



# OPEN Explainable AI analysis for smog rating prediction

Yazeed Yasin Ghadi<sup>1</sup>, Sheikh Muhammad Saqib<sup>2</sup>✉, Tehseen Mazhar<sup>3,9</sup>✉, Ahmad Almogren<sup>4</sup>, Wajahat Waheed<sup>5</sup>, Ayman Altameem<sup>6</sup> & Habib Hamam<sup>7,8,10,11</sup>

Smog poses a direct threat to human health and the environment. Addressing this issue requires understanding how smog is formed. While major contributors include industries, fossil fuels, crop burning, and ammonia from fertilizers, vehicles play a significant role. Individually, a vehicle's contribution to smog may be small, but collectively, the vast number of vehicles has a substantial impact. Manually assessing the contribution of each vehicle to smog is impractical. However, advancements in machine learning make it possible to quantify this contribution. By creating a dataset with features such as vehicle model, year, fuel consumption (city), and fuel type, a predictive model can classify vehicles based on their smog impact, rating them on a scale from 1 (poor) to 8 (excellent). This study proposes a novel approach using Random Forest and Explainable Boosting Classifier models, along with SMOTE (Synthetic Minority Oversampling Technique), to predict the smog contribution of individual vehicles. The results outperform previous studies, with the proposed model achieving an accuracy of 86%. Key performance metrics include a Mean Squared Error of 0.2269, R-Squared ( $R^2$ ) of 0.9624, Mean Absolute Error of 0.2104, Explained Variance Score of 0.9625, and a Max Error of 4.3500. These results incorporate explainable AI techniques, using both agnostic and specific models, to provide clear and actionable insights. This work represents a significant step forward, as the dataset was last updated only five months ago, underscoring the timeliness and relevance of the research.

**Keywords** Machine learning, Random forest, Explainable boosting classifier, SMOTE, Explainable-AI

An international research-based investigation has revealed that air pollution is a major cause of different ailments. In particular, pollution claims one out of four lives attributable to lung cancer, one out of five cardiovascular disease-related fatalities, and one out of five stroke-related fatalities on a global basis. These statistics explain why sound forecasting tools can be essential for anticipating the nature and scope of air quality and its health effects to better prepare for crises. The use of such a forecast could go a long way in improving the management and policy intervention in public health to reduce the effects of pollution-borne diseases<sup>1</sup>. This is because people are shifting from rural areas into urban areas which have been on the rise in recent years. UN reported that 56.15% of the world's population lived in urban areas of the world as of 2020<sup>2</sup>. This number is however believed to grow sharply in the next couple of years as it is estimated that by 2050 seventy percent of the global population will be living in urban areas<sup>3</sup>.

The last update for the dataset MY1995-2023 Fuel Consumption Ratings<sup>4</sup> can be traced back five months ago and except for the project XAI Applications Based on Vehicle Characteristics for Reducing CO<sub>2</sub> Emissions<sup>5</sup>, no more work has been done on the data. This line of work is focused on predicting carbon dioxide emissions given the characterizing parameters of vehicles which is one way of minimizing the carbon footprint. On the other hand, the proposed research moves to Smog Ratings which offers an assessment of the bad effect vehicle emissions on the quality of the air people breathe within that given region.

<sup>1</sup>Department of Computer Science and Software Engineering, Al Ain University, 12555 Abu Dhabi, United Arab Emirates. <sup>2</sup>Department of Computing and Information Technology, Gomal University, Dera Ismail Khan 29050, Pakistan. <sup>3</sup>School of Computer Science, National College of Business Administration and Economics, Lahore 54000, Pakistan. <sup>4</sup>Department of Computer Science, College of Computer and Information Sciences, King Saud University, 11633 Riyadh, Saudi Arabia. <sup>5</sup>Department of Electrical and Computer Engineering, Purdue University, Indiana 46323, USA. <sup>6</sup>Department of Natural and Engineering Sciences, College of Applied Studies and Community Services, King Saud University, 11543 Riyadh, Saudi Arabia. <sup>7</sup>Faculty of Engineering, Uni de Moncton, Moncton, NB E1A3E9, Canada. <sup>8</sup>School of Electrical Engineering, University of Johannesburg, Johannesburg 2006, South Africa. <sup>9</sup>Department of Computer Science and Information Technology, School Education Department, Government of Punjab, Layyah 31200, Pakistan. <sup>10</sup>International Institute of Technology and Management (IITG), Av, Grandes Ecoles, BP 1989, Libreville, Gabon. <sup>11</sup>Bridges for Academic Excellence, Spectrum, Tunisa, Tunisia. ✉email: saqibsheikh4@gu.edu.pk; tehseenmazhar719@gmail.com

With the enhancement of urbanization processes, the problem of air pollution control and, in particular, the emissions of transport become critical. Cars are one of the leading emitters of unhealthy pollutants like nitrogen oxides (NO<sub>x</sub>) and non-methane organic gases (NMOGs) which produce smog. This smog also interferes with human health outcomes as well as climate equilibrium. In this respect, the appreciation of environmental performance of vehicles can be acknowledged significant for minimizing the country's propensity to pollution.

This research necessarily draws out vehicle attributes as an indicator in assigning Smog Ratings, which quantify smog forming emissions from one to the best of eight. As part of the study, Random Forest (RF) and Explainable-Boosting-Classifer (EBC) models are used to determine the factors affecting air quality moving forward to enhance smog forecast and contribute more effectively to policymaking for environmental defense and manage crises.

In order to make the interpretation of the results to be more transparent, this research uses other XAI techniques in addition to checking the performance of the RF and EBC models. To improve the credibility and comprehensibility of the results, both the agnostic and the model-specific reasoning techniques are applied. Agnostic models are thus a general look at how the predictions are undistorted across every instance, irrespective of the model in place. On the other hand, model-specific methods operate at fine-grained level and aim to show how much specific feature contributed to final prediction and thus offer more specific insight into the model behavior. Through the combination of these techniques, the study hopes to enhance the credibility and readability of the smog forecast results such that consumers can place their confidence to some extent on the model.

Despite the potential of machine learning models to handle complex datasets and improve air quality predictions, several limitations persist in the literature. Challenges such as inadequate data preprocessing, poor handling of class imbalance, suboptimal data splitting strategies, and insufficient hyperparameter tuning undermine the reliability and generalization of these models. Many studies report high accuracy for majority classes while underperforming on minority classes, leading to a false perception of model effectiveness. Furthermore, reliance solely on accuracy as a performance metric often masks issues like overfitting and poor generalizability on unseen data. This study aims to bridge these gaps by incorporating robust evaluation metrics such as precision, recall, and F1-score to ensure balanced performance across all classes, particularly for imbalanced datasets.

By addressing these gaps, the proposed work not only achieves an accuracy exceeding 86% but also prioritizes stability and fairness, ensuring the model's performance is both reliable and interpretable for real-world applications. Moreover, the use of XAI techniques directly addresses the need for interpretability in existing studies. While many traditional approaches have focused on improving predictive accuracy, they often fall short in explaining the reasoning behind the predictions. This research's focus on transparent and interpretable models fills a critical void in the literature, ensuring the outputs are not only accurate but also actionable and understandable for stakeholders.

## Literature review

However, as smog prediction for vehicles is a fairly new discipline there is no directly related literature available. However, this research seeks to review similar studies and their drawbacks with a view of establishing the need for the current work. Therefore, this study aims at filling the gap in existing models or methodologies where there is compromise in the three parameters of accuracy, interpretability, and real time when making air quality prediction and environmental impact assessments.

Machine learning and deep learning play essential roles across various fields, utilizing techniques like transfer learning and fuzzy convolutional neural networks and unified frameworks to address diverse challenges. In sentiment analysis, deep learning achieves significant accuracy improvements through Few-Shot Learning, enabling models to generalize from minimal labeled data for nonverbal communication sentiment classification<sup>6</sup>. Similarly, fuzzy convolutional neural networks enhance the classification of mechanical faults by integrating multi-source sensing data, achieving superior accuracy in industrial diagnostics<sup>7</sup>. In healthcare, machine learning drives innovation in detecting cardiovascular diseases (CVD). The Attention-Based Cross-Modal (ABCM) transfer learning framework integrates diverse data sources, such as clinical records and medical imagery, achieving high accuracy and recall rates. This approach not only improves early detection but also supports informed decision-making, showcasing the transformative potential of deep learning in addressing complex challenges<sup>8</sup>. For multimodal sentiment analysis, the Context-Sensitive Multi-Tier Deep Learning Framework (CS-MDF) effectively extracts context-sensitive features, outperforming state-of-the-art methods across multiple datasets<sup>9</sup>. Additionally, a novel unified framework for natural language processing tasks, such as emotion and hate speech detection, outperformed task-specific models like BERT, achieving superior accuracy, F1-scores, and AUC-ROC<sup>10</sup>.

Moreover, "DeepXplainer," a hybrid deep learning technique combining CNNs and XGBoost, enhances lung cancer detection while providing interpretable predictions using SHAP for explainability, demonstrating the potential of interpretable AI in critical applications<sup>11</sup>. The integration of Explainable AI (XAI) and the Internet of Medical Things (IoMT) revolutionizes healthcare by improving illness detection efficiency, transparency, and reducing medical costs<sup>12</sup>. Additionally, a hybrid ResNet U-Net with Transformer blocks excels in glioma brain tumor segmentation from MRI, achieving a remarkable 98% accuracy and dice scores of 0.91, 0.89, and 0.84 for the whole, core, and enhancing tumor regions, respectively, on the BraTS2019 dataset<sup>13</sup>. The "BC2" hybrid model integrates CNN and LGBM for breast cancer detection, achieving exceptional performance (accuracy: 98.29%, precision: 98.71%, recall: 98.71%, F1-score: 98.71%). SHAP was used to explain predictions, enhancing transparency and aiding medical professionals in managing breast cancer patients effectively<sup>14</sup>. The "AC" (Accurate Cardiac Classification) model integrates Convolutional Neural Networks and the Light Gradient Boosting Model for precise cardiovascular disease detection. Its performance is enhanced by Explainable AI

(XAI) through SHAP's local and global explanations, providing clarity on the model's decision-making process. This XAI approach improves the model's comprehensibility, making it more effective for healthcare professionals in diagnosing and treating patients<sup>15</sup>. The work<sup>16</sup> introduces a dynamic feature-gated attention mechanism integrated into a 3D U-Net architecture for glioma segmentation in MRI scans. While not explicitly described as "explainable AI" (XAI), the attention mechanism helps the model focus on important features, enhancing segmentation accuracy and providing insights into feature importance. This adaptive approach could be seen as a form of explainability, as it allows the model to adjust attention weights based on the data, offering clearer insights into the decision-making process for tumor segmentation.

There are several approaches developed for the PM<sub>2.5</sub> prediction, including the use of statistical models and machine learning algorithms, and deep neural networks have recently attracted the researcher's attention. These networks, through using multiple layers and more samples, appropriately confined higher informational deservedness than BP for obtaining higher identification accuracy, and thus more suitable in air pollution prediction<sup>17</sup>. Several modes have indeed been considered for this end. For example, one work<sup>18</sup> compared the performance of PM<sub>2.5</sub> prediction at 12 stations in Beijing by using ARIMA, FB-Prophet, LSTM, and CNN. Among these models, LSTM was best than others with Mean absolute error of 13.2 and root mean square of 20.8. Likewise, another study<sup>19</sup> also validates LSTM and DAE for predicting the PM<sub>2.5</sub> levels at 25 monitoring stations in Seoul, South Korea but, LSTM has been glanced accurate as the study projected. A different approach was adopted in<sup>20,21</sup> wherein bidirectional LSTM model was used to forecast the concentration of PM<sub>2.5</sub> in Beijing with MAE=7.53, RMSE=9.86 and SMAPE=0.1664. Moreover, in<sup>10</sup>, the authors investigated the accuracy of LSTM, LSTM-FC and ANN models for establishing the air quality and meteorological data that had been retrieved from the past with higher MAE = 23.97% and RMSE = 35.82%. Unfortunately, these models do not pay attention to the spatial dimension of concentration of pollutants. Data containing spatial information presents a major aspect in the modeling process because pollutants can transfer from one place to another.

Previous research has explored various methods, including statistical models, mathematical models, and machine learning (ML) approaches, to classify and predict air pollution levels. For instance, work<sup>22</sup> applied Recurrent Neural Networks (RNN) for ozone classification, achieving 81% accuracy. In another study<sup>23</sup>, RNN was used to analyze PM<sub>2.5</sub> and PM<sub>10</sub>, reaching 95% accuracy. Work<sup>24</sup> utilized Support Vector Regression and ensemble models to classify PM<sub>10</sub>, achieving 96% accuracy. Similarly, work<sup>25</sup> employed multiple ML models for air pollution classification, with logistic regression delivering 93% accuracy. However, this study overlooked important aspects such as parameter tuning. A large portion of previous work has relied on Artificial Neural Networks (ANN), hybrid ANNs, and other ML models for air quality forecasting.

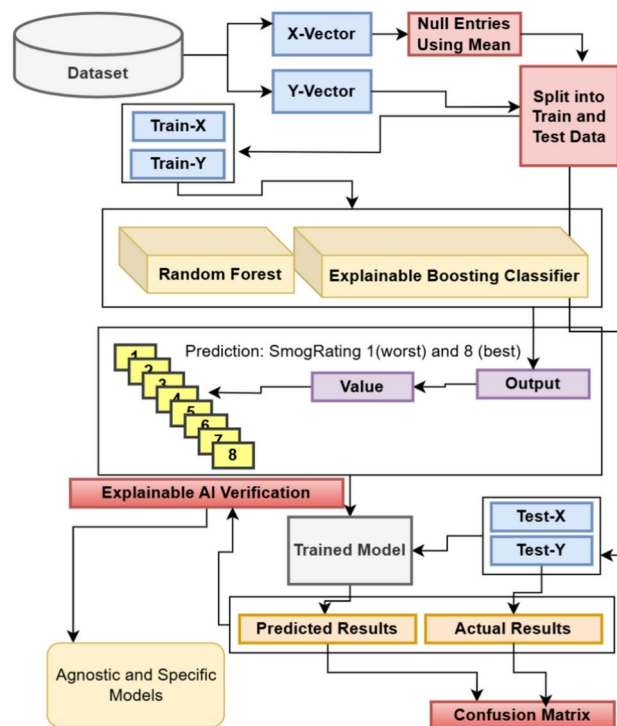
While ML models offer great potential due to their learning capabilities and ability to handle complex data, there are several challenges that hinder their performance. Data preprocessing, class imbalance, data splitting, and hyperparameter tuning are often poorly addressed, limiting the models' accuracy and generalization. Specifically, many studies report high accuracy for the more frequent classes in the dataset but struggle to achieve similar results for less represented classes. This leads to an illusion of high model performance, masking underlying issues with the data and model evaluation. Moreover, the optimization of ML models in terms of accuracy, sensitivity, and stability is heavily dependent on proper data handling and hyperparameter tuning, both of which have been inadequately addressed in many studies. Consequently, there remains a gap in the findings of existing ML-based air pollution studies, mainly due to suboptimal data management and model optimization<sup>26,27</sup>.

In the case of the proposed work the accuracy level is over 86%, so to properly evaluate the work, it is critical to mention some shortcomings. First of all, as a rule, higher accuracy looks rather promising, but, unfortunately, it does not always mean a high performance of a certain model on the presence of specifics such as imbalanced data or not enough tuned model. A model which has a very high accuracy could be a victim of overfitting the majority class or in other way the model can perform poorly when tested for unseen datasets. Thus, while using accuracy, other measures of evaluating the performance of the model should be also included and these include; precision, recall, and the F1-score where the data under consideration is imbalanced and has variations in distribution.

## Proposed work

The processing schema presented in Fig. 1 starts with data preparation, for which a dataset containing a features vector (X-vector) and a target vector (Y-vector)—in the case of smog rating will be used. Another data preprocessing step is related to the presence of some NA values in the dataset, or null values, and, therefore, they are filled by their mean value of the feature to cover the data set. The dataset is then divided into two subsets: There are two sets of data: training data; which are train X and train Y which are used to build the model and testing data: test X and test Y which are used to assess the model. This division is crucial to get the correct training and testing of the model as well. Next comes model training, where the training data is fed into two models: Random forest and Explanatory boosting classifier. The Random Forest is the type of ensemble learning method based on multiple functional decision trees to improve the predictive capacities. Likewise, EBC model, which is also an ensemble technique, adds decision trees one at a time and enhances the performance hence more interpretability. On the other hand, after training the models are employed in the prediction and explanation processes. With the Test-X data, the Random Forest and EBC models provide smog ratings of Test-X on a scale of 1 to 8, the worse smog being 1 and the better smog being 8.

To see if there is accuracy in our results, Explainable AI (XAI) methods are used to analyze and compare the algorithm's predicted values (Test-Y). This step assists in understanding the rationale of the models and their output in order to spot ingrown denial and unmask bias or mistakes. Last, model evaluation is performed employing confusion matrix which gives further details in regards to the accuracy of given models calculating the number of correct and misclassified values in each smog rating class. As well, the evaluation identifies the



**Fig. 1.** Proposed work for SmogRating prediction.

agnostic and special models for interpretation. Agnostic models provide the predictions without consideration of what goes on in the model to make these predictions while specific models involves optimization of the particular structure and parameter of the model, to provide a detailed answer as to why the particular decision was made.

### Dataset preparation

The dataset<sup>4</sup> consists of 27,000 rows and 15 columns, with a CSV file size of 2673KB. It measures the environmentally friendly performance of a vehicle with a keen emphasis on the Smog Rating that reflects the respiratory harming potential of a vehicle. The Smog Rating calculates the relative smog impacts of vehicles concerning nitrogen oxides and non-methane organic gases on a Smog Index ranging from 1 to 8. A rating of 8 shows the cleanliness of the automobile with less emission at all while a 1 shows high levels of emission of pollutants. Smog Rating is the opposite of scale, and a higher score is preferred, particularly where there is degradation of the environment within cities.

From the dataset, some observations emerge. For instance, all 2017 Acura models have a consistent Smog Rating of 6, reflecting moderate air quality impact. Despite variations in engine size, fuel consumption, and CO<sub>2</sub> emissions, their smog-forming emissions remain similar. In contrast, most 2023 Volvo models have a Smog Rating of 5, slightly worse than the Acura models. However, two exceptions, the XC60 B6 AWD and XC90 B6 AWD, achieve a Smog Rating of 7 due to advanced emissions-reduction technologies, despite their higher fuel consumption.

Smog Rating is affected by several factors. There is still much reliance on engines, which have elaborate systems such as the catalytic converter that greatly reduce emissions. Another factor is fuel type as cars that use hybrid or electricity have much lower emissions to those that use fuel. Engine displacement and power are other important parameters; the greater the engine displacement; the higher emissions of smog forming substances, despite advances in technology. Consumer ratings depend on vehicle class as well, since compact vehicles and hybrids are cleaner than, for instance, large SUVs and high-performance automobiles.

The Smog Rating extends the CO<sub>2</sub> Rating, which quantifies a vehicle's climate footprint over its lifetime by factoring in greenhouse emissions. While the CO<sub>2</sub> Rating aims at global warming, the Smog Rating is focused on short-term issues connected with air pollution and is thus important for large cities.

The major challenges for the dataset involve transforming it into a machine-readable format by addressing missing values, normalizing, and scaling. The following techniques were applied to continuous, categorical, and discrete columns to prepare the dataset for machine learning: handling missing data through imputation, applying normalization to continuous features to ensure they are within a consistent scale, and encoding categorical variables to make them interpretable by machine learning algorithms. These steps were essential to ensure the dataset was clean, consistent, and ready for model training.

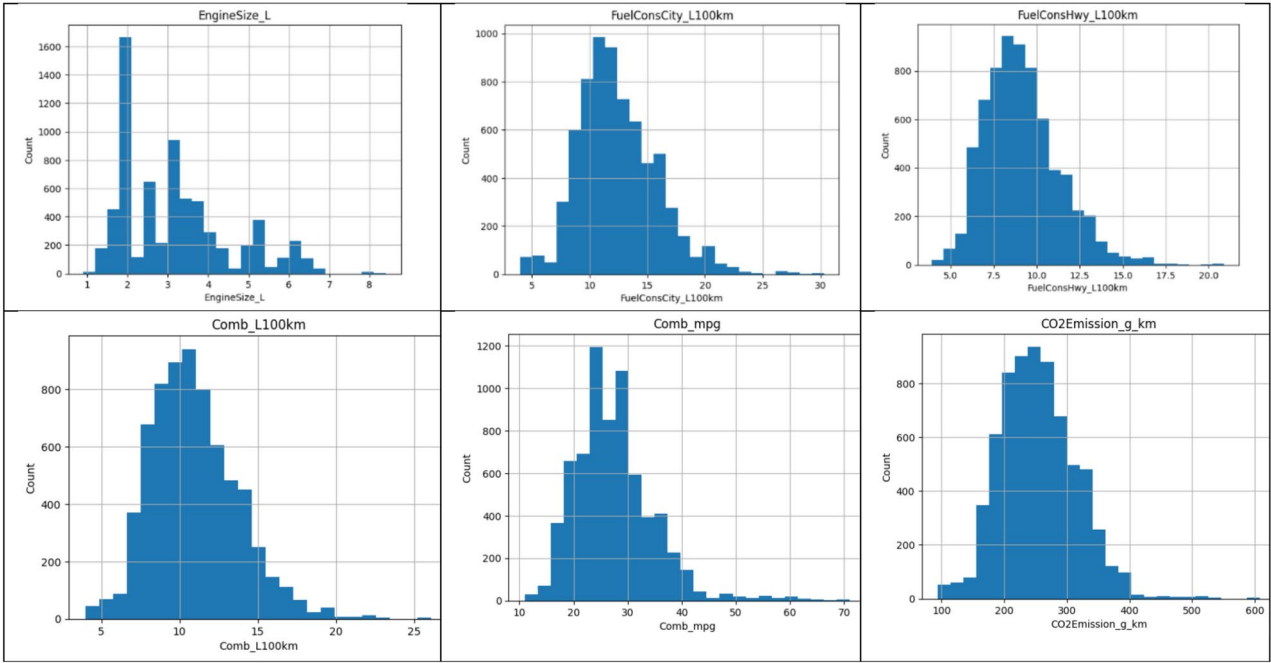
There are fifteen features in the dataset, however, they are classified into discrete, continuous and categorical data. These are the discrete variables; Model Year; Cylinders; CO<sub>2</sub> Rating; and Smog Rating, some of which is captured in Table 1 below.

ModelYear	Cylinders	CO2Rating	SmogRating
2017	4	7	6
2017	6	4	6
2017	6	5	6
2017	6	6	6
2017	6	4	6

**Table 1.** Discrete features in dataset.

EngineSize_L	FuelConsCity_L100km	FuelConsHwy_L100km	Comb_L100km	Comb_mpg	CO2Emission_g_km
2.4	9.4	6.8	8.2	34	192
3.5	12.6	9	11	26	259
3.5	12.2	9	10.7	26	251
3	9.1	9	9	31	210
3.5	11.1	10.8	11	26	261

**Table 2.** Continuous features in dataset.



**Fig. 2.** Normal distribution of continuous features.

The continuous variables including EngineSize\_L, FuelConsCity\_L100km, FuelConsHwy\_L100km, Comb\_L100km, Comb\_mpg and CO2Emission\_g\_km are described in Table 2 with example data.

All continuous features from Table 2 shown to be normally distributed using appropriate statistical methods and visualizations in Fig. 2.

There are several categorical variables in the dataset, such as, models, vehicle classes, transmits, and fuel types, which are crucial when studying vehicles. These features are described in Table 3, while the encoded value is included in Table 4. Of them, the ‘Make’ feature contains 74 subcategories, while an option named Model has been found to contain as many as 2078 subcategories. The VehicleClass feature includes 30 categories, reflecting various types of vehicles, and Transmission accounts for 26 different categories. Lastly, FuelType has 4 categories, representing the range of fuel options used by the vehicles. This detailed categorization underscores the dataset’s complexity and richness in capturing diverse vehicle attributes.

The target variable, SmogRating, originally consists of 8 classes, though 4 are missing in the dataset. This challenge was addressed by handling 7 classes within the 1–8 rating range as provided as shown in Fig. 3.

This dataset offers a rich resource for analyzing vehicle emissions, highlighting the relationships between vehicle features, environmental impact, and performance over nearly three decades.

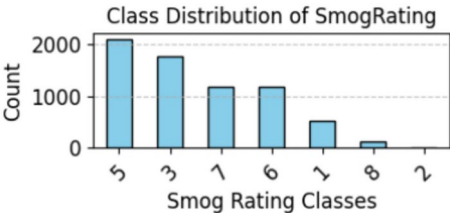


Make	Model	VehicleClass	Transmission	FuelType
Acura	ILX	Compact	AM8	Z
Acura	MDX SH-AWD	SUV—Small	AS9	Z
Acura	MDX SH-AWD elite	SUV—Small	AS9	Z
Acura	MDX HYBRID AWD	SUV—Small	AM7	Z
Acura	NSX	Two-seater	AM9	Z

**Table 3.** Categorical features in dataset.

Make	Model	VehicleClass	Transmission	FuelType
1	1055	1	10	4
1	1228	19	17	4
1	1230	19	17	4
1	1226	19	9	4
1	1313	27	11	4

**Table 4.** Encoding of Categorical Features.



**Fig. 3.** Target variable distribution.

**Embedding techniques**

Embedding techniques in machine learning refer to methods used to represent categorical variables or textual data as numerical vectors, which are suitable for training models. In this case, categorical data was transformed using Label Encoding, where each category is assigned a unique integer. This method allows models to interpret categorical features effectively. For continuous numerical data, Standard Scaling is often employed to normalize values, ensuring that features are on a comparable scale, improving model performance. SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance class distributions, generating synthetic examples to prevent model bias towards the majority class.

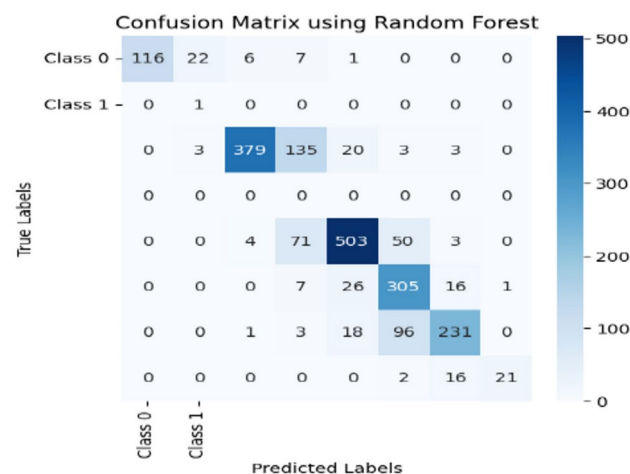
**Parameters for fine tuning of model**

In fine-tuning the model, various steps and parameters were leveraged to enhance the model’s accuracy and generalization. Random Forest Regressor was employed as the base model, with the data being split into training and test sets using the train\_test\_split function (test\_size=0.3, random\_state=42) to ensure that the model could be evaluated on unseen data. The model was initialized with RandomForestRegressor(random\_state=42) to ensure reproducibility. Hyperparameters such as the number of trees (n\_estimators), maximum depth (max\_depth), and minimum samples required to split a node (min\_samples\_split) were tuned to prevent overfitting and improve model performance. The model was trained on the balanced dataset, which was achieved using SMOTE (Synthetic Minority Over-sampling Technique) to address any class imbalance. After training, the model’s performance was evaluated using metrics like Mean Squared Error (MSE), R-Squared (R2 Score), Mean Absolute Error (MAE), Explained Variance Score (EVS), and Max Error, which provide insights into the model’s accuracy and prediction capabilities. Finally, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) were employed to explain the model’s predictions, helping interpret which features contributed the most to the model’s output, thereby enhancing model transparency and trust.

**Prediction using machine learning model on unbalanced data**

A complete flow chart for the Machine Learning Model on Unbalanced Data is provided in Table 5. When the preprocessed dataset was applied to the Random Forest model, the results were not satisfactory, as highlighted in the evaluation metrics. The overall accuracy achieved was 75%, as shown in Table 6, which falls short of expectations for reliable model performance. Furthermore, the confusion matrix in Fig. 4 illustrates that the diagonal, representing correctly classified instances, is not strong, indicating significant misclassifications across several classes.

Class	Precision	Recall	F1-score	Support
1	1.00	0.76	0.87	152
2	0.04	1.00	0.07	1
3	0.97	0.70	0.81	543
4	0.00	0.00	0.00	0
5	0.89	0.80	0.84	631
6	0.67	0.86	0.75	355
7	0.86	0.66	0.75	349
Accuracy	0.75			



When the preprocessed dataset was applied to the Explainable Boosting Classifier, the results were not satisfactory, as depicted in Table 7, where the accuracy achieved was 71%. Also, the confusion matrix in Fig. 6 reveals a weak diagonal showing that the classifier had difficulty in identifying the exact class of several instances. Moreover, Fig. 7 also reveals low values of the coefficient of determination which represents the extent of accuracy of the line prediction of the model; besides this, there are low values of correlation coefficients from the regression line showing that the model cannot predict the actual values as shown below. Such results imply

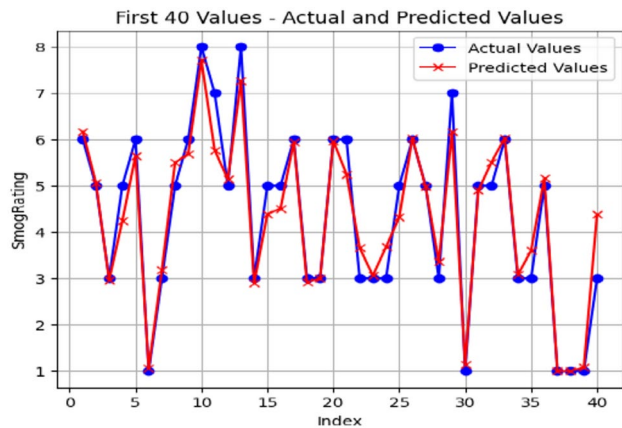


Fig. 5. First 40 actual and predicted values from RF on unbalanced dataset.

Class	Precision	Recall	F1-score	Support
1	0.89	0.80	0.84	152
2	1.00	1.00	1.00	1
3	0.74	0.72	0.73	543
5	0.68	0.72	0.70	631
6	0.69	0.63	0.66	355
7	0.65	0.69	0.67	349
8	0.67	0.62	0.64	39
Accuracy	0.71			

Table 7. Explainable-boosting-classifier results based on unbalanced dataset.

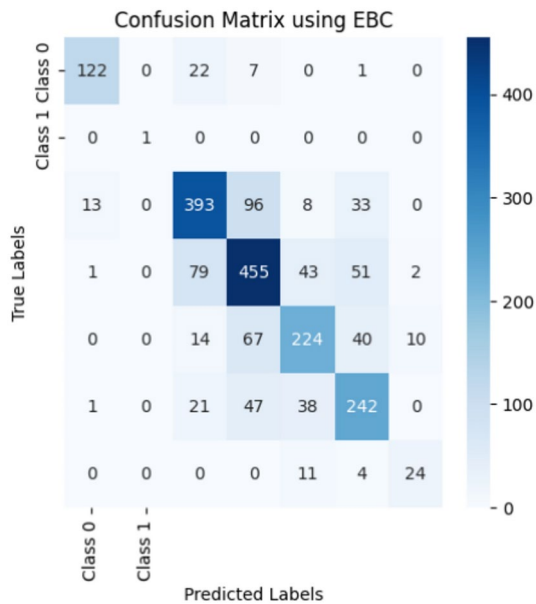


Fig. 6. Explainable-boosting-classifier confusion matrix based on unbalanced dataset.

that the used model could provide suboptimal performance, and further enhancement may be necessary for the preprocessing stage, feature extraction or selection, or model identification.

Prediction using machine learning model on balanced dataset

Complete flow chart for Machine Learning Model on Unbalanced Data is provide in Table 8.



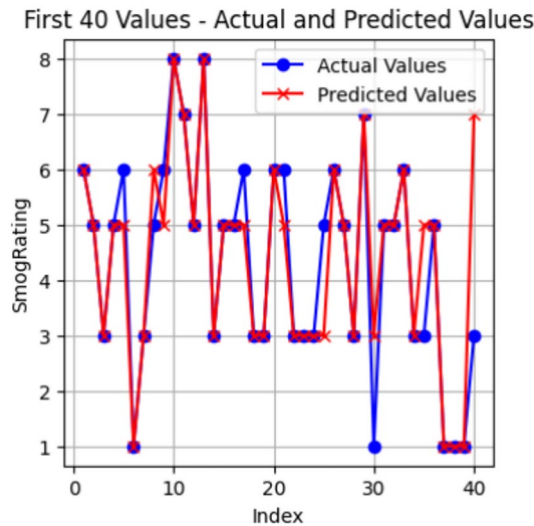


Fig. 7. First 40 Actual and predicted values from EBC on unbalanced dataset.

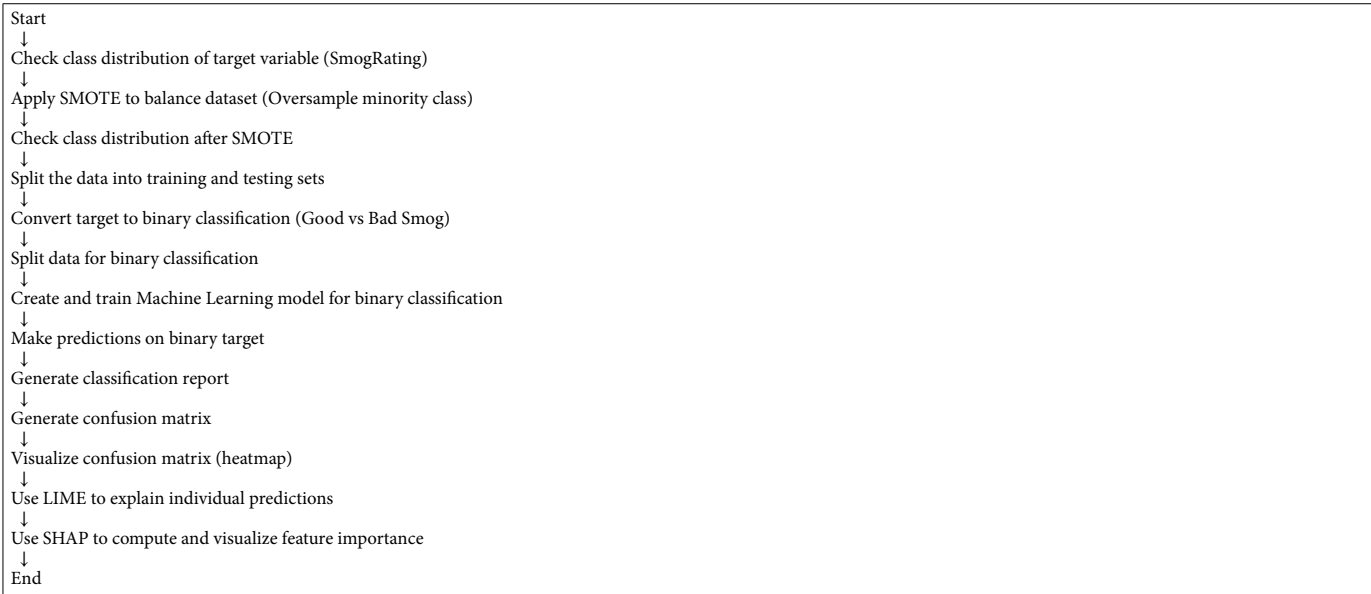


Table 8. Flow chart for machine learning model on balanced data.

The use of SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the class distribution in the dataset. Before balancing, classes like SmogRating 5 had 2106 instances, while others such as SmogRating 2 and SmogRating 8 were severely underrepresented with only 7 and 117 instances, respectively. After SMOTE was applied, synthetic samples were generated for the underrepresented classes, resulting in a balanced distribution where each class had 2,106 instances. The class distribution before and after balancing can be visualised in Table 9, which highlights the shift from imbalance to equal representation across all classes. Figure 8 shows the unbalanced and balanced class.

After applying the balanced dataset to the Random Forest model, the results were the best, with an accuracy of 86%, as shown in Table 9. The confusion matrix in Fig. 9 exhibits a strong diagonal, indicating a high level of correct classifications across all classes. Additionally, Fig. 10 displays the first 40 actual and predicted values, with a near-perfect overlap, highlighting the model's accurate predictions for the sample dataset. These results demonstrate the effectiveness of using a balanced dataset to improve model performance. Table 9 shows the EBC Results based on Balanced Dataset. Figure 11 illustrates the EBC Confusion Matrix based on Balanced Dataset and Fig. 12 shows the First 40 Actual and Predicted Values from EBC on Balanced Dataset.

After applying a balanced dataset on Explainable-Boosting-Classifier (EBC), the achieved results are best with an accuracy of 86% as shown in Table 7 and the diagonal of Fig. 10 is showing best. Figure 11 shows the first

Class	Precision	Recall	F1-Score	Support
1	1.00	0.94	0.97	667
2	0.96	1.00	0.98	662
3	0.96	0.66	0.78	617
4	0.00	0.00	0.00	0
5	0.86	0.77	0.81	646
6	0.74	0.89	0.81	569
7	0.89	0.82	0.85	651
Accuracy	0.86			

Table 9. Random-forest results based on balanced dataset.

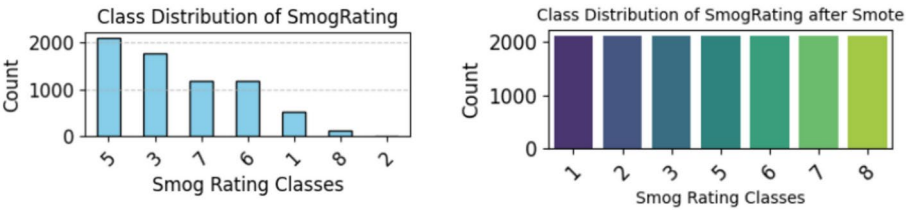


Fig. 8. Unbalanced and balanced classes.

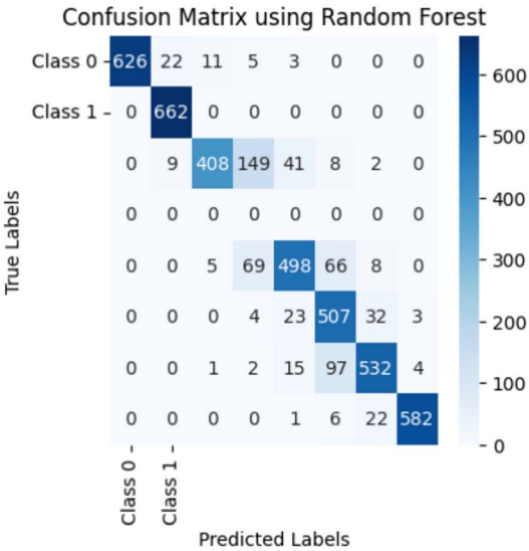


Fig. 9. Random-forest confusion matrix based on balanced dataset.

40 actual and predicted values best overlapping representing samples for all datasets. Table 10 shows the results of the EBC balanced dataset.

Results and discussion with explainable AI

The results demonstrate a significant performance improvement when using the balanced dataset for fog prediction with the Random Forest model. For the unbalanced dataset, the models (RF and EBC) were trained on train X (4830 samples, 14 features) and tested on test X (2070 samples, 14 features), with corresponding target variables train y (4830 samples) and testy (2070 samples). The models achieved reasonable scores for the unbalanced dataset but showed limitations, particularly in terms of explained variance and mean absolute error. However, when the dataset was balanced using SMOTE, the models' performance improved dramatically. In the case of the balanced dataset, the training data consisted of train X (10,319 samples, 14 features) and train y (10,319 samples), with testing data of test X (4423 samples, 14 features) and testy (4423 samples). The models achieved much higher R-squared values, lower errors, and greater overall accuracy with the balanced dataset. These results demonstrate the effectiveness of balancing the dataset in improving models' performance and highlighting the importance of using balanced data for tasks that require accurate and reliable predictions

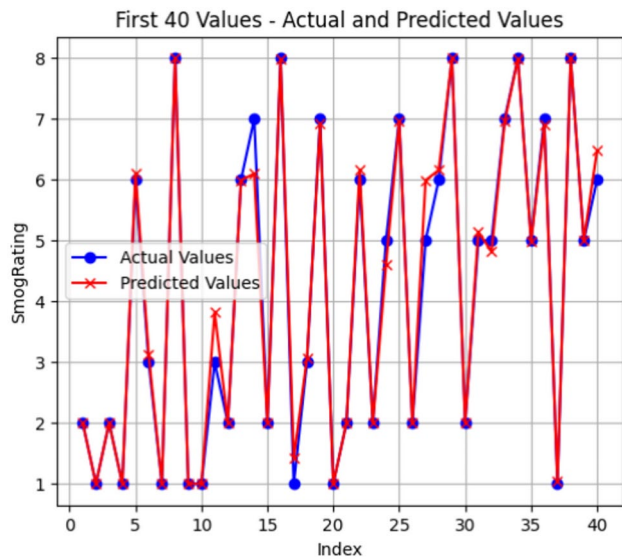


Fig. 10. First 40 actual and predicted values from random-forest on balanced dataset.

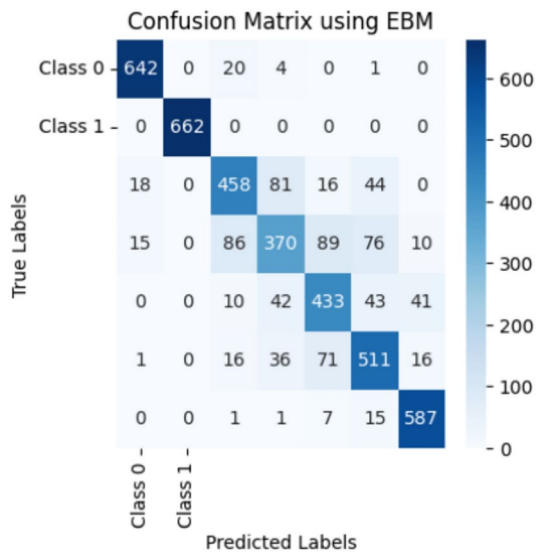


Figure11. EBC confusion matrix based on balanced dataset.

as shown in Table 11 (visually shown in Fig. 13) and Table 12 (visually shown in Fig. 14). Consequently, for explainable AI, the balanced dataset results are considered more representative and will be discussed in detail due to their superior predictive capability.

Explainable AI analysis on RF

Actual and predicted values from TextX are shown in Table 13. For example, record 9512 from TextX has an actual smog rating of 2, and the predicted value is also round as 2. Similarly, for record 11,210, the actual smog rating is 6, and the predicted value is also 6 according to the Random Forest (RF) model on the balanced dataset. These results demonstrate the model’s ability to make accurate predictions, with only a slight difference observed in some cases. The process through which these values are calculated is further explained in the following Explainable AI analysis.

Analysis with LIME explainer on RF

Here, the class labels 1–8 are presented in Table 14, and the few records from Table 14 are explained below. For instance, record 9512 from Xtest has an actual smog rating of 2, and the predicted value is also 2. Using LIME (Local Interpretable Model-agnostic Explanations), this prediction is achieved, as shown in Fig. 15. In this case, the most influential positive feature is ModelYear, which has a significant impact on the prediction.

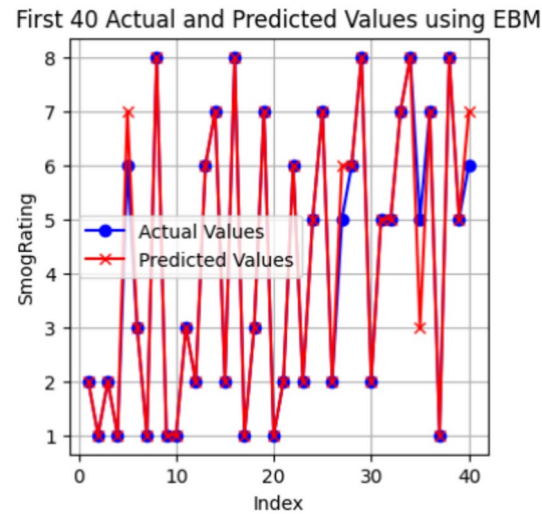


Fig. 12. First 40 actual and predicted values from EBC on balanced dataset.

Smog rating	Precision	Recall	F1-score	Support
1	0.95	0.96	0.96	667
2	1.00	1.00	1.00	662
3	0.77	0.74	0.76	617
5	0.69	0.57	0.63	646
6	0.70	0.76	0.73	569
7	0.74	0.78	0.76	651
8	0.90	0.96	0.93	611
Accuracy	0.83			

Table 10. EBC results based on balanced dataset.

Metric	Unbalanced dataset	Balanced dataset
Mean Squared Error (MSE)	0.3896	0.2269
R-Squared (R2 Score)	0.8775	0.9624
Mean Absolute Error (MAE)	0.3656	0.2104
Explained Variance Score (EVS)	0.8775	0.9625
Max Error	4.4200	4.3500

Table 11. Results from RF using balanced and unbalanced data.

In 14, the explanation provided by LIME (Local Interpretable Model-agnostic Explanations) shows the contribution of different features towards the prediction of the smog rating for record 9512. The figure is divided into negative and positive impacts on the predicted value, where the positive side shows features that increase the predicted smog rating, and the negative side shows those that decrease it.

The most influential feature in this case is ModelYear, which has a positive impact on the prediction. This suggests that a higher ModelYear (newer models) contributes to a high smog rating. Other features, such as VehicleClass, Co2Emission, and EngineSize, also contribute to the prediction with a negative impact.

Figure 16 visualizes how each feature influences the prediction of smog ratings for record 4064 from testX, highlighting the relative importance of different variables in determining the final predicted value of 4, while the actual value is 3. The figure illustrates the contributions of various features, with positive and negative impacts on the predicted smog rating. Each feature's contribution is shown in terms of how it either increases or decreases the predicted smog rating, providing insights into the model's decision-making process.

The Fig. 17 illustrates a LIME XAI model predicting a value of 6, closely matching the actual value of 6. Key influencing features include VehicleClass (importance score of 0.48), EngineSize\_L (score of 0.32), and CO2Rating (score of 0.29), all of which positively contribute to the predicted value. In contrast, Make (score of -0.20), ModelYear (score of -0.15), and Comb\_L100km (score of -0.18) have negative impacts. These feature

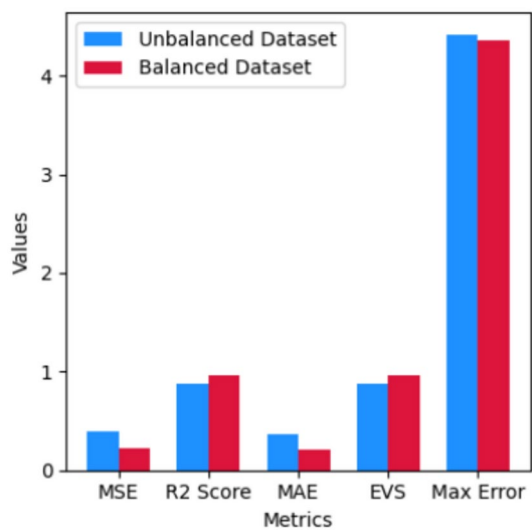


Fig. 13. Comparison of metrics for RF using balanced and unbalanced data.

Metric	Unbalanced dataset	Balanced dataset
Mean Squared Error (MSE)	1.3469	0.7755
R-Squared (R2 Score)	0.5763	0.8715
Mean Absolute Error (MAE)	0.5710	0.3278
Explained Variance Score (EVS)	0.5770	0.8721
Max Error	6.0000	6.0000

Table 12. Results from EBC using Balanced and Unbalanced Data.

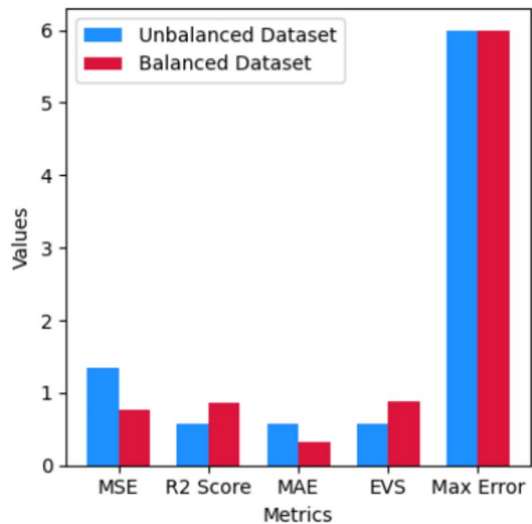


Fig. 14. Comparison of metrics for EBC using balanced and unbalanced data.

importance scores reveal how each vehicle attribute affects the model's prediction, providing transparency in how the prediction is derived.

Analysis with SHAP explainer on RF

To verify the results of the Random Forest model, SHAP (SHapley Additive exPlanations) analysis has been applied to records 9512 and 4064 from the test set. Figure 18 illustrates the predicted value for record 9512, which is 2, while Fig. 19 shows the predicted value for record 4064, which is 3.82 (rounded to 4). In both figures,

Serial No	Record No	Actual	Predicted smog rating
1	9512	2	2.0 round off 2
2	8014	1	1.0 round off 1
3	8596	2	2.0 round off 2
4	1714	1	1.0 round off 1
5	11,210	6	6.11 round off 6
6	6205	3	3.12 round off 3
7	8233	1	1.0 round off 1
8	13,783	8	8.0 round off 8
9	7665	1	1.0 round off 1
10	4064	3	3.82 round off 4
11	7273	1	1.0 round off 1
12	9994	2	2.0 round off 2
13	11,039	6	5.99 round off 6
14	5304	7	6.1 round off 6
15	10,122	2	2.0 round off 2

**Table 13.** First fifteen actual and predicted values from TestX using RF.

Serial no	Record no	Actual	Predicted smog rating
1	9512	2	2
2	8014	1	1
3	8596	2	2
4	1714	1	1
5	11,210	6	7
6	6205	3	3
7	8233	1	1
8	13,783	8	8
9	7665	1	1
10	7273	1	1
11	4064	3	3
12	9994	2	2
13	11,039	6	6
14	5304	7	7
15	10,122	2	2

**Table 14.** First fifteen actual and predicted values from TestX using EBM.

the base value is 4.596, and the positive and negative impacts of each feature on the final prediction are displayed, highlighting how individual features influence the model’s output. These visualizations offer insights into the model’s decision-making process, confirming the predictions made by the RF.

This SHAP Fig. 20 illustrates the contributions of various features to the model’s predicted value of 5.99, compared to the base value of 4.506. Features that positively influence the prediction are shown in blue, such as CO2Rating (contributing + 1.2) and VehicleClass (contributing + 0.9), while features that negatively influence the prediction are shown in pink, such as Make (contributing – 0.5) and ModelYear (contributing – 0.3). The position of each feature along the axis indicates its impact on the final prediction: features to the right increase the predicted value, and those to the left decrease it. This visualization helps understand how specific vehicle characteristics affect the model’s output, enhancing transparency and interpretability.

*Analysis with SHAPASH explainer on RF*

As shown in Fig. 21, two records have been selected from the SHAPash SmartExplainer (the full analysis is provided on GitHub). The predicted values for these records are 5.7 (rounded to 6) and 5.98 (rounded to 6), respectively. The impact of each feature on these predictions is visualized in Figs. 22 and 23, where the influence of positive and negative features on the final predictions is depicted. These figures provide further insights into how different features contribute to the model’s decision-making process for these specific records.

**Explainable AI analysis on EBC.**

Actual and predicted values from TextX are shown in Table 13. For example, record 9512 from TextX has an actual smog rating of 2, and the predicted value is also 2. Similarly, for record 11,210, the actual smog rating is



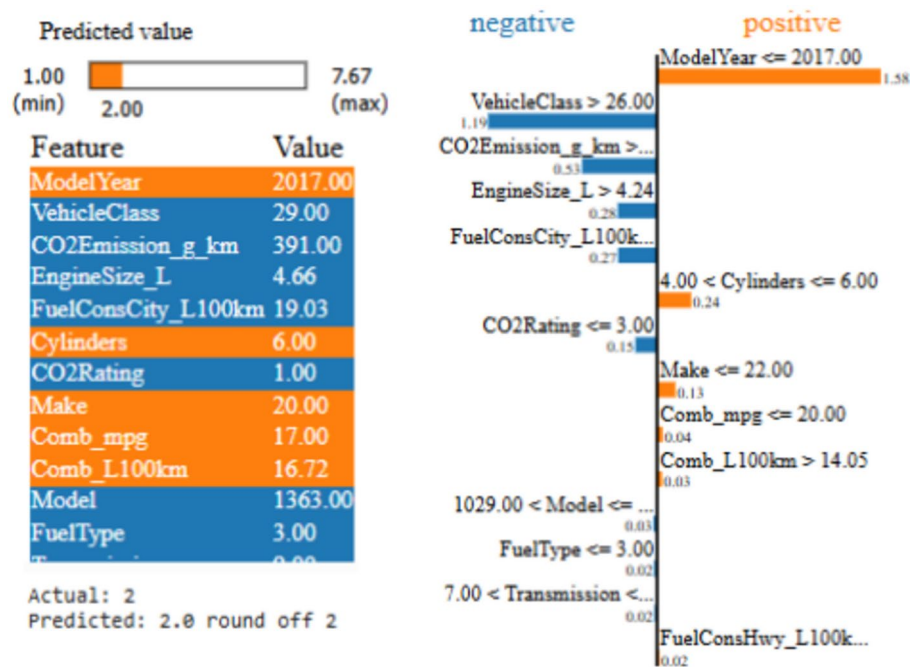


Fig. 15. Explanation for Record 9512 using LIME on RF.

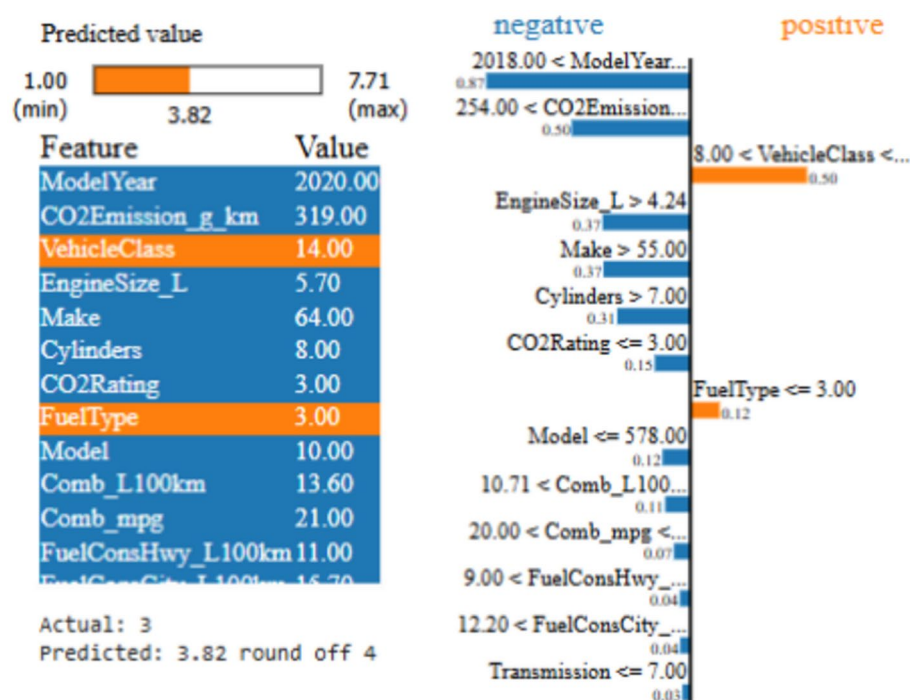


Fig. 16. Explanation for Record 4064 using LIME on RF.

6, and the predicted value is 7, which is nearly close to the actual rating of 6 according to the EBC model on the balanced dataset. These results demonstrate the model's ability to make accurate predictions, with only a slight difference observed in some cases. The process through which these values are calculated is further explained in the following Explainable AI analysis.

#### Analysis with LIME explainer on EBC

Here, the class labels 1–8 are presented in Table 14, and the few records from Table 14 are explained below. For instance, record 8014 from Xtest has an actual smog rating of 1, and the predicted value is also 1. Using LIME

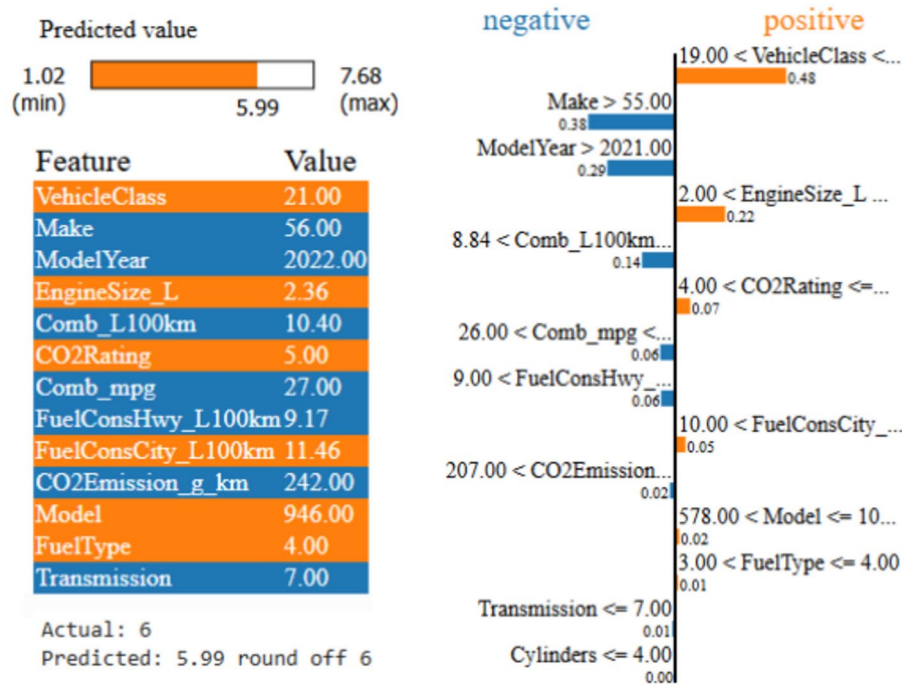


Fig. 17. Explanation for Record 11039 using LIME on RF.



Fig. 18. Explanation for Record 9512 using SHAP on RF.



Fig. 19. Explanation for Record 4064 using SHAP on RF.



Fig. 20. Explanation for Record 11039 using SHAP on RF.

(Local Interpretable Model-agnostic Explanations), this prediction is achieved, as shown in Fig. 24. In this case, the most influential negative feature is ModelYear, which has a significant impact on the prediction.

In Fig. 24, the explanation provided by LIME (Local Interpretable Model-agnostic Explanations) shows the contribution of different features towards the prediction of the smog rating for record 8014. The figure is divided into negative and positive impacts on the predicted value, where the positive side shows features that increase the predicted smog rating, and the negative side shows those that decrease it.

The most influential feature in this case is ModelYear, which hurts the prediction. This suggests that a higher ModelYear (newer models) contributes to a lower smog rating (better air quality). Other features, such as

_index	_predi	ModelYear	Make	Model	VehicleClass	EngineSize_L	Cylinders	Transmissior	FuelType
14	5.73	2017	3	678	6	5.2	12	6	4
17	5.7	2017	3	1906	27	6	12	9	4
59	5.5	2017	9	478	18	6	12	16	4
73	7.44	2017	10	42	17	2	4	16	1
80	6	2017	10	53	1	3	6	25	4
96	5.98	2017	10	105	1	4.4	8	16	4
144	6.11	2017	11	736	19	2.5	4	4	3
150	6.34	2017	15	338	1	2	4	16	4

Fig. 21. Screen Shot from SmartExplainer on RF.

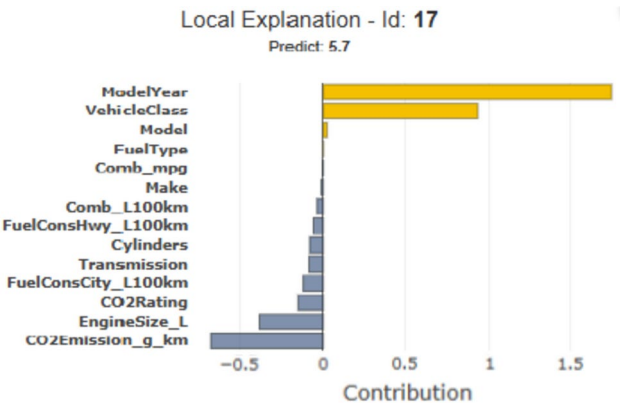


Fig. 22. Contribution of each feature for Prediction using SmartExplainer on RF for ID-17.

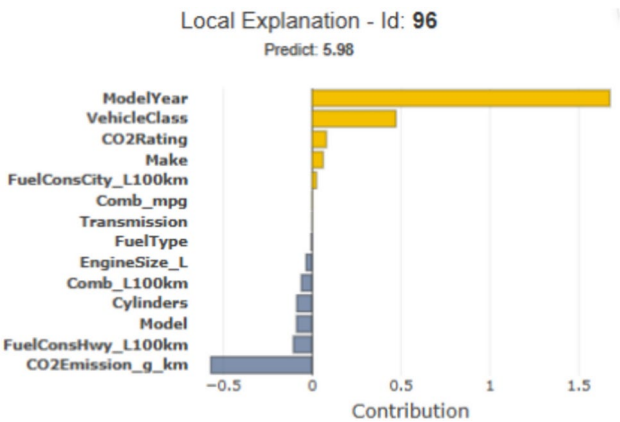


Fig. 23. Contribution of each feature for Prediction using SmartExplainer on RF for ID-96.

VehicleClass, FuelType, and Transmission, also contribute to the prediction, but with lesser impacts compared to ModelYear.

Figure 25 visualizes how each feature influences the prediction of smog ratings for record 11,210 from test X, highlighting the relative importance of different variables in determining the final predicted value of 7, while the actual value is 6. The figure illustrates the contributions of various features, with positive and negative impacts on the predicted smog rating. Each feature's contribution is shown in terms of how it either increases or decreases the predicted smog rating, providing insights into the model's decision-making process.

Analysis with INTERPRETER explainer on EBC

Figure 26 provides a local explanation for the 5th record in the dataset using the Explainable Boosting Classifier (EBC). The chart demonstrates the contribution of various features to the prediction of SmogRating for this

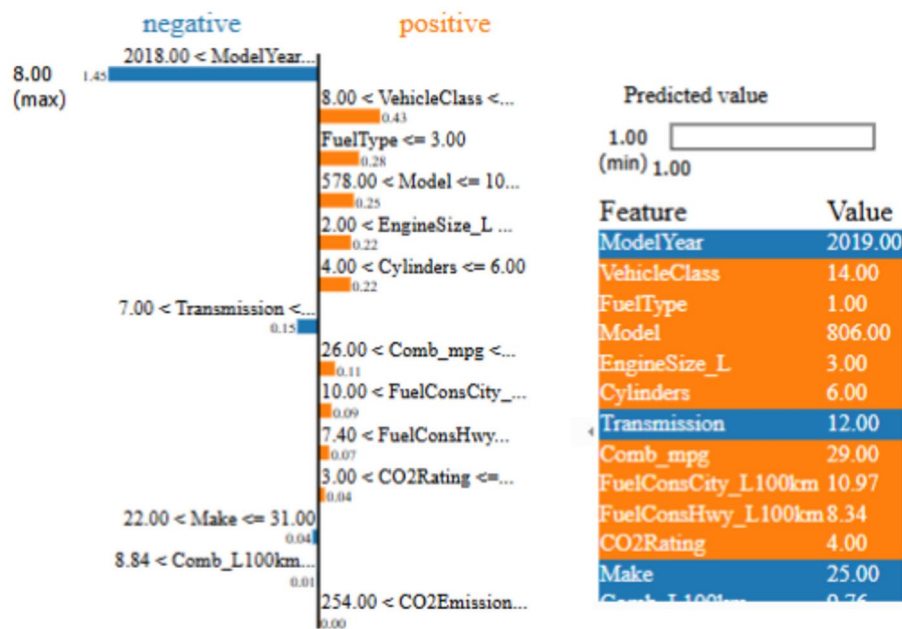


Fig. 24. Explanation for Record 8014 using LIME on EBC.

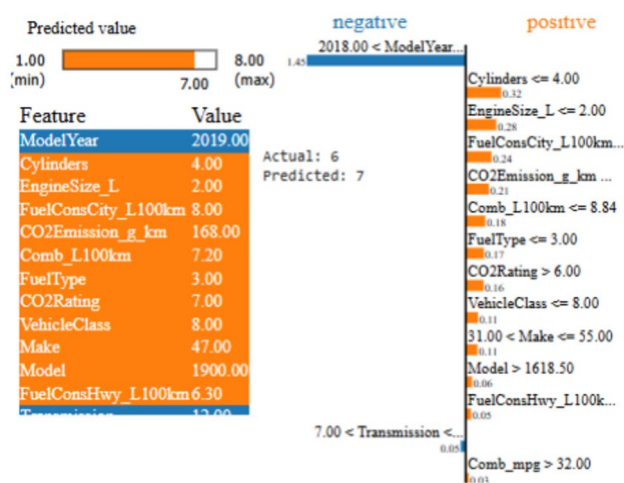


Fig. 25. Explanation for Record 11210 on EBC using LIME.

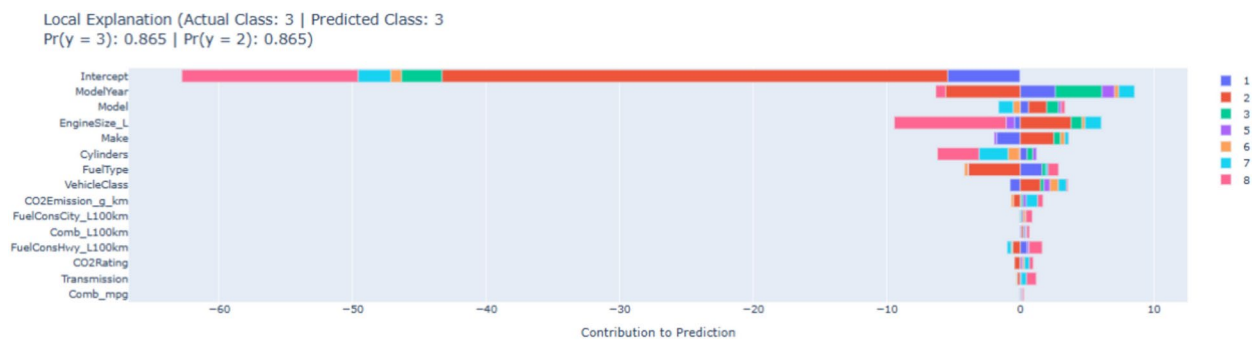
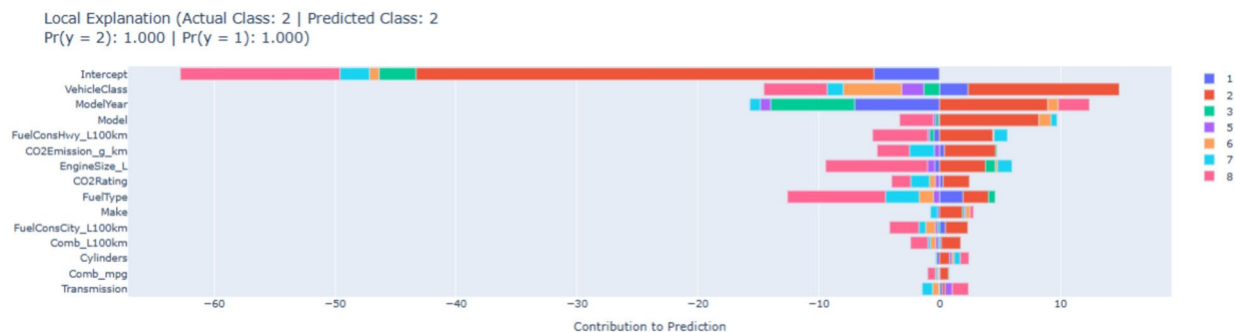
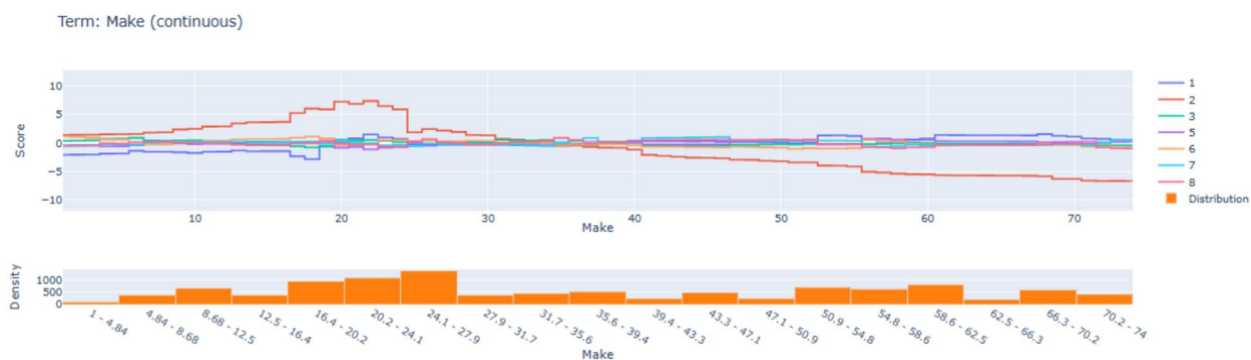


Fig. 26. Explanation for Record-5 on EBC using Local Interpreter.



**Fig. 27.** Explanation for Record-20 on EBC using Local Interpreter.



**Fig. 28.** Global Interpretation of Feature 'Make' Impact on All Classes in EBC Using Global Interpreter.



**Fig. 29.** Global Interpretation of Feature 'Model' Impact on All Classes in EBC Using Global Interpreter.

specific record. The actual SmogRating is 3, and the predicted class is also 3, with a probability of 0.865. Features such as "ModelYear," "EngineSize\_L," and "Make" significantly influence the prediction, either positively or negatively, as represented by the respective bar lengths and directions. This detailed breakdown helps in understanding the individual feature impact for this specific case. Similarly, Fig. 27 provides the local explanation for record 20.

The Fig. 28 shows the global impact of the "Model of Vehicles" feature on all smog rating classes (1 to 8). The X-axis represents vehicle models, and the Y-axis shows their contribution (score) to the prediction. Lines above 0 indicate a positive impact, increasing the likelihood of a class, while lines below 0 indicate a negative impact, reducing it. The variation in the lines demonstrates how different vehicle models influence predictions for each class differently, highlighting both positive and negative effects. Similarly Fig. 29 showing the impact of feature 'Make' for all smog ratings prediction.

Figure 30 shows the global importance of all features, with longer bar of ModelYear indicating features that contribute more significantly to the model's predictions. It highlights which features are most influential in driving the model's decisions.



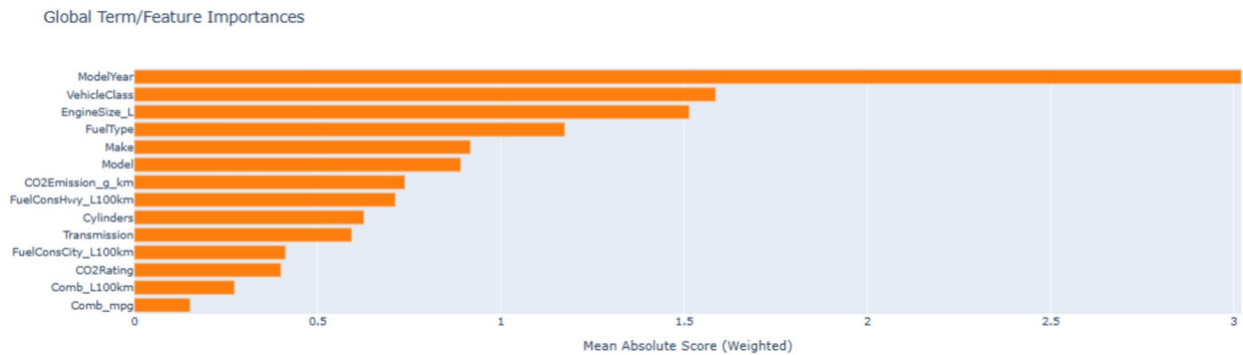


Fig. 30. Global Explanation of the Importance of All Features on EBC Using a Global Interpreter.

Model	MSE	R-Squared	MAE	EVS	Max Error	RMSE
Work <sup>18</sup>	–	–	13.2	–	–	20.8
Work <sup>20</sup>	–	–	7.53	–	–	9.86
Work <sup>21</sup>	–	–	23.97	–	–	35.82
Work <sup>28</sup>	–	–	14.6344	–	–	24.22874
Work <sup>29</sup>	5.5	0.73	1.73			2.35
Proposed work (RF using SMOTE)	0.2269	0.9624	0.2104	0.9625	4.3500	0.4763

Table 15. Performance metrics of the smog rating prediction model.

Comparison with similar studies

To compare the performance of different models, various metrics were used to assess accuracy and prediction capabilities. The metrics include Mean Squared Error (MSE), R-Squared (R<sup>2</sup> Score), Mean Absolute Error (MAE), Explained Variance Score (EVS), Max Error, and Root Mean Squared Error (RMSE). These metrics offer a comprehensive view of how well the models performed in terms of error minimization and predictive power.

The Proposed Work (RF using SMOTE) yields significantly better results than the LSTM model used in<sup>18</sup>. The Proposed Work has much lower MAE (0.2104) and MSE (0.2269), indicating higher accuracy and precision in its predictions compared to the LSTM’s MAE of 13.2 and RMSE of 20.8, which are considerably higher. Also, Proposed Work yields a high coefficient of determination at approximately 0.9624 and an Explained Variance Score (EVS) of about 0.9625 which underscore the qualifier factor that the Proposed Work explains the major variance of data well.

The results of the Proposed Work show that the models implemented with the RF using SMOTE outperform the models used in<sup>20</sup> and<sup>21</sup>. The Proposed Work yields an MSE of 0.2269 and an R<sup>2</sup> Score of 0.9624 which shows the model is good fit in predicting smog ratings by using characteristics of vehicle. Likewise, MAE of 0.2104 and the EVS of 0.9625 prove high levels of accuracy and dependability of the model. Compared with the models in<sup>20</sup> and<sup>21</sup> focusing on PM2.5 predictions (Bi-LSTM and LSTM-FC), the errors are larger with MAE = 7.53 and RMSE = 9.86. This means that the Proposed Work yields more accurate predictions and superior performance in general.

The Proposed Work using Random Forest (RF) with SMOTE has better accuracy than the CNN-LSTM model employed in<sup>28</sup>. The Proposed Work’s MSE of 0.2269, R<sup>2</sup> Score of 0.9624 and MAE of 0.2104 signify high model fitting and accurate prediction. A high value of Explained Variance Score (EVS) of 0.9625 and a low Max Error of 4.3500 substantiates the above finding of the proposed model. The proposed model outperforms<sup>29</sup> in terms of predictive accuracy and error metrics. It uses Random Forest with SMOTE and Explainable Boosting Classifier for improved prediction of vehicle-induced smog, achieving an impressive R-Squared of 0.9624 and an accuracy of 86%. In comparison,<sup>29</sup>’s Extra Tree model, while effective, shows lower performance with R<sup>2</sup> of 0.71 and higher error rates (MSE of 5.91, RMSE of 2.43). Additionally, the proposed model integrates both agnostic and specific explainability techniques like SHAP and LIME, offering a more comprehensive understanding of the model’s decisions, whereas<sup>29</sup> limits its explainability to SHAP for the Extra Tree model. All of the performance measures are reported in Table 15.

Table 16 shows the comparisons of accuracy of other predictive models, investigated in prior studies and the present work. According to the research reported by<sup>22</sup>, the high accuracy achieved was 81%, indicating that recurrent neural networks were best suited to learning by sequentially modelling their input data. Linear Regression (LR), studied by<sup>30</sup>, showed a more modest accuracy of 68%, which is typical for simpler models when dealing with complex datasets. Stochastic Gradient Descent (SGD), also from<sup>30</sup>, resulted in an accuracy of 66%, reflecting its performance in optimizing linear models, but with less accuracy compared to more advanced methods. Random Forest Regression (RFR), as evaluated in<sup>30</sup>, performed better with an accuracy of 72%, highlighting its strength in capturing non-linear patterns and interactions in data. Similarly, Decision Tree



Model	Accuracy (%)
Recurrent Neural Networks (RNN) <sup>22</sup>	81
Linear Regression (LR) <sup>30</sup>	68
Stochastic Gradient Descent (SGD) <sup>30</sup>	66
Random Forest Regression (RFR) <sup>30</sup>	72
Decision Tree Regression (DTR) <sup>30</sup>	71
Support Vector Regression (SVR) <sup>30</sup>	70
Seasonal Autoregressive Integrated Moving Average (SARIMA) <sup>25</sup>	43
Hybrid Artificial Neural Network (Hybrid ANN) <sup>31</sup>	70–83
XGBoost <sup>32</sup>	83
Proposed work	86

**Table 16.** Comparison of proposed model with similar studies.

Regression (DTR), also from<sup>30</sup>, showed an accuracy of 71%, indicating good performance but with potential overfitting issues common in decision trees. Support Vector Regression (SVR), reported in<sup>30</sup>, achieved an accuracy of 70%, performing comparably to DTR but slightly lower. Seasonal Autoregressive Integrated Moving Average (SARIMA), studied by<sup>25</sup>, had the lowest accuracy at 43%, likely due to its limitations in handling non-linear and complex data patterns typically found in environmental data. Hybrid Artificial Neural Networks (Hybrid ANN) as pointed out by<sup>31</sup>, exhibited mean accuracy of between 70 and 83% which indicates that there is variation ability between the various models and possibly with the kind of data fed into the model. Last but not least, the Proposed Work fetched a very high accuracy of 86% the highest compared to all the other models that have been used for smog rating prediction and indeed proves the efficacy of the proposed method. Thus, it shows that the proposed approach is the more accurate and reliable solution rather than the models presented in the literature. The proposed model outperforms the model in<sup>32</sup> with an accuracy of 86% compared to 83%. It utilizes Random Forest and Explainable Boosting Classifier models with SMOTE, offering more precise predictions for vehicle smog contributions, making it more reliable for environmental analysis. Unlike<sup>32</sup>, which focuses on general air quality predictions, the proposed model is tailored to individual vehicle contributions, providing actionable insights for environmental management. Additionally, the proposed model incorporates comprehensive explainable AI techniques, enhancing transparency and clarity in predictions, making it a more timely and specialized solution.

**Potential for real-time application**

The proposed model demonstrates strong performance with real-time or larger datasets, as it leverages the SMOG dataset and achieves an accuracy of 86% through the use of Random Forest and Explainable Boosting Classifier models, augmented by SMOTE for dataset balancing. The integration of Explainable AI (XAI) techniques like LIME, SHAP, and SHAPash adds transparency, allowing for a better understanding of the model’s decision-making process. The model is generalizable, enabling its application to real-world scenarios such as estimating the environmental impact of both new and existing vehicles, as well as aiding policymakers in shaping low-emission vehicle policies. Future enhancements could involve integrating real-time traffic and meteorological data, as well as exploring more complex deep learning techniques like convolutional and recurrent neural networks to improve accuracy and resilience. The work lays a solid foundation for further advancements in vehicle smog prediction and broader environmental applications.

**Conclusion**

This work also presents a new method for predicting the contribution of specific vehicles to smog levels using state-of-art artificial neural networks. Here, using the SMOG dataset comprising of vehicle characteristics and their corresponding Smog Ratings, we proposed Random Forest and Explainable Boosting Classifier models to build and assess. When using SMOTE to balance the given dataset, performances are increased to 86% in both cases.

As for the interpretability, and to increase the level of trust in the models, the Explainable AI (XAI) approaches were used. Feature importance was calculated using LIME, SHAP, and SHAPash to show how the models arrived at their decision to predict Smog Ratings. These tools offered useful information with regard to the surmise of parameters that held influence over a vehicle’s smog effects and allow decisions and interventions to be based on this information.

Thus, the implication of the study is generalizable in nature. Firstly, the models can help to estimate the environmental effect of both new and current vehicular structures and, in this way contribute to the provision of environmentally friendly transport systems solutions. Secondly, findings from the XAI can inform policymakers and regulators on how to design appropriate policies aimed at boosting the uptake of low-emission vehicles. Finally, the method proposed in this research allows the application of the developed methodology to other fields, for example, energy consumption predictions and climate simulation.

The developments made in this study lay down a solid foundation for vehicle smog prediction, however, future research possibility exists. The models could be further improved by the integration of traffic information in real-time as well as meteorological conditions. Furthermore, the application of complicated deep learning

framework like convoluted and recurrent neural nets may prove the forecast to be even more accurate and resilient.

In conclusion, this study points out the role of Machine learning and Explainable AI as solution models to the problem of air pollution. Thus, this research enriches the stream of works providing accurate and interpretable predictions in making the world cleaner and people healthier.

## Data availability

The data used to support the findings of this study are available from the corresponding authors upon request.

Received: 17 December 2024; Accepted: 3 March 2025

Published online: 08 March 2025

## References

- Anenberg, S., Miller, J., Henze, D., and Minjares, R. A global snapshot of the air pollution-related health impacts of transportation sector emissions in 2010 and 2015. *Int. Counc. Clean Transp.*, p. 55, 2019, [Online]. Available: [https://www.theicct.org/sites/default/files/publications/Global\\_health\\_impacts\\_transport\\_emissions\\_2010-2015\\_20190226.pdf](https://www.theicct.org/sites/default/files/publications/Global_health_impacts_transport_emissions_2010-2015_20190226.pdf).
- The World Bank, "Urban population (% of total population)," *World Bank Data*, pp. 1–19, 2022, [Online]. Available: <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS>.
- "Department of Economic and Social Affairs," *United Nations*, 2018. <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html>.
- "<https://www.kaggle.com/datasets/ahmeterdemyenay/my1995-2023-fuel-consumption-ratings/data>."
- "XAI Applications Based on Vehicle Characteristics for Reducing CO2 Emissions," <https://www.kaggle.com/code/ahmeterdemyenay/xai-applications-based-on-vehicle-characteristics>, 2024.
- Meena, G., Mohbey, K. K. & Lokesh, K. FSTL-SA: Few-shot transfer learning for sentiment analysis from facial expressions. *Multimed. Tools Appl.* <https://doi.org/10.1007/s11042-024-20518-y> (2024).
- Gokul, S., Madhorubagan, G. E., & Sasipriya, M., Fault prediction using fuzzy convolution neural network on IOT environment with heterogeneous sensing data fusion. *Bonfring Int. J. Netw. Technol. Appl.*, **11**(1), 01–05 (2024). <https://doi.org/10.9756/bijnta/v11i1/bij24001>.
- Karthikeyan, N. K., A novel attention-based cross-modal transfer learning framework for predicting cardiovascular disease. *Comput. Biol. Med.*, **170**, 2024, <https://doi.org/10.1016/j.combiomed.2024.107977>.
- Paul, A. & Nayyar, A. A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis. *Multimed. Tools Appl.* **83**(18), 54249–54278. <https://doi.org/10.1007/s11042-023-17601-1> (2024).
- Prakash, V. J., Vija, S. A. A. A unified framework for analyzing textual context and intent in social media. *ACM Trans. Intell. Syst. Technol.* (2024). <https://doi.org/10.1145/3682064>.
- Wani, N. A., Kumar, R., and Bedi, J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Comput. Methods Programs Biomed.* **243** (2024). <https://doi.org/10.1016/j.cmpb.2023.107879>.
- Wani, N. A., Kumar, R., Bedi, J., & Rida, I. Explainable AI-driven IoMT fusion: Unravelling techniques, opportunities, and challenges with explainable AI in healthcare. *Inf. Fusion*, **110** (2024). <https://doi.org/10.1016/j.inffus.2024.102472>.
- Rasool, N., Iqbal Bhat, J., Ahmad Wani, N., Ahmad, N. & Alshara, M. TransResUNet: Revolutionizing glioma brain tumor segmentation through transformer-enhanced residual UNet. *IEEE Access* **12**, 72105–72116. <https://doi.org/10.1109/ACCESS.2024.3402947> (2024).
- Wani, N. A., Kumar, R., and Bedi, J. Harnessing fusion modeling for enhanced breast cancer classification through interpretable artificial intelligence and in-depth explanations. *Eng. Appl. Artif. Intell.* **136** (2024). <https://doi.org/10.1016/j.engappai.2024.108939>.
- Wani, N. A., Bedi, J., Kumar, R., Khan, M. A. & Rida, I. Synergizing fusion modelling for accurate cardiac prediction through explainable artificial intelligence. *IEEE Trans. Consum. Electron.* <https://doi.org/10.1109/TCE.2024.3419814> (2024).
- Rasool, N., Bhat, J. I., Wani, N. A., & Miglani, A. FGA-net: Feature-gated attention for glioma brain tumor segmentation in volumetric MRI images (2024).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**(7553), 436–444 (2015).
- Garg, S., Jindal, H. Evaluation of time series forecasting models for estimation of PM2.5 levels in air. In *2021 6th International Conference Converging Technology I2CT 2021*, 2021, <https://doi.org/10.1109/I2CT51068.2021.9418215>.
- Rahmadeyan, A., Mustakim, Erkamim, M., Ahmad, I., Sepriano, & Aziz, S., "Air pollution prediction using long short-term memory variants," *Lect. Notes Data Eng. Commun. Technol.*, **211**, pp. 122–132, 2024. [https://doi.org/10.1007/978-3-031-59707-7\\_11](https://doi.org/10.1007/978-3-031-59707-7_11).
- Jeya, S. and Sankari, L. Air pollution prediction by deep learning model. In *Proceedings of International Conference Intelligent Computing Control System ICIACS 2020*, pp. 736–741, 2020, <https://doi.org/10.1109/ICIACS48265.2020.9120932>.
- Zhao, J., Deng, F., Cai, Y. & Chen, J. Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. *Chemosphere* **220**, 486–492. <https://doi.org/10.1016/j.chemosphere.2018.12.128> (2019).
- "Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network," *J. Artif. Intell. Data Min.*, **4**(1), (2016). <https://doi.org/10.5829/idosi.jaidm.2016.04.01.06>.
- Biancofiore, F. et al. Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmos. Pollut. Res.* **8**(4), 652–659. <https://doi.org/10.1016/j.apr.2016.12.014> (2017).
- Siwek, K. & Osowski, S. Improving the accuracy of prediction of PM 10 pollution by the wavelet transformation and an ensemble of neural predictors. *Eng. Appl. Artif. Intell.* **25**(6), 1246–1258. <https://doi.org/10.1016/j.engappai.2011.10.013> (2012).
- Mangayarkarasi, R. et al. COVID19: Forecasting air quality index and particulate matter (PM2.5). *Comput. Mater. Contin.* **67**(3), 3363–3380. <https://doi.org/10.32604/cmc.2021.014991> (2021).
- Chen, S., Mihara, K. & Wen, J. Time series prediction of CO2, TVOC and HCHO based on machine learning at different sampling points. *Build. Environ.* **146**, 238–246. <https://doi.org/10.1016/j.buildenv.2018.09.054> (2018).
- Sahoo, D., Hoi, S. C. H., and Li, B. Large scale online multiple kernel regression with application to time-series prediction. *ACM Trans. Knowl. Discov. Data*, **13**(1), (2019). <https://doi.org/10.1145/3299875>.
- Huang, C. J. and Kuo, P. H. A deep cnn-lstm model for particulate matter (Pm2.5) forecasting in smart cities. *Sensors (Switzerland)*, **18**(7), (2018). <https://doi.org/10.3390/s18072220>.
- Matará, C., Osano, S., Yusuf, A. O. & Aketch, E. O. Prediction of vehicle-induced air pollution based on advanced machine learning models. *Eng. Technol. Appl. Sci. Res.* **14**(1), 12837–12843. <https://doi.org/10.48084/etasr.6678> (2024).
- Srivastava, C., Singh, S., and Singh, A. P. Estimation of air pollution in Delhi using machine learning techniques. In *2018 International Conference Computing Power Communication Technology GUCON 2018*, pp. 304–309, 2019, <https://doi.org/10.1109/GUCON.2018.8675022>.
- Russo, A. & Soares, A. O. Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. *Math. Geosci.* **46**(1), 75–93. <https://doi.org/10.1007/s11004-013-9483-0> (2014).

32. Chakraborty, S., Misra, B. & Dey, N. Explainable artificial intelligence (XAI) for air quality assessment. *Front. Artif. Intell. Appl.* **383**, 333–341. <https://doi.org/10.3233/FAIA231451> (2024).

## Acknowledgements

This work was supported by King Saud University, Riyadh, Saudi Arabia, through Researchers Supporting Project number (RSP2025R498).

## Author contributions

All authors have equally contributed.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.M.S. or T.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025