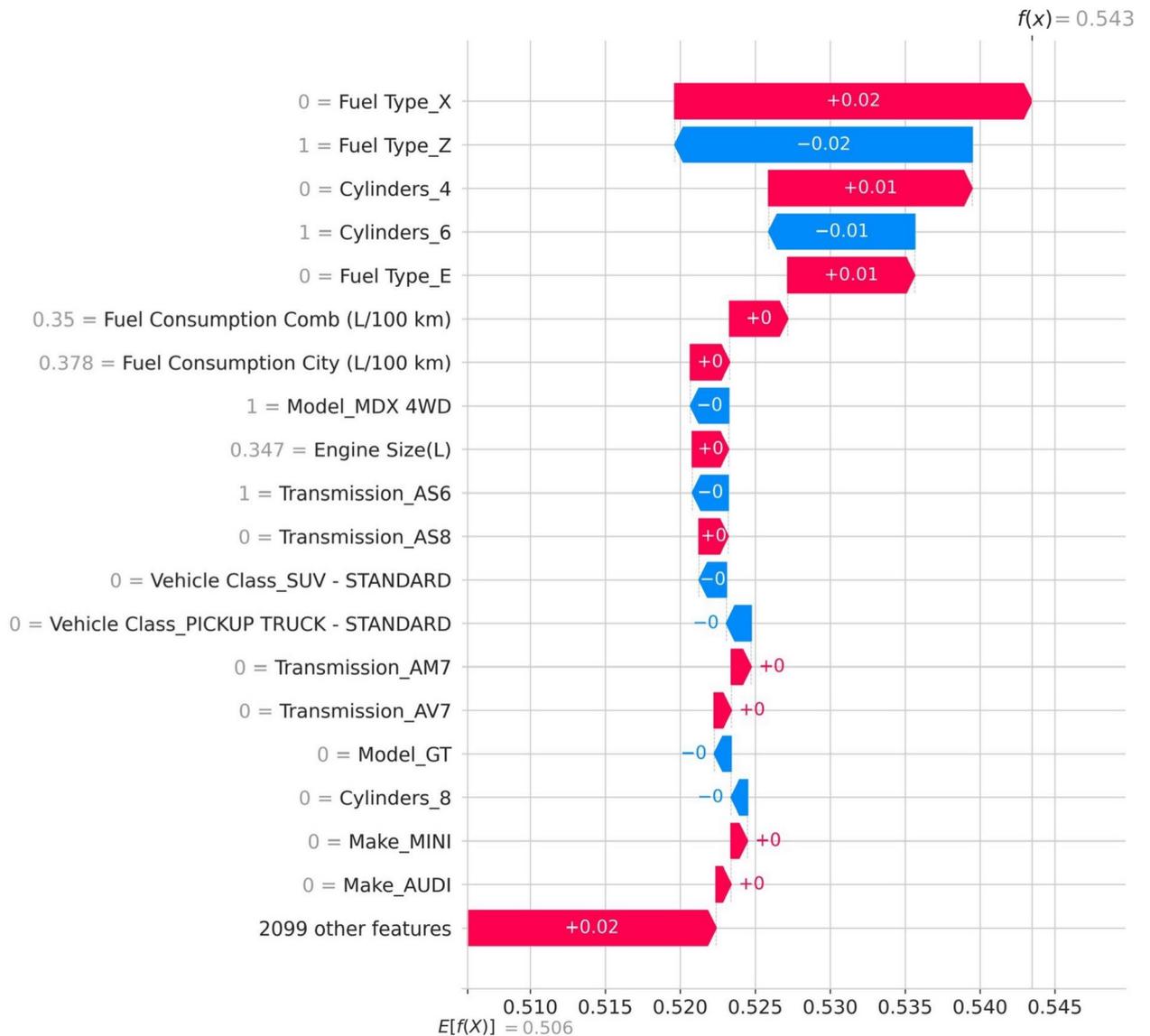
**Fig. 15.** SHAP Summary Plot: Visualization of feature importance, with Fuel Consumption Combined (L/100 km) as the highest ranking feature. Red represents a greater impact on CO<sub>2</sub> emissions, and blue represents a lower impact.

and *Transmission\_AS6*, varies depending on the vehicle configuration, highlighting the importance of feature interactions in the model's predictions.

*SHAP waterfall plot: Instance-Level explanation for CO<sub>2</sub> emission prediction*

To explain the contribution of each feature to the prediction of a specific vehicle (index 0), we used SHAP to generate a waterfall plot, as shown in Figure 16. The SHAP waterfall plot provides an instance-level explanation of how different features contribute to the final CO<sub>2</sub> emission prediction for a specific data point. The key elements in the SHAP waterfall plot are as follows:

- **Baseline Value**  $E[f(x)] = 0.506$ : The expected value of the prediction, which is the mean CO<sub>2</sub> emission prediction across all instances in the dataset. This baseline served as the starting point of the plot.
- **Model Output**  $f(x) = 0.543$ : The final predicted value for this particular instance (vehicle), after accounting for the contributions of all features. This represents the CO<sub>2</sub> emission prediction for the 0th index in the dataset.
- **Feature Contributions**: Each horizontal bar represents the contribution of a feature to the final prediction. Features that increase the predicted value are shown in red (positive contributions), whereas features that de-



**Fig. 16.** SHAP Waterfall Plot: Detailed analysis showing the relative contributions of each characteristic to variations from the base value. Highlights the significant impact of fuel consumption in the city, combined fuel consumption, and highway fuel consumption on the model's CO<sub>2</sub> emissions prediction.

crease the prediction are shown in blue (negative contributions). For example, *Fuel Type\_X* contributes +0.02 to the final prediction, slightly increasing the CO<sub>2</sub> emission prediction. On the other hand, *Cylinder\_6* has a negative contribution (-0.01), reducing the prediction.

- **Cumulative Impact:** The contributions from all features were accumulated, starting from the baseline  $E[f(x)] = 0.506$ , and the final prediction was reached at  $f(x) = 0.543$ . This small contribution of each feature adds to shifting the prediction from the baseline value to the final result.
- **Many Small Contributions:** The plot shows that a large number of smaller features (grouped as “2099 other features”) have a combined contribution of +0.02, which also affects the prediction. These grouped features individually have small impacts but collectively influence the final output.

The SHAP waterfall plot visually demonstrates the effect of each feature on prediction. Features such as *Fuel Type\_X* and *Fuel Consumption Comb* push the prediction upward, while features such as *Fuel Type\_Z* and *Cylinders\_6* have a negative impact. The cumulative effect of all these feature contributions led to a final prediction of 0.543 from the baseline value of 0.506. By visualizing the individual feature contributions, the SHAP waterfall plot provides an interpretable and transparent view of how the model arrives at specific CO<sub>2</sub> emission predictions. This transparency aids in understanding the features that significantly influence the prediction of each vehicle.

*SHAP force plot: Instance-Level explanation for CO<sub>2</sub> emission prediction*

The SHAP Force Plot in Figure 17 provides a detailed explanation of the individual predictions made by the model for the 0th index in the dataset. It visualizes the contributions of each feature, demonstrating whether they push the prediction higher or lower than the baseline.

- **Baseline Value**  $E[f(x)]$ : The baseline value, representing the expected value of the model's output without any feature information, is  $E[f(x)] = 0.50$ . This is the mean prediction of CO<sub>2</sub> emissions across all instances.
- **Model Output**  $f(x) = 0.543$ : The final predicted value for the 0th index is  $f(x) = 0.543$ , which is slightly higher than the baseline, indicating a higher predicted CO<sub>2</sub> emission for this instance.
- **Positive Contributions (Red)**: Features such as *Fuel Consumption City (L/100 km)* and *Fuel Type\_E* significantly increased the prediction, contributing positively to the final output. These are represented by red sections on the left side of the plot.
- **Negative Contributions (Blue)**: Features such as *Fuel Type\_Z* and *Cylinders\_6* decrease the prediction, pulling it lower than it would otherwise. These are represented by the blue sections on the right hand side of the plot.
- **Cumulative Impact**: The plot illustrates the transition from the baseline value to the final predicted value by balancing the positive and negative contributions. The overall effect of these features led to a final predicted value of 0.543, as shown on the right end of the force plot.

This visualization provides an intuitive understanding of the model's decision-making process, highlighting how various features interact to produce the final prediction.

*SHAP dependence plots: Analysis of feature impact on CO<sub>2</sub> emissions*

The SHAP dependence plots in Figure 18 provide a detailed visualization of how specific features affect predicted CO<sub>2</sub> emissions and how they interact with other variables in the model.

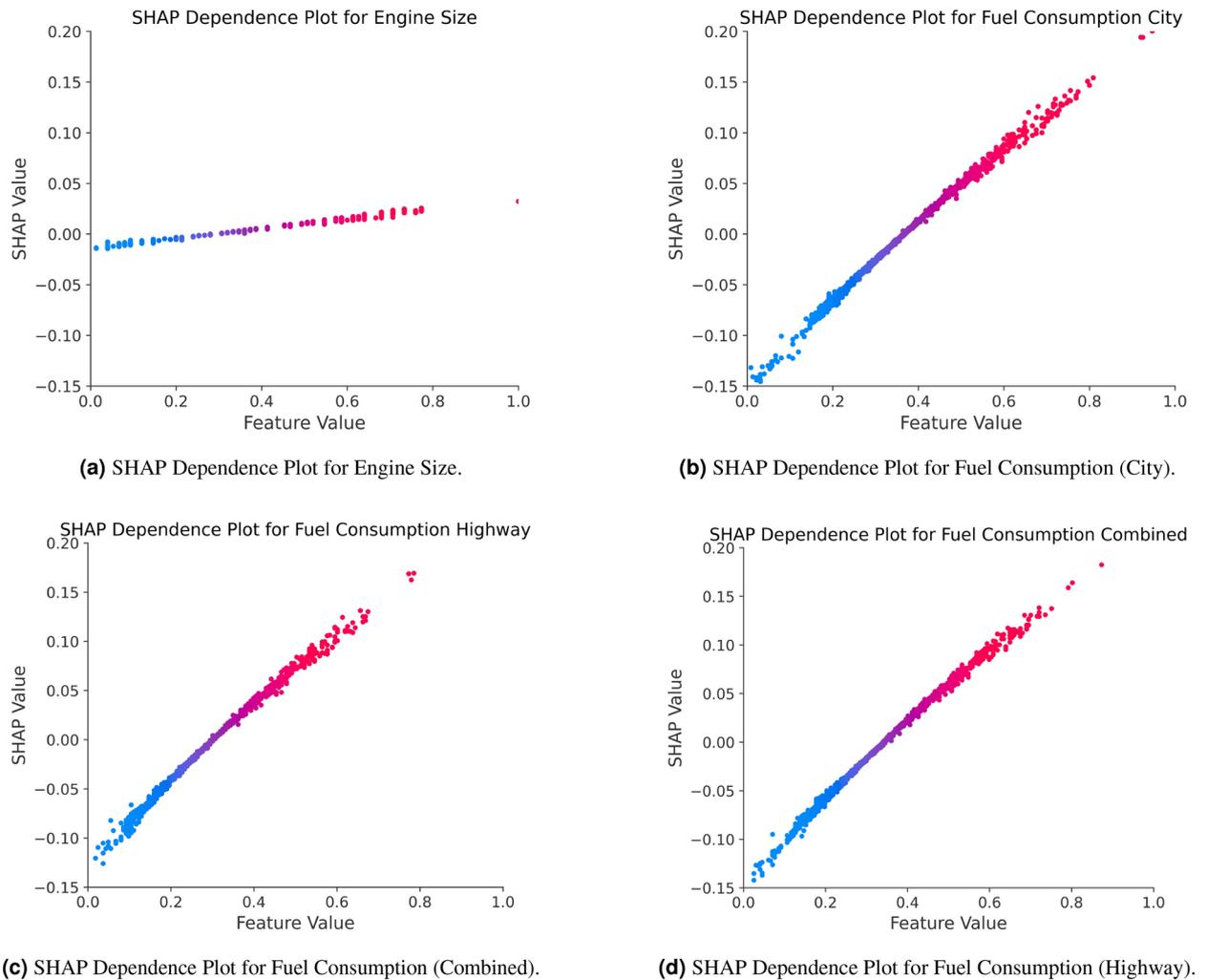
- **Engine Size**: The dependency plot in Figure 18a demonstrates a positive relationship between engine size and CO<sub>2</sub> emissions, suggesting that larger engines contribute to higher emissions. However, the narrower range of SHAP values compared to other features indicates that engine size has a relatively lower impact on CO<sub>2</sub> emissions. This aligns with the nature of the data, where fuel consumption metrics (City, Combined, and Highway) are more directly tied to emissions, making them more influential predictors in the model.
- **Fuel Consumption (City)**: The dependency plot in Figure 18b reveals that fuel consumption under city driving conditions is strongly correlated with CO<sub>2</sub> emissions. Higher fuel consumption in urban environments leads to increased CO<sub>2</sub> predictions, making city fuel efficiency an important target for emission reduction strategies.
- **Fuel Consumption (Combined)**: The combined fuel consumption dependence plot in Figure 18c shows a similar trend, where increased overall fuel consumption results in higher CO<sub>2</sub> emissions. This emphasizes the importance of improving fuel efficiency across different driving conditions to reduce environmental impact.
- **Fuel Consumption (Highway)**: The highway fuel consumption plot in Figure 18d also demonstrates a positive relationship with CO<sub>2</sub> emissions. While highway driving is generally more fuel-efficient than city driving, fuel consumption in this condition still contributes significantly to emissions, highlighting the role of efficiency improvements in reducing emissions.

Overall, the SHAP dependence plots provide a comprehensive and insightful analysis of how engine size and fuel consumption metrics, specifically in city, combined, and highway driving conditions, play significant roles in predicting vehicle CO<sub>2</sub> emissions. The lower variation and smaller magnitude of SHAP values for Engine Size suggest that it has a weaker effect on the target variable when compared to the Fuel Consumption features. While fuel consumption metrics exhibit a wider range of SHAP values, underscoring their dominant influence on the model's predictions, the smaller range observed for engine size indicates its relatively lower importance in comparison. This distinction highlights that fuel efficiency improvements, particularly in urban and combined driving conditions, are critical drivers for reducing CO<sub>2</sub> emissions. Nonetheless, engine size remains a contributing factor that cannot be entirely overlooked.

These findings highlight the potential for substantial reduction in CO<sub>2</sub> emissions through targeted improvements in vehicle design and fuel efficiency. By focusing on developing engines with smaller displacements, optimizing fuel consumption across different driving conditions, and employing advanced technologies to



**Fig. 17.** SHAP Force Plot: In-depth analysis of individual predictions illustrating how each feature affects the model's output. Emphasizes the notable impact of city fuel consumption, combined fuel usage, and highway fuel consumption on CO<sub>2</sub> emissions predictions.



**Fig. 18.** SHAP Dependence Plots for various features: (a) Engine Size, (b) Fuel Consumption (City), (c) Fuel Consumption (Combined), and (d) Fuel Consumption (Highway).

enhance efficiency, manufacturers can make significant strides toward minimizing emissions. This not only aids in complying with environmental regulations but also supports the broader goal of mitigating climate change and achieving a sustainable future. Ultimately, SHAP analysis provides actionable insights that can inform both policymakers and automotive engineers, encouraging the adoption of strategies that prioritize energy efficiency and emission reduction in the transportation sector. This will lead to the development of greener vehicles and cleaner environments, contributing meaningfully to global efforts to combat climate change.

### Comparison with previous work

This analysis expands on the findings of the literature review by comparing the proposed Custom Deep Learning Model with existing models for estimating CO<sub>2</sub> emissions. As shown in Table 7, many studies cited in the literature review used models such as BiLSTM and XGBoost for this task.

Our proposed method outperforms existing methods for predicting CO<sub>2</sub> emissions. It had the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of the models in the table, implying a better fit between the anticipated and actual CO<sub>2</sub> emission data. Furthermore, the proposed model had a high R-squared value, indicating a significant connection between the predictions and actual data. While admitting potential constraints resulting from differences in the datasets and assessment measures used in previous studies, the proposed approach outperformed these important parameters. Furthermore, it has SHAP values that improve its interpretability. Using SHAP values, we obtained insights into how different features influence the model's predictions, improving its overall transparency and dependability. In conclusion, the high performance and interpretability of our custom deep learning model make it an effective tool for predicting CO<sub>2</sub> emissions.

### Discussion

The proposed deep learning model outperformed existing carbon dioxide (CO<sub>2</sub>) emission prediction models with a low Mean Squared Error (MSE) of 0.0002, Root Mean Squared Error (RMSE) of 0.0142, and high R-squared value of 0.9938. These indicators show a better match between the model predictions and actual

Author	Model Used	MSE	RMSE	R-Squared	MAPE Percentage	XAI USE (SHAP)
Al-Nefaie et al. 2023	BiLSTM	0.001177	0.0343	0.9378	-	×
Mądział et al. 2023	Gradient-Boosting	0.7780	0.877	0.6126	-	✓ (SHAP)
Ziółkowski et al. 2021	MLP	-	1.0598	0.9861	5.06% to 10.88%	×
Pandey et al. 2023	GB	-	0.0604	0.8468	-	✓ (SHAP)
Tena-Gago et al. 2023	UWS-LSTM	0.0261	0.1616	0.975	-	×
Çınarlar et al. 2024	XGBoost	0.0011	0.0333	0.9886	5.76%	×
Wen et al. 2021	GBR	-	0.95	0.99	-	×
Wei et al. 2021	CatBoost	-	0.039	0.83	-	×
Ullah et al. 2023	LightGBM	-	16.34	0.9800	-	×
Hien et al. 2022	Univariate Polynomial Regression	-	1.287	0.986	-	×
Li et al. 2023	Ensemble Model	-	2.01	0.88	-	×
<b>Proposed Method</b>	<b>CarbonMLP</b>	<b>0.0002</b>	<b>0.0142</b>	<b>0.9938</b>	<b>2.59%</b>	<b>✓ (SHAP)</b>

**Table 7.** Comparison with previous work.

CO<sub>2</sub> emissions, indicating a significant correlation with the real-world data. While highlighting the possible limitations derived from the differences in previous research approaches, the model's clear outperformance establishes it as a powerful tool for precise CO<sub>2</sub> emission prediction. Our key contribution is the development of a model that not only predicts CO<sub>2</sub> emissions but also provides interpretability by utilizing SHAP values. This transparency helps us to better understand which elements, such as vehicle characteristics, fuel consumption, and engine size, have the greatest impact on CO<sub>2</sub> emissions. By identifying these main factors, the model can inform vehicle-related CO<sub>2</sub> emissions reduction plans. For example, insights gained from the interpretability features of the model can be used to target efforts aimed at improving engine performance, promoting fuel-efficient vehicles, and encouraging environmentally responsible driving practices. This study improves CO<sub>2</sub> emission forecasts and provides useful information to environmental authorities, regulators, and the automobile sector. These stakeholders can use the model's capabilities to create and implement successful plans to reduce CO<sub>2</sub> emissions from automobiles, thus paving the way for a more maintainable environment.

## Conclusion and future work

This study explored the efficacy of a deep learning model for predicting carbon dioxide (CO<sub>2</sub>) emissions from vehicles. The proposed model demonstrated outstanding performance, outperforming the previous techniques in terms of accuracy and interpretability. In summary, this study developed a deep learning model with eXplainable AI (XAI) integration to estimate CO<sub>2</sub> emissions from vehicles using a Multilayer Perceptron (MLP) architecture. The model was trained using a dataset consisting of 7,385 rows and 12 columns, which included vehicle characteristics such as Make, Model, Vehicle Class, Engine Size (L), Cylinders, Transmission, Fuel Type, Fuel Consumption ratings (City, Highway, and Combined), and CO<sub>2</sub> emissions (g/km). The model was extremely successful, as evidenced by a high R-squared value of 0.9938 and low Mean Squared Error (MSE) of 0.0002. Further studies using assessment metrics and visualizations validated the capacity of the model to represent the complex links between vehicle features and CO<sub>2</sub> emissions. Additionally, SHapley Additive exPlanations (SHAP) values were applied to obtain the influence of different features on the model's CO<sub>2</sub> emission predictions of the model. Despite these optimistic findings, this study has some limitations that require further investigation. The proposed model's performance was influenced by the dataset used. If trained using data from a specific region or vehicle type, its applicability to different populations or geographical locations may be limited. Furthermore, this study's emphasis on CO<sub>2</sub> emissions, although important, addresses only one aspect of the environmental challenge. Including other pollutants, such as nitrogen oxides (NOx) and particulate matter (PM), would provide a more complete view of the environmental impact of a vehicle. In addition, the model analysis was constrained to the features found in the training data. A more in-depth look at the external elements that may influence CO<sub>2</sub> emissions, such as driving behavior, weather conditions, and road infrastructure, could provide useful insights for future model modifications.

Future endeavors can build on this study and explore new avenues for creating a more sustainable environment. Expanding the dataset to cover a broader and more diverse range of geographical regions would improve the model's generalizability. Incorporating more contaminants into the analysis of CO<sub>2</sub> emissions would provide a more comprehensive understanding of the environmental impact of vehicles. Exploring sophisticated interpretability techniques may yield further insights into the decision-making process of the model. However, the most significant future advancements will involve real-world applications. Consider vehicles outfitted with sensors that measure CO<sub>2</sub> emissions based on the same parameters that the model identifies as important. These real-time data can be fed back into the model to drive continual improvement and inform drivers of eco-driving advice. Our journey toward a sustainable environment goes beyond the conceptual framework. Future studies should consider incorporating additional variables, such as driving habits, road conditions, and environmental factors. Real-time monitoring systems and vehicle sensor data are essential for improving pollution forecasts. Moreover, investigating sophisticated deep learning architectures and ensemble approaches may enhance the prediction capabilities of the model. By taking these steps, we can capitalize on the power of this science to create a cleaner and more sustainable future for our planet. This study lays the groundwork for more precise

and understandable CO<sub>2</sub> emissions forecasting models for vehicles. This knowledge will enable public officials, automobile manufacturers, and drivers to develop a more sustainable future for transportation by resolving restrictions and exploring new opportunities. The proposed approach, which combines advanced modeling with interpretability, makes a substantial contribution to establishing sustainable transportation systems, protecting the environment, and reducing vehicle emissions.

### Data availability

The dataset used in this study is publicly available on Kaggle: <https://kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles>. It provides comprehensive vehicle information, including the make, model, engine details, transmission type, fuel consumption rates, and CO<sub>2</sub> emissions. The data has been taken and compiled from the Canadian government's official open data website, which can be accessed at: <https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64>. Researchers interested in utilizing this dataset for further analysis can refer to the official source for the latest data.

Received: 23 June 2024; Accepted: 17 January 2025

Published online: 29 January 2025

### References

1. Canada, E. & Change, C. Greenhouse gas emissions (2024). Last accessed 2024/10/14.
2. World Health Organization. Ambient air pollution (2022).
3. Boschee, P. Comments: Nuclear energy gains (some) favor. *Journal of Petroleum Technology*[SPACE]<https://doi.org/10.2118/0723-0010-jpt> (2023).
4. US Environmental Protection Agency (EPA). Greenhouse gas emissions from a typical passenger vehicle (n.d.). Last accessed 2024/04/28.
5. International Council on Clean Transportation. Global transport sector CO<sub>2</sub> emissions up 2.8% in 2021 (2021).
6. The World Counts. Cars impact on the environment (2022).
7. Fleck, A. World car free day. *Statista* (2023). Data Journalist: anna.fleck@statista.com.
8. Hao, H., Cheng, X., Wang, Y., Liu, J. & Wei, Z. A review of progress in life cycle assessment of alternative transportation fuels. *Renewable and Sustainable Energy Reviews* **131**, 110008 (2020).
9. Gnann, E., Schulte, M., Munzer, A. & Breuer, V. Life cycle assessment of hydrogen fuel cell vehicles with a focus on tank-to-wheel emissions. *International Journal of Hydrogen Energy* **45**, 22122–22132 (2020).
10. Zheng, M., Zhao, R., Liu, Z., Li, X. & Li, Z. Downsizing and turbocharging gasoline engines: Will it reduce CO<sub>2</sub> emissions over the entire life cycle?. *Energy Conversion and Management* **248**, 114800 (2021).
11. Zhang, S., Li, J., Ma, X. & Wang, W. A review of the impacts of automatic transmission types on fuel economy and emissions. *Renewable and Sustainable Energy Reviews* **133**, 110223 (2020).
12. Liu, Z., Wang, Y., Tang, L. & Li, Y. Eco-driving behavior recognition based on long short-term memory neural network. *IEEE Transactions on Intelligent Transportation Systems* **21**, 3682–3691 (2020).
13. Zhang, J., Liu, J., Lv, W., Wang, X. & Li, P. Real-time prediction of urban traffic CO<sub>2</sub> emissions using a deep learning approach. *Transportation Research Part D: Transport and Environment* **161**, 102722 (2022).
14. Wang, X., Liu, D., Hao, H., Wang, X. & Liu, J. Impacts of ambient temperature and humidity on near-road NO<sub>x</sub> and CO emissions from gasoline vehicles. *Environmental Science & Pollution Research International* **27**, 30791–30802 (2020).
15. Wang, J., Li, L., Peng, C. & Li, Z. A review of aerodynamic drag reduction technologies for electric vehicles. *Journal of Power Sources* **510**, 120820 (2021).
16. Ones, D. & Patel, R. Real-time prediction of vehicle emissions using machine learning and sensor data fusion. *Transportation Research Part C: Emerging Technologies* **78**, 102–115 (2020).
17. European Environment Agency (EEA). Emission factors for road transport (eft). <https://www.eea.europa.eu/publications/emep-eea-guidebook-2019/emission-factors-database> (2021). Accessed March 10, 2024.
18. Smith, A., Johnson, B. & Williams, C. Machine learning models for predicting vehicle CO<sub>2</sub> emissions. *Environmental Science and Technology* **45**, 567–578 (2021).
19. Dong, M., Guo, W. & Han, X. Artificial intelligence, industrial structure optimization, and CO<sub>2</sub> emissions. *Research Square*[SPACE]<https://doi.org/10.21203/rs.3.rs-2954106/v1> (2023).
20. Zhong, J., Zhong, Y., Han, M., Yang, T. & Zhang, Q. The impact of ai on carbon emissions: evidence from 66 countries. *Applied Economics*[SPACE]<https://doi.org/10.1080/00036846.2023.2203461> (2023).
21. Moraliyage, H. et al. A robust artificial intelligence approach with explainability for measurement and verification of energy efficient infrastructure for net zero carbon emissions. *Sensors* **22**, 9503. <https://doi.org/10.3390/s22239503> (2022).
22. Sharma, N., Singh, U. & Mahapatra, S. S. Prediction of cost and emission from indian coal-fired power plants with CO<sub>2</sub> capture and storage using artificial intelligence techniques. *Frontiers of Engineering Management* **6**, 573–582. <https://doi.org/10.1007/s11708-017-0482-6> (2019).
23. Chen, S. et al. Can artificial intelligence achieve carbon neutrality? evidence from a quasi-natural experiment. *Frontiers in Ecology and Evolution*[SPACE]<https://doi.org/10.3389/fevo.2023.1151017> (2023).
24. Saggari, A. & Nigam, B. Maximising net zero in energy-intensive industries: An overview of ai applications for greenhouse gas reduction. *Journal of Cleaner Production*[SPACE]<https://doi.org/10.3233/jcc230003> (2023).
25. Verdecchia, R., Sallou, J. & Cruz, L. J. A systematic review of green ai. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*[SPACE]<https://doi.org/10.1002/widm.1507> (2023).
26. Chadha, A. S., Shinde, Y., Sharma, N. & De, P. K. Predicting CO<sub>2</sub> emissions by vehicles using machine learning. *Springer*[SPACE][https://doi.org/10.1007/978-981-19-2600-6\\_14](https://doi.org/10.1007/978-981-19-2600-6_14) (2022).
27. Manvitha, M. S., Pujitha, M. V., Prasad, N. H. & Anju, B. Y. A predictive analysis on CO<sub>2</sub> emissions in automobiles using machine learning techniques. *IEEE International Conference on Data Science and Internet of Things*[SPACE]<https://doi.org/10.1109/idiot56793.2023.10053539> (2023).
28. Li, X., Ren, A. & Li, Q. Exploring patterns of transportation-related CO<sub>2</sub> emissions using machine learning methods. *Sustainability*[SPACE]<https://doi.org/10.3390/su14084588> (2022).
29. Kadam, P. & Vijayumar, S. Prediction model: CO<sub>2</sub> emission using machine learning. *IEEE International Conference on Inventive Computation Technologies (ICICT)*[SPACE]<https://doi.org/10.1109/iict.2018.8529498> (2018).
30. Kangralkar, S. & Khanai, R. Machine learning application for automotive emission prediction. *IEEE International Conference on Inventive Computation Technologies*[SPACE]<https://doi.org/10.1109/iict51068.2021.9418152> (2021).
31. Howlader, A. M., Patel, D. & Gammariello, R. Data-driven approach for instantaneous vehicle emission predicting using integrated deep neural network. *Transportation Research Part D: Transport and Environment*[SPACE]<https://doi.org/10.1016/j.trd.2023.103654> (2023).

32. Al-Nefaie, A. H. & Aldhyani, T. H. H. Predicting CO<sub>2</sub> emissions from traffic vehicles for sustainable and smart environment using a deep learning model. *Sustainability*[SPACE]<https://doi.org/10.3390/su15097615> (2023).
33. Hassan, M. A., Salem, H. A., Bailek, N. & Özgür Kişi. Random forest ensemble-based predictions of on-road vehicular emissions and fuel consumption in developing urban areas. *Sustainability*[SPACE]<https://doi.org/10.3390/su15021503> (2023).
34. Khiari, J. & Olaverri-Monreal, C. Uncertainty-aware vehicle energy efficiency prediction using an ensemble of neural networks. *IEEE Transactions on Intelligent Transportation Systems*[SPACE]<https://doi.org/10.1109/ITS.2023.3268032> (2023).
35. Moradi, E. & Miranda-Moreno, L. A mixed ensemble learning and time-series methodology for category-specific vehicular energy and emissions modeling. *Sustainability* **14**, 1900 (2022).
36. Nassef, A. M., Abdelkareem, M. A. & Rezk, H. Application of artificial intelligence to predict CO<sub>2</sub> emissions: Critical step towards sustainable environment. *Sustainability*[SPACE]<https://doi.org/10.3390/su15097648> (2023).
37. Tena-Gago, D., Golkarnarenji, G., Martinez-Alpiste, I., Wang, Q. & Calero, J. M. A. Machine-learning-based carbon dioxide concentration prediction for hybrid vehicles. *Sensors*[SPACE]<https://doi.org/10.3390/s23031350> (2023).
38. Huang, S., Xiao, X. & Guo, H. A novel method for carbon emission forecasting based on ekc hypothesis and nonlinear multivariate grey model: evidence from transportation sector. *Environmental Science and Pollution Research*[SPACE]<https://doi.org/10.1007/s11356-022-20120-5> (2022).
39. Bilotta, S. & Nesi, P. Estimating CO<sub>2</sub> emissions from iot traffic flow sensors and reconstruction. *Sensors*[SPACE]<https://doi.org/10.3390/s22093382> (2022).
40. Li, M. et al. Explaining a machine-learning lane change model with maximum entropy shapley values. *IEEE Transactions on Intelligent Vehicles*[SPACE]<https://doi.org/10.1109/tiv.2023.3266196> (2023).
41. Maździel, M. Liquified petroleum gas-fuelled vehicle CO<sub>2</sub> emission modelling based on portable emission measurement system, on-board diagnostics data, and gradient-boosting machine learning. *Energies* **16**, 2754, <https://doi.org/10.3390/en16062754> (2023).
42. Ziolkowski, J., Oszczypała, M., Małachowski, J. & Szkutnik-Rogoź, J. Use of artificial neural networks to predict fuel consumption on the basis of technical parameters of vehicles. *Energies* **14**, 2639. <https://doi.org/10.3390/en14092639> (2021).
43. Ullah, I. et al. A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability. *Energy & Environment*[SPACE]<https://doi.org/10.1177/0958305x211044998> (2021).
44. Pandey, D. S., Raza, H. & Bhattacharyya, S. Development of explainable ai-based predictive models for bubbling fluidised bed gasification process. *Fuel*[SPACE]<https://doi.org/10.1016/j.fuel.2023.128971> (2023).
45. Hien, N. L. H. & Kor, A.-L. Analysis and prediction model of fuel consumption and carbon dioxide emissions of light-duty vehicles. *Applied Sciences*[SPACE]<https://doi.org/10.3390/app12020803> (2022).
46. Xia, Y. et al. Rapid assessments of light-duty gasoline vehicle emissions using on-road remote sensing and machine learning. *Science of the Total Environment*[SPACE]<https://doi.org/10.1016/j.scitotenv.2021.152771> (2021).
47. Cinarer, G., YESILYURT, M. K., Agbulut, U., Yilbasi, Z. & Kılıç, K. Application of various machine learning algorithms in view of predicting the CO<sub>2</sub> emissions in the transportation sector. *Sustainable Transportation and Energy Transition*[SPACE]<https://doi.org/10.2516/stet/2024014> (2024).
48. Wen, H.-T., Lu, J.-H., Lu, J. & Jhang, D.-S. Features importance analysis of diesel vehicles' nox and CO<sub>2</sub> emission predictions in real road driving based on gradient boosting regression model. *International Journal of Environmental Research and Public Health*[SPACE]<https://doi.org/10.3390/ijerph182413044> (2021).
49. Wei, N. et al. Applying machine learning to construct braking emission model for real-world road driving. *SSRN Electronic Journal*[SPACE]<https://doi.org/10.2139/ssrn.4079725> (2022).
50. Mohammad, F., Saleem, K. & Al-Muhtadi, J. Ensemble-learning-based decision support system for energy-theft detection in smart-grid environment. *Energies*[SPACE]<https://doi.org/10.3390/en16041907> (2023).
51. Grace Imson, M. How to convert mpg to liters per 100km. <https://www.wikihow.com/Convert-MPG-to-Liters-per-100km> (2022). Last accessed: September 29, 2024.
52. Bhandari, P. How to find the mean | definition, examples & calculator (2023). Published on October 9, 2020. Revised on June 21, 2023. Last accessed: September 29, 2024.
53. Holdsworth, J. & Scapicchio, M. What is deep learning? <https://www.ibm.com/topics/deep-learning#:~:text=Deep learning is a subset,applications in our lives today> (2024). Updated: 17 June 2024.
54. Soori, M., Arezoo, B. & Dastres, R. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics* **3**, 54–70. <https://doi.org/10.1016/j.cogr.2023.04.001> (2023).
55. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. <https://doi.org/10.1038/nature14539> (2015).
56. Jin, Y. et al. Carbon emission prediction models: A review. *Science of The Total Environment* **927**, 172319. <https://doi.org/10.1016/j.scitotenv.2024.172319> (2024).
57. Zuo, Z., Guo, H. & Cheng, J. An lstm-stripat model analysis of china's 2030 co2 emissions peak. *Carbon Management* **11**, 577–592 (2020).
58. Liu, B., Wang, S., Liang, X. & Han, Z. Carbon emission reduction prediction of new energy vehicles in china based on gra-bilstm model. *Atmospheric Pollution Research* **14**, 101865 (2023).
59. Mahesh, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*[Internet] **9**, 381–386 (2020).

## Acknowledgements

The authors would like also to acknowledge the Ubiquitous, Cloud and Human-Computer Interaction (UCH) Research Group, Department of Computer Science, American International University-Bangladesh (AIUB) for supporting this research.

## Author contributions

G.M.I.A.: Conceptualization, Methodology, Data curation, Writing - Original Draft Preparation, Software, Visualization, and Investigation. S.A.T.: Conceptualization, Methodology, Data curation, Visualization, Writing - Original Draft Preparation. S.K.S.: Writing - Original Draft Preparation, Software. Y.W.: Conceptualization, Supervision, Reviewing. M.R.I.: Conceptualization, Supervision, Reviewing, and Editing. M.F.M.: Supervision, Reviewing, and Editing. K.N.: Supervision, Reviewing, and Editing.

## Declarations

## Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to R.I. or K.N.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

## Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

[onlineservice@springernature.com](mailto:onlineservice@springernature.com)