2021-24-0082 Published 05 Sep 2021

Prediction of NO_x Emissions from Compression Ignition Engines Using Ensemble Learning-Based Models with Physical Interpretability

Harish Panneer Selvam, Shashi Shekhar, and William F. Northrop Univ. of Minnesota-Twin Cities

Citation: Panneer Selvam, H., Shekhar, S., and Northrop, W.F., "Prediction of NO_x Emissions from Compression Ignition Engines Using Ensemble Learning-Based Models with Physical Interpretability," SAE Technical Paper 2021-24-0082, 2021, doi:10.4271/2021-24-0082.

Abstract

On-board diagnostics (OBD) data contain valuable information including real-world measurements of vehicle powertrain parameters. These data can be used to gain a richer data-driven understanding of complex physical phenomena like emissions formation during combustion. In this study, we develop a physics-based machine learning framework to predict and analyze trends in engine-out NO_x emissions from diesel and diesel-hybrid heavy-duty vehicles. This model differs from black-box machine learning models presented in previous literature because it incorporates engine combustion parameters that allow physical interpretation of the results. Based on chemical kinetics and the characteristics of diffusive combustion, NO_x emissions from compression ignition engines primarily depend non-linearly on three parameters: adiabatic flame temperature, the oxygen concentration in the cylinder when the intake valves are closed, and combustion time duration. Here these parameters were calculated from available OBD data. Linearizing a physics-based NO_x emissions prediction model provides an opportunity to evaluate several machine learning regression techniques. The results show that an ensemble learning bagging-type model like random forest regression (RFR) is highly effective in predicting engine out NO_x emissions measured by the on-board NO_x sensor. We also show that real-world OBD data has high heterogeneity with clustered co-occurrences of vehicle parameters. In terms of accuracy, the developed model provides an average R^2 value of 0.72 and mean absolute error (MAE) of 78 ppm for different vehicle OBD datasets, an improvement of 53% and 42% respectively when compared to non-linear regression models, and provides the opportunity to interpret the results because of its linkage to physical parameters. We also perform drop-column feature sensitivity analysis for the RFR Model and compare prediction results with black-box deep neural network and non-linear regression models. Based on its high accuracy and interpretability, the developed RFR model has potential for use in on-board NO_x prediction in engines of varying displacement and design.

Introduction

Air pollution from combustion sources is implicated in more than 100,000 deaths annually in the U.S. alone. A major component of combustion-generated pollution emitted from vehicles, nitrogen oxides (collectively called NO_x) are important to study because they can react in the atmosphere to produce Ozone and acid rain, which in turn causes eutrophication in water-based ecosystems. NO_x emissions are heavily regulated by the U.S. Environmental Protection Agency (EPA) and more than half of all NO_x present in the air originate from automotive sources, mainly diesel-powered vehicles. NO is primarily formed during the diffusion burning portion of combustion in compression ignition engines due to high temperature (above 1800K) and lean fuel-air ratios. NO_2 is then formed through equilibrium processes in the combustion chamber and in the atmosphere.

In modern diesel vehicles, NO_x is controlled through the use of selective catalytic reduction (SCR) aftertreatment; however, its effective reduction and the required flow rate of diesel exhaust fluid (DEF) highly depends on the NO_x concentration produced by the engine. Predicting and measuring engine-out NO_x is of keen interest to researchers and engine manufacturers alike. Electrochemical NO_x sensors are commonly used in production engines both upstream and downstream of the SCR. Though effective, sensors are expensive and can result in errors, especially at low exhaust temperatures [1]. Predictive models could be an effective way to eliminate the upstream sensor to save cost or for use in validating sensor performance.

Emissions prediction using data-driven methods has been of increasing interest due to the availability of high quality data, the emergence of advanced machine learning models, and faster computational speeds. Data can be obtained either from laboratory engine testing or from instrumented vehicles under real-world driving conditions. Non-linear regression [2], support vector machine [3], adaptive regression splines [4], and decision trees [5] are some of the machine learning regression techniques that are used to predict NO_x emissions using both sources of data. However, these techniques rely on high time resolution (i.e. > 1 Hz) data for accurate predictions

and are seldom capable of handling sparse data that is generally obtained from low-cost loggers used by telematics providers. Furthermore, the lack of a physics basis for these black-box techniques means that each data entry is treated as part of an identical and independent distribution (i.i.d) and they also fail to interpret why a prediction model behaves a certain way for specific types of engines or vehicles.

Black-box neural network models have also been used recently to accurately and conveniently predict emissions from vehicle data. Techniques in this category include artificial neural networks [6], deep neural networks with Bayesian parameter optimization [7], and long short-term memory (LSTM) neural networks [8, 9]. In particular, convolutional neural networks [10] have been used to study emissions based on imagery from computational fluid dynamic simulations and experimental data. In the trade-off between accuracy and interpretability of artificial intelligence models [11], black-box models such as these rank lowest in terms of interpretability while having generally high accuracy. In other words, these techniques, while accurate, cannot explain how NO_x forms inside a compression-ignition engine or be used to diagnose why emissions are elevated under given engine conditions.

Ensemble learning models are a grey-box intermediate between black-box models and physics-based regression models. They have been shown to have better predictive accuracy than popular machine learning models while resulting in better interpretability compared to neural networks. Ensemble learning models such as random forest, XGBoost, GradientBoost, etc, have been successfully employed for regression problems [12] in domains dealing with real-world data. These include areas like chemoinformatics [13], electricity load prediction [14], building energy prediction [15], and vehicle fuel consumption [16, 17]. Random forest models, which are employed in this study, have been explored for applications closely related to vehicle emissions predictions like forecasting combustion profiles for spark ignition engines [18]. Other uses of random forest models include predicting street-level particulate matter and NO_x concentrations [19] and electric vehicle energy consumption [20]. For complex real-world applications, physics-informed models (like [21]) would be imperative in order to understand the physical and chemical causes of observed data. Though ensemble learning models produce good predictive accuracy, physical interpretation is required to understand variations in NO_x formation due to different engine operating conditions observed in on-board diagnostics datasets. Selecting carefully constructed physics-based feature variables based on physical phenomena has potential to result in more effective and efficient predictions when used in random forest models compared to selecting raw attributes. The physics-based features also help handle sparse datasets often obtained through OBD sampling.

In this study, a physics-based random forest regression model is presented that accurately predicts engine-out NO_x emissions from on-board diagnostics datasets collected from different compression-ignition engine powered vehicles. Validation experiments are conducted to compare the proposed model's predictions with that of non-linear regression and deep neural network models, and also to physically interpret the results of the random forest regression model over a range of vehicle operating conditions and duty cycles. A feature importance study is performed to understand variations in the effectiveness of physics-based feature variables for predicting NO_x, and to analyze the variations between different vehicle datasets.

Methodology

Chemical Kinetics-Based NO_x Formation Model

Engine NO_x is mostly formed at high temperature (> 1800 K) in combustion following the well known Extended Zel'dovich Mechanism. Based on previous work [22] and the mechanism [23], engine-out NO_x from a compression-ignition engine is assumed to depend on three main parameters - Adiabatic flame temperature (T_{adiab}), the oxygen concentration in the cylinder with closed intake valves(x_{O_2}), and combustion time duration (t_{comb}). To emulate the chemical kinetics equation for the high temperature NO_x formation mechanism, a chemical kinetics-based NO_x formation model is developed where NO_x emission values depend on the four terms, $\hat{t}_{comb}$, x_{O_2}, $\hat{T}_{adiab}$ and $1/T_{adiab}$ with a time delay of 1 second, as shown in Equation 1. The ground truth data or observed values for engine-out NO_x are obtained from on-board sensors in the Selectic Calatylic Reduction (SCR) system.

$$x_{NO_x,theory} = a * t_{comb}^{b} * {x_{O_2}}^{c} * \hat{T}_{adiab}^{d} * exp\left(\frac{-e}{T_{adiab}}\right)$$
$$NO_{x,theoryppm} = x_{NO_x,theory} * 1000000 \quad (1)$$

where, $x_{NOx,\ theory}$ is the predicted mole fraction of NO_x formed with a delay of 1 second (since a time delay was observed during initial data analysis), $\hat{t}_{comb}$ is the dimensionless form of combustion time duration (t_{comb} in seconds) which is obtained by multiplying t_{comb} by engine speed (*engRPM*), x_{O_2} is the concentration in mole fraction of oxygen in the cylinder when intake valves are closed, $\hat{T}_{adiab}$ is the dimensionless adiabatic flame temperature of the combustion products obtained by dividing T_{adiab} by intake manifold temperature (*intakeT*), a, b, c, d, e are the coefficients to be obtained through a regression analysis. and $NO_{x,\ theoryppm}$ is the predicted NO_x emission value in parts per million.

Taking a natural logarithm of the proposed physics-based NO_x Prediction Model given by Equation 1, Equation 2 is obtained. This provides an interesting opportunity to experiment with different machine learning algorithms like random forest regression (RFR) that are more appropriate for linear equations than on highly non-linear ones like the physics-based NO_x model given in Equation 1.

$$\log\left(x_{NO_x,theory}\right) = \log(a) + b * \log\left(\hat{t_{comb}}\right) + c * \log\left(x_{O_2}\right) + d * \log\left(\hat{T_{adiab}}\right) - \frac{e}{T_{adiab}} \quad (2)$$

Datasets and Predictive Accuracy Metrics

The on-board diagnostics (OBD) datasets used in this study contain 1Hz time resolution data of various engine and vehicle parameters for five different heavy-duty compression-ignition engine-powered vehicles, with either a hybrid or a conventional drivetrain. Since NO_x emissions are formed inside the engine cylinder and feature variables are not dependent on the battery of a hybrid vehicle, the drivetrain of the vehicles is assumed to have negligible impact on the effectiveness of our proposed prediction model. The transit bus (TB) dataset is obtained from Metro Transit, a local public transportation agency in Minneapolis, whereas the remaining four are obtained from the FleetDNA database managed by the National Renewable Energy Laboratory, Colorado [24]. Details of the five datasets are presented in Table 1. The data have been pre-processed to filter out highly erroneous, noisy measurements as well as data files where vehicles have been stationary for prolonged periods of time. However, all the important vehicle operating conditions such as engine idling, acceleration, etc., are represented in the filtered datasets. The terms in Equation 2 were calculated in a similar way for each of these datasets and different ensemble learning models. Their use as part of the random forest regression model, the primary method in this work, is explained in the following sections.

Since there is a need to statistically compare the accuracy of NO_x prediction from different models to ground-truth data, three predictive accuracy metrics are used in the study: Coefficient of determination (R^2 value), root mean square error (RMSE), and mean absolute error (MAE). Adjusted R^2 value, which is generally used to account for overfitting in regression problems, is not required owing to the low number of feature variables used in the model.

Physics-Based Random Forest Regression

Random forest regression (RFR) [25] is a type of ensemble learning model that consists of a set of decision trees (also known as regression trees). The RFR model was chosen because it showed better predictive accuracies than the other ensemble models such as XGBoost and Gradient Boost for predicting NO_x from the obtained datasets. Decision trees, which are binary trees (each parent node splits into two children nodes), help predict a target value using feature variables. Ensemble learning models like random forest involve using multiple low-strength machine learning models in place of a single, one-size-fits-all, high-strength one to obtain accurate predictions in highly heterogenous datasets. An RFR model consists of several such decision trees and predicts values based on a voting procedure. In each decision tree of the RFR model, every data entry or sample is assigned to nodes that are created based on the values of the feature variables. This sampling of entries into nodes is based on finding the minimum sum of square residuals, which is also called variance reduction, to split a parent node. A least-square regression on each leaf node gives the predicted value for the corresponding target variable values.

A sample decision tree of the RFR Model from the scikit-learn python package [26] for predicting NO_x from the TB dataset is visualized in Figure 2. Parameters that can be tuned for this RFR model are n_{trees} (number of decision trees), *maxDepth* (maximum depth of the tree or a threshold for length of the longest branch), and *minSampleSplit* (minimum number of samples in a node that cannot be divided further). As shown in Figure 2, the four feature variables used are: $\log(t_{res}/engRPM)$ (or $\log\left(t_{res}\right)$), $\log(x_{O2})$, $\log(T_{adiab}/intakeT)$

FIGURE 1 Map plot of different trips (represented by different colors) of the vehicle in the transit bus Dataset

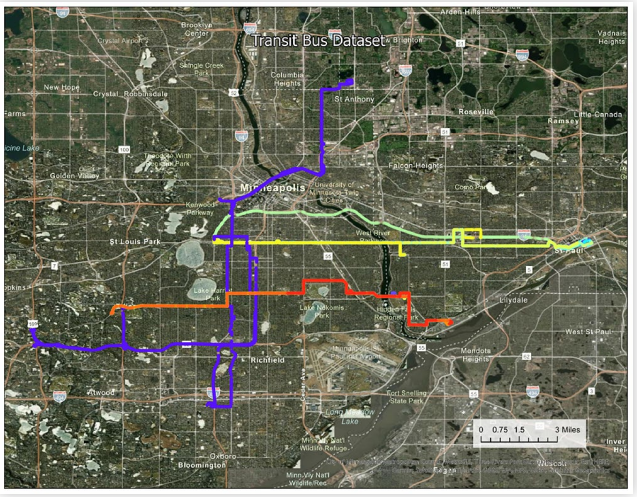
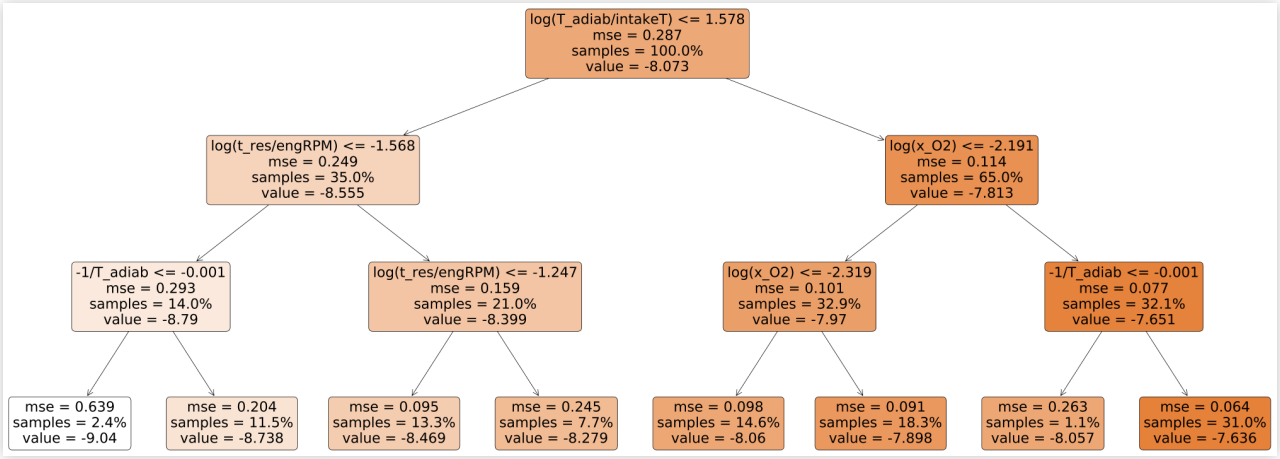


TABLE 1 Engine and route details for the different vehicle OBD datasets used for the ensemble learning regression models

Dataset	Transit Bus (TB)	Food Delivery Truck (FDT)	Yard Tractor (YT)	Drayage Truck (DT)	Refuse Truck (RT)
Engine Model	Cummins ISB6.7	PACCAR PX-7	Cummins QSB6.7	Cummins ISX15	Cummins ISL
Displacement	6.7L	6.7L	6.7L	14.9L	8.9L
Stroke(mm)*					
Bore(mm)	124*107	124*107	124*107	169*137	145*114
Drivetrain	Hybrid	Conventional	Conventional	Conventional	Conventional
Attributes	92	110	133	105	143
Data Entries	99505	81977	54259	183865	147515
No. of Trips	17	207	70	437	80
Route Region	Twin Cities, MN	Denver, CO	New York City, NY	Long Beach, CA	Columbus, OH

FIGURE 2 A sample base decision tree of random forest regression model on the transit bus dataset with n_{trees} = 5, *maxDepth* = 3, *minSampleSplit* = 300



(or $\log\left(T_{adiab}\right)$), and $1/T_{adiab}$. The target variable is $\log(x_{NOx})$. In Figure 2, each node contains information about the feature variable used for splitting, mean square error (*mse*) for samples in the node, percentage of the samples present in the node (*sample*), and the target variable value (*value*).

For the large vehicle datasets used here, more aggressively chosen parameters are required for accurate prediction, hence the RFR Model parameters chosen for NO_x predictions are: n_{trees} = 25, *maxDepth* = 20, and *minSampleSplit* = 15. These values are the result of tuning based on a drop-column sensitivity analysis of the RFR Model, and were chosen as a tradeoff between variance and bias.

Results

Prediction Results for Different Datasets

An RFR model with the aforementioned parameters is used to predict NO_x emissions for the five OBD datasets under consideration. The predictive accuracies of the five datasets, transit bus (TB), food delivery truck (FDT), yard tractor (YT), drayage truck (DT), and refuse truck (RT) are presented in Table 2. A scatter plot of the NO_x values prediction using the RFR model for the TB dataset that contains measurements obtained from on-board sensors on a real-world vehicle trip is shown in Figure 3. The scatter plots for predictions of the other datasets are provided in Appendix A.

TABLE 2 Predictive acccuracy metrics for NO_x prediction for different vehicle datasets using random forest regression

Dataset	R^2	RMSE	MAE
TB	0.7930	68.83	45.38
FDT	0.7831	127.00	70.45
YT	0.6103	129.78	84.54
DT	0.5129	213.21	62.76
RT	0.8754	223.17	130.18

K-Fold Cross Validation Predictions

Cross Validation, or out-of-sample testing, is used to check the effectiveness of machine learning models on data entries it was not trained on. The random forest regression model is trained on 80% of the dataset and uses the entire dataset to obtain a prediction. To assess the generalization of the model and whether it performs as well on the testing data as it has on the training data, a K-Fold Cross Validation is used. Due to the availability of relatively large datasets consisting of different vehicle operating conditions, K=10 was chosen.

In a 10-fold cross validation prediction, the dataset is split into 10 equal-sized sets with entries chosen in random, the model is trained on 9 sets and tested on one. The process is

FIGURE 3 Scatter plot of NO_x prediction using Random Forest Regression Model for the Transit Bus dataset

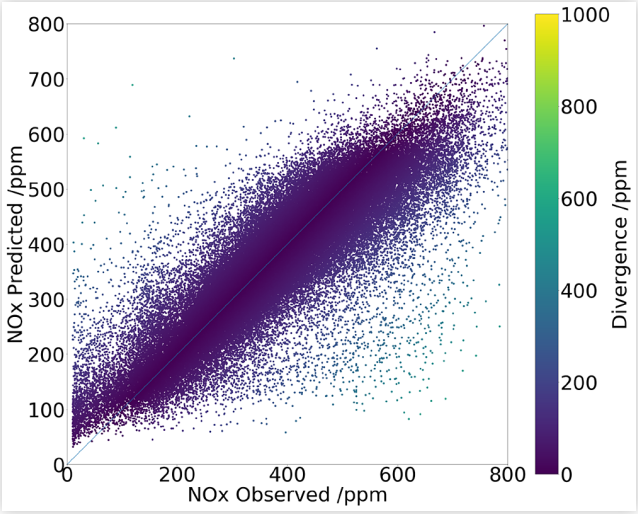


TABLE 3 Predictive acccuracy metrics for 10-fold cross-validated NO_x prediction for different vehicle datasets using random forest regression

Dataset	R^2	RMSE	MAE
TB	0.6933	83.63	54.99
FDT	0.7278	141.72	79.21
YT	0.4510	153.25	99.14
DT	0.4489	222.49	66.91
RT	0.8565	238.41	139.21

then rotated until all 10 sets are used as testing data once. In this way, the predicted value for each entry in the dataset is computed using an estimator fitted on its corresponding training set. This is a popular method to check for model overfitting [27]. The accuracy metrics of the cross-validated predictions for the five datasets are presented in Table 3.

By comparing predictive accuracies in Table 2 with the accuracy metrics of the cross-validated predictions in Table 3, we observed that the R^2, RMSE, and MAE values do not reduce significantly during the cross-validation analysis. Hence, the RFR model used to predict NO_x values from the five datasets is not significantly overfitting for out-of-sample data and hence is a mathematically appropriate method for accurately predicting NO_x emissions.

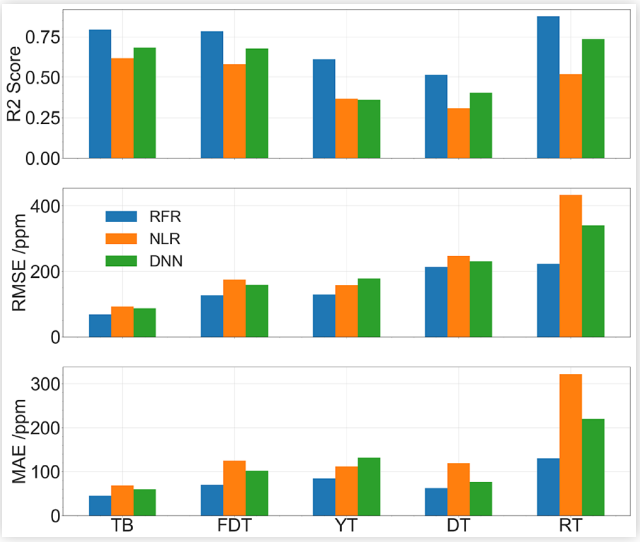
Comparison of Proposed Model with Non-Linear Regression and Deep Neural Network

Ensemble machine learning models like Random Forest Regression are shown to be effective for NO_x prediction. The predictive accuracy metrics of the RFR model for the five datasets are compared with results from a physics-based non-linear regression model and a deep neural network. These comparisons are presented in Figure 4. The non-linear regression (NLR) model utilized the *curve_fit* function coupled with Equation 2, and the deep neural network (DNN) utilized the *MLPRegressor* function (multi-layer perceptron), both from the Scikit-Learn python package. We observe that the RFR model consistently produces better R^2, RMSE, and MAE metrics for all the five vehicle datasets when compared to the corresponding NLR and DNN model predictions. The RT dataset has the highest R^2 score for the RFR model among all the datasets, and also has the highest percentage increase in R^2 score when compared to its corresponding NLR prediction, while the TB dataset has lowest RMSE and MAE error among all the datasets. The NLR model for the YT, DT, and RT datasets is observed to be highly ineffective in predicting NO_x emissions.

Feature Importance Analysis

Different ensemble models such as random forest regression, AdaBoost Regression, bagging regression, extra trees regression, gradient boost regression, and XGBoost regression from the Scikit-Learn Ensemble python package were

FIGURE 4 Comparison of predictive accuracy metrics between Random Forest Regression (RFR), Non-Linear Regression (NLR), and Deep Neural Network (DNN) NO_x prediction models used on different vehicle datasets



used to predict NO_x values. A feature importance (FI) study was performed to analyze how effective each of the feature variables is in predicting the target variable. FI for regression trees and random forest models describes the sum of decrease in the node impurity (in this case, the mean square error) weighted by the probability of a sample reaching that node (number of samples in a node divided by total number of samples). It is calculated using the formula given in Equation 3 [28].

$$ni_j = w_j * C_j - w_{left,j} * C_{left,j} - w_{right,j} * C_{right,j}$$
$$fi_i = \frac{\sum_{j:node\ \ j\ \ splits\ \ on\ \ feature\ \ i} ni_j}{\sum_{k:all\ \ nodes} ni_k} \quad (3)$$

where ni_j is the importance of node j, w_j, $w_{left,j}$, $w_{right,j}$ are the weights or probability of samples ending in the node j, the left child of node j, and right child of node j respectively. C_j are the impurities (in this case, MSE) of node j, the left child of node j and right child of node j respectively. fi_i is the feature importance of a feature 'i'.

The FIs of the four terms in Equation 2 for the different ensemble models and the TB dataset are plotted in Figure 5. We note that the FIs are fairly consistent between the different models for a given dataset. For the TB dataset, the combination of $\log\left(\hat{T}_{adiab}\right)$ and $-1/T_{adiab}$ shows the highest feature importance.

Figure 6 shows the FIs for the three terms (here, $\log\left(\hat{T}_{adiab}\right)$ and $-1/T_{adiab}$ are combined for convenient interpretation) for the five vehicle datasets used in this study. We notice that the FIs vary significantly between the different vehicle datasets. The differences between the datasets responsible for the variation include engine model, vehicle purpose/driving pattern, engine duty cycle, and environmental factors (like humidity, elevation, etc). The R^2 score for NO_x prediction using RFR model for the corresponding datasets are also plotted on

FIGURE 5 Feature Importance study for the Transit Bus dataset for different ensemble learning models

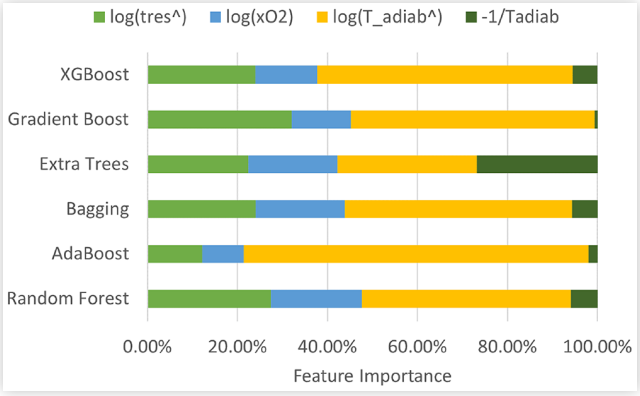


FIGURE 6 Feature Importance variations and corresponding R^2 score for the random forest regression NO_x prediction using the five vehicle datasets

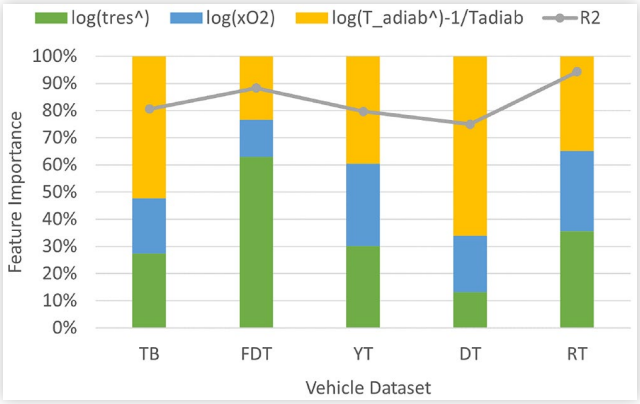


Figure 6. We notice that the R^2 score correlates inversely with the the FI of the T_{adiab} related terms. The lower the FI of $\log\left(\hat{T}_{adiab}\right)-1/T_{adiab}$, higher is the R^2 score and vice versa.

Discussion

Physical Interpretation

From the results, we observe that ensemble learning models, especially random forest regression, are highly effective at accurately predicting NO_x values given an on-board diagnostics dataset. The four feature variables are computed and carefully selected to emulate the general kinetics equation of NO_x formation. Grey-box models are generally difficult to interpret; however, RFR models combined with physics-based calculated parameters are shown here to be an outlier among models due to their advantageous trade-off between interpretability and accuracy. We observed that the predictive accuracies of the RFR model were significantly better than the previously proposed NLR and DNN models (Figure 4), and produced accurate results for the five diverse vehicle datasets used (Table 1). Using a feature importance study (Figure 6 and Figure 5) and regression tree visualizations (Figure 2), the model was also shown to be physically interpretable.

The effectiveness of the RFR model predictions further strengthens the stated assumption that NO_x emissions formation in a compression-ignition engine in real-world conditions is highly dependent on the operating conditions of the specific engine. Thanks to the vastly different vehicle datasets used in this study, the predictive accuracies of the RFR model changed with the duty cycles of the engine. For example, datasets with the highest and lowest R^2 score among the five datasets were the refuse truck (RT) and drayage truck (DT), respectively. The refuse truck had duty cycles with very limited vehicle operating conditions: short acceleration events between frequent stops on the road, short braking events, and frequent reverse driving events (which are equivalent to short acceleration events for the engine). A random forest regression model is able to effectively sample data entries into leaf nodes that denote the non-linear relationship between the four feature variables describing the variations of NO_x formation in an engine. Hence, high R^2 values are observed for the RT dataset. The RMSE and MAE values are high due to the fact that the magnitude of NO_x emissions from the refuse truck are generally higher than the other vehicles. While the RT and DT operating conditions are somewhat uniform, the DT has a more diverse duty cycle. Here, many vehicle operating conditions are encountered as the vehicle is driven on the busy roadways of Long Beach, CA. The vehicle also transports goods from the coast to facilities 50 miles away on state highways. This justifies the comparatively weak R^2 score for the NO_x prediction using the RFR model, which is still higher than for predictions using the NLR model.

Comparing with DWC Pattern Detection

While vehicle operating conditions are an important factor in the predictive effectiveness of the random forest regression models in NO_x prediction, it is also worthwhile to consider anomalous driving events that could be driving inaccuracy in the model. The Divergent Window Co-occurrence (DWC) Pattern Detection algorithm used previously by the authors [22] is a method to identify repetitive temporal events with high prediction error (or divergence) from the physics-based non-linear regression (NLR) model (given by Equation 2). The algorithm analyzes these divergent data entries with their corresponding raw attributes present in the OBD dataset like engine speed, fuel mass flow rate, etc., and outputs statistically significant co-occurrence patterns of these attributes. The resultant DWC patterns provide insights into vehicle operating conditions or driving patterns that are inadequately described by the physics-based NO_x prediction model. Similarities can be found between these DWC patterns and the leaf nodes in decision trees of the RFR model. Each leaf node represents a specific range of values for a subset of the four terms in Equation 2, similar to that of DWC patterns that look at discretized attribute magnitudes over moving time windows. However, the RFR model uses the four terms directly from the Equation 2, whereas the DWC Pattern-based framework analyzes OBD dataset attributes like engine speed,

exhaust gas recirculation mass flow rate, fuel mass flow rate, etc that are used to calculate these four terms. Another difference is that the RFR model explores all operating conditions whereas the DWC Pattern-based framework focuses only at events that are responsible for high prediction error. Hence, both these methods are effective ways to look at driving patterns from a vehicle dataset.

Future Work

Future work related to ensemble learning approaches for NO_x emissions prediction could include developing a framework to stochastically analyze the regression function at all leaf nodes. Examining random forest regression using different sets of raw attributes from the OBD datasets is also of interest, including comparison of the predictive accuracies and physical interpretation with the model using the four physics-based feature variables. Using each of the attribute sets, it would also be interesting to look at the corresponding DWC Patterns and correlate them with the leaf node functions from the RFR Model. Duty cycle parameters like average power, average engine speed, kinetic intensity, etc could be introduced to quantitatively explain the variation in predictive accuracies among the five vehicle datasets using the RFR model. Furthermore, the ensemble models could be trained on entire trips instead of a random test-train split, to evaluate the effectiveness of the model in predicting NO_x emissions for entirely unseen trips in real-time. Future work will also include predicting vehicle performance parameters other than NO_x emissions such as Fuel economy and vehicle driving range. Predictive analytics like fault diagnostics for different subsystems in the vehicles may also be explored.

Conclusion

In this work, a linearized form of a physics-based NO_x formation equation was used to accurately predict compression-ignition engine-out NO_x emissions values from five vehicle on-board diagnostics (OBD) datasets. Ensemble machine learning models, with focus on random forest regression (RFR), were observed to provide high predictive accuracies. Using OBD datasets from five different compression-ignition engine powered vehicles, the RFR model was independently evaluated to predict NO_x emissions after training the model on 80% of the datasets. Prediction using the RFR model provides on an average around 53% better R^2 score, 27% lower RMSE, and 42% lower MAE error compared to predictions using non-linear regression. The performance and interpretability of the RFR model was validated through several experiments and analyses. A K-fold cross-validation was used to perform out-of-sample testing to prove that the model is not significantly overfitting for the testing dataset. The RFR model predictions were compared with that from non-linear regression and deep neural network models for the five vehicle datasets. A feature importance test was also performed to evaluate the effectiveness of each feature variable for different ensemble learning models like random forest, AdaBoost, Gradient Boost, etc. The variation in feature importances for different datasets were studied, and the R^2 score was observed to be inversely proportionaly to the feature importance of T_{adiab} terms in the linearized physics-based NO_x formation equation.

Although the RFR model is generally a grey-box model, it becomes partially physically interpretable through the use of physics-based terms, and analyses of their feature importance distributions. The high effectiveness of the model further strengthens the assumption that NO_x emissions formation in a diesel engine is a complex set of processes that varies with vehicle operating conditions. These operating conditions can be represented by the leaf nodes of individual regression trees in the RFR model. Furthermore, the non-linear leaf-node functions were identified to be similar in many ways to the divergent window co-occurrence patterns that were used to analyze driving features where the general physics-based non-linear regression model becomes inadequate in explaining NO_x emissions values. It is generally concluded that ensemble learning models such as RFR are more effective than black-box models like neural networks for accurately predicting vehicle emissions from large sparse datasets, while also providing the physical interpretability that non-linear regression models possess.

References

1. Kotz, A.J., Kittelson, D.B., Northrop, W.F., and Schmidt, N., "Realworld NOx Emissions of Transit Buses Equipped with Diesel Exhaust Aftertreatment Systems," *Emission Control Science and Technology* 3, no. 2 (2017): 153-160.
2. Le Cornec, C.M.A., Molden, N., Van Reeuwijk, M., and Stettler, M.E.J., "Modelling of Instantaneous Emissions from Diesel Vehicles with a Special Focus on NO_x: Insights from Machine Learning Techniques," *Science of The Total Environment* 737 (Oct. 2020): 139625.
3. Liu, B., Hu, J., Yan, F., Turkson, R.F. et al., "A Novel Optimal Support Vector Machine Ensemble Model for NOx Emissions Prediction of a Diesel Engine," *Measurement* 92 (Oct. 2016): 183-192.
4. Oduro, S.D., Metia, S., Duc, H., Hong, G. et al., "Multivariate Adaptive Regression Splines Models for Vehicular Emission Prediction," *Visualization in Engineering* 3 (June 2015): 13.
5. Li, Q., Qiao, F., and Yu, L., "A Machine Learning Approach for Light-Duty Vehicle Idling Emission Estimation Based on Real Driving and Environmental Information," *Environment Pollution and Climate Change* 1 (Jan. 2017): 1000106.
6. Agbulut, U., GUrel, A.E., and Saridemir, S., "Experimental Investigation and Prediction of Performance and Emission Responses of a CI Engine Fuelled with Different Metal-Oxide Based Nanoparticles-Diesel Blends Using Different Machine Learning Algorithms," *Energy* 215 (Jan. 2021): 119076.
7. Shin, S., Lee, Y., Kim, M., Park, J. et al., "Deep Neural Network Model with Bayesian Hyperparameter Optimization for Prediction of NO_x at Transient Conditions

in a Diesel Engine," *Engineering Applications of Artificial Intelligence* 94 (Sept. 2020): 103761.

8. Wang, Y., Yu, Y., and Li, J., "Predicting the Transient NO_x Emissions of the Diesel Vehicle Based on LSTM Neural Networks," in *2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, 261-264, Dec. 2020.
9. Yu, Y., Wang, Y., Li, J., Fu, M. et al., "A Novel Deep Learning Approach to Predict the Instantaneous NO_x Emissions from Diesel Engine," *IEEE Access* 9 (2021): 11002-11013.
10. Warey, A., Gao, J., and Grover, R., "Prediction of Engine-Out Emissions Using Deep Convolutional Neural Networks," SAE Technical Paper 2021-01-0414, (2021), https://doi.org/10.4271/2021-01-0414.
11. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A. et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion* 58 (June 2020): 82-115.
12. Mendes-Moreira, J., Soares, C., Jorge, A.M., and Sousa, J.F.D., "Ensemble Approaches for Regression: A Survey," *ACM Computing Surveys* 45, no. 1 (Dec. 2012): 1-40.
13. Kaneko, H. and Funatsu, K., "Applicability Domain Based on Ensemble Learning in Classification and Regression Analyses," *Journal of Chemical Information and Modeling* 54 (Sept. 2014): 2469-2482.
14. Qiu, X., Zhang, L., Ren, Y., Suganthan, P.N. et al., "Ensemble Deep Learning for Regression and Time Series Forecasting," in *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*, 1-6, Dec. 2014.
15. Ahmad, M.W., Mourshed, M., and Rezgui, Y., "Trees Vs Neurons: Comparison Between Random Forest and ANN for High-Resolution Prediction of Building Energy Consumption," *Energy and Buildings* 147 (July 2017): 77-89.
16. Perrotta, F., Parry, T., and Neves, L.C., "Application of Machine Learning for Fuel Consumption Modelling of Trucks," in *2017 IEEE International Conference on Big Data (Big Data)*, 3810-3815, Dec. 2017.
17. Massoud, R., Bellotti, F., Berta, R., Gloria, A.D. et al., "Exploring Fuzzy Logic and Random Forest for Car Drivers' Fuel Consumption Estimation in IoT-Enabled Serious Games," in *2019 IEEE 14th International Symposium on Autonomous Decentralized System (ISADS)*, 1-7, Apr. 2019, ISSN: 2640-7485.
18. Application of Random Forest Machine Learning Models to Forecast Combustion Profile Parameters of a Natural Gas Spark Ignition Engine, vol. Volume 6: Design, Systems, and Complexity of ASME International Mechanical Engineering Congress and Exposition, 11 2020. V006T06A003.
19. Li, Z., Yim, S.H.-L., and Ho, K.-F., "High Temporal Resolution Prediction of Street-Level PM2.5 and NO_x Concentrations Using Machine Learning Approach," *Journal of Cleaner Production* 268 (Sept. 2020): 121975.
20. Ullah, I., Liu, K., Yamamoto, T., Zahid, M. et al., "Electric Vehicle Energy Consumption Prediction Using Stacked Generalization: An Ensemble Learning Approach," *International Journal of Green Energy* 0 (Feb.2021): 1-14. https://doi.org/10.1080/15435075.2021.1881902.
21. Panneer Selvam, H., Li, Y., Wang, P., Northrop, W.F., and Shekhar, S., "Vehicle Emissions Prediction with Physics-Aware AI Models: Preliminary Results," in *PGAI-AAAI-20*, 2021.
22. Panneer Selvam, H., Li, Y., Wang, P., Northrop, W.F., "Vehicle Emissions Prediction with Physics-Aware AI Models: Technical Report," in *Retrieved from the University of Minnesota Digital Conservancy*, 2020 https://hdl.handle.net/11299/216628.
23. Liviu-Constantin, S., and Daniela-Elena, M., "Simplified Mechanism Used to Estimate the NO_x Emission of Diesel Engine," in *Proceedings of the 2nd International Conference on Manufacturing Engineering, Quality and Production Systems*, 978-960, 2010.
24. Fleet DNA, "Commercial Fleet Vehicle Operating Data. Golden, Co: National Renewable Energy Laboratory.," https://www.nrel.gov/transportation/fleettest-fleet-dna.html, 2020, Accessed November 23, 2020.
25. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C. et al., "Random Forest: A classification and Regression Tool for Compound Classification and QSAR Modeling," *Journal of Chemical Information and Computer Sciences* 43 (Nov. 2003): 1947-1958.
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. et al., "Scikit-Learn: Machine learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825-2830.
27. Burman, P., "A Comparative Study of Ordinary Cross-Validation, V- Fold Cross-Validation and the Repeated Learning-Testing Methods," *Biometrika* 76, no. 3 (1989): 503-514.
28. D. D. (https://stats.stackexchange.com/users/182882/david dale), "Summing feature Importance in Scikit-Learn for a Set of Features." Cross Validated, https://stats.stackexchange.com/q/313455 (version: 2020-09-11).

Contact Information

Dr. William Northrop,
Director, TE Murphy Engine Research Laboratory,
University of Minnesota-Twin Cities;
Ph: (612)-625-6854
wnorthro@umn.edu
http:/merl.umn.edu

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1901099. We thank the U.S. Department of Energy's National Renewable Energy Laboratory for their Fleet DNA support and assistance. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Definitions, Acronyms, Abbreviations

Acronyms

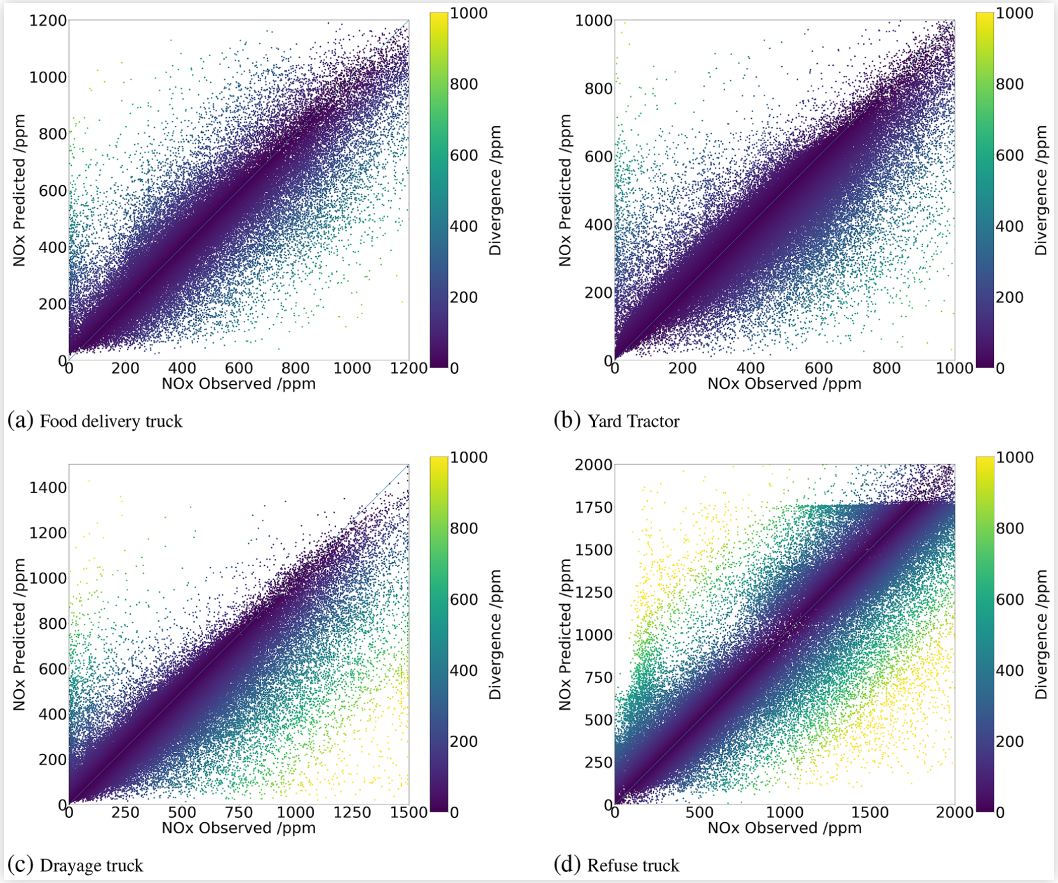
R^2 - Coefficient of determination

DNN - Deep Neural Network
DT - Drayage truck
FDT - Food delivery truck
FI - Feature importance
MAE - Mean absolute error
NLR - Non-linear regression
OBD - On-board diagnostics
RFR - Random forest regression
RMSE - Root mean square error
RT - Refuse truck
TB - Transit bus
YT - Yard tractor

Definitions

$\hat{T}_{adiab}$ - Dimensionless Adiabatic Flame Temperature
$\hat{t}_{comb}$ - Dimensionless combustion duration
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{e}$ - Regression coefficients
$\boldsymbol{C_j}$ - Impurity or mean square error of node j
engRPM - Engine Speed RPM
$\boldsymbol{fi_i}$ - Feature importance of feature variable i
intakeT - Intake Manifold Temperature K
$\boldsymbol{ni_j}$ - Importance of a node j
$\boldsymbol{NO_{x,\,theoryppm}}$ - Predicted NO_x Emission Values ppm
$\boldsymbol{T_{adiab}}$ - Adiabatic Flame Temperature K
$\boldsymbol{t_{comb}}$ - Combustion duration s
$\boldsymbol{w_j}$ - Weight of sample ending in node j
$\boldsymbol{x}_{NOx,\,theory}$ - Predicted NO_x mole fraction
$\boldsymbol{x}_{O_2}$ - Mole fraction of oxygen in the cylinder when intake valves are closed

FIGURE 7 Scatter plot of NO_x prediction using the random forest regression model for the other four datasets



ISSN 0148-7191