**INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND COMPUTER SYSTEMS**
E-ISSN: 2180-0650
VOLUME 12, ISSUE 1, 2026, pp. 16 – 26
DOI: https://doi.org/10.15282/ijsecs.12.1.2026.2.0148

**RESEARCH ARTICLE**

# Prediction of carbon dioxide emissions from motor vehicles using gated recurrent units and density-based spatial clustering of applications with noise

**Mukombo Mukombo*, Muwanei Sinyinda**

School of Engineering and Technology, Mulungushi University, Kabwe 80415, Zambia

**Abstract -** Extreme climatic conditions resulting from climate change pose significant challenges to humans, prompting urgent action across various fronts. As vehicle emissions contribute substantially to carbon dioxide emissions, there is a pressing need for efficient prediction methods to develop informed mitigation strategies. While some research has been conducted on predicting carbon dioxide using machine learning, few studies have explored Recurrent Neural Networks, particularly those that can be modelled with a smaller dataset to improve emissions prediction. This study aimed to improve the prediction of carbon dioxide emissions by employing a new approach. The study was conducted using a Canadian dataset with 7,385 samples featuring motor vehicle parameters. Therefore, this study employs a deep learning model known as Gated Recurrent Units (GRU) for predictive modelling, coupled with another algorithm called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for robust outlier detection and subsequent removal. Other algorithms used for preprocessing included label encoding and min-max normalisation. The study was evaluated using four metrics: root MSE (RMSE), mean square error (MSE), mean absolute error (MAE), and determination coefficient ($R^2$). Compared with the highest-performing results from prior studies, our approach achieved a 4.59% increase in $R^2$ and a 2.14% increase in R, along with reductions in MSE and RMSE of 0.0008766 and 0.0015804, respectively. Thus, it shows superior handling of data irregularities and temporal patterns. These results highlight the potential for improved emissions predictions, which can be instrumental in refining the carbon tax calculations and informing environmental policy. Future work will involve expanding the dataset to include the age of the motor vehicle and using hybrid deep learning models.

**Article History**
Received : 1 December 2024
Revised : 22 September 2025
Accepted : 20 April 2026
Published : 4 May 2026

**Keywords**
*Climate change*
*$CO_2$ emissions*
*Deep learning*
*DBSCAN*
*Gated recurrent units*

## 1. Introduction

The world is grappling with growing concern about climate change, with temperatures rising steadily and triggering environmental, economic, and social consequences [1]. The primary source of this warming trend and subsequent climate alterations is the increase in greenhouse gas (GHG) emissions, particularly carbon dioxide ($CO_2$), methane (CH4), and nitrous oxide ($N_2O$) from various human activities, including industrial processes, deforestation, and energy production, which have been increasing in proportion to rising temperatures[2-4]. These emissions come from many sources, including the transport sector [5]. The transportation sector is among the leading contributors to GHG emissions, accounting for a substantial share of global $CO_2$ emissions [6]. The transport sector, especially motor vehicles, plays a critical role in increasing atmospheric $CO_2$ levels because of its heavy dependence on fossil fuel derivatives. As vehicles combust diesel, gasoline, and other petroleum products, they release $CO_2$ and other pollutants. Thus, they contribute not only to global warming and, subsequently, climate change but also to poor air quality and other health concerns [1, 7-9]. These emissions do not remain confined to the region where they are produced but disperse into the atmosphere, affecting communities, ecosystems, and economies worldwide. Considering these growing challenges, reducing $CO_2$ emissions from motor vehicles has become a pivotal objective of environmental policy and technological innovation. Numerous strategies have been proposed, ranging from promoting electric vehicles and alternative fuels to developing more efficient engines. However, one critical area that demands attention is accurately predicting $CO_2$ emissions from motor vehicles. Accurate prediction models are essential for creating informed policies, improving vehicle technologies, and implementing carbon tax systems that reflect the actual environmental impact [5, 10]. Given the importance of reliable emission data, machine learning has attracted significant interest in developing predictive models of $CO_2$ emissions. These models, which rely on large datasets of vehicle parameters, can identify key factors contributing to emissions and offer insights into how to reduce or mitigate them. Several studies have applied various models to predict $CO_2$ emissions with promising results. For example, Al-Nefaie and Aldyani [9] used a Bidirectional Long Short-Term Memory (BiLSTM) model to predict $CO_2$ emissions from motor vehicles, achieving reasonable accuracy with an RMSE of 0.0356 and an $R^2$ score of 93.55. However, while these results are encouraging, they have limitations. Most models are designed for larger datasets and do not perform well on much smaller datasets. Additionally, many studies rely on traditional preprocessing methods that may overlook important patterns in the data, such as outliers [9-12].

To address these gaps, this study employs two key methodologies: a deep learning model known as GRU and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for preprocessing. The GRU belongs to the recurrent neural network (RNN) family and is known for its efficiency in handling time-series data and its ability to perform well with smaller datasets. Compared to other models, such as BiLSTM, GRU simplifies computation by using fewer gating mechanisms, making it particularly suitable for smaller datasets while still effectively capturing temporal patterns. In contrast, DBSCAN is a powerful clustering algorithm that excels at identifying outliers and handling noise. It offers

***CORRESPONDING AUTHOR | M. Mukombo** | ✉ mmukombo06@gmail.com
© 2026 The Author(s). Published by Universiti Malaysia Pahang Al-Sultan Abdullah Press. This is an open access article under the CC BY-NC 4.0 license

significant advantages over traditional clustering methods, such as K-means, as it does not require a predefined number of clusters and is robust to varying densities. Combining DBSCAN and GRU allows us to build a robust predictive model that handles datasets effectively and manages outliers to improve overall accuracy. By improving input data quality and refining the learning process, we expect this model to outperform existing approaches and provide more reliable predictions of motor vehicle $CO_2$ emissions. This study makes the following major contributions.

1. Using DBSCAN for preprocessing ensured that outliers were effectively managed, improving the overall quality of the input data. This step goes beyond traditional methods, such as simple normalisation or basic clustering, by specifically addressing the challenge of noisy data in emission prediction.
2. The GRU model demonstrated superior performance in capturing complex temporal patterns within smaller datasets, outperforming previous models such as BiLSTM. This study establishes GRU as an effective tool for emission prediction, particularly when available data are limited.
3. This study integrates artificial intelligence (AI) with environmental science and policy development. Bridging these disciplines not only improves technological approaches to emission prediction but also provides practical applications for policymaking, such as in carbon tax frameworks.

The remainder of this paper is organised as follows: Section 2 presents the related works. Section 3 describes the methodology, including the dataset, preprocessing steps, and model development. Section 4 discusses the results, followed by a comparative analysis of the previous studies. Finally, Section 5 concludes the paper and outlines future work.

## 2. Related Works

Some researchers have used AI to improve $CO_2$ emissions prediction. Machine Learning (ML), deep learning (DL), and ensemble learning models were utilised on a large dataset from the European Union. The researchers chose AI because they observed that government-provided $CO_2$ figures did not properly account for motor vehicles on the roads and that the statistical methods used were less accurate. The study's findings could predict temperature increases and help shape key policies, such as adopting electric vehicles. As further work, the authors proposed validating their DL and ML ensemble techniques on other research problems [13]. Another study aimed at assessing the environmental impact of intelligent systems in transportation by developing a vehicle emissions model with a high degree of accuracy. Thus, the authors developed and evaluated a deep learning-based vehicle emissions model (DL-VEM) to estimate the instant $CO_2$ emissions of taxicabs, focusing on establishing a correlation between driving conditions and emissions. One key finding was a 24.94% increase in emissions when petrol was used instead of compressed natural gas [14]. In this study, the authors reviewed strategies for managing transportation emissions and addressing their impact. They investigated the impact of taxi trips on $CO_2$ emissions using two datasets: taxi trip records and a fuel-economy dataset. The study identified five clusters of motor vehicles associated with emissions, thus aiding in the categorisation of vehicles by emission levels. Motor vehicles that are categorised as high emitters can be targeted for further scrutiny and possible sanctions [15].

A methodology for estimating emissions using onboard diagnostic data in real driving conditions, utilising machine learning and artificial neural networks, was proposed. This research began because researchers found no models to estimate emissions without large measurement campaigns. The main contribution of this study is its potential to aid vehicular homologation and emissions inventories. They also proposed adjusting the model for parameters such as vehicle age, driving style, and weather conditions in future work [16]. There is a need to gather knowledge on $CO_2$ emissions from the transport sector in several countries so countries can realign their future energy strategies and policy agendas by exploring transportation-related emissions patterns. We used data from several countries and applied ordinary least squares regression (OLS), SVM, and GBR for $CO_2$ forecasting. One finding indicates a close relationship between $CO_2$ emissions and several other factors. Identifying these factors can help decision-makers develop strategies to curb near-term growth in transport-related $CO_2$ emissions [17]. Ensemble learning models were used to predict $CO_2$ emissions from light-duty vehicles by leveraging vehicle specifications. The dataset was sourced from the Government of Canada and included features such as cylinders, engine size, and fuel economy. CatBoost outperformed the other algorithms used in this study, demonstrating its robustness across various traffic conditions. This study contributes to the body of knowledge by estimating emissions and exploring their possible application in motor vehicle homologation and vehicular emissions inventories. The researchers proposed improving model performance by expanding the dataset features to increase accuracy [18].

An investigation of the rise in $CO_2$ emissions in Canada focused on transport-related emissions. Machine learning and deep learning models were used to investigate increases in $CO_2$ emissions due to factors such as transportation, economic development, and population growth, with a particular focus on transport-related emissions. This study recommends using alternative motor vehicle fuels and fuels with low carbon content [19]. A random forest model was used to develop a $CO_2$ emissions prediction application using a dataset consisting of fuel features. The application has the potential to provide insights into various aspects, including identifying car manufacturers with the lowest-emission vehicles. This could further stimulate competition among manufacturers and force them to produce environmentally friendly vehicles [20]. This study is one of the first endeavours to develop a methodology for microscale $CO_2$ emission models of liquefied petroleum gas (LPG). This study relied on data acquired from road tests using portable emission measurement systems and onboard diagnostic interfaces. This study is important because it can be used to analyse continuous $CO_2$ emissions and create emission maps for environmental analysis in built-up areas [21]. Table 1 presents an overview of previous studies related to this research.

Table 1. Summarised related works

| Reference | Model(s) Used | Dataset | Key findings | Gaps |
|---|---|---|---|---|
| [13] | Neural Network and Random Forest ensemble | EU dataset | Enhanced accuracy of $CO_2$ emissions predictions, shaped policy recommendations like the adoption of electric vehicles | Their proposed DL and ML ensemble techniques need to be validated on other research problems. |
| [14] | DL-VEM (LSTM) | Taxi cab driving data | Found a 24.94% increase in emissions with petrol compared to compressed natural gas | Limited to specific driving conditions |
| [15] | Optimised hierarchical clustering algorithm | Taxi trip records, fuel economy data | Identified high emitters and aided in the categorisation and sanctioning of emitters. | Not explicitly stated |
| [17] | OLS, SVM and GBR | Transportation data across various countries | Identified influential factors in transportation related to $CO_2$ emissions | Not explicitly stated |
| [19] | ANN, SVM, Decision Tree and GBR | Canadian dataset | Recommended adoption of low-carbon fuel types | Not explicitly stated |
| [18] | Catboost, Histboost, SVR and Ridge | Canadian government dataset | Catboost outperformed other algorithms in $CO_2$ predictions | Limited application to diverse traffic conditions |
| [20] | Random forest | fuel features dataset | Developed an app to predict $CO_2$ emissions and identify low-emitting vehicles | Application limited by available fuel features |
| [21] | Gradient Boosting | On-board diagnostic and road tests | Developed a microscale $CO_2$ emission model for LPG vehicles | focused only on LPG vehicles |
| [8] | LSTM, BiLSTM and Rough K-means | Canadian dataset with motor vehicle parameters | Developed an original strategy for $CO_2$ emissions forecasting | issues with handling a smaller dataset and prudent outlier detection |
| [22] | Random forest | Motor vehicle features | Identified vehicles with high $CO_2$ emissions target scrutiny | Limited number of features in the dataset |

AI techniques have been applied to predict $CO_2$ emissions from motor vehicles. Equipped with a Canadian motor vehicle dataset from kaggle.com, LSTM and BiLSTM deep learning models coupled with a rough K-means clustering algorithm were used in this study. Consequently, the study developed an original strategy for accurate $CO_2$ emissions forecasting, providing policymakers with a tool to create policies [8]. An experiment was conducted to predict motor vehicle $CO_2$ emission ratings based on various factors and to identify vehicles with $CO_2$ emissions beyond a certain range. The random forest performed best. This study has the potential to identify and design motor vehicles with low $CO_2$ emissions [22]. The reviewed studies demonstrate various applications of machine learning and deep learning models for $CO_2$ emissions prediction, each with strengths depending on the dataset characteristics. Commonly used datasets, such as transportation data and vehicle features, are often tied to specific regions or fuel types, limiting the model's applicability on a broader scale. Although some models have achieved commendable accuracy, many of these approaches lack versatility across different datasets and environmental contexts. A recurring challenge that we observed was handling outliers, especially within smaller datasets, where irregularities can skew predictions. This highlights the need for robust methods that can adapt to temporal patterns and variations in emissions data and effectively identify anomalies without relying on predetermined assumptions about data distribution. By focusing on models that balance computational efficiency with the ability to recognise complex dependencies in time-series data, we address gaps in previous studies. The GRU-DBSCAN approach not only enhances prediction accuracy but also goes further to ensure the model's adaptability across diverse real-world scenarios, setting a strong foundation for reliable, generalizable $CO_2$ emissions forecasting across different conditions

## 3. Materials and Methods

### 3.1 Dataset

In this study, we used a publicly available dataset from kaggle.com to train the model. The dataset was sourced from the Canadian Government. The dataset spans seven years and contains 7385 samples and 12 features. The dataset's 12 features are Make, Model, Vehicle Class, Engine Size (*L*), Cylinders, Transmission, Fuel Type, Fuel Consumption City (*L*/100 km), Fuel Consumption Highway (*L*/100 km), Fuel Consumption Combined (*L*/100 km), Fuel Consumption Combined (mpg), and $CO_2$ Emissions (g/km). The 'Make' feature includes about 42 motor vehicle types, ranging from Acura to Volvo. The data is textual. The vehicle 'Model' feature uses different names by manufacturer but includes key aspects such as four-wheel drive (4WD/4×4), flexible-fuel vehicles, all-wheel drive, long wheelbase, short wheelbase (SWB), and extended wheelbase. The 'Vehicle Class' mostly deals with the shape of the vehicle and varies from the compact class

to passenger vans. 'Engine size (*L*)' has the displacement measurements of the engine, with the smallest engine being 0.9 L and the largest being 8.4 L. The number of cylinders in the engine falls under the feature named 'cylinders'. The fewest cylinders were three, while the largest engine had 16 cylinders. Another feature is '' Transmission, which has several options: automatic (A), automatic with shift (AS), manual (M), and continuously variable. The transmission options have varying numbers of gears, with our dataset ranging from 3 to 10 gears. The dataset includes five 'Fuel Types' as a feature: diesel (D), premium gasoline (Z), ethanol-E85 (E), natural gas (N), and regular gasoline (X). Other features include 'Fuel Consumption City (L/100 km)', 'Fuel Consumption Hwy (L/100 km)', 'Fuel Consumption Comb (L/100 km), and 'Fuel Consumption Combination (mpg)'. Of the four fuel consumption features, the combination in miles per gallon is an integer, whereas the others consist of decimal numbers/floats. The last feature, '$CO_2$ Emissions (g/km)', is an integer and represents the amount of $CO_2$ released into the atmosphere from the engine exhaust.

### 3.2 Preprocessing

Data preprocessing is an important step in machine learning because it ensures the data are in a form suitable for analysis and modelling. Good data preprocessing improves the quality and reliability of results [23]. Under preprocessing, we utilised label encoding and min-max normalisation. The following paragraphs briefly explain each preprocessing technique used. We used label encoding, or ordinal encoding, to convert categorical data into numerical data. Typically, a random number is assigned to each unique category in the feature of concern without distorting the data set. In other words, it converts a label into an integer [24]. The dataset used had several categorical data under the labels of 'Make', " 'Model'," ' Transmission', "' Vehicle Class ', and 'Fuel type '; hence, the decision to use label encoding. The technique is easy to interpret and use and works well with a smaller number of unique categorical values, as in our features [25]. Min-max normalisation is a commonly used preprocessing method. It transforms numerical variables to a common scale while preserving differences in value ranges and preventing information loss. The scaling was in the range of zero (0) to one (1). Min-max normalisation was used because it optimises machine learning processes, such as gradient descent, enabling faster convergence. It also improves the algorithm's performance and speed. Min-max normalisation also scales features without losing information, which is a key advantage [26].

### 3.3 Outlier Detection using DBSCAN

Outliers can significantly affect the accuracy of any machine learning model, making it essential to remove them from the dataset before beginning the modelling process. In this study, we selected DBSCAN as our preferred outlier-detection algorithm. DBSCAN is a widely used clustering algorithm, particularly suited for outlier detection because of its density-based approach. The algorithm begins by selecting a random point in the dataset and identifying all other points within a specified radius, known as epsilon ($\varepsilon$), around it. Once the number of points within this radius meets or exceeds a predetermined threshold (min_samples), a cluster is formed. DBSCAN then continues to include neighbouring points, expanding each cluster until no more points meet the criteria for addition. Points that do not belong to any cluster are labelled as outliers or noise and are assigned to cluster -1 [27-28]. DBSCAN was an ideal choice for this study because of several key advantages. DBSCAN relaxes the requirement to specify the number of clusters, unlike K-means, making it adaptable to datasets with unknown or complex structures. Moreover, DBSCAN effectively detects outliers in datasets with irregularly shaped clusters. Its computational efficiency and scalability are additional benefits, and the algorithm is highly customisable, offering parameters that can be adjusted to optimise performance for different datasets [8, 16].

### 3.4 Statistical Insights

Before model training and testing, it is important to emphasise the value of providing statistical insights into the dataset. Prioritisation through statistical analysis not only reveals data characteristics, interdependencies, and distribution, but also ensures that subsequent modelling efforts are based on a solid understanding of the dataset, leading to more robust and accurate predictive models. Previous researchers who worked on the same dataset, such as Natarajan et al. [18] and Al-Nefaie and Aldhyani [9], have already conducted comprehensive statistical analysis. Rather than duplicating their efforts, we examined the dataset before and after preprocessing, including label encoding, min-max normalisation, and outlier detection using DBSCAN. We focused on examining the number of instances, identifying duplicates, and handling missing values. **Error! Reference source not found.** presents key insights. A correlation plot visually represents the strength and direction of associations between features. It shows the correlations among the features selected for modelling. **Error! Reference source not found.**shows a correlation plot of the features under consideration. After careful consideration, we included engine size, cylinders, the four fuel consumptions, and $CO_2$ emissions as our target features in the model.

Table 2. Dataset insight

| Details | Value |
|---|---|
| Number of instances before preprocessing | 7385 |
| The number of duplicates and outliers removed | 1114 |
| Number of empty values | 0 |
| Number of instances after preprocessing | 6271 |

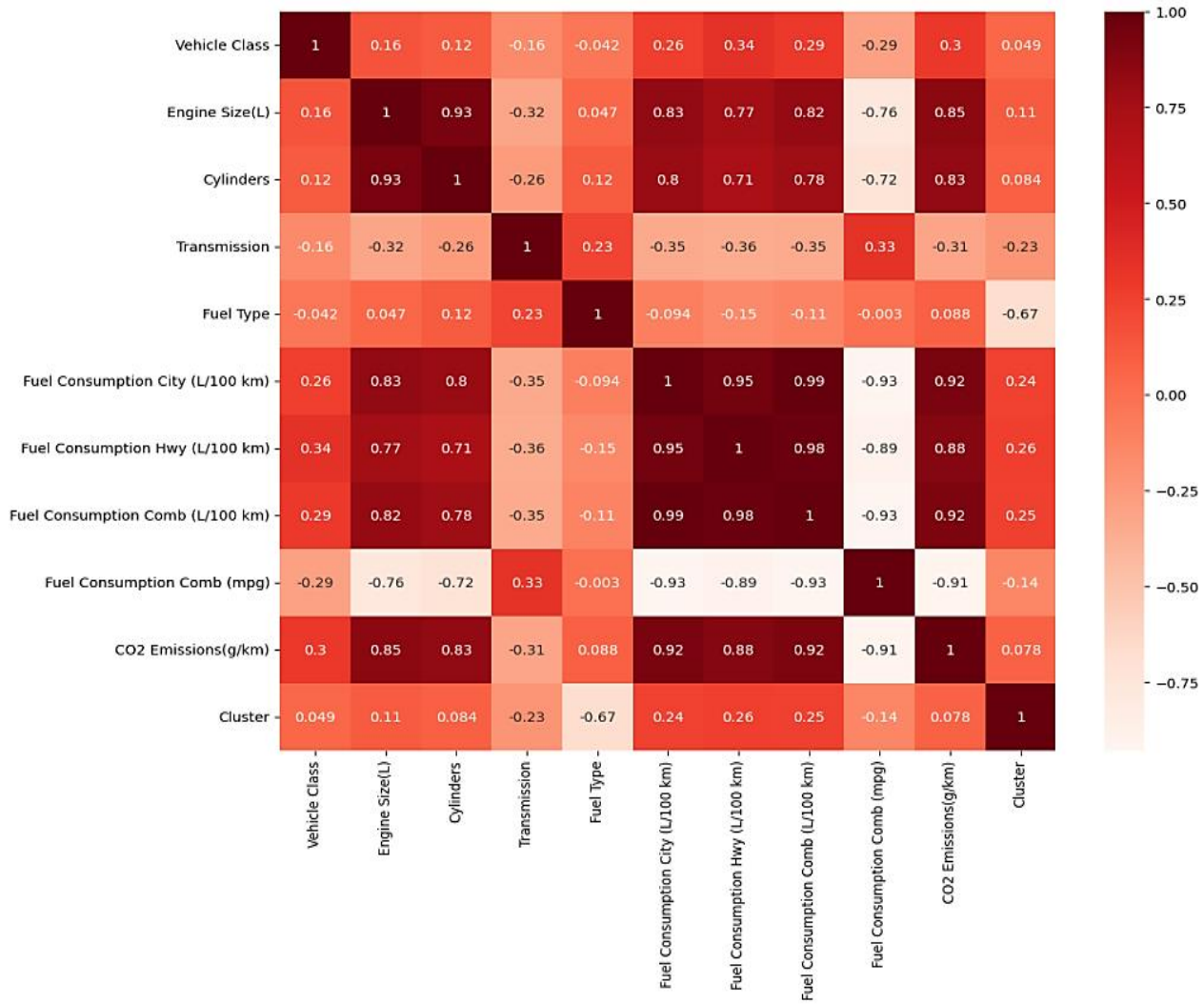



Figure 1. Correlation plot of the features in the dataset

### 3.5 Prediction Model

The GRU is a variant of the RNN designed to overcome challenges in learning and retaining long-term dependencies in sequential data. The GRU addresses vanishing gradients during training [29], a common limitation of traditional RNNs [30]. In our study, we employed GRU because it performs better with smaller datasets [31]. The GRU algorithm uses gating mechanisms to selectively update and reset the hidden state, allowing it to capture relevant information across sequences. Key components of a typical GRU cell are the update gate ($z$), the reset gate ($r$) and the new memory content ($h_t$). The update gate controls how much of the previous hidden state ($h_{t-1}$) is passed to the current time step. The reset gate [32] determines how much past information to forget. The reset gate modulates the previous hidden states, and the input at the current step combination comprises the memory content [33]. The operations of a typical GRU cell are defined mathematically by Eqs. (1)-(4). **Error! Reference source not found.** shows the GRU unit architecture. It illustrates the information flow through the two gates and the hidden state by showing how information is updated and retained at each step. We used several structures and hyperparameters to train the model. **Error! Reference source not found.** lists the most important parameters employed.

$$z_t = \sigma(W_z.[h_{t-1}, x_t]) \tag{1}$$

$$r_t = \sigma(W_r.[h_{t-1}, x_t]) \tag{2}$$

$$h_t = \tanh(r_t * [h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_t - 1 + z_t * h_t \tag{4}$$

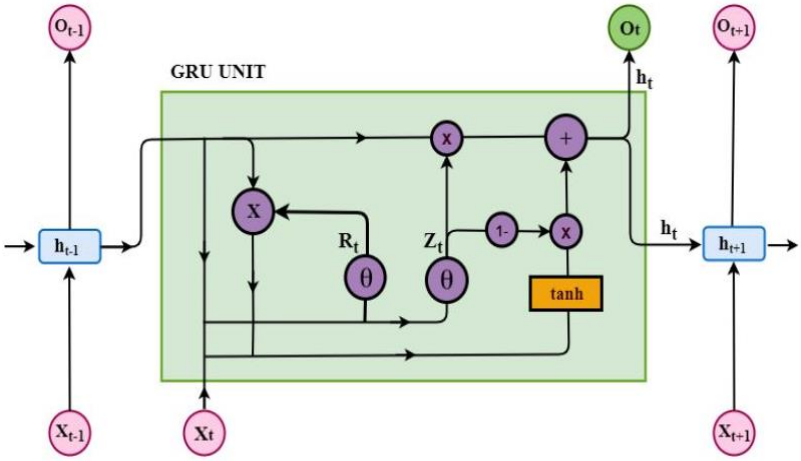



Figure 2. Typical GRU architecture and how information flows through the cell

Table 3. GRU model parameters

| Parameter | Value |
|---|---|
| GRU First Layer | 100 |
| GRU Second Layer | 100 |
| GRU Third Layer | 200 |
| Dense First Layer | 100 |
| Dense Second Layer | 1 |
| Optimizer | Adam (learning rate = 0.005) |
| Loss function | MSE |
| Metric | MAE |
| Epoch | 100 |
| Batch size | 20 |
| Execution Environment | TPU (Google Colab) |

**3.6 Evaluation Metrics**

We used several metrics to comprehensively evaluate the model. The evaluation metrics used included $R^2$, MAE, RMSE, R%, and MSE. The coefficient of determination (R-squared) in Eq. (5) quantifies the proportion of variance in the dependent variable explained by the independent variables. The closer the score is to 1, the better the reading [34].

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

MAE in Eq. (6) represents the average difference between the prediction and actual values. This provides an easily interpretable measure of the model accuracy [35]. The lower the MAE score, the better the results. The mathematical formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

RMSE metric in Eq. (7) is similar to MAE, although it assigns weights that are higher to large errors. If the RMSE is lower, then the model is better able to fit the dataset. The mathematical formula is as follows

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

The MSE evaluation metric in Eq. (8) is the average of the squared difference between the predicted and actual values. If the MSE is low, the model fits the data better [36]. The formula for finding the MSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

## 4. Results and Discussion

One objective of this study was to use better preprocessing methods coupled with an effective deep learning model that would perform better on a smaller dataset while avoiding overfitting. This study used a deep learning model to predict $CO_2$ emissions using a dataset from the Canadian vehicle database. The experiment was conducted in a Google Colab environment using Google Chrome. Google Colab was connected to a Python 3 Google computing backend (TPU-Tensor Processing Unit). At the time of execution, 12.67 GB of Random Access Memory was allocated, along with approximately 107.72 GB of disk space. The dataset was first pre-processed using label encoding for categorical features and scaling for numerical features, followed by min-max normalisation. DBSCAN was employed for outlier detection and clustering. The dataset was then split into training (80%) and testing (20%) datasets, and a variant of the RNN model called GRU was defined and trained on the training data. The model was evaluated using several metrics, namely $R^2$, MAE, MSE, and RMSE, on both the training and test data. Pearson's correlation coefficient ($R$) was used to assess the correlations. We used approximately 80% of the dataset samples to train the GRU model. This step is crucial for shaping the model's effectiveness before it is exposed to unseen data. **Error! Reference source not found.** presents the results obtained during the training phase. We used approximately 20% of the dataset samples to test the model. This step was performed after the model trained successfully. **Error! Reference source not found.** presents the results from model testing.

Table 4. GRU model training results

| Evaluation Metric | Score |
|---|---|
| R2 (%) | 97.97 |
| MAE | 0.009812 |
| MSE | 0.000378 |
| RMSE | 0.019433 |
| R (%) | 99.00 |

Table 5. GRU model testing results

| Evaluation metric | Score |
|---|---|
| R2 (%) | 98.14 |
| MAE | 0.0095097 |
| MSE | 0.000392 |
| RMSE | 0.019796 |
| R (%) | 99.09 |

Table 6. Comparison between results from previous studies and our model

| Model | RMSE score | Ref |
|---|---|---|
| ANN | 1.286 | [11] |
| SVR | 2.752 | [11] |
| LSTM | 0.1648 | [10] |
| SVR | 0.71 | [12] |
| BiLSTM | 0.03560 | [8] |
| GRU | 0.019796 | Proposed |

The findings of this study clearly show that we achieved our objective of improving results by using an alternative preprocessing methodology coupled with an effective GRU model to predict $CO_2$ emissions from motor vehicles. The high $R^2$ scores (97.97% for training and 98.14% for testing) indicate that the model captures a significant proportion of the variance in the dependent variable, demonstrating its effectiveness in capturing the complex relationships between the input and target features. Additionally, lower MAE, MSE, and RMSE values further validate the model's accuracy in predicting $CO_2$ emissions. A comparison with previous studies shows that the proposed method outperforms existing methodologies. **Error! Reference source not found.**-7 confirm our claim. Our model achieved higher scores than those of some previous studies, indicating its superiority. Specifically, our model showed higher Pearson correlations and R2 scores, indicating a stronger correlation and a better fit to the data.

In this study, the GRU outperformed the SVR, ANN, LSTM, and BiLSTM. GRU's superior performance of GRU is attributable to its ability to capture temporal dependencies while maintaining computational efficiency, making it particularly suitable for datasets with fewer samples. Unlike LSTM and BiLSTM, which also capture sequential dependencies, GRU has a simpler architecture that lets it process data with reduced computational load and a smaller parameter set. This makes it well-suited to scenarios with limited data, as is common in emissions datasets were collecting diverse, extensive data can be challenging. Choosing DBSCAN for outlier detection before model training significantly affected the results. Unlike traditional clustering methods such as K-means or Rough K-means, which assume spherical clusters, DBSCAN identifies data points that do not conform to the majority trend, regardless of cluster shape. This flexibility is critical for $CO_2$ emissions data, where outliers can arise from unique driving conditions, vehicle

configurations, or other atypical patterns. While K-means-based methods are prone to misclassifying such data points, DBSCAN is density-based and thus adept at isolating noise, improving the overall robustness of the model.

Table 7. Comparison between our model and the best model from previous studies

| Evaluation metric | GRU (Proposed) | BiLSTM [8] |
|---|---|---|
| R2 | 98.14 | 93.55 |
| MAE | 0.0095097 | Not Reported |
| MSE | 0.000392 | 0.0012678 |
| RMSE | 0.019796 | 0.03560 |
| R | 99.09 | 96.95 |

In comparison, other models from previous studies struggled with smaller datasets because they relied on larger data volumes to learn generalised patterns effectively. Although LSTM and BiLSTM are effective for time-series data, they are more prone to overfitting with limited data because their more complex architectures require extensive training data to avoid capturing noise. The combination of GRU and DBSCAN addresses this challenge by integrating a sequential model that is inherently resistant to overfitting with an outlier-detection approach that robustly handles anomalies, ensuring the model adapts well to smaller datasets. As shown in the tables above, our approach improved $R^2$ by 4.59% and R by 2.14%, while reducing MSE and RMSE by 0.0008766 and 0.0015804, respectively. Some developing countries have implemented taxation on motor vehicle emissions. The carbon tax amount is primarily calculated based on engine size (displacement). As a practical implementation of our research, we propose calculating the carbon tax based on $CO_2$ emissions rather than the current method. The system can utilise our proposed model to predict $CO_2$ emissions. We can then calculate the vehicle carbon tax based on the entered motor vehicle parameters. Integrating carbon tax calculations based on emission levels is a tangible application of our research findings. This facilitates the implementation of environmental policies aimed at reducing $CO_2$ emissions.

## 5. Conclusions

This study addresses the urgent need for accurate prediction of $CO_2$ emissions from motor vehicles to mitigate climate change, leveraging the GRU model with DBSCAN for improved outlier detection and data handling. Motor vehicle emissions substantially contribute to global $CO_2$ levels, and current emission prediction methods often lack robustness, particularly with smaller datasets. We modelled our approach on a Canadian dataset with 7,385 samples and used label encoding, Min-Max normalisation, and DBSCAN preprocessing to ensure high data quality before applying the GRU model. The GRU model achieved strong results, with an $R^2$ of 98.14%, MAE of 0.0095097, MSE of 0.000392, R of 99.09%, and RMSE of 0.019796, thus demonstrating improved accuracy and resilience. Compared with prior models, our model reduced MSE and RMSE by 0.0008766 and 0.0015804, respectively, while increasing $R^2$ by 4.59% and R by 2.14%. These results underscore the model's potential to inform more precise carbon tax calculations based on predicted $CO_2$ emissions rather than simpler parameters such as engine size (displacement) alone, offering a practical pathway to incorporate emissions data into policymaking. Future research can explore hybrid deep learning models and incorporate additional vehicle features, such as age or year of manufacture, to increase the model's applicability and predictive accuracy. This study and its extension align with our motivation to develop comprehensive and adaptable models that can support sustainable environmental policies and help curb emissions through informed regulatory strategies.

**Acknowledgements**

The authors thank the School of Engineering and Technology, Mulungushi University, Zambia, for providing the infrastructure needed to carry out this research.

**Funding**

This study was not supported by any grants from funding bodies in the public, private, or not-for-profit sectors.

**Declaration of Competing Interest**

The authors declare no conflict of interest.

**Credit Authorship Contribution Statement**

M. Mukombo: Methodology; Data curation; Writing - original draft; Resources
M. Sinyinda: Conceptualization; Formal analysis; Supervision; Writing – reviewing & editing

**Availability of Data and Materials**

Data sharing is not applicable to this article, as this study did not generate or analyse new datasets.

**Ethics Declarations**

This study did not involve human participants or animal subjects. Therefore, this study did not require ethical approval.

**Generative Artificial Intelligence Declarations**

No generative artificial intelligence tools were used to prepare this manuscript. All intellectual content is original and is solely attributable to the authors.

**References**


[1] H. Jin, "Prediction of direct carbon emissions of Chinese provinces using artificial neural networks," *PLOS ONE*, vol. 16, no. 5, Art. no. e0236685, 2021, https://doi.org/10.1371/journal.pone.0236685.

[2] I. H. Shah, M. A. Manzoor, W. Jinhui, *et al.*, "Comprehensive review: Effects of climate change and greenhouse gases emission relevance to environmental stress on horticultural crops and management," *Journal of Environmental Management*, vol. 351, Art. no. 119978, 2024, https://doi.org/10.1016/j.jenvman.2023.119978.

[3] T. R. Bajracharya, S. R. Shakya, and A. Sharma, "Climate change and sustainable energy systems," in *Handbook of Energy and Environmental Security*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 531–545, https://doi.org/10.1016/B978-0-12-824084-7.24001-0.

[4] A. Goswami, H. S. Kapoor, R. K. Jangir, *et al.*, "Impact of economic growth, trade openness, urbanization and energy consumption on carbon emissions: A study of India," *Sustainability*, vol. 15, no. 11, Art. no. 9025, 2023, https://doi.org/10.3390/su15119025.

[5] S. Aminzadegan, M. Shahriari, F. Mehranfar, and B. Abramović, "Factors affecting the emission of pollutants in different types of transportation: A literature review," *Energy Reports*, vol. 8, pp. 2508–2529, 2022, https://doi.org/10.1016/j.egyr.2022.01.161.

[6] M. Emami Javanmard, Y. Tang, Z. Wang, and P. Tontiwachwuthikul, "Forecast energy demand, $CO_2$ emissions and energy resource impacts for the transportation sector," *Applied Energy*, vol. 338, Art. no. 120830, 2023, https://doi.org/10.1016/j.apenergy.2023.120830.

[7] V. L. Bartlett, H. Doernberg, M. Mooghali, *et al.*, "Published research on the human health implications of climate change between 2012 and 2021: A cross-sectional study," *BMJ Medicine*, vol. 3, no. 1, Art. no. e000627, 2024, https://doi.org/10.1136/bmjmed-2023-000627.

[8] H. Boogaard, A. P. Patton, R. W. Atkinson, *et al.*, "Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis," *Environment International*, vol. 164, Art. no. 107262, 2022, https://doi.org/10.1016/j.envint.2022.107262.

[9] A. H. Al-Nefaie and T. H. H. Aldhyani, "Predicting $CO_2$ emissions from traffic vehicles for sustainable and smart environment using a deep learning model," *Sustainability*, vol. 15, no. 9, Art. no. 7615, 2023, https://doi.org/10.3390/su15097615.

[10] R. Zhang, Y. Wang, Y. Pang, *et al.*, "A deep learning micro-scale model to estimate the $CO_2$ emissions from light-duty diesel trucks based on real-world driving," *Atmosphere*, vol. 13, no. 9, Art. no. 1466, 2022, https://doi.org/10.3390/atmos13091466.

[11] O. S. Azeez, B. Pradhan, H. Z. M. Shafri, *et al.*, "Modeling of CO emissions from traffic vehicles using artificial neural networks," *Applied Sciences*, vol. 9, no. 2, Art. no. 313, 2019, https://doi.org/10.3390/app9020313.

[12] A. S. Chadha, Y. Shinde, N. Sharma, and P. K. De, "Predicting $CO_2$ emissions by vehicles using machine learning," in *Data Management, Analytics and Innovation*. Singapore: Springer, 2023, pp. 197–207, https://doi.org/10.1007/978-981-19-2600-6_14.

[13] S. Shah, S. Thakar, K. Jain, B. Shah, and S. Dhage, "A comparative study of machine learning and deep learning techniques for prediction of $CO_2$ emission in cars," 2022, arXiv:2211.08268.

[14] T. Jia, P. Zhang, and B. Chen, "A microscopic model of vehicle $CO_2$ emissions based on deep learning: A spatiotemporal analysis of taxicabs in Wuhan, China," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18446–18455, 2022, https://doi.org/10.1109/TITS.2022.3151655.

[15] M. Ghahramani and F. Pilla, "Analysis of carbon dioxide emissions from road transport using taxi trips," *IEEE Access*, vol. 9, pp. 98573–98580, 2021, https://doi.org/10.1109/ACCESS.2021.3096279.

[16] N. D. Rivera-Campoverde, J. L. Muñoz-Sanz, and B. del Valle Arenas-Ramirez, "Estimation of pollutant emissions in real driving conditions based on data from OBD and machine learning," *Sensors*, vol. 21, no. 19, Art. no. 6344, 2021, https://doi.org/10.3390/s21196344.

[17] X. Li, A. Ren, and Q. Li, "Exploring patterns of transportation-related $CO_2$ emissions using machine learning methods," *Sustainability*, vol. 14, no. 8, Art. no. 4588, 2022, https://doi.org/10.3390/su14084588.

[18] Y. Natarajan, G. Wadhwa, K. R. Sri Preethaa, and A. Paul, "Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms," *Electronics*, vol. 12, no. 10, Art. no. 2288, 2023, https://doi.org/10.3390/electronics12102288.

[19] R. Abdulmalik and G. Srivastava, "Forecasting of transportation-related $CO_2$ emissions in Canada with different machine learning algorithms," *Advances in Artificial Intelligence and Machine Learning*, vol. 3, no. 3, pp. 1295–1312, 2023, https://doi.org/10.54364/AAIML.2023.1176.

[20] H. Mangla, A. Sharma, and S. Shivangi, "Fuel emission detection using machine learning," *International Journal of Advances in Engineering and Management*, vol. 3, no. 6, pp. 2471–2476, 2021.

[21] M. Mądziel, "Liquified petroleum gas-fuelled vehicle $CO_2$ emission modelling based on portable emission measurement system, on-board diagnostics data and gradient-boosting machine learning," *Energies*, vol. 16, no. 6, Art. no. 2754, 2023, https://doi.org/10.3390/en16062754.

[22] S. D. Bappon, A. Dey, S. M. Sabuj, and A. Das, "Toward a machine learning approach to predict the $CO_2$ rating of fuel-consuming vehicles in Canada," in *Proceedings of the 25th International Conference on Computer and Information Technology (ICCIT)*, 2022, pp. 384–389, https://doi.org/10.1109/ICCIT57492.2022.10054732.

[23] O. Sami, Y. Elsheikh, and F. Almasalha, "The role of data preprocessing techniques in improving machine learning accuracy for predicting coronary heart disease," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 816–824, 2021, https://doi.org/10.14569/IJACSA.2021.0120695.

[24] M. K. Dahouda and I. Joe, "A deep-learned embedding technique for categorical features encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, https://doi.org/10.1109/ACCESS.2021.3104357.

[25] N. Kosaraju, S. R. Sankepally, and K. Mallikharjuna Rao, "Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—A practical review study on heart disease prediction dataset using Pearson correlation," in *Proceedings of International Conference on Data Science and Applications*. Singapore: Springer, 2023, pp. 369–382, https://doi.org/10.1007/978-981-19-6631-6_26.

[26] S. Sinsomboonthong, "Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification," *International Journal of Mathematics and Mathematical Sciences*, vol. 2022, Art. no. 3584406, 2022, https://doi.org/10.1155/2022/3584406.

[27] M. Z. Rodriguez, C. H. Comin, D. Casanova, *et al.*, "Clustering algorithms: A comparative approach," *PLOS ONE*, vol. 14, no. 1, Art. no. e0210236, 2019, https://doi.org/10.1371/journal.pone.0210236.

[28] H. E. G. Lopes and M. de S. Gosling, "Cluster analysis in practice: Dealing with outliers in managerial research," *Revista de Administração Contemporânea*, vol. 25, no. 1, Art. no. e200081, 2021, https://doi.org/10.1590/1982-7849rac2021200081.

[29] G. Yigit and M. F. Amasyali, "Simple but effective GRU variants," in *Proceedings of the 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, 2021, pp. 1–6, https://doi.org/10.1109/INISTA52262.2021.9548535.

[30] M. Hasnain, S. R. Jeong, M. F. Pasha, and I. Ghani, "Performance anomaly detection in web services: An RNN-based approach using dynamic quality of service features," *Computers, Materials & Continua*, vol. 64, no. 2, pp. 729–752, 2020, https://doi.org/10.32604/cmc.2020.010394.

[31] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, *et al.*, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, vol. 137, Art. no. 106024, 2019, https://doi.org/10.1016/j.cie.2019.106024.

[32] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022, https://doi.org/10.5194/gmd-15-5481-2022.

[33] A. Zameer, F. Jaffar, F. Shahid, *et al.*, "Short-term solar energy forecasting: Integrated computational intelligence of LSTMs and GRU," *PLOS ONE*, vol. 18, no. 10, Art. no. e0285410, 2023, https://doi.org/10.1371/journal.pone.0285410.

[34] T. O. Kvalseth, "Note on the $R^2$ measure of goodness of fit for nonlinear models," *Bulletin of the Psychonomic Society*, vol. 21, no. 1, pp. 79–80, 1983, https://doi.org/10.3758/BF03329960.

[35] N. Haut, W. Banzhaf, and B. Punch, "Correlation versus RMSE loss functions in symbolic regression tasks," in *Genetic Programming Theory and Practice XIX*, L. Trujillo, S. M. Winkler, S. Silva, and W. Banzhaf, Eds. Cham, Switzerland: Springer, 2023, pp. 31–55, https://doi.org/10.1007/978-981-19-8460-0_2.

[36] G. S. Handelman, H. K. Kok, R. V. Chandra, *et al.*, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *AJR American Journal of Roentgenology*, vol. 212, no. 1, pp. 38–43, 2019, https://doi.org/10.2214/AJR.18.20224.