ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

An Enhanced Light Gradient Boosting Regressor for Virtual Sensing of CO, HC and NOx

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Giovannardi E., Brusa A., Petrone B., Cavina N., Corti E., Barichello M. (2023). An Enhanced Light Gradient Boosting Regressor for Virtual Sensing of CO, HC and NOx. NEW YORK : Institute of Electrical and Electronics Engineers Inc. [10.1109/MetroAutomotive57488.2023.10219122].

Availability:
This version is available at: https://hdl.handle.net/11585/945039 since: 2024-05-14

Published:
DOI: http://doi.org/10.1109/MetroAutomotive57488.2023.10219122

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

10 November 2026

An Enhanced Light Gradient Boosting Regressor for Virtual Sensing of CO, HC and NOx

1st Emanuele Giovannardi
Industrial Engineering Department
University of Bologna
Bologna, Italy
emanuele.giovannard2@unibo.it

2nd Alessandro Brusa
Industrial Engineering Department
University of Bologna
Bologna, Italy
alessandro.brusa6@unibo.it

3rd Boris Petrone
Industrial Engineering Department
University of Bologna
Bologna, Italy
boris.petrone2@unibo.it

4th Nicolò Cavina
Industrial Engineering Department
University of Bologna
Bologna, Italy
nicolo.cavina@unibo.it

5th Enrico Corti
Industrial Engineering Department
University of Bologna
Bologna, Italy
enrico.corti2@unibo.it

6th Massimo Barichello
Head of Powertrain Development Testing
Ferrari S.p.A
Maranello, Modena, Italy
massimo.barichello@ferrari.com

***Abstract*—The present study introduces a novel methodology that utilizes Light Gradient Boosting Regressors to predict engine-out emissions of NOx, HC, and CO. The accuracy of the proposed models is evaluated on different types of homologation cycles. The dataset used in this study is derived from a set of 48 experimental driving cycles, including RDE, WLTC, NEDC, ECE, US06, and HWFET. The experimental driving cycles are performed on a roll bench using a spark-ignited, naturally aspirated, V12 engine-equipped vehicle. A three-second sliding window is incorporated in the models to capture the dynamic behavior of pollutant emissions. The performance of the LightGBR models is assessed using the mean absolute percentage error (MAPE), which is found to be 5% for CO, 5.4% for HC, and 7.4% for NOx. The results demonstrate the efficacy of the proposed methodology, which can be used to estimate the impact of powertrain calibration changes on pollutant emissions in a virtual environment, thereby reducing the number and the cost of the experimental tests.**

***Index Terms*—Virtual sensing, Emissions modelling, Data-driven, Machine Learning**

I. Introduction

The increasing production of pollutant emissions by modern internal combustion engines in the automotive field has prompted the implementation of more stringent rules and regulations worldwide to minimize their impact on the environment. This has resulted in the development of homologation cycles to test the emissions and fuel consumption under various operating conditions, requiring automakers to improve the engine calibrations to comply with emissions limits during a wide variety of maneuvers [1]. The resulting need for the experimental testing has significant time and cost implications in the development phase of the engine and after-treatment system, influencing the overall design process.

Models and simulations can help to reduce the experimental testing, and the 0-D modeling and the artificial intelligence methods based on machine learning and deep learning algorithms are becoming increasingly popular [2]. All these methods are typically suitable for the implementation in real-time hardware and they have been widely used for various applications in autonomous driving, vehicle control, smart connections, virtual sensing, and anomaly detection, including the emission modeling [3]–[9]. Machine learning models based on support vector machines (SVM) [10], ensemble of tree-based models (random forest or gradient boosted forests) [11], and neural networks have been largely documented in the literature for emission forecasts [12]. This paper presents a methodology for predicting CO, HC, and NOx emissions using a Light Gradient Boosting Regressor [13], trained and validated on a set of driving cycles performed on a roll bench. An earlier study published by the authors [14], shows an interesting application of Light Gradient Boosting Regressor-based (LightGBR), data-driven model for the offline prediction of NOx engine-out emissions in an internal combustion engine using some ECU channels as inputs. It compares multiple regressors from machine learning and deep learning algorithms with LightGBR, proving the proposed methodology represents the most accurate solution. The models are tested by estimating the NOx emissions during two Real Driving Emission (RDE) cycles and under steady-state conditions. This method is a successful demonstration of the effectiveness of data-driven models in real-world industrial applications. However, expanding this methodology to include other pollutant species and homologation cycles is needed for assessing its broad applicability. This paper presents an application that involves the development of engine surrogate models to estimate pollutant emissions. The study is based on the previous work of the authors [14], which effectively predicted NOx emissions using a feature selection process based on Features Importance Permutation [15] and a sliding window approach [16] to capture the dynamic behaviour of pollutant emissions. The current work aims to extend this methodology to other pollutants (HC and CO) and it makes the proposed approach more robust and reliable on different homologation

cycles, beyond the limited set of the previous study.

II. Experimental Setup

For this activity, an experimental campaign is conducted to collect the data needed for both the models training and validation. The tests are carried out on a laboratory roll bench on a vehicle equipped with a spark ignition, naturally aspirated, V12 engine. The main characteristics of such engine are listed in the Table I. For all the tests, the vehicle is automatically driven to follow the speed profile imposed for different homologation cycles, while the emissions are measured immediately upstream of the catalyst through the installation of specific measurement devices. The homologation cycles under study are coming from different legislations (from Europe and USA) and are in particular the Real Driving Emissions (RDE), the Worldwide Harmonized Light Vehicles Test Cycle (WLTC), the New European Driving Cycle (NEDC), the Economic Commission for Europe (ECE), the Federal Test Procedure 75 (FTP-75), the United States 2006 (US06), and the Highway Fuel Economy Test (HWFET).

TABLE I: Engine specifications

Engine Specifications	
Engine Type	V12
Displacement [cc]	6495.6 cc
Aspiration	Naturally Aspirated
Combustion System	DI Spark-ignition
Number of cylinders	12 (6 per bank)
Valves per cyl [#]	4 (2 int + 2 exh)
Bore x Stroke [mm]	94.0 x 78.0

A. Chemiluminescent Detector analyzer

The concentration of NOx emissions is measured using a Chemiluminescent Detector analyzer (CLD) which exploits the reaction between nitric oxide (NO) and ozone (O3) to generate electronically excited NO2 molecules. These molecules emit visible radiation upon returning to equilibrium, with intensities proportional to the concentration of NO in the gas. The CLD can evaluate NOx levels in exhaust gases by measuring the emitted light.

B. Flame Ionization Detector analyzer

The Flame Ionization Detector (FID) is utilized to measure HC emissions and can be used to quantify the amount of hydrocarbons in the exhausts. The FID relies on a process known as hydrogen flame ionization, which generates ions proportional to the amount of carbon atoms in a sample when hydrocarbons are injected into a hydrogen flame. Due to its sensitivity to nearly all HC compounds, it is frequently used to detect exhaust gases from engines. The FID has a broad linear range of up to seven orders of magnitude, making it useful for samples containing a variety of different chemicals because it generates a signal proportional to the flow of carbon atoms through it, regardless of the chemical species' composition.

C. Non-Dispersive Infra-Red analyzer

A Non-Dispersive Infra-Red analyzer is utilized to measure the concentration of various gases such as CO2 and CO. This analyzer employs the principle that a molecule absorbs infrared light at a particular range of frequencies, which is dependent on its bond energy and the mass of its atoms, and the amount of energy absorbed is proportional to its concentration. The NDIR uses this principle to detect several molecules in exhaust gases, with CO2 and CO being measured at wavelengths of 4.2 and 4.6 m, respectively.

III. Methodology

The inputs of the model are selected between the available Engine Control Unit (ECU) signals and actuations to make them compatible with the on-board implementation. The sampling frequency of ECU channels can vary. For example, some of them are sampled at a set frequency (often 10 Hz, 100 Hz, or 1000 Hz), while others are sampled at a frequency directly proportional to the engine speed. Conversely, emissions are measured externally and sampled uniformly at 10 Hz. Thus, the post-processing techniques are mainly employed to resample all the channels to 10 Hz frequency. The Feature Importance Permutation algorithm is adopted to determine the most relevant features for the data-driven model. This is further complemented with a manual refinement based on the physical domain knowledge and the practical experience. For each pollutant species, a unique set of features is selected and reported in Table II. Given the dynamic nature of the investigated phenomenon, the output values are not solely determined by the current input, but also by their previous values. A sliding window of fixed width is applied to each input before feeding it to the model to account for such temporal dependence. The width of the sliding window specifies the number of contiguous samples from each input channel that are utilized to predict emissions at a certain time.

$$y(n) = f(x(n),\, x(n-1), ...,\, x(n-w)) \tag{1}$$

The function in (1) defines the calculated emissions at sample *n* and relies on the inputs from sample *n-w* up to sample *n*, where *w* denotes the window width as measured in number of samples. The previous research [14] reported a sensitivity analysis which demonstrated that increasing the window width led to improve the accuracy, yet also resulted in a concomitant increase in the computational costs. Accordingly, distinct input channels and sliding window widths were chosen for each of the models generated. Based on the findings of this analysis, the optimal balance is achieved with a window width of 30 samples, which corresponds to a time frame of 3 seconds.

A. Dataset description

The dataset used for this study is selected from a wider group of 47 experimental tests. From those acquisitions, the ECU channels shown in Table II and the corresponding measured emissions are collected and arranged in a tabular form.

TABLE II: List of features used for each pollutant species estimation

NOx	CO	HC
Engine rpm	Engine rpm	Engine rpm
Engine load	Engine load	Engine load
AFR (both banks)	AFR (both banks)	AFR (both banks)
Spark advance (SA)	Spark advance (SA)	Spark advance (SA)
IVO angle	Injection time	Injection time
EVC angle	Injection pressure	Injection pressure
Injection pressure	Water temperature	Water temperature
Exhaust temperature	Exhaust temperature	Exhaust temperature

The whole list of cycles is presented in Table III, and as can be seen, it is made up of several sorts of homologation cycles, such as RDE, WLTC, NEDC, ECE, HWFET, US06, and FTP75. A special mention should be made for RDE cycles, because, as representative of real driving conditions, they should be conducted on real roads. Nevertheless, in this case, the speed profiles acquired under real-world driving conditions are reproduced by a "virtual driver" on the roll bench. This is done for two major reasons:

- The emission measuring instruments in a laboratory environment are more accurate than the Portable Emission Measurement Systems (PEMS)
- Only at the roll bench the emissions upstream of the catalyst (engine-out) can be measured.

In this study, three different pollutant species are considered, and three separate LightGBR-based models are developed, one for each species. These models are trained using the features listed in Table II and a subset of 15 cycles is selected for training. NEDC, US06, and HWFET cycles are intentionally excluded from the training set. However, such cycles are later considered in the test set.

The remaining 32 cycles are being selected to validate the models by handling diverse homologation cycles and including those that were not considered in the training set. These cycles are being chosen to verify the robustness of the models and evaluate their performance under operating conditions that have not been encountered during the training phase. Table III provides a summary of the cycles in the dataset and the split between the train and the test sets.

TABLE III: List of driving cycles in the dataset, divided in training set and test set

Driving Cycle	Legislation	Duration	Number in Training Set	Number in Test Set
RDE	Europe	90 min	3	6
WLTC	Europe	30 min	5	9
NEDC	Europe	20 min	0	2
ECE	Europe	15 min	1	0
FTP75	America	40 min	6	7
US06	America	20 min	0	4
HWFET	America	25 min	0	4

IV. Results

The results of the model predictions are reported for all the three considered pollutant species (HC, CO and NOx). The models' output is in the form of a time series representing the concentration (expressed as ppm) of each pollutant during the driving cycle.

The results are initially presented in terms of the correlation between the cumulated, experimental emissions and the corresponding calculated values to give an overview of the results on the many experimental driving cycles. This provides a comprehensive assessment of the overall accuracy and reliability of the model. The calculated emissions are calculated as shown in (2):

$$m_p = \int_{t_{start}}^{t_{end}} \dot{m}_p(t)\, dt \tag{2}$$

where m_p is the mass of pollutant, t_{start} and t_{end} are the time instants when the driving cycle begins and ends respectively, and $\dot{m}_p$ is the pollutant mass flow computed as:

$$\dot{m}_p(t) = \rho_p\, k_p(t)\, Q_{exh}(t)\, C_p(t) \tag{3}$$

where the subscript p is referred to each pollutant species (CO, HC, NOx), ρ_p is the density of pollutant, k_p is the dry to wet emission correction factor, Q_{exh} is the exhaust volumetric flow calculated by the ECU, and C_p is the pollutant concentration in the exhausts either measured experimentally or predicted by the LightGBR-based model.

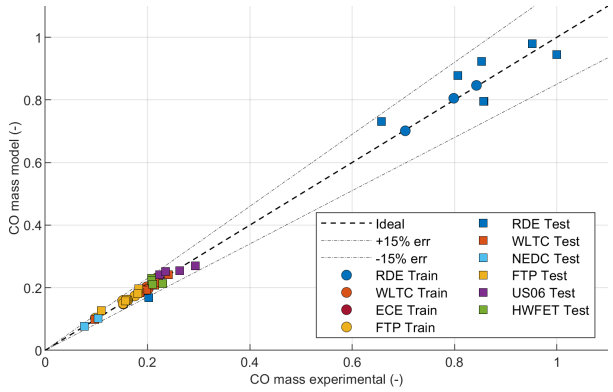
The scatter plots in Fig. 1 compares the experimental and the estimated cumulative mass of emissions for CO (Fig. 1a), HC (Fig. 1b) and NOx (Fig. 1c). Regardless of the pollutant species considered, the model is consistent with the experimental mass of emissions as can be seen looking at the dotted line that indicates the boundary for the 15% of relative error. The scatter plot represents all the cycles in the dataset, including both the training set and the test set. As expected, the correlation between experimental and calculated masses of pollutants is higher for the cycles included in the training set. This can also be seen in Fig. 2, where a bar plot reports the Mean Average Percentage Error (MAPE) of the three models on the training set and on the test set, calculated as the relative error between the actual value A_t and the forecast F_t averaged on the total number of samples n (4).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_t - Ft}{A_t} \right| \tag{4}$$

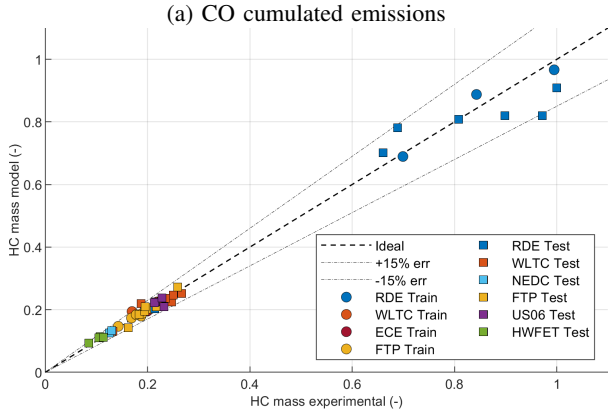
Specifically, for the test set results, the MAPE is 5.0% for CO, 5.4% for HC, and 7.4% for NOx. It is interesting to note that the error is slightly higher for NOx compared to HC and CO. Further analysis may be necessary to better understand this difference in performance.

The relative errors for each type of driving cycle, as presented in Fig. 3, demonstrate that the models' error is generally higher on the RDE cycles. This is likely due to the fact that RDE cycles represent a more realistic driving scenario that involves a wider range of maneuvers performed under highly dynamic conditions.

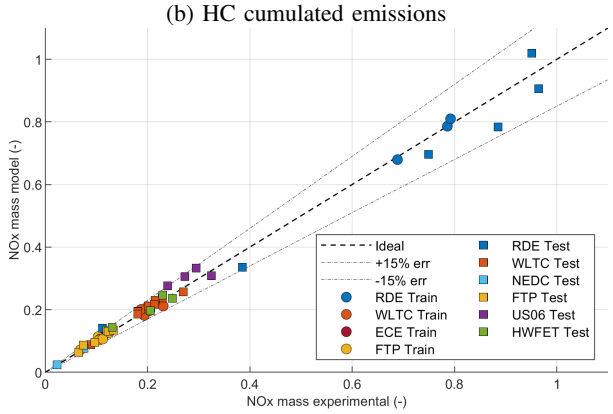
The HWFET and US06 cycles, which are not included in the model training, also exhibit slightly higher relative errors



(a) CO cumulated emissions



(b) HC cumulated emissions



(c) NOx cumulated emissions

Fig. 1: Comparison between modeled and experimental cumulated emissions over different cycles (data normalized)

than the other cycles. This is not unexpected, since the models have not been specifically trained on these types of driving cycles, and therefore may not perform as well on them as on the cycles included in the training.

It is important to note that these observations hold true for all three models.

V. Conclusions

The ultimate objective of this study was to assess the viability of a data-driven approach as an alternative technique for reducing the duration and expenses involved in the engine development phase and for enabling the real-time estimation

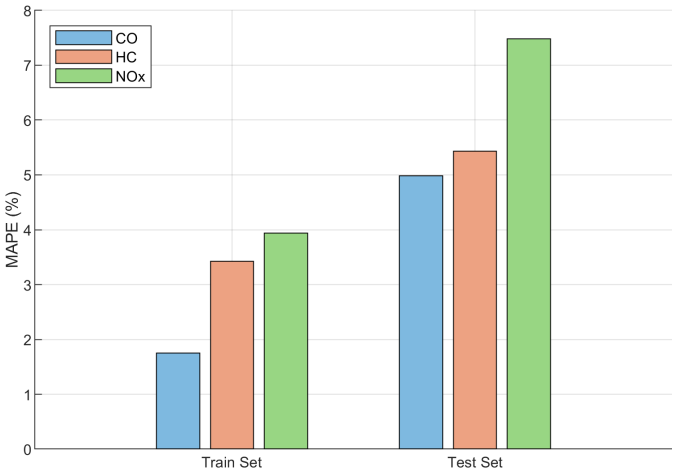
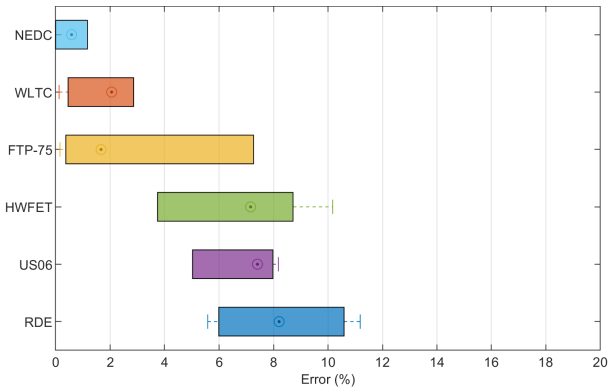


Fig. 2: MAPE of the LightGBR models applied to the train and the test set, divided per pollutant species.

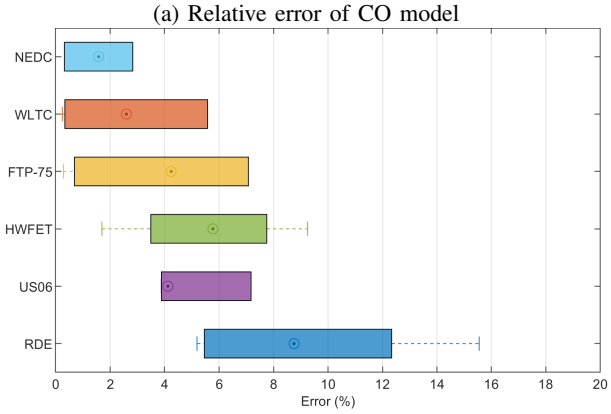
of emissions in scenarios where physical sensors cannot be installed. The primary focus of this investigation was to determine whether this approach could reduce the number of experimental tests required to calibrate the engine for emission purposes by replacing them with simulated cycles. The study extended a previously introduced methodology that involved the use of a LightGBR model to estimate pollutant emissions by testing the model on multiple homologation driving cycles of various types. The results showed that the LightGBR model provided accurate predictions of emissions mass for all the three pollutant species examined (HC, CO, and NOx) with a relative error on cumulated emissions lower than 20% even in the worst cases.

When comparing various pollutant species, it was found that the error is slightly greater for NOx (MAPE = 7.5%) in contrast to HC (MAPE = 5.4%) and CO (MAPE = 5.0%). On the other hand, when taking cycle types into account, the error is generally higher for the RDE cycles. This is because RDE cycles involve a wider range of maneuvers performed under highly dynamic conditions. Moreover, the HWFET and US06 cycles, which are not included in the model training, show slightly higher relative errors. This outcome was expected as the models have not been specifically trained for these types of driving cycles. Nonetheless, the overall accuracy of the models also in this complex and challenging scenario is an indication of the good robustness of the results.

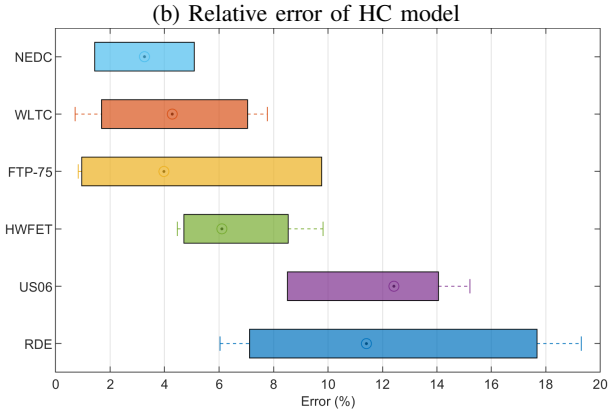
The primary implication is that the methodology presented in the study can effectively estimate pollutant emissions on different homologation driving cycles with a high degree of confidence in the results. Consequently, the impact on pollutant emissions of the engine calibration procedure can be modeled and calculated in a virtual environment, eliminating the need for a specific experimental campaign or reducing the number of tests required. Moreover, the availability of models suitable for the real-time calculation of pollutant emissions represents a strategic tool for future on-board estimations.



(a) Relative error of CO model



(b) Relative error of HC model



(c) Relative error of NOx model

Fig. 3: Relative error distribution on each model grouped by driving cycle

REFERENCES

[1] EU, "Commission regulation 2017/1154," *Official journal of the European Union*, vol. 1154, 2017.

[2] H. Wei, "Analysis on the applications of ai in vehicles and the expectation for future," 2020, doi: 10.1109/ISCTT51595.2020.00095.

[3] M. Fischer, "Transient nox estimation using artificial neural networks," *IFAC Proceedings Volumes*, vol. 46, pp. 101–106, 2013, doi: 10.3182/20130904-4-JP-2042.00006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1474667016383501

[4] V. Karri and T. N. Ho, "Predictive models for emission of hydrogen powered car using various artificial intelligent tools," *Neural Computing and Applications*, vol. 18, 2009, doi: 10.1007/s00521-008-0218-y.

[5] T. Donateo and R. Filomena, "Real time estimation of emissions in a diesel vehicle with neural networks," vol. 197, 2020, doi: 10.1051/e3sconf/202019706020.

[6] A. Jaworski, M. Mądziel, and K. Lejda, "Creating an emission model based on portable emission measurement system for the purpose of a roundabout," *Environmental Science and Pollution Research*, vol. 26, 07 2019, doi: 10.1007/s11356-019-05264-1.

[7] A. Brusa, N. Cavina, N. Rojo, J. Mecagni, E. Corti, V. Ravaglioli, M. Cucchi, and N. Silvestri, "Development and experimental validation of an adaptive, piston-damage-based combustion control system for si engines: Part 1—evaluating open-loop chain performance," *Energies*, vol. 14, 2021, doi: 10.3390/en14175367.

[8] A. Brusa, N. Cavina, N. Rojo, J. Mecagni, E. Corti, D. Moro, M. Cucchi, and N. Silvestri, "Development and experimental validation of an adaptive, piston-damage-based combustion control system for si engines: Part 2-implementation of adaptive strategies," *Energies*, vol. 14, 2021, doi: 10.3390/en14175342.

[9] F. Shethia, J. Mecagni, A. Brusa, and N. Cavina, "Development and software-in-the-loop validation of an artificial neural network-based engine simulator," 2022, doi: 10.4271/2022-24-0029.

[10] X. Niu, C. Yang, H. Wang, and Y. Wang, "Investigation of ann and svm based on limited samples for performance and emissions prediction of a crdi-assisted marine diesel engine," *Applied Thermal Engineering*, vol. 111, 2017, doi: 10.1016/j.applthermaleng.2016.10.042.

[11] N. Papaioannou, X. Fang, F. Leach, A. Lewis, S. Akehurst, and J. Turner, "A random forest algorithmic approach to predicting particulate emissions from a highly boosted gdi engine," 2021, doi: 10.4271/2021-24-0076.

[12] S. Khurana, S. Saxena, S. Jain, and A. Dixit, "Predictive modeling of engine emissions using machine learning: A review," vol. 38, 2020, doi: 10.1016/j.matpr.2020.07.204.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," vol. 2017-December, 2017.

[14] A. Brusa, E. Giovannardi, M. Barichello, and N. Cavina, "Comparative evaluation of data-driven approaches to develop an engine surrogate model for nox engine-out emissions under steady-state and transient conditions," *Energies*, vol. 15, no. 21, p. 8088, 2022, doi: 10.3390/en15218088.

[15] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, 2010, doi: 10.1093/bioinformatics/btq134.

[16] S. Elsayed, D. Thyssens, A. Rashed, H. S. Jomaa, and L. Schmidt-Thieme, "Do we really need deep learning models for time series forecasting?" *arXiv preprint arXiv:2101.02118*, 2021.