Modelling of CO2 Emission Prediction for Dynamic Vehicle Travel Behavior Using Ensemble Machine Learning Technique

Navarajan Subramaniam
Faculty of Built Environment & Surveying
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
Navarajan16@yahoo.com

Norhakim Yusof
Faculty of Built Environment & Surveying
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
norhakim@utm.my

2021 IEEE 19th Student Conference on Research and Development (SCOReD) | 978-1-6654-0193-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/SCOReD53546.2021.9652757

***Abstract*—Urban growth in most developing countries mainly results from vast economic development. As, consequences, capital cities have become the center of many activities. A large amount of population become permanently resides in capital cities thereby raising a need for living space, social activity areas as well as transportation. One of the major challenges in urbanizing cities is poor air quality due to transportation emission particularly CO2 from vehicles. Continuous CO2 emission could lead to irreversible air pollution which causes a significant negative impact on the environment and human health. To date, most studies have employed a specific emission factor to estimate CO2 emission from vehicles. However, the emission factor varies based on vehicle type and climate. Therefore, this study aims to develop a vehicle travel CO2 model using the ensemble technique by incorporating with large volume of data collected from laboratory. The advantage of this study may assist the urban transportation planner to design smart transportation planning that enables them to respond to the current carbon footprint map.**

Keywords—Ensemble machine learning technique, CO2 emission, Big Data,

I. INTRODUCTION

Road transport is a significant source of air pollutants which brings harmful effects to health and the environment. Greenhouse gas emissions from road transport such as carbon dioxide (CO2) and nitrogen oxides (NOx) affect the regional and local air quality [1-7]. This is because the transportation sector plays a vital role in driving the national economy and improving the livelihood of society. Consistently CO2 discharge could prompt irreversible greenhouse gas pollution (GHG), which causes enormous adverse effects on the climate and human wellbeing. In the urban area, passenger vehicle (i.e., personal car and taxi) is commonly used to travel within the city areas. To date, various efforts towards estimating CO2 emission have been studied by considering different emission factors. However, these studies have improved the estimated emissions but cannot generalize well with different types of vehicles and the country's climate. A recent study has shown that CO2 emission prediction could benefit from the advance of machine learning (ML) approaches [8]. ML can learn complex and non-linear relationships that are difficult to model using other techniques such as standard statistics [9]. Statistics are commonly used to collect and interpret data to uncover patterns and trends. However, these techniques will have difficulties predicting CO2 emission in a complex urban environment due to the data inconsistencies and non-linearity [4, 6]. Because of the complexity of the problem, this study will aim to develop a vehicle travel CO2 model using the ensemble technique by incorporating the large volume of data collected from laboratory test using standard driving cycle pattern. Using this new fundamental the discovered patterns can be used to represent the vehicles' activities in road networks for uncovering the dynamic vehicle behaviour in urban area.

This paper is arranged as follows: Section II Analysis of data, Section III CO2 Prediction Model, Section IV Results and Discussion, and finally, the paper concludes in Section V.

II. ANALYSIS OF DATA

A. Data Description

The data for this research is obtained from the conducted test in dynamometer from laboratory to measure the fuel consumption and tailpipe emission of two passenger cars by utilizing New European Driving Cycle (NEDC) standard emission. Governing authorities frequently employ drive cycles in regulations to manage emissions and fuel usage. Vehicle manufacturers and automotive suppliers also utilize drive cycles to evaluate their products. Local governments in Malaysia continue to use NEDC for legislative objectives and local manufacturers and suppliers for evaluation. [10]. The measurements were carried out with a modern passenger car propelled with a conventional internal combustion engine. The tests were conducted at an emission test facility that included a climatic chamber, a chassis dynamometer, and an emission monitoring system that met European legislative standards for passenger car type approval [11]. The 1.3 cc engine was used which is complying with the Euro 5 emission standards because technically, 1.3 cc engines consume lower fuel and reduce CO2 emission. However, the driving behavior in Malaysia causes the pattern of CO2 emission irregularly. Besides, urban areas are a pioneer of these types of vehicles. As a result, we chose a 1.3cc motor vehicle as the subject of our research and the methodology for measuring exhaust emissions. The present legislative procedure for type approval of passenger cars was used to calculate fuel consumption. [10].

B. *Data Analysis*

Based on the analysis, Fig 1 shows that the higher distribution of CO2 emission occurs between 500 ppm to 10000 ppm. This is because when the vehicle travels at a lower speed or keeps idling on the road causes the CO2 emission to be emitted in high proportion.

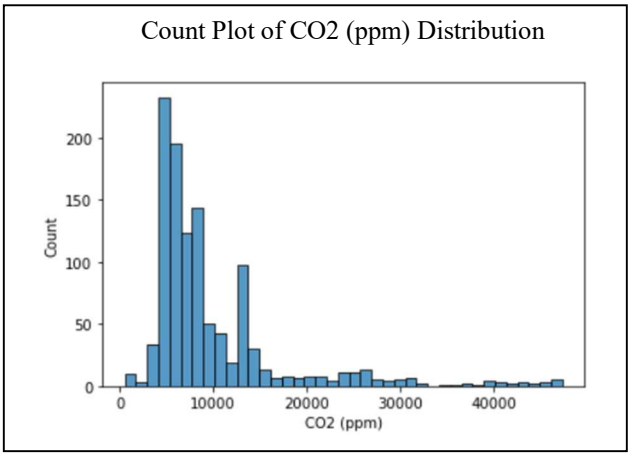

Fig. 1. CO2 distribution based on 1.3cc engine

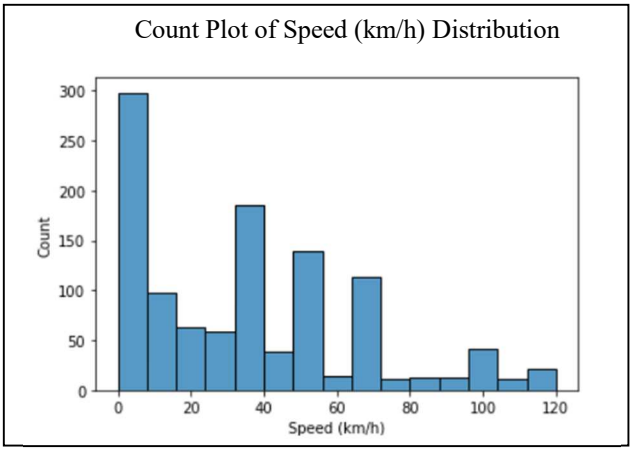

Fig. 2. Vehicle speed distribution ranging from 0 to 120 (km/h)

As shown in Fig. 2, low vehicle speed is more dominant than the higher speed for this laboratory test. The majority of the vehicle speed falls in the range of 0 to 70 km/h, whereas the remaining speed falls in the range of 80 to 120 km/h. The reason for this vehicle speed division is because low speed represents the urban driving condition at the main road, roundabout, and traffic light, whereas high speed represents the driving condition at the expressway. Using these two pieces of information, we can observe that the CO2 emission is highly influenced by the car speed, as depicted in Fig 3. Fig 3 shows the relationship between vehicle speed and CO2 emission.

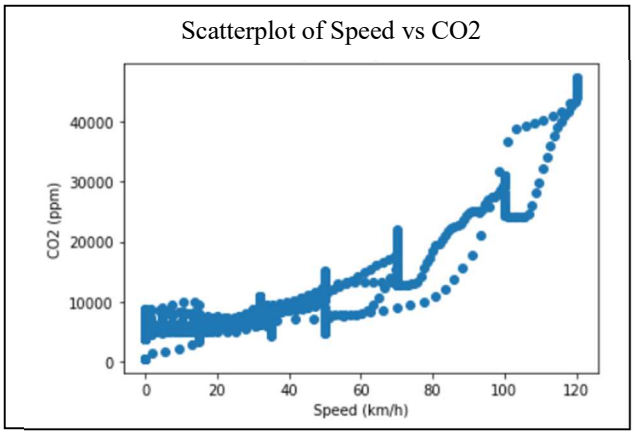

Fig. 3. Relationship between speed and CO2 emission

The scatter plot in Fig 3 explains that the CO2 emission increases exponentially when the speed of the vehicle increases. Therefore, vehicle speed becomes the input for the model to predict the CO2 emission in the urban area.

III. CO2 PREDICTION MODEL

A. *Ensemble Technique*

In this study, the Gradient Boosting Regression (GBR) algorithm is used to develop the CO2 prediction model for the urban area. GBR is an ensemble technique in the machine learning paradigm, in which numerous models, frequently referred to as "weak learners," are trained to tackle the same issue and then aggregated to get better results [12-15]. The core idea is that by correctly combining weak models, the model becomes more robust and capable of accurately predicting. Moreover, the idea of the ensemble method is to reduce bias and variance from the weak learners. Most of the time, these basic models perform poorly on their own, either due to a strong bias (low degree of freedom models) or too much variance to be robust (high degree of freedom models). Fig. 4 below shows the algorithm for GBR [13].

Algorithm:

1. $F_0(x) = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, \rho)$

2. For m = 1 to M do:

3. $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]$

4. $\boldsymbol{a}_m = \arg\min_{\boldsymbol{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(x_i; a)]^2$

5. $\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L[\tilde{y}_i - F_{m-1}(x_i) + \rho h(x_i; a_m))$

6. $F_m(x) = F_{m-1}(x) + \rho_m h(x; \boldsymbol{a}_m)$

Fig. 4. Gradient Boosting Algorithm

B. Model Development

Fig 5 shows the flowchart of the CO2 prediction model development.

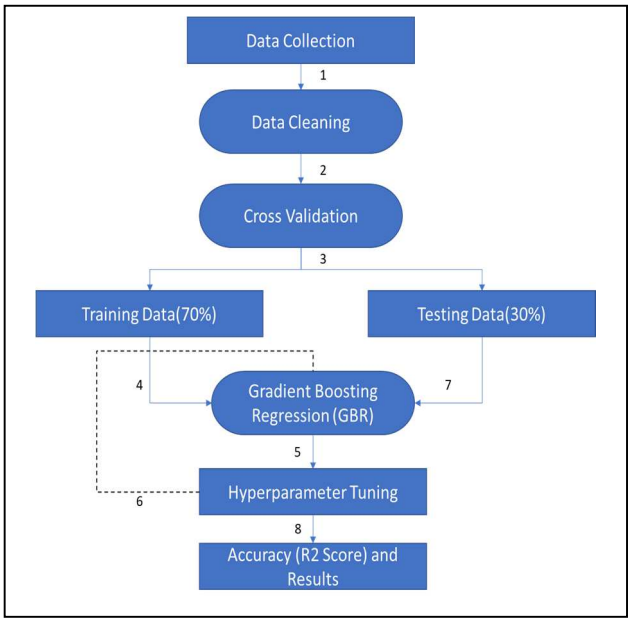

Fig. 5. Flowchart of the CO2 prediction model

As mentioned in Section II Part A, the data is collected from the laboratory. The raw data is cleaned and pre-processed before developing the model. Besides, the data has been normalized to ensure the values are within the standard scale across all data. Fig 6 shows the comparison of data transformation using different methods.

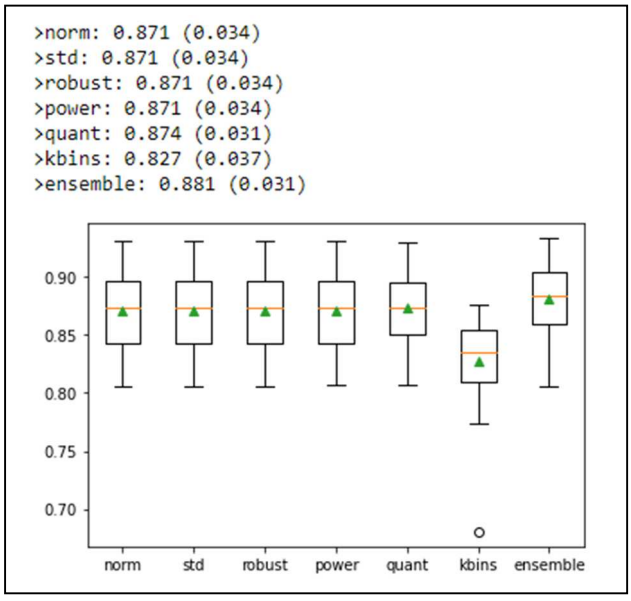

Fig. 6. Comparison of data transformation

From Fig. 6, every method gives almost the same accuracy for data transformation except for kbins. Therefore, we use the normalization (norm) method to transform the data. Although ensemble provides the highest accuracy, it applies to specific datasets only. On the other hand, normalization can be generalized very well to any dataset, and the process to build machine learning will be easy.

```
folds=2, accuracy=0.897560 (0.508950,0.999631)
folds=3, accuracy=0.897559 (0.508950,0.999627)
folds=4, accuracy=0.897560 (0.508950,0.999631)
folds=5, accuracy=0.897560 (0.508950,0.999631)
folds=6, accuracy=0.897559 (0.508950,0.999627)
folds=7, accuracy=0.897560 (0.508950,0.999629)
folds=8, accuracy=0.897560 (0.508950,0.999629)
folds=9, accuracy=0.897560 (0.508950,0.999631)
folds=10, accuracy=0.897560 (0.508950,0.999631)
```

Fig. 7. Accuracy of 10-fold cross-validation using the training set

The following process is to cross-validate (CV) the data. Cross-validation is important because it can assess the effectiveness of the model on unseen data. The grid search CV, in the python module, was used to identify the best hyperparameter using 10-folds cross-validation on the training set. Fig. 7 shows the accuracy of 10-fold cross-validation that can reach the maximum of 89%.

The CO2 prediction model is also tuned with several parameters. Fig. 8, 9, and 10 show the selected best hyperparameters for the model to obtain the highest accuracy. These hyperparameters play a major role in the generalization of the GBR model on unseen data. Therefore, it is important to tune these hyperparameters carefully.

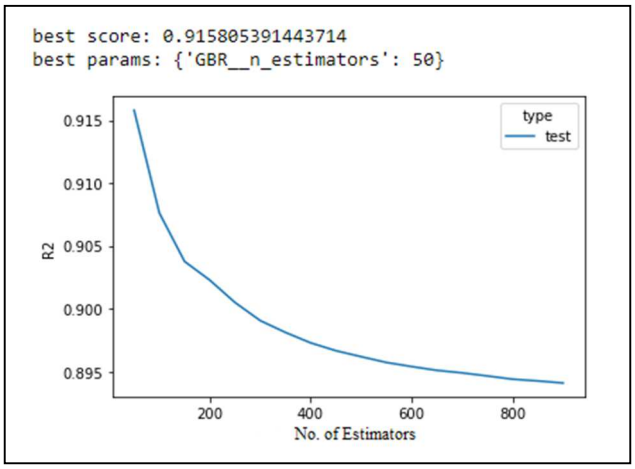

Fig. 8. Number of estimators based on their best score (R^2)

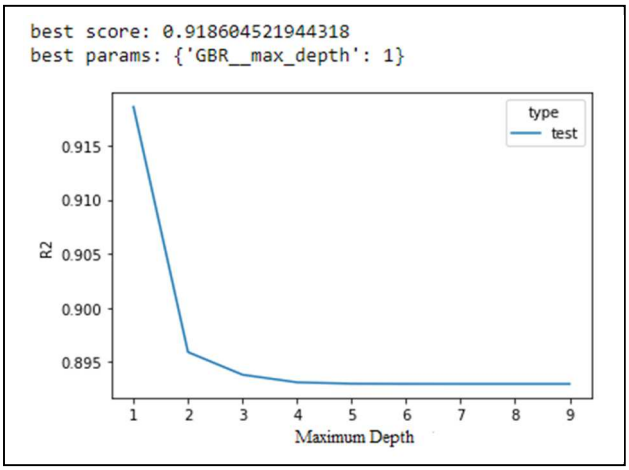

Fig. 9. Maximum depth values based on their best score (R^2)

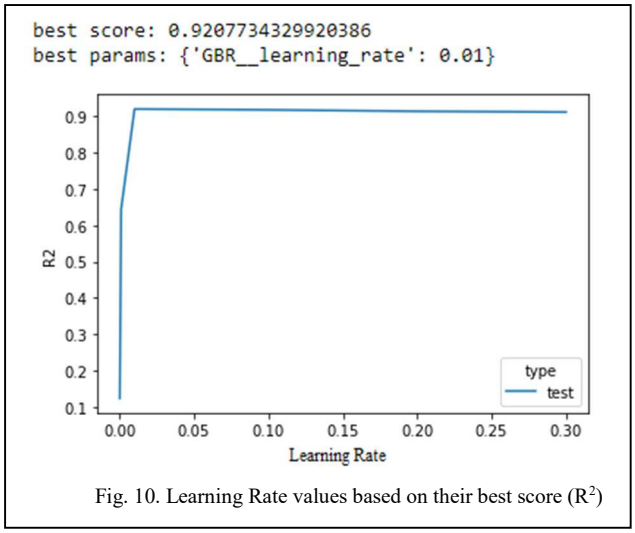

Fig. 10. Learning Rate values based on their best score (R^2)

Fig. 10. Learning rate values based on their best score (R^2)

After tuning the hyperparameters of the CO2 prediction model, we were able to achieve up to 92 % of accuracy. As for comparison, we have compared our GBR model with other ensemble models consisting of Multilayer Perceptron (MLP) and Xtreme Gradient Boosting Regression (XGBR). Fig. 11 shows the accuracies and errors from these three different models.

The trained model using the GBR algorithm has the highest R2 score and the lowest MSE error from the comparison. This means that the model is generalized very well to the input dataset.

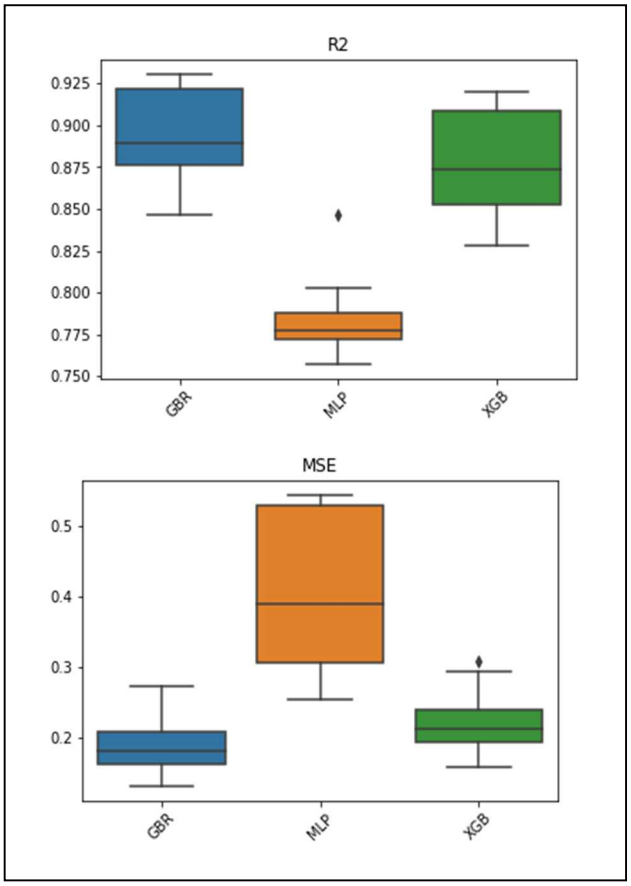

Fig.11.Comparison among CO2 prediction model using GBR,MLP and XGB algorithm

IV. RESULTS AND DISCUSSION

Once the GBR model has completed the tuned with the best hyperparameters, the test set is fed into the model to predict the CO2 emission. The testing set is also used to validate the internal part of the model. The model is validated through regression evaluation metrics. The R^2, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) is calculated for obtaining the prediction accuracy [16-18]. Finally, the CO2 prediction model has achieved 91% of accuracy and RMSE 0.05276 (Fig. 12). However, the scatter output is because the sudden changes of vehicle speed represent events like deceleration or acceleration and harsh breaking especially in urban driving behaviour cause the changes just took a few second, so it recorded as individual point which scattered along urban driving period. Nevertheless, the developed model has sufficient information to learn from the training data.

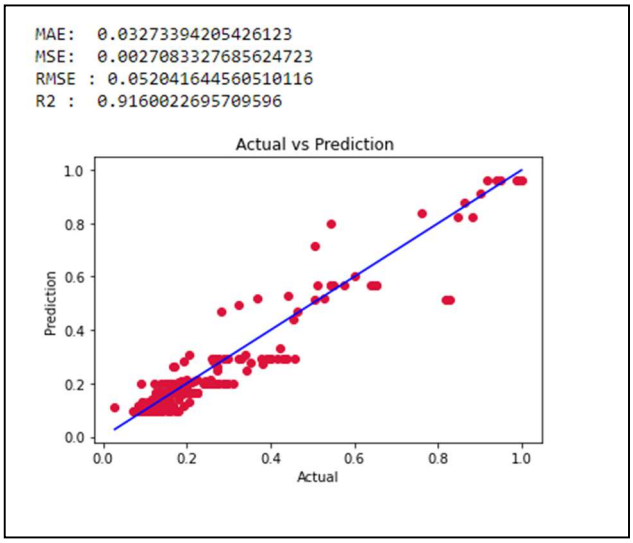

Fig.12. The comparison between actual and predicted CO2 emission

V. CONCLUSION AND FUTURE WORK

This study has proposed an approach to predict CO2 emission using the NEDC standard driving pattern obtained from the laboratory test. Specifically, optimized Gradient Boosting Regression (GBR) has been used by selecting the best hyperparameters to predict the CO2 emission. The result demonstrates that tuning the GBR with 50 estimators, 1 maximum depth, and 0.01 learning rate can get the best accuracy.

However, future work will focus on fuel consumption as another input to make the model more robust in each situation and visualize the CO2 emission in the geographical map that can depict the CO2 patterns produced by the passenger vehicles. Additionally, the developed model also will be used in fieldwork to demonstrate the real-time spatial pattern of CO2 emission in the urban area.

ACKNOWLEDGEMENT

This work is supported by Universiti Teknologi Malaysia (UTM) under Fundamental Research Grant Scheme (R.J130000.7852.5F151 & reference no. PY/2019/01089) from the Ministry of Higher Education (MOHE).

[1] J. Seo, J. Park, J. Park, and S. Park, "Emission factor development for light-duty vehicles based on real-world emissions using emission map-based simulation," *Environ Pollut,* vol. 270, p. 116081, Feb 1 2021, doi: 10.1016/j.envpol.2020.116081.

[2] A. K. Agarwal and N. N. Mustafi, "Real-world automotive emissions: Monitoring methodologies, and control measures," *Renewable and Sustainable Energy Reviews,* vol. 137, 2021, doi: 10.1016/j.rser.2020.110624.

[3] H. Fujii, K. Iwata, A. Chapman, S. Kagawa, and S. Managi, "An analysis of urban environmental Kuznets curve of CO2 emissions: Empirical analysis of 276 global metropolitan areas," *Applied Energy,* vol. 228, pp. 1561-1568, 2018, doi: 10.1016/j.apenergy.2018.06.158.

[4] T. Jia, Q. Li, and W. Shi, "Estimation and analysis of emissions from on-road vehicles in Mainland China for the period 2011–2015," *Atmospheric Environment,* vol. 191, pp. 500-512, 2018, doi: 10.1016/j.atmosenv.2018.08.037.

[5] F. Li, T. Zhou, and F. Lan, "Relationships between urban form and air quality at different spatial scales: A case study from northern China," *Ecological Indicators,* vol. 121, 2021, doi: 10.1016/j.ecolind.2020.107029.

[6] R. Smit, P. Kingston, D. W. Neale, M. K. Brown, B. Verran, and T. Nolan, "Monitoring on-road air quality and measuring vehicle emissions with remote sensing in an urban area," *Atmospheric Environment,* vol. 218, 2019, doi: 10.1016/j.atmosenv.2019.116978.

[7] S. Sun, W. Jiang, and W. Gao, "Vehicle emission trends and spatial distribution in Shandong province, China, from 2000 to 2014," *Atmospheric Environment,* vol. 147, pp. 190-199, 2016, doi: 10.1016/j.atmosenv.2016.09.065.

[8] L. Chen, Z. Wang, S. Liu, and L. Qu, "Using a chassis dynamometer to determine the influencing factors for the emissions of Euro VI vehicles," *Transportation Research Part D: Transport and Environment,* vol. 65, pp. 564-573, 2018/12/01/ 2018, doi: https://doi.org/10.1016/j.trd.2018.09.022.

[9] M. Grote, I. Williams, J. Preston, and S. Kemp, "Including congestion effects in urban road traffic CO2 emissions modelling: Do Local Government Authorities have the right options?," *Transportation Research Part D: Transport and Environment,* vol. 43, pp. 95-106, 2016, doi: 10.1016/j.trd.2015.12.010.

[10] M. A. Abas, S. Rajoo, and S. F. Zainal Abidin, "Development of Malaysian urban drive cycle using vehicle and engine parameters," *Transportation Research Part D: Transport and Environment,* vol. 63, pp. 388-403, 2018, doi: 10.1016/j.trd.2018.05.015.

[11] G. Fontaras, V. Franco, P. Dilara, G. Martini, and U. Manfredi, "Development and review of Euro 5 passenger car emission factors based on experimental results over various driving cycles," *Science of The Total Environment,* vol. 468-469, pp. 1034-1042, 2014/01/15/ 2014, doi: https://doi.org/10.1016/j.scitotenv.2013.09.043.

[12] D. Hirasen, V. Pillay, S. Viriri, and M. Gwetu, "Skeletal Age Estimation from Hand Radiographs Using Ensemble Deep Learning," in *Pattern Recognition*, Cham, E. Roman-Rangel, Á. F. Kuri-Morales, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. A. Olvera-López, Eds., 2021// 2021: Springer International Publishing, pp. 173-183.

[13] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics,* vol. 29, no. 5, pp. 1189-1232, 2001. [Online]. Available: http://www.jstor.org/stable/2699986.

[14] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis,* vol. 38, no. 4, pp. 367-378, 2002/02/28/ 2002, doi: https://doi.org/10.1016/S0167-9473(01)00065-2.

[15] N. Dahiya, B. Saini, and H. D. Chalak, "Gradient boosting-based regression modelling for estimating the time period of the irregular precast concrete structural system with cross bracing," *Journal of King Saud University - Engineering Sciences,* 2021, doi: 10.1016/j.jksues.2021.08.004.

[16] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai," *Aerosol and Air Quality Research,* vol. 20, no. 1, pp. 128-138, 2020, doi: 10.4209/aaqr.2019.08.0408.

[17] O. Azeez, B. Pradhan, H. Shafri, N. Shukla, C.-W. Lee, and H. Rizeei, "Modeling of CO Emissions from Traffic Vehicles Using Artificial Neural Networks," *Applied Sciences,* vol. 9, no. 2, 2019, doi: 10.3390/app9020313.

[18] S. E. K. Kutsev Bengisu Altug˘, "Predicting Tailpipe NOx Emission using Supervised Learning Algorithms," *IEEE Xplore,* 2019.