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 GRU and DBSCAN

Mukombo Mukombo[1*], Muwanei Sinyinda[1]

[1] School of Engineering and Technology, Mulungushi University, Kabwe 80415, Zambia

ABSTRACT - Extreme climatic conditions resulting from climate change pose significant challenges to mankind, prompting urgent action across various fronts. With vehicle emissions contributing substantially to carbon dioxide emissions, there is a pressing need for efficient prediction methods for informed mitigation strategies. While some research has been carried out on predicting carbon dioxide using machine learning, there have been few studies that explore Recurrent Neural Networks particularly those that can be modelled with a smaller dataset to improve emissions prediction. The study aimed to improve the prediction of carbon dioxide emissions by employing a new approach. The study was conducted on a Canadian dataset with 7,385 samples featuring motor vehicle parameters. Therefore, this research employs a deep learning model known as Gated Recurrent Units (GRU) for predictive modelling coupled with another algorithm called the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for robust outlier detection and subsequent removal. Other algorithms that were used for preprocessing include label encoding and min-max normalisation. The study was evaluated using four metrics, namely 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, alongside reductions in MSE and RMSE by 0.0008766 and 0.0015804, respectively. Thus, showing superior handling of data irregularities and temporal patterns. These results highlight the potential for improved emissions predictions, which can be instrumental in refining carbon tax calculations and informing environmental policy. Future works will involve expanding the dataset to include the age of the motor vehicle and using hybrid deep learning models.

ARTICLE HISTORY
Received : 01 December 2024
Revised : 22 September 2025
Accepted : 20 April 2026
Published : 04 May 2026

KEYWORDS
Climate Change
CO_2 emissions
Deep learning
DBSCAN
Gated Recurrent Units

1.0 INTRODUCTION

The world is currently grappling with the growing concern of climate change with temperatures rising steadily thus triggering a number of environmental, economic and social consequences [1]. The primary source of this warming trend and subsequently climate alterations are increases in greenhouse gas (GHG) emissions, particularly carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O) from various human activities which include industrial processes, deforestation and energy production which have been increasing in proportion to rising temperatures[2-4]. There are many sources of these emissions of which the transport sector is one of the many which are human-induced [5]. Among the leading contributors to GHG emissions is the transport sector, which accounts for a substantial share of CO_2 emissions globally [6]. The transport sector especially motor vehicles, plays a critical role in increasing atmospheric CO_2 levels due to its heavy dependency on fossil fuels derivatives. As vehicles combust diesel, gasoline and other petroleum products, they release CO_2 and other pollutants thus contributing 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 rather disperse into the atmosphere, affecting communities, ecosystems and economies worldwide.

In light of these growing challenges, reducing CO_2 emissions from motor vehicles has become a pivotal objective for environmental policy and technological innovation. Numerous strategies have been proposed, ranging from the promotion of electric vehicles and alternative fuels to the development of more efficient engines. However, one critical area that demands attention is the ability to accurately predict 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 actual environmental impact [5][10]. Given the importance of reliable emission data, machine learning has seen significant interest in developing predictive models for CO_2 emissions. These models, which rely on large datasets of vehicle parameters, can help identify key factors contributing to emissions and offer insights into how these emissions can be reduced or mitigated. 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 for the prediction of 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 are not without limitations. Most models have been designed

***CORRESPONDING AUTHOR | M. Mukombo** | ✉ mmukombo06@gmail.com

for larger datasets and do not perform well when applied to 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 DBSCAN for preprocessing. 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 like BiLSTM, GRU simplifies the computational process by using fewer gating mechanisms, which makes it particularly suitable for smaller datasets while still capturing temporal patterns effectively. DBSCAN, on the other hand, is a powerful clustering algorithm that excels in identifying outliers and handling noise in datasets. It offers significant advantages over traditional clustering methods like K-means, as it does not require a predefined number of clusters and is robust to varying densities.

The combination of DBSCAN and GRU allows us to build a robust predictive model that not only handles datasets effectively but also manages outliers in a way that improves overall accuracy. By improving the quality of the input data and refining the learning process, we expect this model to outperform existing approaches and offer more reliable predictions for motor vehicle CO2 emissions.

This paper presents the following major contributions:

1. The use of DBSCAN for preprocessing ensures that outliers are 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 demonstrates superior performance in capturing complex temporal patterns within smaller datasets, outperforming previous models such as BiLSTM. This work establishes GRU as an effective tool for emission prediction, particularly when the available data is limited in size.
3. This research 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 policy-making, such as in carbon tax frameworks.

The rest of the paper is organized 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 previous studies. Section 5 concludes the paper and outlines the future works.

2.0 RELATED WORKS

Some researchers conducted a study that employed AI to enhance the prediction of CO_2 emissions. Machine Learning (ML), deep learning (DL) and ensemble learning models were utilised on a large dataset from the European Union. The research elected to use AI as they observed that the CO_2 figures that were made available by the governments were not properly incorporating motor vehicles on the roads and that the employed statistical methods were less accurate. The study's findings had the potential to foretell any rise in temperature and help shape key policies such as the adoption of electric vehicles among others. As a further work, a proposal was made that there is a need to validate their proposed DL and ML ensemble techniques on other research problems [13].

Another study aimed at assessing the impact on the environment of intelligent systems in transportation by developing a vehicle emissions model with a high degree of accuracy was undertaken. Thus, a deep learning-based vehicle emissions model (DL-VEM) was modelled and evaluated as a way of estimating the instant CO_2 emissions of taxicabs with a focus on establishing a correlation between driving conditions and emissions. One important finding of this study was that there was a 24.94% increase in emissions if petrol was used instead of compressed natural gas [14]. In this study, the authors reviewed strategies for managing issues associated with transportation emissions and how to address the impact. They investigated the impact of taxi trips on CO_2 emissions using two sets of datasets namely the taxi trips records and a fuel economy dataset. The study resulted in the identification of five clusters of motor vehicles associated with emissions thus aiding in categorising vehicles by emission levels. Motor vehicles that were categorised as high emitters could be targeted for further scrutiny and possible sanctions [15].

A methodology for estimating emissions using onboard diagnostics data in real driving conditions utilising machine learning and artificial neural networks was proposed. This research was initiated because it was discovered that there were no models for the estimation of emissions without larger measurement campaigns. The research's main contribution is its potential to aid vehicular homologation and emissions inventories. They further proposed to adjust their model for various parameters like the age of the vehicle, driving styles and weather conditions as part of future works [16].

There was a need to garner knowledge on CO_2 emissions produced by the transport sector in several countries so that countries could realign their future energy strategies and policy agendas by exploring patterns of emissions related to transportation. Several countries' data was utilised by employing ordinary least square regression (OLS), SVM and GBR in CO_2 forecasting. One of the findings indicates that there is a close relationship between CO_2 emissions and several other factors. Pinpointing influential factors can enable decision-makers to develop strategies aimed at curbing transport-related CO_2 emissions growth in the near future [17]. Ensemble learning models were employed in CO_2 emissions

prediction of light-duty vehicles in leveraging vehicle specifications in CO_2 forecasting. The dataset was sourced from the Government of Canada and included features such as cylinders, engine size and fuel economy. Catboost outperformed other algorithms which were used in the study thus demonstrating robustness in various traffic conditions. This study's contribution to the body of knowledge is its ability to estimate emissions and its possible application in motor vehicle homologation and vehicular emissions inventories. The researchers of this particular study proposed to enhance the model performance by expanding dataset features in a bid to increase accuracy [18].

An investigation of the rise in CO_2 emissions in Canada was conducted by focusing on transport-related emissions. Machine learning and deep learning models were utilised to investigate the increase of CO_2 emissions due to various factors like transportation economic development and population growth with a particular focus on transport-related emissions. This study recommends the usage of alternative motor vehicle fuel types as well as fuels that have low carbon content [19]. A random forest model was utilised in developing a CO_2 emissions prediction application by using a dataset consisting of fuel features. The application developed has the potential to provide insights into various aspects including the identification of car manufacturers with the lowest emissions. This could further stimulate competition among manufacturers and force them to make environmentally friendly vehicles [20]. This study is one of the first endeavors to develop a methodology in microscale CO_2 emission models of liquefied petroleum gas (LPG). This study relied on data that was acquired from road tests by utilising portable emission measurement systems and on-board diagnostics interfaces. This study is important as it can be used to analyse continuous CO_2 emissions and in the creation of emission maps for environmental analysis in built-up areas [21].

Table 1 presents an overview of the previous studies that are 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	There is a need to validate their proposed DL and ML ensemble techniques 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 also 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 were applied to the prediction of 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. Consequently, as a result of this study, an original strategy for accurate CO_2 emissions forecasting was developed thus giving policymakers a tool that they can employ to create policies [8]. An experiment was conducted

to predict CO_2 emission ratings of motor vehicles based on various factors and identification of vehicles with CO_2 emissions beyond a certain range. The best-performing model was random forest. This work has the potential to identify and design motor vehicles that have 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 particular strengths depending on dataset characteristics. Commonly used datasets, such as transportation data and vehicle features, often relate to specific regions or fuel types, thus limiting the model's applicability on a broader scale. While some models have achieved commendable accuracy, many of these approaches lack versatility across different datasets and environmental contexts. A recurring challenge that we observed is handling outliers, especially within smaller datasets where irregularities can skew predictions. This highlights a need for robust methods that can adapt to temporal patterns and variations in emissions data, as well as effectively identify anomalies without relying on predetermined assumptions about data distribution. By focusing on models that inherently balance computational efficiency with the ability to recognize complex dependencies in time-series data; we address gaps found in previous studies. The GRU-DBSCAN approach does not only enhance prediction accuracy but 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

3.0 METHODS AND MATERIAL

3.1 Dataset

In this study, a publicly available dataset from kaggle.com was used. The source of the dataset is the Canadian government. The dataset was built over seven years and has a total number of 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 Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg) and CO_2 Emissions (g/km).

The 'Make' feature has about 42 different motor vehicle types ranging from Acura to Volvo. The data is textural. The vehicle 'Model' feature has varying names according to the manufacturers but features the following key aspects: four-wheel drives (4WD/4×4), flexible-fuel vehicles (FFV), all-wheel drives (AWD), long wheelbase (LWB), short wheelbase (SWB) and extended wheelbase (EWB). The 'Vehicle Class' mostly deals with the shape of the vehicle and varies from the compact class to van passengers. 'Engine size (L)' has the displacement measurements of the engine with the smallest engine being 0.9L and the largest being 8.4L. The number of cylinders in the engine falls under the feature named 'cylinders'. The least number of cylinders is 3 with the largest engine having 16. Another feature is 'Transmission', which has several options namely automatic (A), automatic with shift (AS), manual (M) and continuously variable (AV). The transmission options are equipped with varying numbers of gears, with our dataset ranging from 3 to 10 gears.

The dataset has five ‘Fuel Types’ as a feature, namely 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 in integers while the others consist of decimal numbers/floats. The last feature is ‘CO_2 Emissions (g/km)’ which is in integer form is 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 as it ensures that the data is in a form that is suitable for analysis as well as modelling. Good data preprocessing improves the results' quality and reliability[23]. Under preprocessing, we utilised label encoding and Min-Max normalisation. The following paragraphs under preprocessing explain in brief each of the preprocessing techniques that were used.

Label encoding or ordinal encoding was used in the conversion of categorical data into numerical data. Typically, a random number is assigned to each unique category in the feature of concern without distorting the dataset. In other words, it is the conversion of a label into an integer[24]. The dataset which was 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 like in the case of our features [25].

Min-Max normalisation is one of the commonly used preprocessing methods. It was used to transform numerical variables to a common scale whilst maintaining the differences in the range of values and preventing loss of information. The scaling was in the range of zero (0) and one (1). Min-Max normalisation was used due to its ability to optimize machine learning processes like gradient descent thus enabling convergence to occur faster. It also improves the algorithm’s performance and speed. Min-Max normalisation also scales features without losing the 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 due to its density-based approach. The algorithm initiates by selecting a random point within the dataset and then identifies all other points located within a specified radius, known as epsilon (ε), around this point. 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 that were assigned to cluster -1 [27][28].

DBSCAN was an ideal choice for this study due to several key advantages. DBSCAN relaxes the requirement for cluster number specification, 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 customizable, offering parameters that can be adjusted to optimize performance for different datasets [8][16].

3.4 Statistical Insights

Before embarking on model training and subsequent model testing, it is important to emphasize the significance of providing statistical insights into the dataset. Prioritization through statistical analysis doesn't just bring out data characteristics, interdependencies and distribution but it equally ensures that subsequent modelling efforts are based on a solid understanding of the dataset which leads 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 [8] have already conducted comprehensive statistical insights. Rather than duplicating their efforts, we conducted an insight into the dataset before and after preprocessing which included label encoding, min-max normalisation and the detection of outliers using DBSCAN. The focus was on examining the number of instances, identification of duplicates and handling of empty values. **Error! Reference source not found.** below presents the key insights.

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

A correlation plot represents visually, the strength and direction of association between various features. It is a display of the correlation between various features that were elected to be incorporated in modelling. **Error! Reference source not found.**shows the correlation plot of the features under consideration. After careful consideration, we incorporated engine size, cylinders and the four fuel consumptions as our variables and CO_2 emission as our target feature in our model.

3.5 Prediction Model

GRU is a variant of RNN designed to overcome challenges in learning and retaining long-term dependencies in sequential data among others. GRU addresses the problem of vanishing gradient during training[29], which is a common limitation that traditional RNNs possess[30]. In our study, we employed GRU because it performs better with a smaller dataset[31]. The GRU algorithm implements smart gating mechanisms to selectively update and reset the hidden state, allowing for the capture of 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 extent to which the hidden state from the previous time step (h_{t-1}) is controlled by the update gate and thus passed to the current time step. The amount of past information that should be forgotten is determined by the reset gate[32]. The reset gate modulates previous hidden states and the input at a current step combination makes up the memory content[33]. The operations of a typical GRU cell are defined mathematically in (1) to(4) below:

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

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

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

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

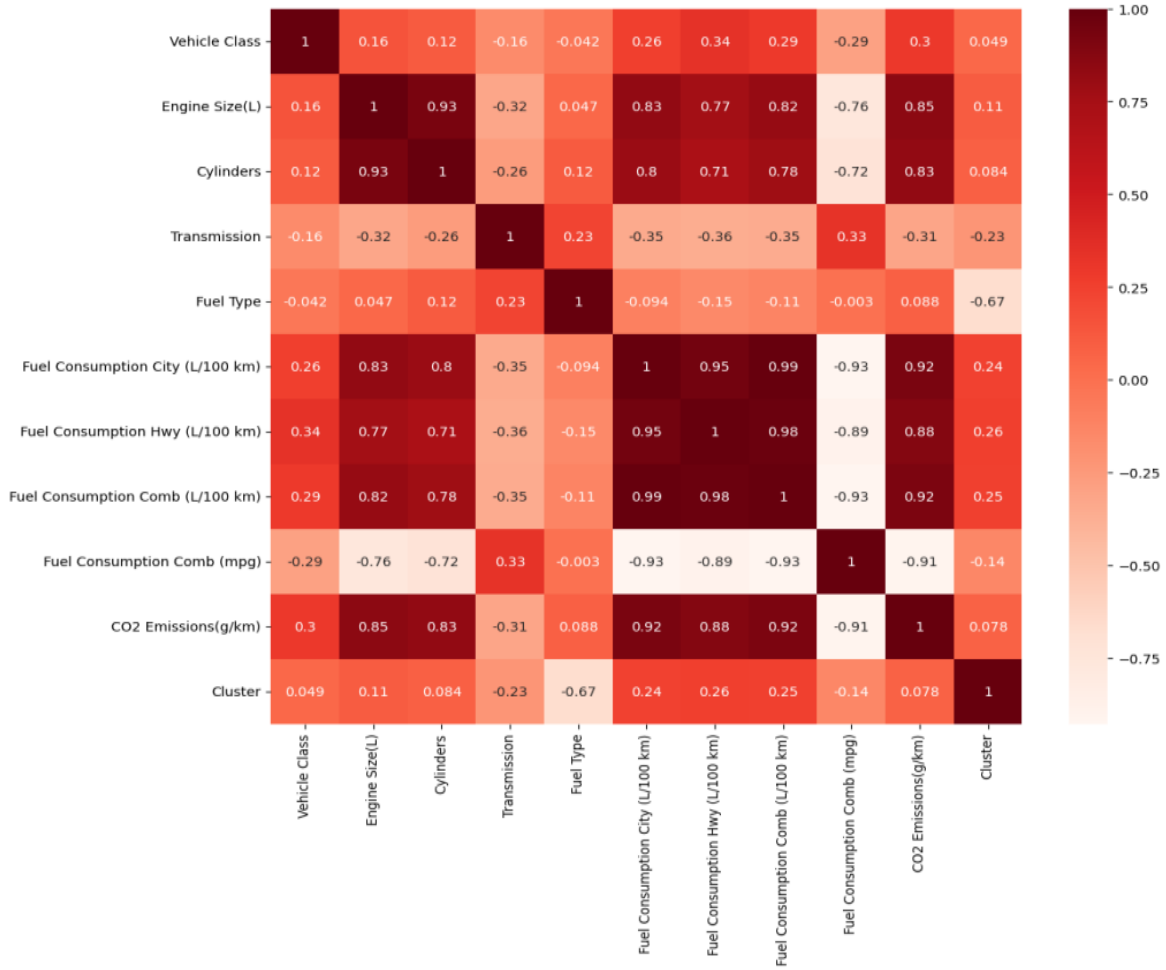

Figure 1. Correlation plot of the features in the dataset

Error! Reference source not found. below shows a GRU unit architecture. It shows the information flow through the two gates and the hidden state by illustrating how information is updated and retained at each step.

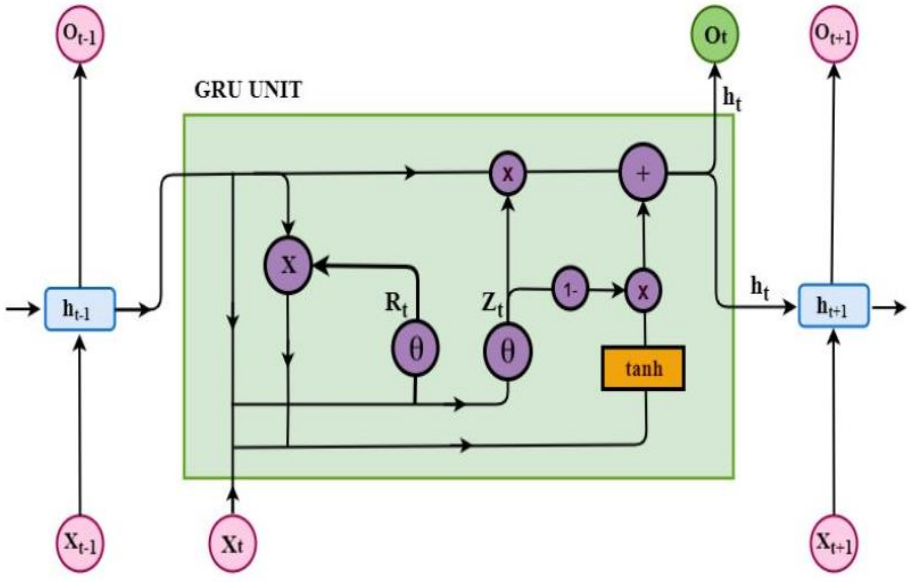

Figure 2. Typical GRU architecture and how information flows through the cell

Several structures and hyperparameters are used in the training of the model. **Error! Reference source not found.** below shows the most important parameters that were employed.

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

To comprehensively evaluate the model, several metrics were employed. The evaluation metrics that were utilised include R^2, MAE, RMSE, R% and MSE.

The coefficient of Determination (R-squared) in (5) is used in regression models to quantify the proportion of the variance in independent variables that is explained by the independent variables. The closer the score is to 1 or 100%, the better the reading[34].

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

MAE in (6) represents a measure of the average between the prediction values and the actual ones. It provides an easily interpretable measure of the accuracy of the model [32]. The lower the MAE score, the better the result. The mathematical formula is:

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

RMSE metric in (7) is similar to MAE though it gives weights that are higher to large errors. If the RMSE is lower, then the better the model was able to fit a dataset [32], [35]. 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 (8) is the average of the squared difference between the predicted values and the actual ones. If the MSE is low, then the better the model fits the data [36]. The formula for finding MSE is as follows:

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

4.0 RESULTS AND DISCUSSION

One of the objectives of this study was to employ better preprocessing methods coupled with an effective deep learning model that would perform better on a smaller dataset whilst avoiding suffering from overfitting. The experiment is anchored on the usage of a deep learning model in the prediction of CO_2 emissions using a dataset from the Canadian vehicle database. The experiment was carried out on a Google Colab environment using a Google Chrome browser. Google Colab was connected to a Python 3 Google computing backend (TPU-Tensor Processing Unit). At the time of execution of the experiment, Random Access Memory (RAM) amounting to 12.67 GB was allocated alongside disk space of about 107.72 GB.

Preprocessing of the dataset was carried out first using label encoding of categorical features and then scaling of numerical features soon followed using Min-Max normalisation. DBSCAN was employed for outlier detection and clustering purposes. The dataset was then split into training (80%) and testing (20%) and then a variant of the RNN model called GRU was defined and trained on the trained data. The model was evaluated using several metrics namely R2, MAE, MSE and RMSE both the training and test data. Pearson's correlation coefficient (R) was also used to assess correlations.

About 80% of the dataset samples were used for training the GRU model. This is a crucial step in shaping the effectiveness of the model before exposing it to unseen data. **Error! Reference source not found.** below shows the results that were obtained during the training phase.

Table 4. GRU model training results

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

Approximately 20% of the dataset samples were used in testing the model. This step is undertaken after successful training of the model. **Error! Reference source not found.** below shows the results that were obtained during the testing of the model.

Table 5. GRU model testing results

Evaluation Metric	Score
R2 (%)	98.14
MAE	0.0095097
MSE	0.000392
RMSE	0.019796
R (%)	99.09

4.1 Discussion

The findings of this study clearly show that our objective of scoring better results by employing an alternate preprocessing methodology coupled with an effective GRU model in the prediction of CO_2 emissions from motor vehicles was fulfilled. The high scores of R^2 (97.97% for training and 98.14% for testing) indicate that the model captures a significant proportion of the variance in the dependent variable, thus demonstrating effectiveness in capturing the complex relationships between the input features and that of the target feature. Additionally, lower values of MAE, MSE and RMSE go further to validate the model's accuracy in predicting CO_2 emissions.

A comparison of these results with previous studies shows that this experimentation outperforms existing methodologies. **Error! Reference source not found.** and **Error! Reference source not found.** below confirm our claim. Our model achieved higher scores compared to some previous studies thus indicating its superiority. Specifically, our model exhibits higher Pearson's correlations and R2 scores suggesting a much stronger correlation and better fit to the data.

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

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 this study, GRU outperformed SVR, ANN, LSTM and BiLSTM. GRU's superior performance is attributable to its ability to capture temporal dependencies whilst maintaining computational efficiency, making it particularly suitable for datasets with fewer samples. Unlike LSTM and BiLSTM, which also capture sequential dependencies, GRU has a much simpler architecture that allows it to process data with reduced computational load and a smaller parameter set. This makes it well-suited for scenarios where data availability is limited, as is common in emissions datasets where collecting diverse, extensive data can be challenging.

The choice of DBSCAN for outlier detection before model training significantly impacted the results. Unlike traditional clustering methods like K-means or Rough K-means, which assume spherical clusters, DBSCAN effectively identifies data points that don't conform to the majority trend, independent of cluster shape. This flexibility is critical in CO_2 emissions data, where outliers can arise from unique driving conditions, vehicle configurations, or other non-typical patterns. While K-means-based methods are prone to misclassify such data points, DBSCAN is density-based and thus adept at isolating noise, improving the overall robustness of the model.

In comparison, other models from previous studies struggle with smaller datasets because they rely on larger data volumes to learn generalized patterns effectively. LSTM and BiLSTM, while effective for time-series data, are more prone to overfitting with limited data, as their more complex architectures demand extensive training data to avoid capturing noise. The combination of GRU and DBSCAN addresses this challenge by integrating a sequential model inherently resistant to overfitting with an outlier detection approach that robustly handles anomalies, ensuring the model's adaptability to smaller datasets. As evidenced from the tables above our approach scored fairly well with a 4.59% increase in R^2 and a 2.14% increase in R while reductions were recorded in MSE and RMSE by 0.0008766 and 0.0015804 respectively.

Some developing countries have implemented a form of taxation on motor vehicle emissions. Mostly, the amount of carbon tax is calculated from the engine size (displacement). As a practical implementation of our research, we propose calculating carbon tax based on CO_2 emissions rather than the current method. The system can utilise our proposed model to predict the amount of CO_2 emissions. Subsequently, it can calculate the vehicle carbon tax based on entered motor vehicle parameters. The integration of carbon tax calculations based on emissions levels is a tangible application of our research findings. This facilitates the implementation of environmental policies which are aimed at reducing CO_2 emissions.

5.0 CONCLUSIONS

This research responds to the urgent requirement for accurate prediction of CO_2 emissions from motor vehicles to mitigate climate change, leveraging the GRU model with DBSCAN for enhanced outlier detection and data handling. Motor vehicle emissions are one of the substantial contributors to global CO_2 levels and the current emissions prediction methods often lack robustness, particularly with smaller datasets. Our approach was modelled on a Canadian dataset with 7,385 samples and utilised 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. When compared to prior models, our model showed reductions in MSE and RMSE by 0.0008766 and 0.0015804 respectively while a 4.59% increase in R^2 and a 2.14% increase in R was observed. These results underscore the potential for this model to inform more precise carbon tax calculations, which can be based on predicted CO_2 emissions rather than simpler parameters like engine size (displacement) alone, offering a practical pathway to incorporate emissions data into policy-making.

Future research can explore hybrid deep learning models and incorporate additional vehicle features, such as the 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, adaptable models that can support sustainable environmental policies and help curb emissions through informed regulatory strategies.

ACKNOWLEDGEMENTS

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

AUTHORS CONTRIBUTION

M. Mukombo (Methodology; Data curation; Writing - original draft; Resources)

M. Sinyinda (Conceptualization; Formal analysis; Supervision)

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] Jin H. Prediction of direct carbon emissions of Chinese provinces using artificial neural networks. PLoS One. 2021;16(5):e0236685.

[2] Shah IH, Manzoor MA, Jinhui W, Li X, Hameed MK, Rehaman A, et al. Comprehensive review: Effects of climate change and greenhouse gases emission relevance to environmental stress on horticultural crops and management. J Environ Manage. 2024;351:119978.

[3] Bajracharya TR, Shakya SR, Sharma A. Climate change and sustainable energy systems. In: Handbook of Energy and Environmental Security [Internet]. Elsevier; 2022 [cited 2024 Feb 29]. p. 531–545. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128240847240010

[4] Goswami A, Kapoor HS, Jangir RK, Ngigi CN, Nowrouzi-Kia B, Chattu VK. Impact of economic growth, trade openness, urbanization and energy consumption on carbon emissions: A study of India. Sustainability. 2023;15(11):9025.

[5] Aminzadegan S, Shahriari M, Mehranfar F, Abramović B. Factors affecting the emission of pollutants in different types of transportation: A literature review. Energy Rep. 2022;8:2508–2529.

[6] Emami Javanmard M, Tang Y, Wang Z, Tontiwachwuthikul P. Forecast energy demand, CO2 emissions and energy resource impacts for the transportation sector. Appl Energy. 2023;338:120830.

[7] Bartlett VL, Doernberg H, Mooghali M, Gupta R, Wallach JD, Nyhan K, et al. Published research on the human health implications of climate change between 2012 and 2021: A cross-sectional study. BMJ Med. 2024;3(1):e000627.

[8] Boogaard H, Patton AP, Atkinson RW, Brook JR, Chang HH, Crouse DL, et al. Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis. Environ Int. 2022;164:107262.

[9] Al-Nefaie AH, Aldhyani THH. Predicting CO2 emissions from traffic vehicles for sustainable and smart environment using a deep learning model. Sustainability. 2023;15(9):7615.

[10] Zhang R, Wang Y, Pang Y, Zhang B, Wei Y, Wang M, et al. A deep learning micro-scale model to estimate the CO2 emissions from light-duty diesel trucks based on real-world driving. Atmosphere. 2022;13(9):1466.

[11] Azeez O, Pradhan B, Shafri H, Shukla N, Lee CW, Rizeei H. Modeling of CO emissions from traffic vehicles using artificial neural networks. Appl Sci. 2019;9(2):313.

[12] Chadha AS, Shinde Y, Sharma N, De PK. Predicting CO2 emissions by vehicles using machine learning. In: Data Management, Analytics and Innovation [Internet]. Singapore: Springer; 2023 [cited 2024 Mar 4]. p. 197–207. Available from: https://link.springer.com/10.1007/978-981-19-2600-6_14

[13] Shah S, Thakar S, Jain K, Shah B, Dhage S. A comparative study of machine learning and deep learning techniques for prediction of CO2 emission in cars [Internet]. arXiv; 2022 [cited 2023 Nov 1]. Available from: http://arxiv.org/abs/2211.08268

[14] Jia T, Zhang P, Chen B. A microscopic model of vehicle CO2 emissions based on deep learning: A spatiotemporal analysis of taxicabs in Wuhan, China. IEEE Trans Intell Transp Syst. 2022;23(10):18446–18455.

[15] Ghahramani M, Pilla F. Analysis of carbon dioxide emissions from road transport using taxi trips. IEEE Access. 2021;9:98573–98580.

[16] Rivera-Campoverde ND, Muñoz-Sanz JL, Arenas-Ramirez BDV. Estimation of pollutant emissions in real driving conditions based on data from OBD and machine learning. Sensors. 2021;21(19):6344.

[17] Li X, Ren A, Li Q. Exploring patterns of transportation-related CO2 emissions using machine learning methods. Sustainability. 2022;14(8):4588.

[18] Natarajan Y, Wadhwa G, Sri Preethaa KR, Paul A. Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms. Electronics. 2023;12(10):2288.

[19] Abdulmalik R, Srivastava G. Forecasting of transportation-related CO2 emissions in Canada with different machine learning algorithms. Adv Artif Intell Mach Learn. 2023;3(3):1295–1312.

[20] Mangla H, Sharma A. Fuel emission detection using machine learning. Int J Innov Res Technol. 2021;3(6):1–6.

[21] Mądziel M. Liquified petroleum gas-fuelled vehicle CO2 emission modelling based on portable emission measurement system, on-board diagnostics data and gradient-boosting machine learning. Energies. 2023;16(6):2754.

[22] Bappon SD, Dey A, Sabuj SM, Das A. Toward a machine learning approach to predict the CO2 rating of fuel-consuming vehicles in Canada. In: 2022 25th International Conference on Computer and Information Technology (ICCIT). IEEE; 2022. p. 384–389.

[23] Sami O, Elsheikh Y, Almasalha F. The role of data preprocessing techniques in improving machine learning accuracy for predicting coronary heart disease. Int J Adv Comput Sci Appl. 2021;12(6).

[24] Dahouda MK, Joe I. A deep-learned embedding technique for categorical features encoding. IEEE Access. 2021;9:114381–114391.

[25] Kosaraju N, Sankepally SR, Mallikharjuna Rao K. Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models. In: Proceedings of International Conference on Data Science and Applications. Singapore: Springer; 2023. p. 369–382.

[26] Sinsomboonthong S. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalisation methods for artificial neural network classification. Int J Math Math Sci. 2022;2022:1–9.

[27] Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LDF, et al. Clustering algorithms: A comparative approach. PLoS One. 2019;14(1):e0210236.

[28] Lopes HEG, Gosling MDS. Cluster analysis in practice: Dealing with outliers in managerial research. Rev Adm Contemp. 2021;25(1):e200081.

[29] Yigit G, Amasyali MF. Simple but effective GRU variants. In: 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA). IEEE; 2021. p. 1–6.

[30] Hasnain M, Jeong SR, Pasha MF, Ghani I. Performance anomaly detection in web services: An RNN-based approach using dynamic quality of service features. Comput Mater Contin. 2020;64(2):729–752.

[31] Carvalho TP, Soares FAAMN, Vita R, Francisco RDP, Basto JP, Alcalá SGS. A systematic literature review of machine learning methods applied to predictive maintenance. Comput Ind Eng. 2019;137:106024.

[32] Hodson TO. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci Model Dev. 2022;15(14):5481–5490.

[33] Zameer A, Jaffar F, Shahid F, Muneeb M, Khan R, Nasir R. Short-term solar energy forecasting: Integrated computational intelligence of LSTMs and GRU. PLoS One. 2023;18(10):e0285410.

[34] Kvalseth TO. Note on the R^2 measure of goodness of fit for nonlinear models. Bull Psychon Soc. 1983;21(1):79–80.

[35] Haut N, Banzhaf W, Punch B. Correlation versus RMSE loss functions in symbolic regression tasks [Internet]. arXiv; 2022 [cited 2023 Nov 14]. Available from: https://arxiv.org/abs/2205.15990

[36] Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. AJR Am J Roentgenol. 2019;212(1):38–43.