Prediction of Vehicle-induced Air Pollution based on Advanced Machine Learning Models

Caroline Matara
Department of Civil & Construction Engineering, University of Nairobi, Kenya | School of Civil and Resource Engineering, Technical University of Kenya, Kenya
caroltransy@gmail.com (corresponding author)

Simpson Osano
Department of Civil & Construction Engineering, University of Nairobi, Kenya
sosano@uonbi.ac.ke

Amir Okeyo Yusuf
Department of Chemistry, University of Nairobi, Kenya
ayusuf@uonbi.ac.ke

Elisha Ochungo Aketch
Department of Civil, Faculty of Engineering and Technology (FoET), Multimedia University, Kenya
elishaketch1@gmail.com

Received: 27 November 2023 | Revised: 11 December 2023 and 19 December 2023 | Accepted: 19 December 2023

 DOI: https://doi.org/10.48084/etasr.6678

ABSTRACT

Vehicle-induced air pollution is an important issue in the 21st century, posing detrimental effects on human health. Prediction of vehicle-emitted air pollutants and evaluation of the diverse factors that contribute to them are of the utmost importance. This study employed advanced tree-based machine learning models to predict vehicle-induced air pollutant levels, with a particular focus on fine particulate matter ($PM_{2.5}$). In addition to a benchmark statistical model, the models employed were Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGBoost), Extra Tree (ET), and Random Forest (RF). Regarding the evaluation of $PM_{2.5}$ predictions, the ET model outperformed the others, as shown by MAE of 1.69, MSE of 5.91, RMSE of 2.43, and R^2 of 0.71. Afterward, the optimal ET models were interpreted using SHAP analysis to overcome the ET model's lack of explainability. Based on the SHAP analysis, it was determined that temperature, humidity, and wind speed emerged as the primary determinants in forecasting $PM_{2.5}$ levels.

Keywords-air pollutants; machine learning; SHAP analysis

I. INTRODUCTION

Air pollution has become a prominent concern in the 21st century posing a significant threat to human health [1]. Vehicle emissions contribute significantly to atmospheric pollution in urban areas, often serving as the primary source of ultrafine particles and chemical pollutants, including carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and Total Volatile Organic Compounds (TVOCs) [2-3]. In general, increased air pollution can be attributed to increased motorization, energy consumption, and urbanization [4]. Exposure to elevated levels of automobile exhaust pollution adversely affects human health. Numerous studies have shown a link between exposure to air pollution on heavily trafficked roads and various harmful health effects, including elevated risk of mortality [5-6], higher rates of cardiopulmonary mortality [7], higher incidence of coronary heart disease [8], impaired respiratory function such as wheezing and reduced peak respiratory flow during infancy [9], lung cancer [10], pulmonary edema [11], allergic alveolitis [12], and development of chronic bronchitis and asthma [13]. Higher incidences of these health effects have been observed in cities that rely on automobiles for daily transportation due to inadequate public transport [14]. It is important to recognize that some of the other factors that contribute to these emissions include the presence of old and deteriorated engines, the use of improper fuel grades, the lack of routine maintenance, engine degradation over time, excessive vehicle use, mishandling of lubricants, and limitations in achieving optimal fuel combustion [15]. In addition to these factors, meteorological factors have been proven to influence the concentration and spread of air pollutants.

Air pollution modeling can describe the causal relationship between emissions, meteorology, atmospheric concentrations, deposition, and other aspects. Air quality modeling aims to forecast the dispersion, spatial distribution, and concentration of atmospheric pollutants using mathematical or statistical methods that replicate physical and chemical processes [16]. Currently, there is widespread adoption of Machine Learning (ML) techniques for air quality forecasting. ML is a computational approach to extracting information from data, characterized by its ability to autonomously adjust its algorithms and models in response to new datasets, requiring minimal initial configuration [17]. The typical steps in this process involve data collection, screening model evaluation, analysis using a complex model and algorithm, and finally evaluation and verification of the model's output. Although ML has proven to be effective in making predictions, one of its major disadvantages is its black-box nature. Therefore, the post hoc SHapley Additive exPlanations (SHAP) [18] technique has been used to explain models from both a global and a local perspective. SHAP is normally used to make an ML model more explicable by visualizing its output. Due to the high efficiency of SHAP in interpreting various ML models, it has been used in several fields, such as the safety assessment of infrastructure projects [19], clinical medicine and healthcare modeling [20-21], transportation and traffic safety [22-25], and economic risk analysis [26].

ML techniques have been used by many studies in air quality modeling and prediction, facilitated by the transition from manual to automated methods. In [27], roadside air pollution in Lisbon, Portugal, was modeled by training ML models with data from meteorological sensors and mobile monitoring stations, showing that the Random Forest (RF) algorithm was more effective in forecasting pollutant concentrations. In [28], data from fixed monitoring stations and meteorological sensors were used to perform deep learning-based prediction, showing that a Support Vector Regression (SVR) model was the best in predicting pollutant concentrations. In [29], air pollution levels in Delhi, India, were predicted using data from fixed monitoring stations and meteorological sensors, showing that a Multi-Layer Perceptron (MLP) Neural Network (NN) was best suited to predict pollutant concentrations. In [16], ML algorithms were used to predict traffic-related particulate matter pollution in Sao Paulo, Brazil, using data from fixed monitoring stations and traffic sensors and showing that SVR was the best in predicting pollutant concentrations. Some studies used sophisticated methods, such as ensemble learning classification systems, to predict air quality [30]. Artificial Neural Networks (ANNs) have also been popular techniques in predicting air pollution [31-33]. In general, ML algorithms can be effective in simulating and forecasting levels of roadside air pollution and can be used to create more precise and effective air quality monitoring systems.

This study used different ML algorithms, including Extreme Gradient Boosting (XGBoost) [34], RF [35], Extra Tree (ET) [36], Gradient Boosting (GB) [37], and Light Gradient Boosting Machine (LGBM) [38], which were optimized using a Bayesian optimization approach [39]. These models were trained and evaluated using data on vehicle-induced air pollutants, specifically $PM_{2.5}$, collected from the JKIA-Westlands expressway corridor in Kenya. Additionally, a statistical multivariate linear regression model was used as a benchmark model. This study used Bayesian optimization and ML regression models in conjunction with SHAP analysis to determine the optimal regression model for the dataset. This amalgamation was expected to provide a reliable and accurate method for evaluating vehicle-induced air pollutants. Figure 1 shows the complete research process.

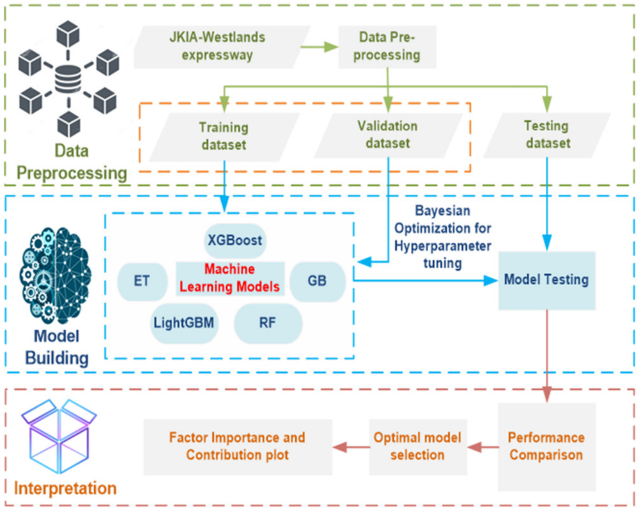

Fig. 1. Research framework.

II. METHODOLOGY

A. Site and Data Collection

Data were collected at three locations along the JKIA-Westlands Expressway in Kenya, as shown in Figure 2: Westlands, Bellevue, and City Cabanas. The Westlands location has mixed commercial and residential land use whereas City Cabanas is a predominantly industrial area and has commercial land use. The Bellevue station is close to the city center and characterized by residential and commercial land use. Data were collected 24/7 during August 2022, December 2022, and February 2023. Traffic volume was collected using manual traffic counts that were recorded every 15 minutes and summarized as hourly traffic volume. Various vehicle classifications were recorded: motorcycles, cars, minibuses, buses, light-goods vehicles, medium-goods vehicles, heavy-goods vehicles, and articulated trucks. In addition to documenting traffic volume, the study simultaneously collected data on air pollutant concentrations, average vehicle speed, and meteorological data, such as humidity, wind speed, and temperature. Air quality was measured using Open-Seneca sensors for each second. Air quality data includes the concentrations of Particulate Matter ($PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, PM_{10}), Typical Particle Size (TPS), Number of Concentrations of particles ($NC_{1.0}$, $NC_{2.5}$, $NC_{4.0}$, NC_{10}), TVOC, and equivalent CO_2. The air quality data was then averaged to the hourly data so that a comparison could be made between air quality, hourly traffic, and meteorological data. Regarding air quality data, this study focused exclusively on $PM_{2.5}$.

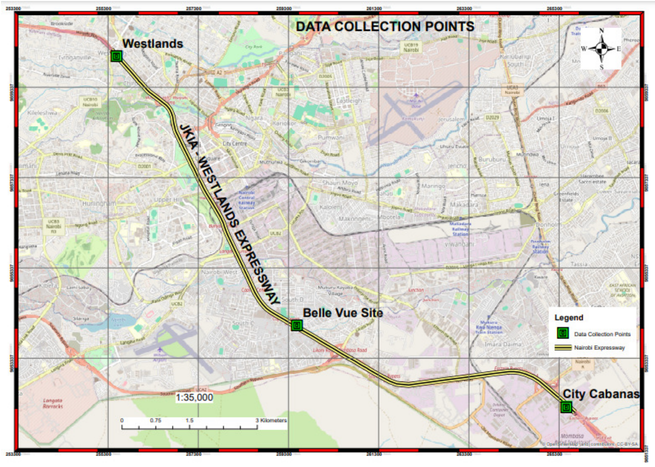

Fig. 2. Data collection sites along the JKIA-Westlands Expressway (ArcGIS 10.5.1).

B. Model Development

A statistical multiple variable Linear Regression (LR) model and five advanced ML models (GB, XGBoost, RF, ET, and LGBM) were used to forecast $PM_{2.5}$. Python 3.7.1 was used to implement the models and Bayesian optimization was used to tune their hyperparameters. Figure 3 illustrates the procedure involved in developing the ML models.

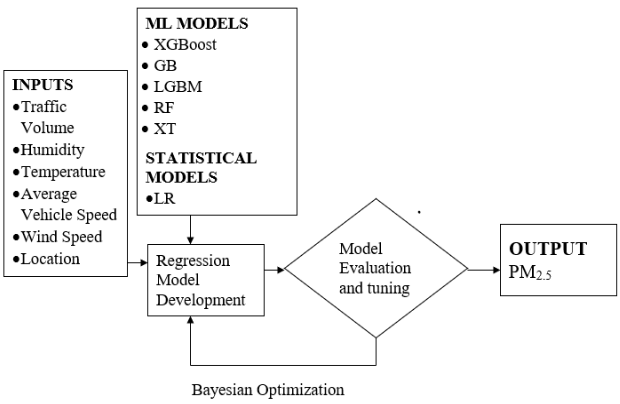

Fig. 3. Flow diagram of ML modeling, showing input and output variables.

C. Gradient Boosting (GB)

GB is an ML algorithm that gradually merges multiple weak learners into a single robust one. The following are the necessary steps in GB: Start by making a simple tree with only one root node that is the first impression of each sample. Next, use the flawed nodes to build a new tree. Sort the branches by how quickly they learn (the value is usually between 0 and 1). This learning rate will be used as input to the tree for the forecast. Then, combine the newly constructed tree with the older trees to make a prediction. If the fit has not improved after a certain number of trees was added, return to the second step. The combined set of trees is the prediction model.

D. Random Forest (RF)

The RF regression model is a collection of multiple decision trees that function as parallel estimators. The result is determined by aggregating the majority vote of the results obtained from each decision tree. The efficiency of an RF model depends on the utilization of uncorrelated decision trees. The training phase for each decision tree is improved by incorporating bootstrapping and feature randomness. The bootstrapping process involves the random selection of samples from a given training dataset with replacement. On the other hand, feature randomness is achieved by randomly selecting a subset of features for each decision tree within the RF model. Consequently, the base estimators exhibit independence and identical distribution, which leads to improved performance when subjected to the bagging technique.

E. Extreme Gradient Boosting (XGBoost)

XGBoost regression employs the technique of gradient descent on decision trees to iteratively generate a series of models. These models are subsequently combined sequentially, with each model aiming to rectify the errors made by the previous models. The ultimate objective is to produce a final model that is optimized for the given task. XGBoost demonstrates high efficiency in terms of computational resource utilization and processing speed.

F. Light Gradient Boosting Machine (LGBM) Regression

This framework uses tree-based learning to achieve efficient and distributed boosting. LGBM employs a leaf-wise growth strategy to construct the tree, in which a tree is generated for each individual sample. The selection of the leaf is determined by the maximum potential for growth inhibition. The leaf-wise algorithm exhibits less loss compared to the level-wise tree algorithm due to the fixed nature of the leaf. Therefore, the growth of trees in a leaf-wise manner results in heightened complexity in the model and occasionally results in overfitting when applied to datasets of limited size. LGBM aims to decrease complexity through the use of gradient-based one-side sampling and exclusive feature bundling.

G. Multivariate Linear Regression (MLR)

MLR is a widely used technique in supervised ML that examines the relationship between a group of independent variables or features and a singular outcome variable. This technique uses mathematical equations to represent and predict outcomes, particularly in cases where the correlation coefficient between the variables suggests a statistically significant association.

H. Extra Tree (ET) Regression

The ET algorithm was created to mitigate the risk of overfitting the dataset. Like RF, ET uses a random subset of features to train each base estimator. The selection of the optimal feature and value for node partitioning is carried out randomly.

I. Bayesian Optimization

Bayesian optimization strategy was used to fine-tune the hyperparameters in the models. This algorithm uses ideas from the Bayesian theorem and is recognized as a prominent method for achieving global optimization. Bayesian optimization has been extensively used in various disciplines [40]. Optimization of the hyperparameters of the model involves maximization or minimization of the objective function. Consequently, the algorithm identifies the suitable combination of hyperparameters that guarantees that the model's performance is maximized to its utmost potential. This study uses R^2 as the evaluation metric.

J. SHapley Additive exPlanations (SHAP)

SHAP is a post hoc evaluation of ML models based on the game theory [18]. SHAP employs an additive factor attribution method to generate a coherent model. SHAP values improve the model's transparency and give insight into the functioning of the prediction model. Through the SHAP feature importance plot, significant features are selected based on the Shapley values. The feature effects and feature importance are combined in the summary plot, which indicates the correlation between a feature's value and its impact on the prediction. In this study, SHAP calculates the contribution of each feature to the prediction, which seeks to explain the prediction of an instance of air quality.

III. RESULTS AND DISCUSSION

Table I presents the descriptive statistics of multiple input factors.

TABLE I. DESCRIPTIVE STATISTICS OF INPUT FACTORS

Factors	Statistics			
	Mean	**St. Dev.**	**Min**	**Max**
Humidity	35.52	12.95	13.33	68.16
Temperature	26.75	4.53	16.86	42.07
Average traffic volume	1377.05	653.16	340	3211
Average vehicle speed	44.41	7.71	22.70	60.18
Wind speed	4.61	1.83	0.95	9.75
Location	0.98	0.81	0	2

The dataset used to predict air pollutant concentrations was divided into two subsets: a training-validation set, which accounted for 70% of the total data, and a testing/holdout set, which was 30% of the total data. The training-validation dataset was used for the development of the ML models and Bayesian optimization. The objective of Bayesian optimization was to determine the optimal hyperparameter configuration for different ML models to maximize the R^2 value within a specified sample space. Table II presents the optimal hyperparameter values obtained for $PM_{2.5}$.

TABLE II. ML ALGORITHMS WITH THEIR OPTIMAL HYPERPARAMETERS FOR THE ESTIMATION OF $PM_{2.5}$

Models	Hyperparameters	Range	Optimal values
LGBM	{(learning rate), (*n*_estimators)}	{(0.01-0.1), (50-500)}	{0.01, 92}
GB	{(learning rate), (*n*_estimators)}	{(0.01-0.2), (50-500)}	{0.08, 100}
RF	{(max_depth)}	{2-16)}	{7}
XGBoost	{(learning rate), (*n*_estimators)}	{(0.01-0.2), (50-500)}	{0.07, 160}
ET	{(max_depth)}	{2-16)}	{11}

A. Prediction of $PM_{2.5}$

Following the determination of the optimal hyperparameters, the holdout or testing data was used to compare the performance of the models. The dataset was divided into partitions after randomization, with 40% and 50% of the data allocated for testing. This analysis showed that the metric values for model testing remained consistent within the 95% confidence interval. This precautionary step was implemented to mitigate the risk of overfitting. No anomalies were observed in the performance metrics, regardless of the size of the test data. Table III presents the efficiency metrics for $PM_{2.5}$ prediction, using both training and test datasets. The ET model exhibited superior performance compared to the others, as shown by its lower Mean Absolute Error (MAE) of 1.69, Mean Squared Error (MSE) of 5.91, Root Mean Squared Error (RMSE) of 2.43, and higher R^2 of 0.711. The linear regression model showed the poorest performance, having 3.57 MAE, 19.11 MSE, 4.37 RMSE, and 0.064 R^2. Prediction error plots were used, as shown in Figure 4, to evaluate the efficiency of the ML regression models in predicting $PM_{2.5}$ and their ability to make predictions on unobserved data.

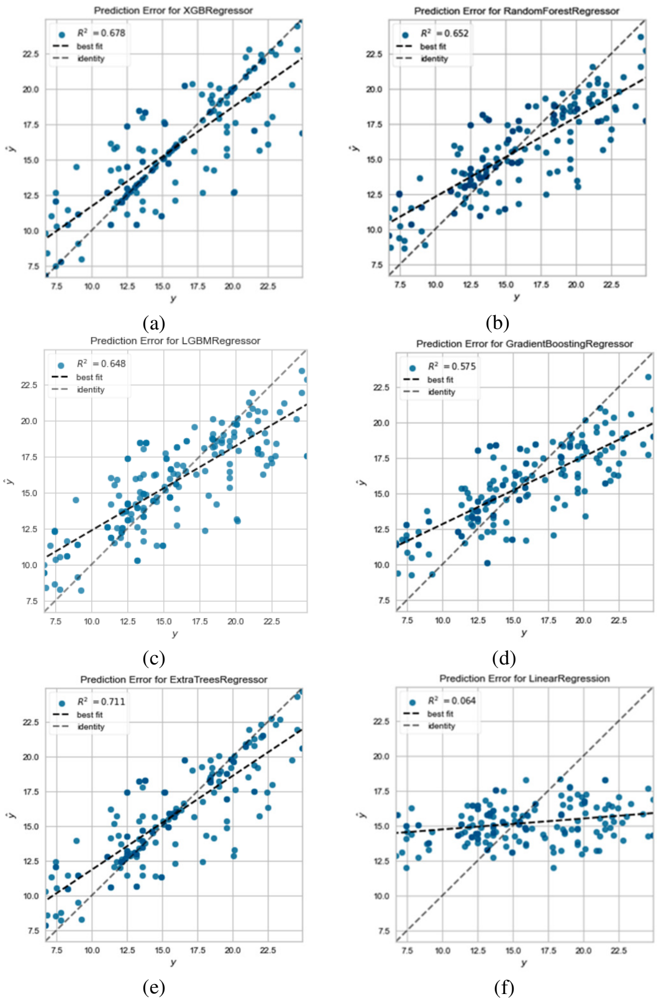

Fig. 4. Prediction Error for $PM_{2.5}$: (a) XGBoost, (b) RF, (c) LightGBM, (d) GB, (e) ET, and (f) LR.

TABLE III. PERFORMANCE USING ML AND STATISTICAL MODELS FOR PREDICTING $PM_{2.5}$

Training dataset				
Model	**MAE**	**MSE**	**RMSE**	**R^2**
RF	1.73	5.5	2.35	0.73
XGBoost	1.36	4.78	2.18	0.77
LGBM	1.73	5.56	2.35	0.73
GB	2.03	6.89	2.62	0.67
ET	1.48	4.92	2.22	0.76
LR	3.53	19.1	4.37	0.09
Testing dataset				
Model	**MAE**	**MSE**	**RMSE**	**R-Square**
RF	2.06	7.11	2.66	0.652
XGBoost	1.64	6.57	2.56	0.678
LGBM	2.04	7.19	2.68	0.648
GB	2.35	8.69	2.94	0.575
ET	1.69	5.91	2.43	0.711
LR	3.57	19.11	4.37	0.064

B. Global Interpretation by SHAP

The decision to employ ET as a fitting method for $PM_{2.5}$, considering the given factors, was determined based on its R^2 value. The analysis of the ET prediction for $PM_{2.5}$ yields insights into the global factor interpretation, as shown in Figure 5, which illustrates the significance and contribution of the SHAP factors. The mean absolute SHAP value shown in Figure 5(a) signifies the average influence on the magnitude of the model's output.

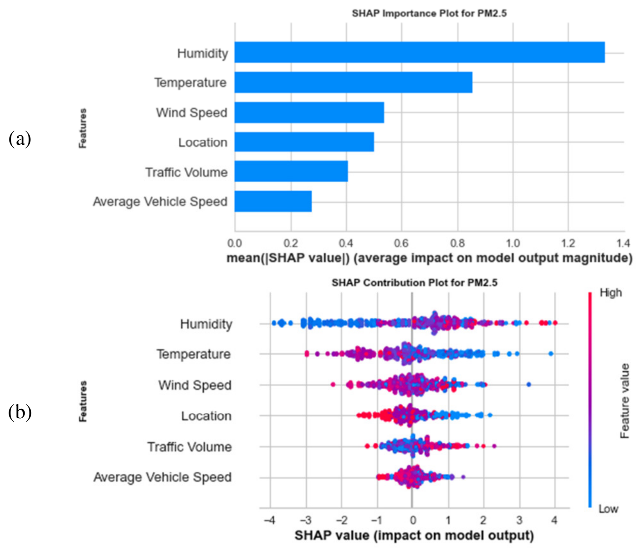

Fig. 5. Global factor interpretation: (a) Factor importance plot, and (b) contribution plot for PM2.5.

Humidity had the highest SHAP significance score of 1.35, while temperature had 0.84 and Wind Speed had 0.55. The dots that exhibit a color gradient from purple to red, descending to the right of the vertical reference line denoting humidity, indicate an elevated level of $PM_{2.5}$ risk. Similarly, the dots that symbolize low-temperature values are situated to the right of the vertical reference line, suggesting a higher likelihood for $PM_{2.5}$ to escalate. The findings indicate higher accuracy and reliability due to the use of ML, which is consistent with the findings from earlier studies on air quality prediction [27-31]. SHAP was useful in determining the degree to which the input parameters had an impact on the forecasts, which is consistent with prior studies that used SHAP analysis to gather local information about the factors causing either greater or lower pollutant concentrations [41-42]. Therefore, SHAP analysis provided an effective method for addressing the issue of limited interpretability inherent in ML regression models.

IV. CONCLUSION

From the analyzed results, the following deductions can be made:

- According to the collected dataset, ET exhibited superior performance compared to the other models, as demonstrated by its lower MAE (1.69), MSE (5.91), RMSE (2.43), and higher R^2 value (0.711). The linear regression model exhibited the poorest performance, as shown by its MAE (3.57), MSE (19.11), RMSE (4.37), and R^2 (0.064).
- In the context of $PM_{2.5}$ humidity, temperature, and wind speed were found to have the most significant influence.
- Bayesian optimization improved the prediction by selecting the important features, which resulted in better predictive results. This is in agreement with [43], who obtained better results using an improved collection of characteristics to predict cardiovascular diseases.
- Undoubtedly, ML is a highly valuable resource with many benefits in different fields, including medicine [43-44], information technology [45], and transportation [23, 25]. The findings indicate that ML has the potential to be used to predict roadside air pollution.

However, it is important to consider several suggestions for future research endeavors:

- Although this study employed various input parameters to predict $PM_{2.5}$, it is important to note that several additional factors could be considered in future investigations.
- This study focused solely on predicting $PM_{2.5}$. However, it should be noted that future studies could potentially explore the prediction of CO, NO_2, TVOCs, and SO_2.

REFERENCES

[1] P. H. Avogbe *et al.*, "Hematological changes among Beninese motor-bike taxi drivers exposed to benzene by urban air pollution," *African Journal of Environmental Science and Technology*, vol. 5, no. 7, pp. 464–472, 2011.

[2] Y. Zhu, W. C. Hinds, S. Kim, and C. Sioutas, "Concentration and size distribution of ultrafine particles near a major highway," *Journal of the Air & Waste Management Association (1995)*, vol. 52, no. 9, pp. 1032–1042, Sep. 2002, https://doi.org/10.1080/10473289.2002.10470842.

[3] S. Bhandarkar, "Vehicular Pollution, Their Effect on Human Heatlh and Mitigation Measures," *Vehicle Engineering*, vol. 1, no. 2, pp. 33–40, 2013.

[4] M. M. Jackson, "Roadside Concentration of Gaseous and Particulate Matter Pollutants and Risk Assessment in Dar-Es-Salaam, Tanzania," *Environmental Monitoring and Assessment*, vol. 104, no. 1, pp. 385–407, May 2005, https://doi.org/10.1007/s10661-005-1680-y.

[5] M. Krzyżanowski, B. Kuna-Dibbert, and J. Schneider, Eds., *Health effects of transport-related air pollution*. Copenhagen, Denmark: World Health Organization Europe, 2005.

[6] N. Künzli *et al.*, "Public-health impact of outdoor and traffic-related air pollution: a European assessment," *The Lancet*, vol. 356, no. 9232, pp. 795–801, Sep. 2000, https://doi.org/10.1016/S0140-6736(00)02653-2.

[7] G. Hoek, B. Brunekreef, S. Goldbohm, P. Fischer, and P. A. van den Brandt, "Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study," *The Lancet*, vol. 360, no. 9341, pp. 1203–1209, Oct. 2002, https://doi.org/10.1016/S0140-6736(02)11280-3.

[8] M. Rosenlund, S. Picciotto, F. Forastiere, M. Stafoggia, and C. A. Perucci, "Traffic-Related Air Pollution in Relation to Incidence and Prognosis of Coronary Heart Disease," *Epidemiology*, vol. 19, no. 1, pp. 121–128, 2008.

[9] E. Nordling *et al.*, "Traffic-Related Air Pollution and Childhood Respiratory Symptoms, Function and Allergies," *Epidemiology*, vol. 19, no. 3, pp. 401–408, 2008.

[10] E. Garshick *et al.*, "Lung Cancer and Vehicle Exhaust in Trucking Industry Workers," *Environmental Health Perspectives*, vol. 116, no. 10, pp. 1327–1332, Oct. 2008, https://doi.org/10.1289/ehp.11293.

[11] A. Ghorani-Azam, B. Riahi-Zanjani, and M. Balali-Mood, "Effects of air pollution on human health and practical measures for prevention in Iran," *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, vol. 21, Sep. 2016, Art. no. 65, https://doi.org/10.4103/1735-1995.189646.

[12] G. C. Kisku, S. Pradhan, A. H. Khan, and S. K. Bhargava, "Pollution in Lucknow City and its health implication on exposed vendors, drivers and traffic policemen," *Air Quality, Atmosphere & Health*, vol. 6, no. 2, pp. 509–515, Jun. 2013, https://doi.org/10.1007/s11869-012-0190-7.

[13] J. A. Araujo *et al.*, "Ambient Particulate Pollutants in the Ultrafine Range Promote Early Atherosclerosis and Systemic Oxidative Stress," *Circulation Research*, vol. 102, no. 5, pp. 589–596, Mar. 2008, https://doi.org/10.1161/CIRCRESAHA.107.164970.

[14] K. A. Salami, "Emission Control Technology by Automotive Industry: Trends and Challenges," *Inaugural lecture series*, vol. 10, pp. 8–9, 2007.

[15] S. Dey and N. S. Mehta, "Automobile pollution control using catalysis," *Resources, Environment and Sustainability*, vol. 2, Dec. 2020, Art. no. 100006, https://doi.org/10.1016/j.resenv.2020.100006.

[16] A. Aggarwal, A. K. Haritash, and G. Kansal, "Air pollution modelling-a review," *International Journal of Advanced Technology Engineering Science*, vol. 2, pp. 255–264, 2014.

[17] A. Wang, J. Xu, R. Tu, M. Saleh, and M. Hatzopoulou, "Potential of machine learning for prediction of traffic related air pollution," *Transportation Research Part D: Transport and Environment*, vol. 88, Nov. 2020, Art. no. 102599, https://doi.org/10.1016/j.trd.2020.102599.

[18] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[19] K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun, "Developing a National Data-Driven Construction Safety Management Framework with Interpretable Fatal Accident Prediction," *Journal of Construction Engineering and Management*, vol. 149, no. 4, Apr. 2023, Art. no. 04023010, https://doi.org/10.1061/JCEMD4.COENG-12848.

[20] S. Lu, R. Chen, W. Wei, M. Belovsky, and X. Lu, "Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions," *AMIA Annual Symposium Proceedings*, vol. 2021, pp. 813–822, Feb. 2022.

[21] P. N. Ramkumar *et al.*, "Association Between Preoperative Mental Health and Clinically Meaningful Outcomes After Osteochondral Allograft for Cartilage Defects of the Knee: A Machine Learning Analysis," *The American Journal of Sports Medicine*, vol. 49, no. 4, pp. 948–957, Mar. 2021, https://doi.org/10.1177/0363546520988021.

[22] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Analysis & Prevention*, vol. 136, Mar. 2020, Art. no. 105405, https://doi.org/10.1016/j.aap.2019.105405.

[23] A. Khattak, P.-W. Chan, F. Chen, and H. Peng, "Prediction and Interpretation of Low-Level Wind Shear Criticality Based on Its Altitude above Runway Level: Application of Bayesian Optimization–Ensemble Learning Classifiers and SHapley Additive exPlanations," *Atmosphere*, vol. 13, no. 12, Dec. 2022, Art. no. 2102, https://doi.org/10.3390/atmos13122102.

[24] H. Qi, Y. Yao, X. Zhao, J. Guo, Y. Zhang, and C. Bi, "Applying an interpretable machine learning framework to the traffic safety order analysis of expressway exits based on aggregate driving behavior data," *Physica A: Statistical Mechanics and its Applications*, vol. 597, Jul. 2022, Art. no. 127277, https://doi.org/10.1016/j.physa.2022.127277.

[25] A. Khattak, P. W. Chan, F. Chen, and H. Peng, "Time-Series Prediction of Intense Wind Shear Using Machine Learning Algorithms: A Case Study of Hong Kong International Airport," *Atmosphere*, vol. 14, no. 2, Feb. 2023, Art. no. 268, https://doi.org/10.3390/atmos14020268.

[26] S. Ben Jabeur, R. Khalfaoui, and W. Ben Arfi, "The effect of green energy, global environmental indexes, and stock markets in predicting oil price crashes: Evidence from explainable machine learning," *Journal of Environmental Management*, vol. 298, Nov. 2021, Art. no. 113511, https://doi.org/10.1016/j.jenvman.2021.113511.

[27] A. Analitis *et al.*, "Prediction of PM2.5 concentrations at the locations of monitoring sites measuring PM10 and NOx, using generalized additive models and machine learning methods: A case study in London," *Atmospheric Environment*, vol. 240, Nov. 2020, Art. no. 117757, https://doi.org/10.1016/j.atmosenv.2020.117757.

[28] U. Pak *et al.*, "Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China," *Science of The Total Environment*, vol. 699, Jan. 2020, Art. no. 133561, https://doi.org/10.1016/j.scitotenv.2019.07.367.

[29] C. Srivastava, S. Singh, and A. P. Singh, "Estimation of Air Pollution in Delhi Using Machine Learning Techniques," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, Greater Noida, India, Sep. 2018, pp. 304–309, https://doi.org/10.1109/GUCON.2018.8675022.

[30] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmospheric Environment*, vol. 80, pp. 426–437, Dec. 2013, https://doi.org/10.1016/j.atmosenv.2013.08.023.

[31] J. Zhang and W. Ding, "Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong," *International Journal of Environmental Research and Public Health*, vol. 14, no. 2, Feb. 2017, Art. no. 114, https://doi.org/10.3390/ijerph14020114.

[32] X. Y. Ni, H. Huang, and W. P. Du, "Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data," *Atmospheric Environment*, vol. 150, pp. 146–161, Feb. 2017, https://doi.org/10.1016/j.atmosenv.2016.11.054.

[33] J. Chen, H. Chen, Z. Wu, D. Hu, and J. Z. Pan, "Forecasting smog-related health hazard based on social media and physical sensor," *Information Systems*, vol. 64, pp. 281–291, Mar. 2017, https://doi.org/10.1016/j.is.2016.03.011.

[34] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, May 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[35] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[36] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *Journal of Cleaner Production*, vol. 203, pp. 810–821, Dec. 2018, https://doi.org/10.1016/j.jclepro.2018.08.207.

[37] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[38] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems*, 2012, vol. 25.

[39] N. A. Alsharif, S. Mishra, and M. Alshehri, "IDS in IoT using Machine Learning and Blockchain," *Engineering, Technology & Applied*

Science Research, vol. 13, no. 4, pp. 11197–11203, Aug. 2023, https://doi.org/10.48084/etasr.5992.

[40] K. Wang and A. W. Dowling, "Bayesian optimization for chemical products and functional materials," *Current Opinion in Chemical Engineering*, vol. 36, Jun. 2022, Art. no. 100728, https://doi.org/10.1016/j.coche.2021.100728.

[41] M. Vega García and J. L. Aznarte, "Shapley additive explanations for NO2 forecasting," *Ecological Informatics*, vol. 56, Mar. 2020, Art. no. 101039, https://doi.org/10.1016/j.ecoinf.2019.101039.

[42] J. Gu, B. Yang, M. Brauer, and K. M. Zhang, "Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models," *Atmospheric Environment*, vol. 246, Feb. 2021, Art. no. 118125, https://doi.org/10.1016/j.atmosenv.2020.118125.

[43] A. K. Dubey, A. K. Sinhal, and R. Sharma, "An Improved Auto Categorical PSO with ML for Heart Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8567–8573, Jun. 2022, https://doi.org/10.48084/etasr.4854.

[44] M. A. Alsuwaiket, "Feature Extraction of EEG Signals for Seizure Detection Using Machine Learning Algorthims," *Engineering, Technology & Applied Science Research*, vol. 12, no. 5, pp. 9247–9251, Oct. 2022, https://doi.org/10.48084/etasr.5208.

[45] S. Nuanmeesri, "A Hybrid Deep Learning and Optimized Machine Learning Approach for Rose Leaf Disease Classification," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7678–7683, Oct. 2021, https://doi.org/10.48084/etasr.4455.