A Novel Approach to Vehicular CO2 Emission Predictive Modelling

Shreejeet Sahay, Pranav M. Pawar, Dipesh N. Sonawane

***Abstract*— Climate change today is a global crisis requiring grave concern to the extent that governments worldwide have realized that if this problem is left unmitigated, catastrophic events may occur, ultimately jeopardizing humanity's survival. Climate change is primarily due to too much presence of greenhouse gases, mainly CO2, in the atmosphere. Vehicular exhaust is one of the main contributors to the emissions of CO2. Although specialized sensors exist for CO2 monitoring, they are inefficient and not highly prevalent. This study suggests a workable, pragmatic, and feasible monitoring system for vehicular CO2 emissions that involves an LSTM network trained and tested based on OBD-II data available in the public domain. Also, this work presents a comparison of the proposed model with a latter-day solution. This proposed system could be deployed on the cloud, with the IoT-based dongles put in the vehicles that can collect in-sensor data from vehicles and send them to the cloud for processing the data, where the deployed model can give real-time predictions of CO2 emissions.**

***Keywords*—Vehicular Emissions, CO2, Climate Crisis, LSTM.**

I. INTRODUCTION

The climate crisis has come to the center stage of international debate, encompassing significant consequences for the health and welfare of the coming generation. Despite all the abnormal influences from natural forces, it is human activity which is the major factor leading to massive climate change over the last century, mainly caused by a sudden spurt in global industrialization. This eventually resulted in the exploitation of natural resources on a large scale. Untimely resolution of these issues could possibly lead to the extinction of our species. Indeed, the tangible impacts of global warming are already evident. These include the melting of glaciers and icecaps, rising seas, heightened CO2 concentration in the atmosphere, tree cutting and desertification, reduced numbers of wildlife species, and water shortages. Moreover, studies suggest that around 3.3 to 3.6 billion people reside in areas with high vulnerability [1]. This highlights the inimical prospective of this occurrence.

Increasingly, greenhouse gas accumulation, particularly carbon dioxide (CO2) serves to spark climate change [2]. See Fig. 1 [3]. In addition, this argument is backed up by empirical facts depicted in Fig. 2 showing global emission of greenhouse gases with respect to different sectors in 2019 till 2022 [4], [22]. There is clear evidence of a significant influence by the transportation industry.

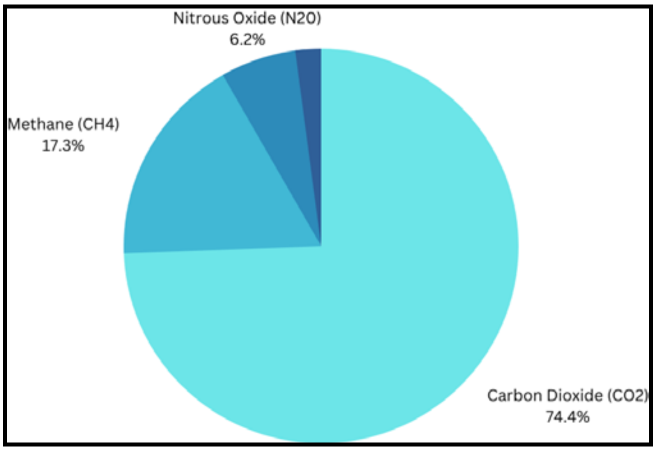


Fig. 1. Global greenhouse emissions by gas.

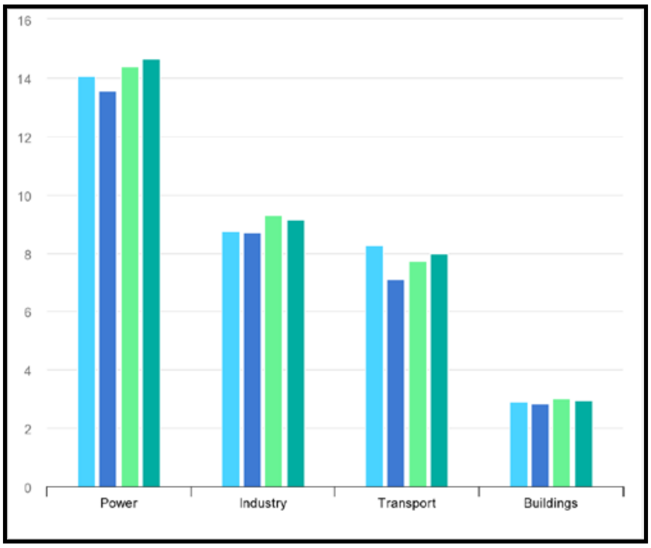


Fig. 2. Global CO2 emissions by sector, 2019-22.

The role of the transport sector in climate change is obvious in India, with about 337 metric tons of CO2 emissions in 2019 through this sector [5], as represented in Fig. 3. Thus, this highlights the transport sector's major contribution to climate change.

Such observation points out the necessity of an extensive monitoring system for automobile CO2 pollution, which is

Manuscript received October 26, 2023.

Shreejeet Sahay is a Data Engineer at Atidiv (India) Private Limited, Pune, Maharashtra, India – 411014 (e-mail: shreejeet.sahay@atidiv.com, sahayshreejeet@gmail.com).

Dr. Pranav M. Pawar is an Assistant Professor, Department of Computer Science at BITS Pilani, Dubai Campus, Dubai International Academic City, P. O. Box No. - 345055, Dubai, UAE. (e-mail: pranav@dubai.bits-pilani.ac.in).

Dipesh Sonawane is a Research Intern at IIT Mandi, Himachal Pradesh, India and is pursuing B.E. in IT at Savitribai Phule Pune University, Pune, Maharashtra, India – 411041 (e-mail: dipeshsonawane2212@gmail.com).

not easy as there is a heterogeneous and huge amount of vehicles that move around the world. The adoption of specialized vehicle emission sensors can be taken into consideration, but in that case, cost and scalability will be a major challenge to deal with. This is where the dilemma arises that propelled the proposed research.

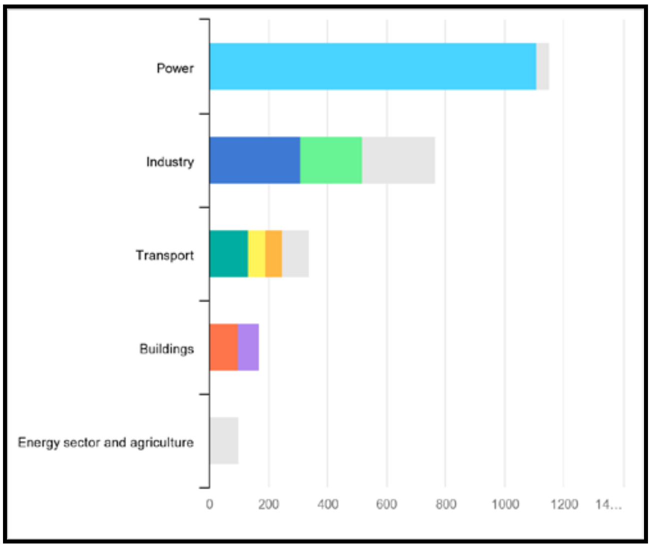


Fig. 3. CO2 Emissions in India by economic sector, 2019.

The existing literature in this regard indicates that numerous machine learning based models on diverse datasets have been successful in estimating CO2 emissions. These techniques are cost-effective and can be replicated. This paper presents a 2-layer Long Short-Term Memory (LSTM) model based on the OBD – II dataset for improving the efficiency of emission monitoring. In this novel approach, a system of dongles for transmitting electronic signal information from an On-Board Diagnostics (OBD) chip [28] to a “cloud”-based solution driven by IoT is proposed. Finally, it gives on-the-spot predictions of anthropogenic carbon. This makes use of modern advances in the field of mobile communications, which makes this easily deployable [6] at a reasonable price.

The proposed model was extensively trained and validated on an open-source dataset extracted from OBD-II data aggregated by P. Rettore et al. [7], [8]. OBD-II provides detailed readings by the sensors within an automobile including Engine Load and Engine RPM, along with other values, and records fault codes for vehicles. Indeed, it is stated that internal combustion engine emission readings correlate directly with sensor readings such as RPM and are therefore useful for inferring a vehicle’s emission characteristics [9] [29]. It is worth noting that machine learning approaches traditionally work on present input-output combinations and do not account for the outcome of previous events when it comes to vehicle telematics data, which is mostly presented in time-series form. Therefore, such methods fail to consider sequential relations among data. This study suggests using a deep learning-based LSTM network model, which is very popular and efficient when forecasting time-series. Moreover, the presented model is compared with the latter-day implementation and delivers better results.

There are five different parts of this paper. Section II discusses the related works studied, which formed part of a thorough literature review for this research. Section III presents an elaborate description of the research methodology used in this study. Section IV summarizes the results obtained and their detailed analysis. Finally, section V explains what this study contributes, recognizes the shortcomings of this study, and suggests directions for further studies.

II. Literature Review

The development in mobile telecommunication technology [6] combined with the continued growth of different aspects of artificial intelligence and related technologies has made deploying data-driven models [22] based on vehicle telematics trouble-free along with ease in access, thus leading to its broad acceptance by the masses. The subsequent paragraphs in this section represent a summary of the relevant literature reviewed during this study.

In [10], light has been thrown upon the the insufficient approximation of CO2 emissions utilization only two features, by training Support Vector Machine (SVM), Artificial Neural Network (ANN), and VT-Micro solutions to predict emissions of CO2, incorporating OBD data such as throttle, speed, and acceleration collected at 30-second intervals to identify environmentally optimal vehicle routes. A linear relationship was formulated in [11] amongst CO2 emissions and vehicle speed and acceleration via the use of regression analysis, inferring that speed exhibited a higher correlation with CO2 emissions compared to acceleration. A blend of OBD-based models and Inductive Loop Detectors (ILDs) is proposed in [12], suggesting the integration of ILD data with vehicle categorization using OBD to prgmatically model CO2 emissions.

Studies have shown that though intensive investigations were made on forecasting CO2 emissions by means of machine learning for environmental purposes, prediction of vehicle exhaust emission still remained untouched area for a significant period of time. For example a linear regression model for projection of CO2 emissions gave RMSE of 0.22 [13]. Moreover, another linear regression model, that was primarily developed to forecast CO2 emissions for various sectors and electric generation attained high R2 accuracy of 0.941 [14]. In addition, a latter-day ensembled method using principal component analysis and support vector machine for NOX prediction was found to be highly effective [15]. Building on this, authors in [16] gathered data from OBD for Maruti Dzire 2019 via an ELM327 microcontroller and Torque app while driving within the city and selected seven features for the training of five conventional machine learning models, thus recommending random forest estimators and decision tree due to their favorable performance over Support Vector Machine, Linear Regression, and K-Nearest Neighbor. A notable drawback of this study was that it had reliance on data collected from a single vehicle, thus limiting its scalability. Furthermore, [17] describes a study where data was collected from one fully hybrid vehicle using a complex Portable Emissions

Measurement System (PEMS) system, that restrained its generalizability to different vehicular makes.

[18] emphasized work suggesting a Boosted and Bagged Decision Trees (BBDT) [22] solution, and it made use of vehicular and environmental data from OBD-II in a light-duty vehicle and a system based on PEMS to model idling emissions of hydrocarbon (HC), nitrogen oxides (NO), carbon monoxide (CO) along with carbon dioxide (CO2). A notable drawback of work in [18] is that the model's training and evaluation were performed only on a single vehicular make data, thus keeping its applicability within bounds. [19] suggested Gradient Boosting Regression (GBR) for CO2 modeling, thereby inferring a relationship between CO2 emissions and vehicle speed that was exponential in nature [22], but the data here was collected in a controlled environment within a laboratory and the solution relied on a single feature.

A major inference from the literature survey above is that the currently available systems for CO2 modeling are unsuitable for a scalable and generalized deployment [22]. Trailblazing systems based upon complex techniques lack scalability despite being accurate. Also, the aforementioned information clearly is in favor of the usage of OBD-II data for predictive modeling of CO2 emissions. Furthermore, platitudinous techniques like ANN, SVM etc cannot work with time-series data, because of their inability to consider sequential interrelation amongst the data points. In such cases, Recurrent Neural Network (RNN)-based approaches [23], [24] such as LSTM [25] ,[26] have shown relatively better performance.

[20] shows a study implying an approach involving the application of a combination of deep learning and machine learning approaches to estimate CO2 emissions in vehicles. [21] proposed a 3-layer LSTM approach [27] outdoing ANN and SVM to estimate CO2 on a time-series OBD-II data. The six features that it used were Speed, Engine RPM, Mileage, Acceleration, Fuel Flow and Throttle. Our work has considered [21] as a foundation, with the 2-layer and six feature LSTM in [22] as our intermediate solution.

III. Research Methodology

This section throws light on this work's methodology of research along with the trends observed in the data in a comprehensive manner. Here, A gives an overview of the proposed solution, B emphasizes the dataset used, C elaborates on the data preparation for the training and evaluation data, D highlights the trends in the data and E explains the LSTM model used for estimation in this study.

A. Proposed Solution Overview

The system architecture of the proposed solution is depicted in Fig. 4. First, raw time-series data are collected from the vehicle’s OBD ports. Then these data points pass through several stages of data preparation so as finally to transform into the supervised learning format for the LSTM model. Then, this information becomes the input of the proposed LSTM solution to foretell the real-time CO2 emissions. The systems depicted here are cloud deployable allowing for constant and real-time prediction and reporting of vehicle’s CO2 emissions via IoT-based dongles.

B. Dataset

The suggested solution’s training and evaluation has been conducted using an openly available dataset collated by Federal University of Minas Gerais Department of Computer Science’s Dr. P. Rettore [22], [7]. The process of collecting the data is described in Table 1 [7]. Here, the data were gathered from 14 drivers and 2 vehicles, with detailed information provided in [8], [30]. Vehicle 1 was driven by ten drivers, while Vehicle 2 was driven by the rest four. Notably, the trips recorded by both vehicles differed: the four drivers of Vehicle 2 were instructed to drive along two distinct routes, while the ten drivers of Vehicle 1 used the vehicle for various ends in their daily routines. To ensure driver privacy, all available data were anonymized, and the start as well as end points of every trip were removed, resulting in a reduced dataset. The CO2 emission readings can be found in the dataset, which are utilized as the true values for training and evaluating the suggested solution.

For simplicity, during the phase of training, we trained the model using data from Vehicle 1 and drivers 1 to 8. We then performed validation of the model using data from Vehicle 1 and drivers 9 and 10. For testing purposes, we evaluated the model on data from Vehicle 2 and driver 11, which were not learnt by the model while it was being trained. This approach ensures that the solution suggested in this paper considers various vehicle types and numbers in its predictions.

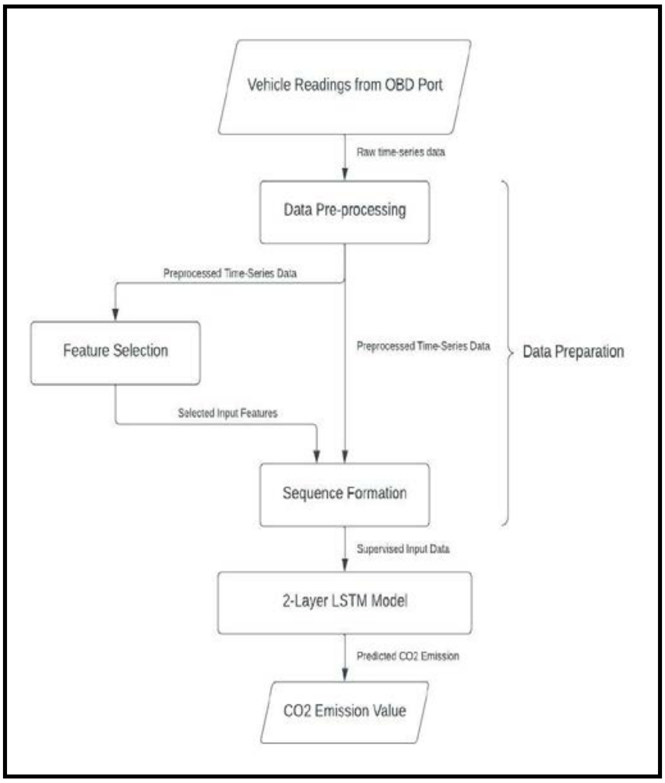


Fig. 4. Flow of the proposed model.

Table 1. Data collection set up

	Vehicle 1	Vehicle 2
Engine	1.0 16v	1.6 16v
Max RPM	7000	7000
Transmissions	5	5
Power	76cv	122cv

Weight	1025 kg	1000 kg
Manufacturer	Renault	Hyundai
Model	Sandero	HB20
Trips	36	8
Trip Time	28 hours	3 hours
Trip Type	Naturalistic	Controlled
Drivers	10	4
Gender	6 M - 4 F	2 M – 2 F
Age	25-61	20-53

C. *Data Preparation*

As depicted in the previous Fig. 4, the data preparation step plays a crucial role in processing and transforming raw time-series data into a format suitable for training and evaluating the LSTM network, specifically in the form of supervised learning data. This step encompasses several processes, including data pre-processing, feature selection, and sequence formation [22], which finally output the supervised learning data, ultimately ensuring that the input data is appropriately processed and structured for the LSTM network. The following paragraphs provide a detailed explanation of each of these processes.

Data Pre-processing. Data pre-processing is a crucial step in the preparation of data for model training and evaluation. In this step, after loading the dataset, the presence of any NaN values was checked, none of which were found. The process of data cleaning was performed that focused on the column of device timestamp. For example, in several instances where the month of September was shortened as 'set' instead of 'Sep,' it was standardized to 'Sep' using the function of replace available in Python's str object. Additionally, the timestamp column's datatype was equalized to datetime64[s] [22].

The column names in the dataset may be tough for noivce to grasp, such as 'OBD_CO2_gkm_Instant,' which represents CO2 emission at an instant in simpler terms. Consequently, the names of the column were modified to make them more comprehensible.

Next, the dataset was divided into two sets [22]: the input set, which contained the input columns, and the output set, which contained the output columns. The input set was then subjected to normalization to confirm that every feature followed a scale that was same. Normalization is essential because features measured on different scales can have unequal contributions to fitting the model, potentially introducing biases. In this study, Min-Max normalization has been utilized for normalizing of data, the mathematical equation for which is-

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where $x_{scaled\ i}$s the normalized feature value [22], for a data point with value x.

Hence, as seen in Fig. 4 earlier, Data Pre-processing takes Raw time-series data and outputs a normalized and cleaned pre-processed time-series data, which is then utilized in feature selection and sequence formation.

Feature Selection. Feature selection is one of the crucial steps of data analysis for high dimensional data because of the presence of irrelevant and redundant data points. It assists in the elimination of such data, thereby improving the solution's performance. Feature selection for this study entailed Principal Component Analysis (PCA) loading scores and correlation values. A biplot of the features using the first two Principal Components (PC) with a percentage variance of 91.21% is shown in Fig. 5. Loading matrix generated through PCA have the weights for numerous features for each of its components. The biplot has x-axis representing PC-1 and y-axis representing PC-2. The vector lengths of all features represented by their corresponding vectors are proportional to their contribution in the PC, ans so it serves as a basis of ranking for feature selection. Vector labels 1 to 6 in Fig. 5 are feature IDs as given in Table II. From the biplot in Fig. 5, it is clearly evident that speed, engine RPM, and mileage contribute the most on PCs. Thus, these 3 features were taken into account for estimating the emissions of CO2 in this study.

Fig. 6 shows the correlation matrix of the feature set, the relational data which can cater to prediction of missing values. The low co-relation scores of the selected attributes or features as seen in the correlation matrix in Fig. 6, clearly imply that they add new information to the set.

To sum up, this step of data preparation selects features from pre-processed time-series data and helps in formulating the optimal feature subset for the model.

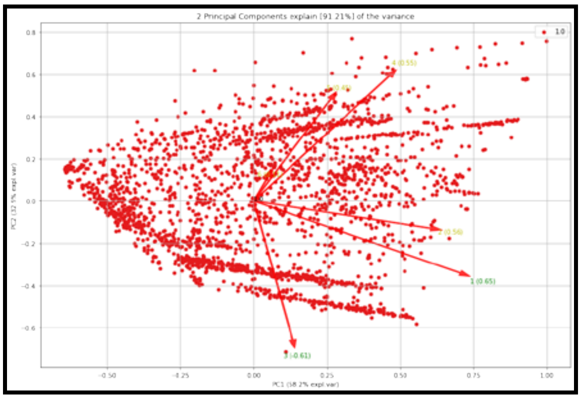
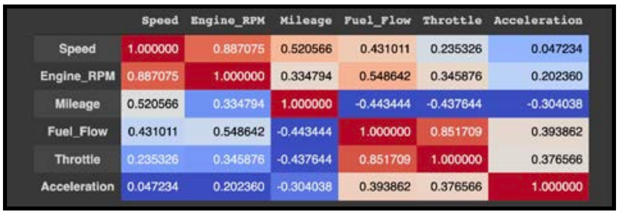


Fig. 5. PCA biplot using first two components.



	Speed	Engine_RPM	Mileage	Fuel_Flow	Throttle	Acceleration
Speed	1.000000	0.887075	0.520566	0.431011	0.235326	0.047234
Engine_RPM	0.887075	1.000000	0.334794	0.548642	0.345876	0.202360
Mileage	0.520566	0.334794	1.000000	-0.443444	-0.437644	-0.304038
Fuel_Flow	0.431011	0.548642	-0.443444	1.000000	0.851709	0.393862
Throttle	0.235326	0.345876	-0.437644	0.851709	1.000000	0.376566
Acceleration	0.047234	0.202360	-0.304038	0.393862	0.376566	1.000000

Fig. 6. Correlation matrix of the feature set.

Table 2. Feature set description

FEATURE	UNITS	RANGE	ID
Speed	*km/h*	0-121	1
Engine RPM	*Revolutions/min*	0-5500	2
Mileage	*km/l*	0-45	3
Fuel Flow	*cc/min*	0-350	4
Throttle	*%*	0-100	5
Acceleration	*km/s^2*	-25-45	6

Sequence Formation. As seen in Fig. 4 earlier, this step uses the pre-processed time-series data and optimal feature subset selected from feature selection process to convert the time-series data to a format suitable for deep learning, i.e., supervised learning format of input and output sequences.

As previously stated during the pre-processing stage, the dataset has been separated into input and output sets, where input set has the three selected input features, i.e., Speed, Engine RPM and Mileage, and output set has one output feature, i.e., instantaneous CO2 emission, or simply put, CO2. Then, min-max normalization is performed on the input set to standardize the input. After this, both input and output sets, being time-series, are converted to supervised learning format, by using sliding window technique, for which a sliding window of breadth as 64 seconds has been used with 50% of overlap, i.e., 1st window has data from 0th till 64th second, 2nd window has data from 32th till 96th second and so on. Fig. 7 depicts the windowing technique used here. This window width of 64 seconds has been used similar to [21] where different window widths were used for the validation set, and 64 gave the least mean squared error. Also, mean of every window is found and is used for the output set.

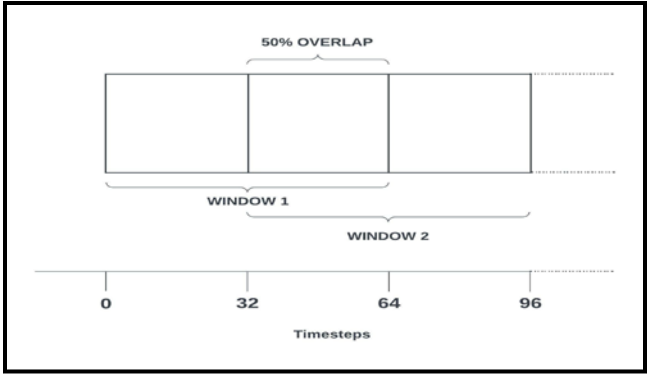


Fig. 7. Sliding window with 50% overlap.

As a result of this step, the input set has 64*3 columns and output set has 1 single column for CO2 emission output. We have performed this step for all training, validation and test sets as mentioned earlier.

D. Trends in the Data

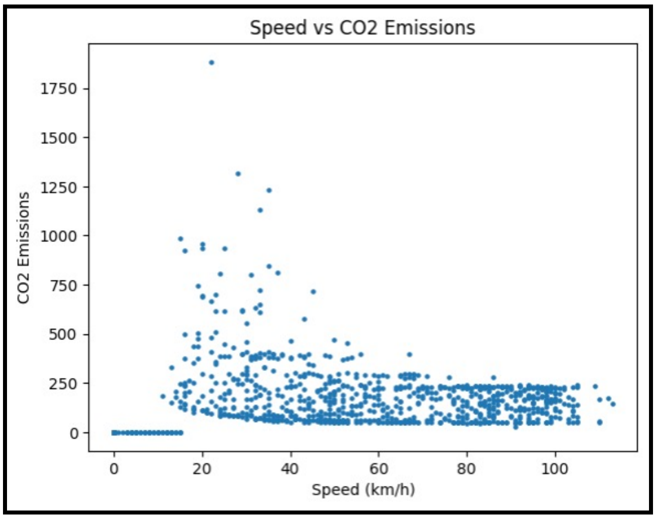


Fig. 8. Speed vs CO2.

Speed vs CO2. See Fig. 8. The “Speed vs. CO2” graph offers invaluable information concerning the correlation between the speed of a vehicle and its related carbon dioxide (CO2) discharge. Most of the data points are all within the speed range of 10 to 100 mph and the CO2 emission range of 100 – 300 g/km, consistent with what one could observe during a normal driving situation.

Typically, it is seen that as the vehicle speed goes upwards, so does the amount of carbon dioxide emitted. This accords with a common perception that more speed would translate to increased usage of energy products such as fuel, which in turn causes the release of more emissions. However, it should be noted that these different data points occur under other influencing factors like engine efficiency, truck type, and road conditions.

However, there is significant data inconsistency while analyzing the sub-segment readings at speeds between 20 and 40mph, as these particular data points diverge from the overall trends. Further research on these high emissions at low speeds is needed, given that some specific driving conditions or engine performance problems might strongly influence air pollution.

It is vital to comprehend how CO2 emissions change according to vehicle speed as it assists in advancing environmental sustainability. This entails that speed control, while moving slowly, might be an essential element leading to reducing carbon emissions. Additional research should be conducted to determine the exact contributors to this observation and what remedial measures are applicable and practical during real-world driving conditions

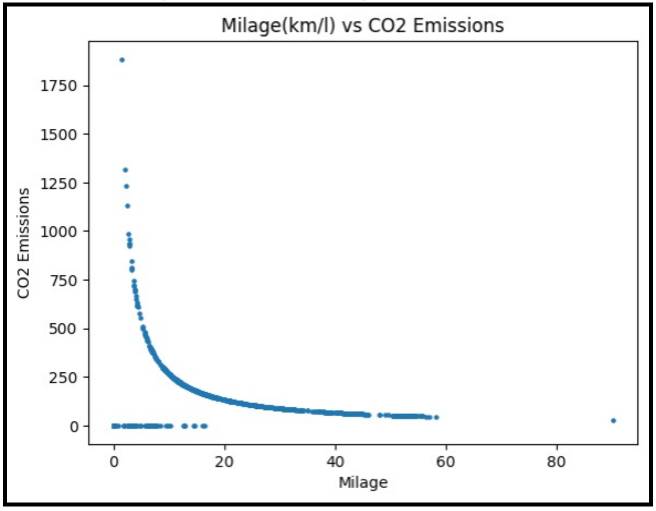


Fig. 9. Mieage vs CO2.

Mileage vs CO2. See Fig. 9. There is an impressive trend in the correlation between a vehicle's carbon dioxide (CO2) emissions and its mileage. This chart depicts different points on different mileage values and corresponding CO2 emissions values.

A steep reduction in CO2 emissions characterizes this trend as the mileage increases. Vehicles that travel farther on each fuel gallon have much lower CO2 emissions. The reverse correlation reflects fuel economy's critical role in improving transportation's environmental impact. The more miles vehicles put on the road, the faster CO2 levels decrease.

A detailed analysis of the graph's behavior reveals that the reduction in CO2 emission begins from high levels around 1250 g/km. When the mileage exceeds a specific limit, usually 35 miles, the CO2 emission reduces following the exponential decay formula to about 150 g per km. Essentially, CO2 emissions at the stage after this crucial distance are more or less constant, and additional miles do not affect the amount of emissions.

Understanding this trend is essential in developing effective means to improve the efficiency of vehicles in the environment. It is essential for fuel efficiency and emphasizes technology, which allows an exponential reduction in CO2 emission as the mile increases. It also highlights the importance of defining the corresponding mileage limits once emissions attain an equilibrium, guiding suitable emission management plans.

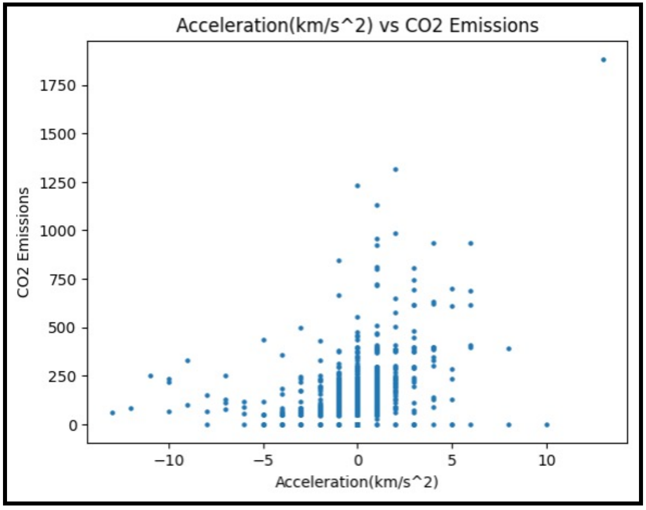


Fig. 10. Acceleration vs CO2.

Acceleration vs CO2. See Fig. 10. "Acceleration VS CO2" is one of the most critical graphs since it offers an extremely crucial analysis of a highly complex situation where the acceleration of vehicles is related to their CO2 emissions. The graph above shows data scattered among various acceleration values on the x-axis and CO2 emissions on the y-axis.

One notable feature of this trend is the dense data points within the acceleration range from −5 to 5 kilometers per second squared (km/s2). The vehicles have different acceleration patterns in this range that are based on actual driving situations and how drivers behave.

However, a striking feature on the graph is the different CO2 emissions across various acceleration rates. The data points' values reveal a significant rise in CO2 amount, even up to 1250 g/km, for acceleration values varying between 2 and 3 km/s^2. This indicates that fast and intense acceleration can lead to significant fuel consumption, which implies high carbon dioxide emissions.

Nevertheless, the existing pattern goes even further. CO2 emission is almost constant above and below 600g/km within the 2-3 km/s2 range, approximately 250 g/km. This gives the impression that the car emitted the same amount throughout this period, and it is probably the result of factors like the engine at its optimal stage performance during moderate acceleration, driving conditions, and the nature of the machine.

In addition, non-linearity behavior complicates the relationship between acceleration and CO2 emissions. Increased emissions typically arise from higher acceleration but not always in the exact degree. These are not the only additional factors because, at certain acceleration levels, the growth of emissions exceeds that expected for the reasons mentioned above.

RPM vs CO2. See Fig. 11. The "RPM vs. CO2" graph mirrors the intricate relationship between RPM, which stands for revolutions per minute, and CO2. In this graph, data points are mainly found between the RPM values of 500 and 3500 on the x–axis, thereby giving an overview of how emissions are affected by engine speed.

A significant feature of this trend is the well-defined peak in CO2 emissions between the revolutions per minute of 2000 and 2500. Vehicles have the highest CO2 emissions within the RPM range, ranging between 1000 and 1250 g/km. The highest acceleration peak is 2-3 km/s^2, which shows that emissions are strongly linked to engine speed and acceleration.

Additionally, beyond the 2000-2500 RPM, CO2 emissions are maintained within the 250 g/km region. It means that the same pattern of emissions exists at different engine speeds, which proves the efficiency and fuel consumption of the engine operation under various conditions.

This emphasizes the idea of non-linearity in the relationship between engine speed, acceleration, and CO2 emissions. The specific range of RPMs (2000 – 2500) produces much more air pollutants than other RPMs, though other RPMs usually generate more emissions generally. These may include engine design, power output, and optimizing fuel efficiency.

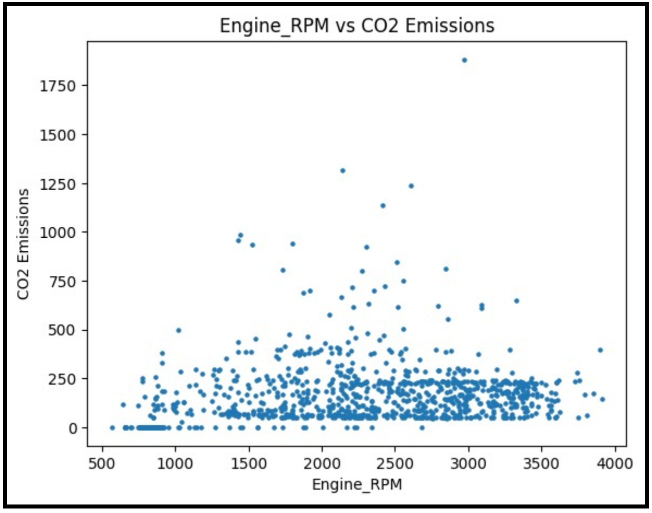


Fig. 11. Engine RPM vs CO2.

Acceleration and RPM Redundancy. The "Acceleration vs. CO2" and "RPM vs. CO2" graphs reflect an intriguing commonality in the behavior of two distinct driving parameters, acceleration (expressed in kilometers per second squared, km/s^2) and engine rpm (measured in revolutions per minute, RPM), concerning their influence on carbon dioxide (CO2) emissions. Both parameters exhibit a non-linear relationship with CO2 emissions, characterized by concentration within specific ranges and peak emissions at particular values.

In the "Acceleration vs. CO2" graph, higher acceleration levels correspond to significantly elevated CO2 emissions, with peak emissions occurring in the 2-3 km/s^2 range. Similarly, the "RPM vs. CO2" graph reveals that within the 2000-2500 RPM range, CO2 emissions reach their zenith. This overlapping pattern between acceleration and RPM indicates that these two variables are proportional and that changes in one can directly affect the other.

Considering these shared characteristics and the high correlation between acceleration and engine speed in the context of CO2 emissions, it becomes evident that either acceleration or RPM can be a representative feature for evaluating and predicting emissions. As either of them could be considered redundant, we have used the results of the PCA to determine the redundant feature, which is acceleration in this case.

E. Proposed LSTM Model

In the proposed methodology, this particular step utilizes the data prepared in the Data Preparation phase for the suggested solution’s training and evaluation. Since the sample of input are in sequential format, conventional techniques of machine learning such as SVM or standard feed-forward neural networks are not suitable for forecasting. Therefore, this study introduces the utilization of LSTM networks, which have specialized connections that facilitate feedback and enable efficient processing of sequential inputs. As discussed earlier in the terminology section of this paper, LSTMs outperform RNNs as they possess the risk of vanishing gradients. LSTMs are especially advantageous for time-series data as they are resilient to duration gaps in the data that are not known, which sets them apart from RNNs and traditional machine learning techniques [21].

LSTM needs input in a 3D shape and not a 2D shape. The 2D shape format refers to the shape of the data as [Number of Rows (r), Number of Columns (c)]. The 3D shape format refers to the shape of the data as [Number of Rows (r), Timestep of the window (t), Number of Features (f)]. It is evident that-

$$c = t \, X \, f \tag{2}$$

The dataset generated by the data preparation step initially exists in a 2D format. However, since this data is sequential in nature, it needs to be transformed into a 3D format to provide essential details regarding the number of features and timesteps involved. This 3D shape is crucial for processing tasks such as training and evaluating sequential data. To achieve this, the set of inputs is reshaped to 3D format from their 2D format using the reshape function in Python. This reshaping process is applied to all the training, validation, and testing input sets.

Hyperparameters like the number of layers, nodes, and others directly influence the complexity as well as performance of the model. The optimum settings regarding the hyperparameters of the suggested LSTM were determined via conducting a random search.

In this study, the suggested optimal architecture for the neural network is a sequential 2-layer LSTM model. The first layer of the LSTM comprises 90 neurons, while the second layer consists of 180 neurons. The model is trained over 120 epochs, utilizing loss function in the form of “mean squared error”, that can be understood from the following equation-

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

where N is the number of values, y_i is the actual value and $\hat{y}_i$ is the predicted value of a variable [22]. Put simply, MSE calculates the average of the squared differences between the predicted and actual values of a variable.

For training, the model utilizes the RMS Prop Optimizer. RMSProp has similarities with algorithm of gradient descent with momentum and limits the oscillations vertically, thereby allowing to increase the learning rate and proposed algorithm could take larger steps in the horizontal direction converging faster. RMSProp's core principle involves maintaining a moving average of the squared gradients and dividing the gradient by the square root of this average. By employing this optimizer, the gradients’s moving average is leveraged for the variance estimation.

Table 3 below shows the algorithm for the proposed LSTM model. It shows the requirements of this algorithm, followed by its steps and the output.

Table 3. Proposed Algorithm

ALGORITHM ALGORITHM FOR THE PROPOSED LSTM MODEL

Require: Batch size=32, Epochs=120, No of input features=3, No of timesteps=64, No of output features=1

```
1. Define Sequential Model
2. model.add(LSTM(90,return_sequences=True,input_shape=(64,3)))
3. model.add(LSTM(180))
4. model.add(Dense(1))
5. model.compile(loss=’mse’,optimizer=’rmsprop’)
6. history=model.fit()
7. model.predict()
8. rmse=sqrt(mean_squared_error(y_predicted,y))
```

Output: Training loss, Validation loss, Evaluation RMSE

IV. RESULTS

In this section, we analyze the outcomes of observations in this study. The methodology we developed was executed in a Google Colab Pro environment, utilizing Python 3 and leveraging the hardware accelerator of Graphics Processing Unit (GPU).

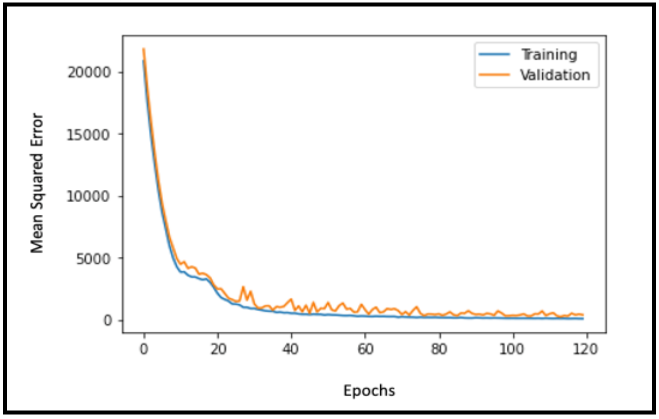


Fig. 12. Training Loss (Mean Squared Error).

The evaluation of the results utilized two metrics: MSE and RMSE. MSE represents Mean Squared Error (Equation (3)), while RMSE denotes Root Mean Squared Error [22]. The concept of MSE can be found in section IIIE. RMSE, on the other hand, is the square root of the average of the squared differences between the predicted and the actual values of a variable. Mathematically, the calculation of RMSE can be expressed as follows-

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{4}$$

where N is the number of values, y_i is the actual value and $\hat{y}_i$ is the predicted value of a variable [22].

The obtained MSE loss curve during the training of the suggested solution is demonstrated in Fig. 12. As previously stated in section IIIB, the model's evaluation is conducted using the data from vehicle 2, which was not utilized during the training phase. This approach ensures the model's robustness and effectiveness, as well as its independence from the specific type and make of the vehicle.

When assessed on the test dataset, the proposed model demonstrated promising results with an MSE of 63.52 and an RMSE of 7.97. Notably, these evaluations were conducted on data from a vehicle that the model had never encountered before. This outcome suggests that the suggested solution is scalable and so it can be employed on a wide range of vehicles, regardless of their type or make. It signifies that a single model is proficient and effective in predicting CO2 emissions using general vehicle characteristics such as Speed, Engine RPM, and Mileage.

The performance of the suggested 2-layer LSTM solution is contrasted with the latter-day implementation [21], as well as a 3-layer Deep Convolutional Neural Network (CNN) and 4-layer DNN models [21] [22]. Table IV presents a comparative analysis of the performance of all four models. It is evident from the table that this study's suggested solution, exhibits strong performance with 7.97 as RMSE.

When assessed on the test dataset, the proposed model demonstrated promising results with an MSE of 63.52 and an RMSE of 7.97. Notably, these evaluations were conducted on data from a vehicle that the model had never encountered before. This outcome suggests that the suggested solution is scalable and so it can be employed on a wide range of vehicles, regardless of their type or make. It signifies that a single model is proficient and effective in predicting CO2 emissions using general vehicle characteristics such as Speed, Engine RPM, and Mileage.

The performance of the suggested 2-layer LSTM solution is contrasted with the latter-day implementation [21], as well as a 3-layer Deep Convolutional Neural Network (CNN) and 4-layer DNN models [21] [22]. Table 4 presents a comparative analysis of the performance of all four models. It is evident from the table that this study's suggested solution, exhibits strong performance with 7.97 as RMSE.

Table 5 shows some of the key differences between the 2-layer LSTM solution proposed in this work, and latter-day solution [21].

Table 4. Comparison of various models of deep learning

MODEL	RMSE
4-layer DNN	64.87
3-layer Deep CNN	17.82
3-layer LSTM	9.30
Proposed Model	7.97

Table 5. Comparison of Latter-Day Solution with Suggested Solution

	LATTER-DAY SOLUTION	SUGGESTED SOLUTION
i.	Utilizes 6 input features- Speed, Acceleration, Throttle, Mileage, Fuel Flow, and Engine RPM.	Utilizes 3 input features- Speed, Engine RPM and Mileage.
ii.	3 layers having 120, 240 and 500 neurons respectively.	2 layers having 90 and 120 neurons respectively.
iii.	RMSE achieved- 9.30	RMSE achieved- 7.97

V. CONCLUSION AND FUTURE SCOPE

This paper depicts a 2-Layer LSTM network-based solution that forecasts CO2 emissions from vehicles, employing three key features from OBD-II data: Speed, Engine RPM, and instantaneous Mileage. The suggested system can be implemented on a cloud platform with IoT dongles inside vehicles. This work emphasizes the robustness and efficiency of the suggested model, backed by strong performance metrics, and tackling existing solutions. It adapts to varying conditions, like road and traffic variations, that may influence OBD-II readings and so CO2 emission predictions.

However, a constraint of this study is the dependence on a limited OBD-II dataset from only two vehicles. Expanding the dataset to inculcate diverse vehicle types and sizes is crucial for greater accuracy and applicability, and thus forms basis for the future scope of this work. This expansion will, in turn, enhance predictions across different vehicle categories, thus strengthening validity, and improving practical utility for policymakers and environmentally conscious individuals. Furthermore, it will support identifying potential biases and constraints specific to certain vehicle types, allowing for model refinements for better and broader applicability and reliability.

REFERENCES

[1] IPCC, 2022: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. In Press.

[2] Susan Solomon, Gian-Kasper Plattner, Reto Knutti, and Pierre Friedlingstein, "Irreversible climate change due to carbon dioxide

emissions," Proceedings of the National Academy of Sciences, vol. 106, no. 6, pp. 1704-1709, 2009, doi: 10.1073/pnas.0812721106.

[3] Hannah Ritchie, Pablo Rosado and Max Roser, "Greenhouse gas emissions," Published online at OurWorldInData.org, 2020, https://ourworldindata.org/greenhouse-gas-emissions.

[4] IEA, Global CO2 emissions by sector, 2019-2022, IEA, Paris https://www.iea.org/data-and-statistics/charts/global-co2-emissions-by-sector-2019-2022, IEA. Licence: CC BY 4.0.

[5] IEA, CO2 emissions from the Indian energy sector, 2019, IEA, Paris https://www.iea.org/data-and-statistics/charts/co2-emissions-from-the-indian-energy-sector-2019.

[6] D. Bandara, M. Amarasinghe, S. Kottegoda, A.L. Arachchi, S. Muramudalige, and A. Azeez, "Cloudbased driver monitoring and vehicle diagnostic with obd2 telematics," volume 6, 2015.

[7] Vehicular data trace of the city of belo horizonte and surroundings, brazil. 2018. http://www.rettore.com.br/prof/vehicular-trace/.

[8] Paulo H.L. Rettore, Andre B. Campolina, Leandro A. Villas, and Antonio A.F. Loureiro, "Identifying relationships in vehicular sensor data: A case study and characterization," DIVANet '16, page 33–40, New York, NY, USA, 2016. Association for Computing Machinery.

[9] Fernando Ortenzi and Maria Costagliola, "A new method to calculate instantaneous vehicle emissions using obd data," SAE Technical Papers, 04 2010.

[10] Weiliang Zeng, Tomio Miwa, and Takayuki Morikawa, "Prediction of vehicle co2 emission and its application to eco-routing navigation," Transportation Research Part C: Emerging Technologies, 68:194 – 214, 2016.

[11] Seth Oduro, Santanu Metia, Hiep Duc, and Quang Ha, "CO2 vehicular emission statistical analysis with instantaneous speed and acceleration as predictor variables," pp. 158–163, 11 2013.

[12] Matt Grote, Ian Williams, John Preston, and Simon Kemp, "A practical model for predicting road traffic carbon dioxide emissions using inductive loop detector data," Transportation Research Part D: Transport and Environment, vol. 63, pp. 809 – 825, 2018.

[13] P. Kadam and S. Vijayumar, "Prediction Model: CO 2 Emission Using Machine Learning," 2018, 3rd Int. Conf. Converg. Technol. I2CT 2018, pp. 1–3, 2018, doi: 10.1109/I2CT.2018.8529498.

[14] T. C. Ho, S. C. Keat, M. Z. M. Jafri, and L. H. San, "A prediction model for CO2 emission from manufacturing industry and construction in Malaysia," Int. Conf. Sp. Sci. Commun. Iconsp., vol. 2015-Septe, pp. 469–472, 2015, doi:10.1109/IconSpace.2015.7283771

[15] B. Liu, J. Hu, F. Yan, R. F. Turkson, and F. Lin, "A novel optimal support vector machine ensemble model for NOX emissions prediction of a diesel engine," Meas. J. Int. Meas. Confed., vol. 92, no. X, pp. 183–192, 2016, doi: 10.1016/j.measurement.2016.06.015.

[16] S. Kangralkar and R. Khanai, "Machine Learning Application for Automotive Emission Prediction," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-5, doi: 10.1109/I2CT51068.2021.9418152

[17] M. Mądziel, A. Jaworski, H. Kuszewski, P. Woś, T. Campisi, and K. Lew, "The Development of CO2 Instantaneous Emission Model of Full Hybrid Vehicle with the Use of Machine Learning Techniques," Energies, vol. 15, no. 1, p. 142, Dec. 2021, doi: 10.3390/en15010142

[18] Li Q, Qiao F, Yu L, "A Machine Learning Approach for Light-Duty Vehicle Idling Emission Estimation Based on Real Driving and Environmental Information," Environ Pollut Climate Change, 2016, 1, p. 106.

[19] N. Subramaniam, and N. Yusof, "Modelling of CO2 Emission Prediction for Dynamic Vehicle Travel Behavior Using Ensemble Machine Learning Technique," 2021 IEEE 19th Student Conference on Research and Development (SCOReD), 2021, pp. 383-387, doi: 10.1109/SCOReD53546.2021.9652757.

[20] Shah Samveg, Thakar Shubham, Jain, Kashish, Shah Bhavya, and Dhage Sudhir, "A Comparative Study of Machine Learning and Deep Learning Techniques for Prediction of Co2 Emission in Cars," 2022, doi: 10.48550/arXiv.2211.08268.

[21] M. Singh, and R. Dubey, "Deep Learning Model Based CO2 Emissions Prediction using Vehicle Telematics Sensors Data," in IEEE Transactions on Intelligent Vehicles, 2021, doi: 10.1109/TIV.2021.3102400.

[22] S. Sahay, and P. Pawar, "An Optimal Approach to Vehicular CO2 Emissions Prediction using Deep Learning," 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2023, pp. 1-5, doi: 10.1109/ESCI56872.2023.10099940.

[23] Thomas G. Dietterich, "Machine learning for sequential data: A review," in Terry Caelli, Adnan Amin, Robert P. W. Duin, Dick de Ridder, and Mohamed Kamel, editors, Structural, Syntactic, and Statistical Pattern Recognition, pp. 15–30, Berlin, Heidelberg, Springer, 2002.

[24] Y. Wang, Y. Liu, M. Wang, and R. Liu, "LSTM Model Optimization on Stock Price Forecasting," 2018, 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2018, pp. 173-177, doi: 10.1109/DCABES.2018.00052.

[25] Sepp Hochreiter, and Jürgen Schmidhuber; "Long Short-Term Memory," Neural Comput, 1997, vol. 9(8), pp. 1735–1780, doi: https://doi.org/10.1162/neco.1997.9.8.1735

[26] J. Xiang, Z. Qiu, Q. Hao, and H. Cao, "Multi-time scale wind speed prediction based on wt-bi-lstm," in MATEC Web of Conferences, vol. 309. EDP Sciences, 2020, p. 05011.

[27] K. Moharm, M. Eltahan and E. Elsaadany, "Wind Speed Forecast using LSTM and Bi-LSTM Algorithms over Gabal El-Zayt Wind Farm," 2020 International Conference on Smart Grids and Energy Systems (SGES), 2020, pp. 922-927, doi: 10.1109/SGES51519.2020.00169.

[28] United States Environment Protection Agency (EPA). Vehicle emissions on-board diagnostics (obd). 2020. https://www.epa.gov/state-and-local-transportation/vehicle-emissions-board-diagnostics-obd.

[29] ISO. Open diagnostic data exchange (odx). OBD-II Exchange.

[30] P. H. L. Rettore, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, "A method of eco-driving based on intra-vehicular sensor data," in 2017, IEEE Symposium on Computers and Communications (ISCC), pp. 1122–1127, 2017.

Shreejeet Sahay completed his B.E. in IT from Savitribai Phule Pune University (formerly University of Pune) in 2019. During his academic journey of undergraduation, he became the recipient of several accolades like Academic Excellence Award, Best Outgoing Student Award, etc awarded by his college, S.K.N. College of Engineering and he was also awarded a rank certificate for securing position amongst toppers, both in his branch of IT as well as all branches, by Savitribai Phule Pune University. He was also awarded Elite certificate by NPTEL for securing position amongst Top 5% in the course "Programming, Data Structures and Algorithms in Python" all over India.

Sahay began his professional journey at Informatica in 2019, and received Most Valuable Player (MVP) within 1 year of the inception of his industrial stint. He has been working as a Data Engineer at Atidiv (India) Private Limited, Pune, Maharashtra, India since July 2022. He has also taken active participation in research, with a total of 4 research papers in his name (2 published, and 2 presented and waiting to be published), prior to this paper, and in addition, he was invited as a reviewer for the International Conference on Computational Intelligence and Network Systems CINS 2023 held at BITS Pilani, Dubai Campus in October, 2023. Furthermore, he frequently visits some of the prestigious institutions like Vishwakarma Institute of Technology, Pune as an external examiner.

Pranav M. Pawar graduated in Computer Engineering from Dr. Babasaheb Ambedkar Technological University, Maharashtra, India, in 2005, received Master in Computer Engineering from Pune University, in 2007 and received PhD in Wireless Communication from Aalborg University, Denmark in 2016, his PhD thesis received nomination for Best Thesis Award from Aalborg University, Denmark. From 2006 to 2007, was working as System Executive in POS-IPC, Pune, India. From Jan 2008 to June 2018, he worked as an Associate Professor in Department of Information Technology, STES's Smt. Kashibai Navale College of Engineering, Pune. He also worked as Associate Professor at MIT School of Engineering, MIT-ADT University, Pune during June 2018 to September 2019. He received Recognition from Infosys Technologies Ltd. for contribution in Campus Connect Program received different funding for research and attending conferences at international level.

From March 2019 to October 2020 he was working as full time Post-doctoral Researcher in Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel in an area of wireless communication and deep learning. He received outstanding postdoctoral fellowship from Israel Planning and

Budgeting Committee (PBC) and Israel Science Foundation. Currently from October 2020 he is working as an Assistant Professor in Department of Computer Science, BITS Pilani, Dubai Campus, Dubai, UAE. He published more than 40 papers at national and international level. He is IBM DB2 and IBM RAD certified professional and completed NPTEL certification in different subjects. His research interests are resource allocation and QoS in wireless networks, machine and deep learning, wireless security and bioinformatics.

Dipesh Nemichand Sonawane is a dedicated and ambitious student pursuing a Bachelor of Engineering in Information Technology with Honors in Data Science and Visualization at Savitribai Pune University (formerly University of Pune). Currently in his fourth year of undergraduate studies, Dipesh has demonstrated a profound passion for the field of computer science, particularly in the realms of deep learning and machine learning. His commitment to academic excellence is exemplified through his role as a research intern at the esteemed Indian Institute of Technology, Mandi, where he specializes in Computer Vision. Moreover, Dipesh's multifaceted talents extend to the field of aerospace engineering, as he serves as an Avionics Engineer and Operations Lead within the STES rocketry team, showcasing his diverse skill set and dedication to innovation. With a strong academic foundation and a deep-rooted interest in cutting-edge technologies, Dipesh is a promising young researcher who continues to make significant contributions to the fields of computer science and engineering.