

2025

Bridging Physics and Data: Hybrid Modelling of NO_x and Soot Emissions in Alternative-Fuel Diesel Engine

Fatih OKUMUŞ

Mersin University, Marine Engineering Department, Mersin, Türkiye, f.okumus@mersin.edu.tr

Follow this and additional works at: <https://commons.yildiz.edu.tr/seatific>

Recommended Citation

OKUMUŞ, Fatih (2025) "Bridging Physics and Data: Hybrid Modelling of NO_x and Soot Emissions in Alternative-Fuel Diesel Engine," *Seatific Journal*: Vol. 5: Iss. 2, Article 4.

<https://doi.org/10.29187/2792-0771.1044>

Available at: <https://commons.yildiz.edu.tr/seatific/vol5/iss2/4>

This Research Article is brought to you for free and open access by YTU Press. It has been accepted for inclusion in Seatific Journal by an authorized editor of YTU Press.

RESEARCH ARTICLE

Bridging Physics and Data: Hybrid Modelling of NO_x and Soot Emissions in Alternative-Fuel Diesel Engine

Fatih OKUMUŞ 

Mersin University, Marine Engineering Department, Mersin, Türkiye

ABSTRACT

In this study, a basic machine learning approach based on fuel type for estimating NO_x and soot emissions in diesel engines was comparatively evaluated against a physics-based model incorporating physical indicators related to the combustion process. Experimental data were obtained from a single-cylinder, four-stroke diesel engine under single and dual fuel conditions using diesel, biodiesel, and ammonia.

Quantities derived from the in-cylinder pressure signal, including IMEP, maximum pressure rise rate, maximum pressure, the crankshaft angle corresponding to maximum pressure, the CA₁₀–CA₉₀ combustion interval, ignition delay and total combustion duration, were integrated into the model as quantitative indicators representing the engine's thermodynamic behaviour and combustion evolution, and the resulting framework is referred to in this study as a physically informed model. With the inclusion of these parameters, physical information reflecting combustion intensity, heat release dynamics, and mixture formation characteristics has been incorporated into the data-driven structure.

Models developed using linear regression, Random Forest, and Gradient Boosting methods were compared using R², RMSE, and MAE metrics. The findings revealed that the baseline model exhibited significant deviations, particularly at high NO_x and medium-high soot levels. The physics-based model significantly reduced error across all algorithms, narrowed the distribution, and largely eliminated systematic bias. The highest accuracy was achieved with the Gradient Boosting method, where the physics-based model elevated the R² value to 0.99 for NO_x predictions and again to 0.99 for soot predictions.

The results obtained demonstrate that integrating in-cylinder combustion indicators into the model significantly improves emission prediction performance and that combining physical processes with data-driven approaches provides a robust and reliable prediction framework.

Keywords: Ammonia, Biodiesel, Machine learning, Emission prediction, Physically informed model**How to cite this article:** OKUMUŞ, F. (2026). Bridging Physics and Data: Hybrid Modelling of NO_x and Soot Emissions in Alternative-Fuel Diesel Engine. *Seatific*, 5(2), 91–111. <https://doi.org/10.29187/2792-0771.1044>

1. Introduction

As long as humanity exists, the demand for energy will continue. A significant portion of this demand is still met today through internal combustion engines and turbine-based energy conversion systems. Although these systems operate on different principles, they share a common challenge in

achieving environmental sustainability, which is the reduction of harmful emissions. Emission species such as carbon dioxide (CO₂), carbon monoxide (CO), nitrogen oxides (NO_x), unburned hydrocarbons (HC), particulate matter (PM), sulfur oxides (SO_x), aldehydes, and soot produced during energy conversion pose serious risks to both the environment and human health. In the maritime sector, the

Received 18 November 2025; revised 31 December 2025; accepted 5 January 2026.
Published online 19 January 2026

E-mail address: f.okumus@mersin.edu.tr (F. OKUMUŞ).

<https://doi.org/10.29187/2792-0771.1044>

2792-0771/© 2026 Published by Yıldız Technical University Press, İstanbul, Türkiye. This is an open access article under the CC BY-NC 4.0 Licence (<https://creativecommons.org/licenses/by-nc/4.0/>).

International Maritime Organization (IMO) has implemented increasingly stringent regulations to limit harmful emissions from marine diesel engines (Göksel, 2022). Therefore, research has increasingly focused on developing and applying alternative fuels.

Alternative fuels offer advantages over fossil fuels because they can be produced from renewable resources and have the potential to reduce the carbon footprint. However, they may also adversely affect combustion characteristics, engine performance, and certain types of emissions. Thus, although alternative fuels present clear environmental advantages, they simultaneously introduce challenges and unfavourable effects in combustion behaviour that warrant thorough investigation.

To investigate these interactions, the literature primarily employs experimental or numerical methods such as Computational Fluid Dynamics (CFD). Experimental studies can provide reliable insights into engine behaviour through direct measurements. Yet, their scope is often limited due to high costs, time requirements, equipment needs, measurement uncertainties, and repeatability issues. Moreover, examining every possible fuel combination experimentally is neither economically nor environmentally sustainable.

In this regard, CFD-based combustion simulations have become powerful tools, as they can analyze in-cylinder flow, turbulence, heat transfer, and chemical reactions in three dimensions. Detailed reaction mechanisms and optimized combustion models enable high-fidelity prediction of emissions such as CO, CO₂, unburned hydrocarbons (HC), methane (CH₄), formaldehyde (HCHO), and nitrous oxide (N₂O). Since these emissions are closely related to combustion efficiency and the air–fuel ratio, modern CFD solvers can represent these effects effectively using chemical kinetic models. This complexity makes modelling difficult (Z. Wang & Yang, 2024). Furthermore, the intricate nature of chemical mechanisms imposes additional limitations. For instance, accurate NO_x prediction typically requires detailed reaction paths such as the Zeldovich mechanism (Barış, 2024), while soot modelling involves processes like nucleation, coagulation, and oxidation occurring at different spatial and temporal scales (Curinao et al., 2024).

Researchers employ engine simulators (Karatuğ, 2025) and CFD-based numerical studies (Kanberoğlu et al., 2025) to analyze combustion processes and develop strategies for minimizing harmful emissions from diesel engines. CFD simulations also depend on numerous user-defined parameters and empirical sub-models, which demand careful calibration. NO_x and soot predictions are highly sensitive to these parameters and must be re-tuned for each new engine

geometry or fuel type (Shaparia et al., 2024). Similarly, mesh resolution and time-step selection critically influence simulation accuracy. High-resolution meshes and small time steps are required to resolve turbulence and temperature gradients accurately, but they significantly increase computational cost (Sofos et al., 2025).

The complexity of modelling increases further with the introduction of alternative fuels. Low-carbon or carbon-free fuels such as biodiesel (Testa et al., 2024) and ammonia (NH₃) (Okumuş et al., 2024) are promising for diesel engines from an environmental perspective. Yet, their distinct physical and chemical properties complicate accurate NO_x and soot predictions.

In the case of biodiesel, the primary modelling challenge arises from compositional variability. The fatty acid methyl ester (FAME) composition of biodiesel varies depending on the feedstock, such as soybean, rapeseed, palm, or waste oil. This variability affects its chemical properties and makes it challenging to develop universally applicable reaction mechanisms and parameter settings in CFD models (Zettervall & Nilsson, 2024). Moreover, biodiesel's high oxygen content increases combustion temperature and promotes NO_x formation while simultaneously reducing soot due to its oxygenated structure. These opposite effects require careful balancing to maintain model accuracy (Yang et al., 2025). Biodiesel's high viscosity also affects spray behavior, atomization quality, and air–fuel mixing, leading to less homogeneous combustion. Since these physical differences directly influence NO_x and soot formation, accurate predictions are only possible with detailed spray and atomization sub-models (Radha Srikakolapu et al., 2025).

Similarly, ammonia presents unique challenges for CFD modelling. It has a long ignition delay and a low flame speed, leading to low-stability combustion zones and prolonged combustion durations, which in turn promote high NO_x formation (Mashruk et al., 2024). In addition to NO and NO₂, unburned NH₃ and N₂O also become significant emission components. Accurately predicting these species requires advanced nitrogen chemistry models beyond the classical Zeldovich approach (Mashruk et al., 2024). Furthermore, since ammonia alone often fails to provide stable combustion, it is typically blended with fuels like hydrogen or biodiesel. Modelling the combined effects of different flame speeds, turbulence intensities, and energy release rates within a single framework further increases the complexity (Sun et al., 2025; Wu et al., 2023).

Consequently, traditional CFD-based methods remain limited in both computational efficiency and

model validation, particularly for multi-fuel diesel engines. At this point, data-driven modelling techniques, especially those based on machine learning (ML), have gained growing attention. ML algorithms can achieve high-accuracy predictions by learning statistical relationships directly from observational data, without explicitly modelling every physical process.

In ML-based emission prediction models, the selection of input variables is critical for both accuracy and generalization. Literature studies employ a wide range of measured and calculated parameters representing actual engine operating conditions. For instance, a deep learning-based NOx prediction study identified engine speed, torque, EGR valve position, exhaust flow, air mass flow, and oil temperature as the most influential inputs (Qiu et al., 2024). In another study using artificial neural networks and OBD data, vehicle speed, engine speed, torque, coolant temperature, air–fuel ratio, and intake airflow were used to predict CO₂, CO, and NOx emissions with high accuracy (Seo & Park, 2023). In a model employing dimensionality reduction, 37 parameters were initially considered, and injection pressure, fuel flow rate, exhaust temperature, in-cylinder pressure, and turbo pressure were found to be dominant (Mohammad et al., 2023). For marine engines, parameter sets such as engine speed, load, fuel flow, air mass flow, and injection pressure are widely used due to their relevance to large-scale systems (M. Wang et al., 2010). Studies addressing both diesel and gasoline engines customize input variables such as air flow, boost pressure, fuel flow, load, cycle number, and ignition timing according to engine type (Tütüncü & Allahverdi, 2009). Collectively, these findings highlight the importance of selecting inputs that are both physically meaningful and information-rich with respect to the target emission type and engine architecture.

In this study, the selected input parameters are directly linked to NOx and soot formation mechanisms not only statistically but also through their physical and thermodynamic foundations. The parameters include IMEP (Indicated Mean Effective Pressure), MPRR (Maximum Pressure Rise Rate), P_{max} (Maximum Cylinder Pressure), P_{max}_CA (Crank Angle at P_{max}), CA₁₀–CA₉₀ intervals, ignition delay, and combustion duration. Each of these indicators represents fundamental dynamics of the combustion process.

IMEP reflects the engine's overall energy efficiency, while MPRR and P_{max} are closely related to combustion intensity and high-temperature regions, which are crucial for thermal NOx formation. The CA₁₀–CA₉₀ interval and combustion duration influence the

heat release rate and mixture homogeneity, affecting soot formation tendencies. Ignition delay determines the premixed fraction and early chemical kinetics, influencing both NOx and soot in opposite directions. Crank angle at P_{max} indicates combustion phasing and forms a strong link between combustion efficiency and emission characteristics.

Despite the growing number of studies employing machine learning techniques for emission prediction in internal combustion engines, most existing models rely predominantly on easily accessible operational parameters or fuel type information, treating the combustion process as a black box. Such approaches often achieve acceptable accuracy within limited operating ranges but lack physical interpretability and robustness, particularly under multi-fuel and alternative fuel conditions. In contrast, the present study explicitly integrates combustion-derived physical indicators obtained from in-cylinder pressure measurements into the machine learning framework. By incorporating parameters such as IMEP, maximum pressure rise rate, combustion phasing, ignition delay, and CA₁₀–CA₉₀ duration, the proposed model embeds thermodynamic and chemical combustion characteristics directly into the data-driven structure. This hybrid, physically informed approach differs fundamentally from conventional fuel-based or purely statistical models by linking emission formation to underlying combustion dynamics. The main contribution of this study is the demonstration that physically informed feature integration significantly enhances prediction accuracy, reduces systematic bias, and improves generalization for both NOx and soot emissions, particularly for biodiesel- and ammonia-based fuel strategies.

2. Materials and methods

2.1. Data acquisition

The experiments were carried out using a four-stroke, single-cylinder, water-cooled, direct-injection diesel engine (Antor 3LD510). The engine was operated under steady-state conditions, which allowed stable combustion and repeatable emission measurements. Single-cylinder diesel engines are widely used in fuel and combustion studies because they eliminate cylinder-to-cylinder variations and reduce cyclic dispersion, thereby improving measurement accuracy and experimental control (Koch et al., 2008). The direct-injection configuration and relatively high compression ratio enable clear observation of ignition delay, combustion phasing, and pressure rise characteristics, which are directly related to NOx and soot formation mechanisms. The experiments

were carried out at a constant speed of 1500 rpm and at three load levels: 50%, 75%, and 100%. The tested fuels included pure diesel (B0), biodiesel–diesel blends (B20, B40, B100), ammonia–diesel, and ammonia–biodiesel dual-fuel mixtures. The experimental dataset consists of 48 data points, obtained from repeated measurements under identical operating conditions. In dual-fuel operation, ammonia was introduced through the intake manifold while diesel or biodiesel–diesel blends were used as pilot fuels.

In-cylinder pressure data were recorded using an AVL QC34D piezoelectric pressure transducer, which offers high dynamic sensitivity (19 pC/bar), excellent linearity, and long service life ($>10^3$ cycles). The pressure signal was conditioned and amplified through an AVL IndiSmart GigaBit charge amplifier, ensuring high signal-to-noise ratio and data reliability. A 360-pulse AVL 365C optical crank angle encoder was employed to provide precise crank-angle synchronization between pressure and piston motion. Together, this instrumentation enabled the accurate capture of in-cylinder pressure fluctuations throughout the combustion cycle.

NO_x emissions were measured using an MRU Vario Plus exhaust gas analyzer equipped with electrochemical sensors, with measurement ranges of 0–5000 ppm for NO and 0–1000 ppm for NO₂ and a sensor sensitivity of 5 ppm. Soot emissions were measured using a Bosch BEA 070 smoke meter operating on the optical absorption principle. In this method, the attenuation of a light beam passing through a known exhaust gas path length is used to determine the smoke absorption coefficient (k), expressed in units of m^{-1} . The smoke meter has a measurement range of 0–9.99 m^{-1} , a resolution of 0.01 m^{-1} , a permissible error limit of $\pm 0.1 m^{-1}$, and an operating temperature range of 5–40 °C.

Further details regarding the experimental procedures, instrumentation, and calibration protocols are available in the referenced research work by the same laboratory group, which provides a comprehensive description of the test setup and measurement system (Kanberoğlu et al., 2025; OKUMUŞ, 2024).

2.2. Data characteristics and distribution

The experimental dataset used in this study consists of a total of 48 individual experiments conducted under different fuel strategies and engine load conditions. All experiments were completed in accordance with the planned test matrix, and no missing or discarded data points are present in the dataset. Each experimental case fully represents its corresponding operating condition and fuel configuration, and no data exclusion or imputation procedures were applied during data processing.

Table 1 presents the statistical characteristics of the experimental dataset used in this study. An examination of the table reveals relatively high standard deviation values for certain fuel-related input variables. This behaviour primarily arises from the experimental design, in which some fuels were intentionally not used under specific operating conditions. For instance, in experiments conducted with pure biodiesel (B100), the diesel fuel input is inherently zero, which naturally increases the dispersion and standard deviation of the corresponding fuel variable across the complete dataset. Similar considerations apply to other fuel-related inputs in dual-fuel operating modes.

A comparable trend is observed in the maximum heat release rate statistics, where the standard deviation appears relatively high. This behaviour is particularly pronounced in experiments with high ammonia substitution ratios. Under such conditions, the heat release profile deviates significantly from the characteristic double-peaked or “hump-shaped” structure typically associated with conventional diesel combustion and transitions toward a single dominant peak. This transformation reflects fundamental changes in ignition behaviour, flame propagation, and combustion phasing caused by the introduction of ammonia, as widely reported in the literature. The resulting variation in HRR curve morphology across different fuel strategies contributes directly to the increased dispersion observed in HRR_{max} values.

2.3. Derived parameters

Indicated Mean Effective Pressure (IMEP) was derived from the in-cylinder pressure data acquired during engine operation. The pressure measurements were recorded with respect to the crank angle and synchronized with piston displacement to generate pressure–volume (P–V) diagrams for each cycle. The indicated work (w_i) was computed by integrating the pressure over the corresponding volume change throughout one complete engine cycle. Subsequently, the IMEP was calculated by normalizing the indicated work with the displacement volume (V_d) of the cylinder. The final expression used for IMEP is as follows (Heywood, 2018):

$$IMEP = \frac{1}{V_d} \oint P dV \quad (1)$$

Where P is the instantaneous cylinder pressure and dV is the differential volume change. The integration was performed numerically over one full engine cycle using the trapezoidal rule. This parameter provides a cycle-averaged measure of the engine’s ability to

Table 1. Statistical summary of the experimental dataset.

	Mean	Median	Min	Max	StdDev	Q1	Q3	Skew
Diesel	500.05	468.50	0.00	1650.00	415.77	144.90	710.50	0.65
Biodiesel	350.51	209.50	0.00	1780.00	406.03	47.55	503.50	1.68
Ammonia	849.45	899.35	0.00	2081.87	668.57	247.78	1341.15	0.17
Nox	1281.10	1333.00	391.00	2441.00	480.15	939.25	1493.75	0.33
Soot	0.76	0.60	0.02	2.76	0.62	0.27	1.10	1.44
IMEP	7.68	7.60	5.53	9.98	1.75	5.63	9.80	0.07
PPRR	10.18	10.11	5.48	15.75	2.87	7.91	12.74	−0.08
PPRR_CA	−6.57	−6.75	−10.00	−2.70	1.70	−7.95	−5.40	0.28
Pmax	83.18	83.34	68.11	99.99	8.59	75.98	88.65	0.21
Pmax_CA	4.59	4.65	2.90	6.60	0.85	4.00	5.10	−0.07
HRR_max	63.91	64.69	30.89	102.88	19.35	52.09	77.29	0.08
HRR_max_CA	−6.88	−7.00	−10.00	−3.00	1.88	−8.25	−5.75	0.36
Burn_Duration	57.53	60.01	39.70	67.17	7.70	53.57	63.40	−0.83

convert combustion pressure into mechanical work, independent of engine size.

Maximum Pressure Rise Rate (MPRR) was calculated to quantify the intensity and aggressiveness of the combustion process. This parameter represents the steepest slope of the in-cylinder pressure curve with respect to the crank angle and is indicative of how rapidly combustion energy is released. MPRR was obtained by computing the first derivative of the in-cylinder pressure P with respect to crank angle θ , as expressed by (Heywood, 2018):

$$\left(\frac{dP}{d\theta}\right)_{\max} \quad (2)$$

Heat Release Rate (HRR) quantifies the rate at which chemical energy is converted into thermal energy during the combustion process. It provides insight into combustion phasing, efficiency, and stability. The apparent heat release rate was calculated using the first law of thermodynamics applied to a closed system, considering only the pressure–volume work and neglecting heat losses. Based on the in-cylinder pressure data and the corresponding volume trace derived from engine geometry, the HRR was computed using the following formulation (Heywood, 2018):

$$\frac{dQ_{\text{net}}}{d\theta} = \frac{\gamma}{\gamma - 1} \cdot P \cdot \frac{dV}{d\theta} + \frac{1}{\gamma - 1} \cdot V \cdot \frac{dP}{d\theta} \quad (3)$$

The cumulative heat release (CHR) was obtained by numerically integrating the instantaneous HRR with respect to the crank angle. The trapezoidal integration method was employed to calculate the total heat released up to each crank angle degree.

The burning duration was calculated based on the CRR curve derived from in-cylinder pressure data. Specifically, the duration was defined as the crank angle interval between the points where 10% and 90%

of the total heat was released, commonly referred to as CA10 and CA90, respectively.

The crank angles corresponding to specific mass fraction burned (MFB) values—such as MFB_01, MFB_10, MFB_50, and MFB_90—were determined based on the CHR curve. The heat release rate was first integrated with respect to crank angle to obtain the CHR. This cumulative energy trace was then normalized by the total released heat to compute the MFB curve. The crank angle positions at which 1%, 10%, 50%, 90%, and other specified fractions of the total heat were released were identified using interpolation.

2.4. Model performance metrics

To evaluate the prediction performance of the developed regression model, two commonly used statistical metrics were employed: Coefficient of Determination (R^2) and Root Mean Square Error (RMSE).

R^2 indicates how much of the variation in the dependent variable is explained by the model. It ranges from 0 to 1, with values closer to 1 indicating a better model fit. It is calculated as:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

Where y_i are the observed values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of the observed values.

RMSE measures the average magnitude of the prediction errors. It represents the square root of the average squared differences between predicted and actual values. Lower RMSE values indicate higher prediction accuracy. The formula is given by:

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \quad (5)$$

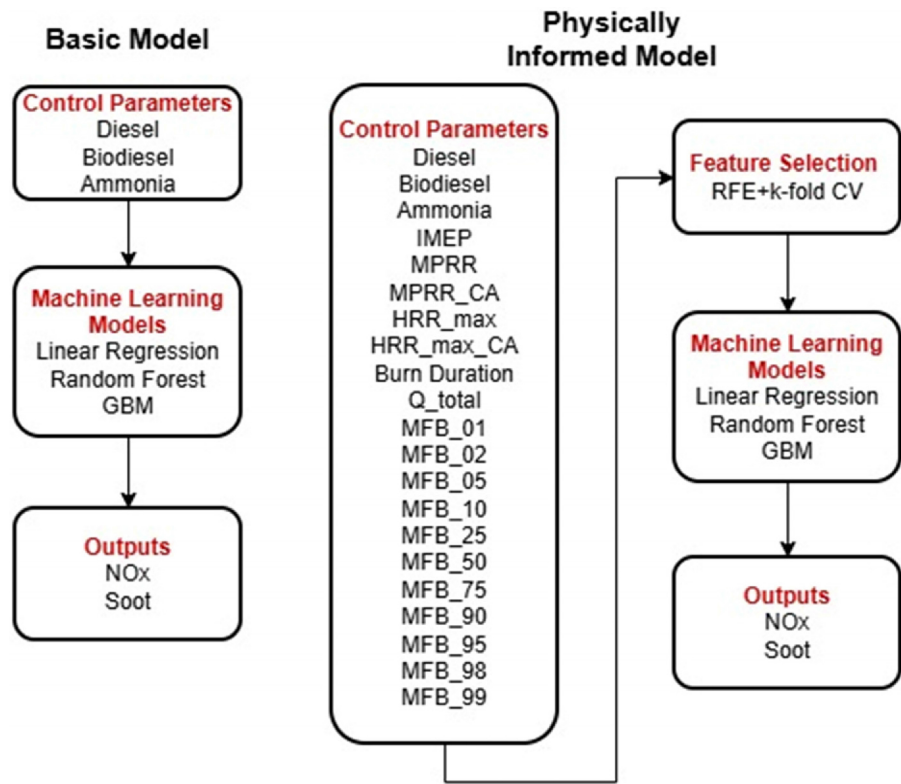


Fig. 1. Schematic representation of the basic and physically informed machine learning models developed for NOx and soot prediction.

These two metrics were calculated for both the training and test datasets to assess the model’s learning and generalization capabilities. A high R^2 value and a low RMSE value indicate that the model performs well in capturing the underlying pattern of the data.

2.5. Feature selection using recursive feature elimination

To enhance the prediction performance of the model and eliminate the influence of irrelevant variables, the Recursive Feature Elimination (RFE) method was employed. RFE is a feature selection technique that evaluates the contribution of each variable to the model and identifies the most influential inputs (C. Chen et al., 2025). In this method, the model is first trained using all variables, and the importance of each feature is assessed. Then, the least important feature is removed, and the process is repeated with the remaining variables. Through this recursive procedure, the optimal subset of features that contribute the most to model performance is determined.

During the implementation of the RFE method, k-fold cross-validation was applied to evaluate the generalization capability of the model (Ait tchak-

oucht et al., 2024). In this technique, the dataset is divided into k equal parts; in each iteration, one part is used for testing while the remaining parts are used for training. The average performance across all folds is calculated to analyze whether the selected features consistently yield high performance during validation.

This combined approach aims to reduce the risk of overfitting and include only statistically meaningful inputs in the model. Consequently, the interpretability of the model is improved, resulting in a more robust and reliable predictive structure.

Fig. 1 schematically illustrates two different machine learning approaches developed for the prediction of NOx and soot emissions: the Basic Model and the Physically Informed Model. In the Basic Model, only the fuel types (diesel, biodiesel, and ammonia) were used as input variables. In contrast, the Physically Informed Model incorporated additional combustion-related parameters derived from in-cylinder pressure, enabling the model to reflect both fuel composition and the underlying combustion dynamics.

To prevent overfitting and identify the most influential variables for each regression algorithm, the Recursive Feature Elimination method was applied in combination with k-fold cross-validation. In addition

to the RFE approach, statistical model selection criteria were employed to determine the optimal feature subset. Among them, the Bayesian Information Criterion (BIC) was utilized to balance model complexity and goodness of fit.

BIC is defined as (Afape et al., 2024):

$$BIC = n \ln(\hat{\sigma}^2) + k \ln(n) \quad (6)$$

As a result, the set of selected input variables differed among the machine learning models (Linear Regression, Random Forest, and GBM), depending on their sensitivity to the physical features. Therefore, the Physically Informed Model provides a more comprehensive and physically consistent framework that couples data-driven prediction with combustion physics.

2.6. Mathematical and theoretical background of regression models

2.6.1. Linear model

Linear regression is one of the most fundamental supervised learning algorithms used to describe the linear relationship between a dependent variable and one or more independent variables. It assumes that the output variable can be expressed as a linear combination of the input features and a residual error term. Despite its simplicity, it constitutes the mathematical foundation for many regression and optimization methods in machine learning. The general mathematical representation of the multiple linear regression model is given by (Sangeetha & Alfia, 2024):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \quad (7)$$

where y_i is the predicted response for the i^{th} observation, x_{ij} denotes the j^{th} input feature, β_j are the regression coefficients representing the contribution of each feature, and ε_i is the random error term. In matrix form, the model can be explicitly written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (8)$$

To predict NOx emissions, a multiple linear regression model was developed using a Best Subset Selection approach. The goal was to identify the optimal combination of independent variables that best describe the variation in NOx while avoiding overfitting.

The dataset was first divided into training (70%) and test (30%) subsets, and all possible feature combinations were evaluated using an exhaustive search method. The Bayesian Information Criterion was adopted as the primary model selection criterion because it provides an effective balance between model accuracy and complexity by penalizing unnecessary variables.

Among the candidate models, the one with the lowest BIC value (−33.74) included eight variables:

$$\begin{aligned} \text{NOx} = & \beta_0 + \beta_1(\text{Diesel}) + \beta_2(\text{Biodiesel}) \\ & + \beta_3(\text{Ammonia}) + \beta_4(\text{IMEP}) + \beta_5(\text{PPRR}) \\ & + \beta_6(\text{PPRR_CA}) + \beta_7(\text{Pmax}) + \beta_8(\text{Pmax_CA}) \end{aligned}$$

Among the 21 input variables tested, the terms Diesel, Biodiesel, Ammonia, IMEP, PPRR, PPRR_CA, Pmax, and Pmax_CA yielded the most suitable combination for achieving an optimal balance between model accuracy and complexity. Although some larger models provided slightly higher R^2 values, the selection was guided by the criterion favouring parsimonious models with strong explanatory power.

Similarly, the same approach was applied for soot prediction using the same modelling framework. The dataset was again divided into training and test subsets, and all possible combinations of variables were examined through exhaustive subset selection. The criterion favoring simpler yet informative structures indicated that the most suitable model included Diesel, Biodiesel, Ammonia, IMEP, and PPRR as predictors. Although several larger models exhibited marginally higher determination coefficients, these were disregarded to prevent overfitting and maintain interpretability.

The soot regression equation can therefore be expressed as:

$$\begin{aligned} \text{Soot} = & \beta_0 + \beta_1(\text{Diesel}) + \beta_2(\text{Biodiesel}) \\ & + \beta_3(\text{Ammonia}) + \beta_4(\text{IMEP}) + \beta_5(\text{PPRR}) \end{aligned}$$

2.6.2. Random forest

Random Forest (RF) regression is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. It was originally introduced by Breiman (2001) as a stochastic extension of bagging (bootstrap aggregating) to reduce the correlation among individual trees through feature randomness (Breiman, 2001).

The Random Forest model can be formally described as an ensemble of K regression trees $h(x; \theta_k)$,

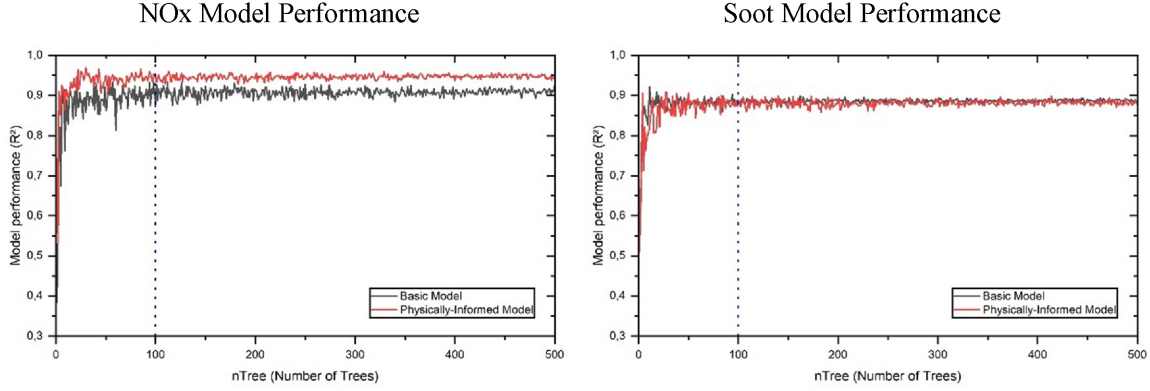


Fig. 2. Effect of tree number on model performance for the basic and physically-informed frameworks.

where each tree is constructed using a random vector θ_k sampled independently from the same distribution. The final prediction of the random forest for an input vector x is given by the average of the individual tree predictions (Breiman, 2001):

$$\hat{f}(x) = \frac{1}{K} \sum_{k=1}^K h(x; \theta_k) \quad (9)$$

The out-of-bag (OOB) samples (the approximately $1 - e^{-1} \approx 1/3$ of observations not included in each bootstrap sample) are used to estimate the generalization error without requiring a separate validation dataset. The OOB prediction for the i^{th} observation is obtained by averaging only over trees where i was not included in the bootstrap sample (Breiman, 2001):

$$\hat{y}_i^{(OOB)} = \frac{1}{|T_i|} \sum_{k \in T_i} h(x_i; \theta_k) \quad (10)$$

where T_i is the set of trees for which observation i is out-of-bag.

The OOB mean squared error (MSE) provides an unbiased estimate of the generalization error (Breiman, 2001):

$$MSE_{OOB} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{(OOB)})^2 \quad (11)$$

To ensure a consistent comparison between algorithms, the same set of input variables selected for the linear regression model was also employed in the Random Forest regression framework. Using identical predictors allows for an unbiased evaluation of how different modelling strategies, one purely linear and another nonlinear ensemble, capture the underlying relationships between the combustion parameters and the target emissions (NOx and soot).

Fig. 2 illustrates the variation of model performance (R^2) with the number of trees for the Basic and Physically-Informed frameworks. In the Random Forest regression stage, the same input variables identified in the linear model were used for both the Basic and Physically Informed frameworks to ensure a consistent comparison. This design allows differences in performance to be attributed solely to the modelling approach rather than to variations in input configuration.

The number of trees (nTree) is a critical hyperparameter in Random Forest models, as it defines the total number of individual decision trees that constitute the ensemble (Manafifard, 2024). Each tree is trained on a random subset of data and features, and the final model output is obtained by averaging the predictions across all trees. Increasing nTree typically reduces model variance and stabilizes prediction accuracy, but beyond a certain point the improvement becomes marginal while computational cost increases (Oshiro et al., 2012).

To identify the optimal nTree value, a parametric scan was conducted over the range of 0–500 trees for both the Basic and Physically Informed models. During this tuning, each configuration was evaluated using the same training dataset and hyperparameter settings. The resulting performance curves (Figure X for NOx and Figure Y for soot) demonstrate that the coefficient of determination (R^2) gradually increases with nTree and then reaches a plateau as the ensemble size grows.

For NOx prediction, the R^2 values stabilized at approximately nTree = 101, whereas for soot prediction, convergence occurred at nTree = 102. These points were therefore selected as the optimal ensemble sizes, balancing predictive accuracy and computational efficiency. The convergence of R^2 beyond these thresholds indicates that additional trees do not meaningfully improve model generalization.

It is also noteworthy that the out-of-bag (OOB) evaluation confirmed the robustness of these settings. Since roughly one-third of the data in each bootstrap sample remains out-of-bag, the OOB estimate provides an internal validation metric equivalent to cross-validation. The stable OOB error trends observed beyond the selected nTree values further verify that both models achieved sufficient ensemble diversity without overfitting.

Overall, this systematic tuning of nTree ensured that the Random Forest models for both NOx and soot captured the nonlinear relationships among fuel type, in-cylinder parameters, and emission formation mechanisms with high accuracy while maintaining computational tractability.

2.6.3. Gradient boosting regression

Gradient Boosting Regression (GBM) is an ensemble learning technique that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. Each tree is trained to correct the residual errors of the previous ensemble, gradually minimizing the loss function. The method was introduced by Friedman (2001) and has since become one of the most powerful regression algorithms due to its high flexibility and accuracy (Friedman, 2001).

The GBM algorithm starts with an initial model that minimizes the chosen loss function (Friedman, 2001):

$$Y = \arg \min_{\hat{Y}_i} \sum_{i=1}^N L(Y_i, \hat{Y}_i) \quad (12)$$

At each iteration, the algorithm computes the pseudo-residuals, which are the negative gradients of the loss function with respect to the model predictions (Friedman, 2001):

$$g_i^{(m)} = - \left[\frac{\partial L(Y_i, F(X_i))}{\partial F(X_i)} \right]_{F(X)=F_{m-1}(X)}, \quad i = 1, 2, \dots, N \quad (13)$$

A new weak learner (typically a regression tree) is then fitted to these residuals, and the optimal step size (learning rate) is determined by minimizing the loss function (Friedman, 2001):

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^N L(Y_i, F_{m-1}(X_i) + \gamma_m h_m(X_i)) \quad (14)$$

The ensemble model is updated as follows (Friedman, 2001):

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(X), \quad m = 1, 2, \dots, M \quad (15)$$

Here, $F_m(X)$ represents the boosted model after m iterations, $h_m(X)$ denotes the m^{th} weak learner, γ_m is the learning rate that scales the contribution of each weak learner, and L represents the loss function. The iterative process continues until the maximum number of boosting stages M is reached or the error improvement becomes negligible.

GBM provides several advantages, including the ability to model complex nonlinear relationships, robustness against overfitting (when regularization is applied), and flexibility in using different loss functions (e.g., mean squared error, absolute error, or Huber loss) (Jia, 2024). However, its sequential nature makes it more computationally expensive compared to parallel ensemble methods such as Random Forest.

The GBM algorithm is highly sensitive to hyperparameters such as the learning rate (shrinkage), the number of trees (n.trees), tree depth (interaction.depth), and the minimum number of observations per terminal node (n.minobsinnode) (Mwita et al., 2023; Okumuş & EkmekçiOğlu, 2021). Therefore, a systematic hyperparameter tuning process was conducted to balance predictive accuracy and the risk of overfitting. The input set was kept identical across all models, consistent with the other algorithms, ensuring that performance differences arose solely from the modelling approach rather than changes in the input features.

For the Standard NOx input configuration, a comprehensive grid search procedure was conducted to identify the optimal gradient boosting hyperparameters, and the corresponding cross validation results are illustrated in Fig. 3. The model exhibited notable sensitivity to the interaction between shrinkage, tree depth, and the number of boosting iterations. A clear trend was observed in which lower shrinkage values benefited from a greater number of trees to stabilize learning, while higher shrinkage levels required controlled complexity to avoid overfitting. Examination of the cross-validation error surface revealed that the most favorable performance was achieved with 6000 trees, an interaction depth of 7, a shrinkage value of 0.10, and a minimum terminal node size of 7. This hyperparameter combination provided the lowest RMSE among all candidates, indicating an effective balance between model flexibility and generalization.

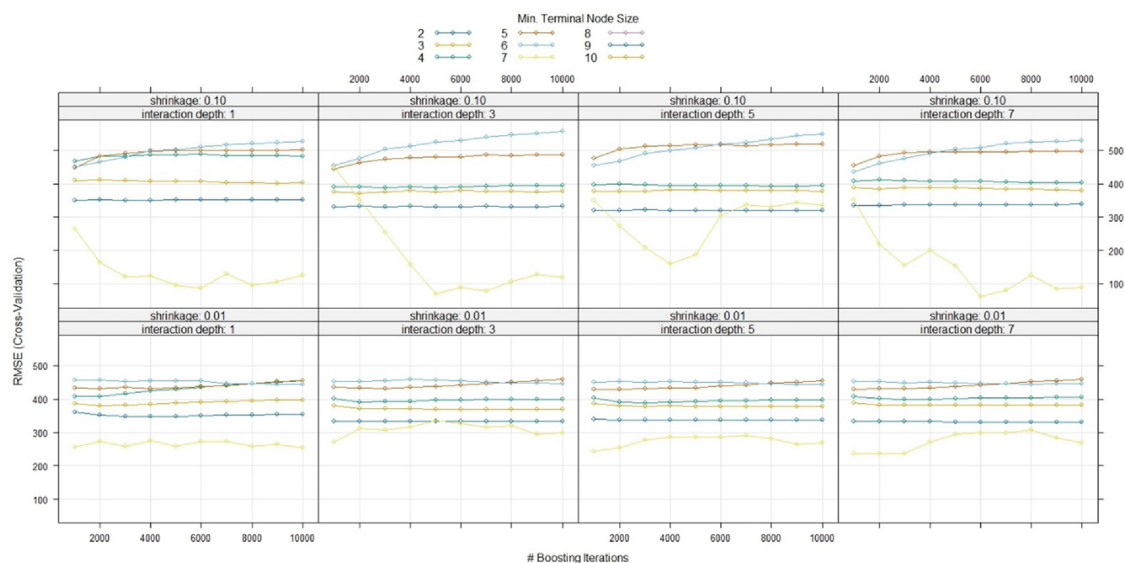


Fig. 3. Grid search results for the basic NOx model.

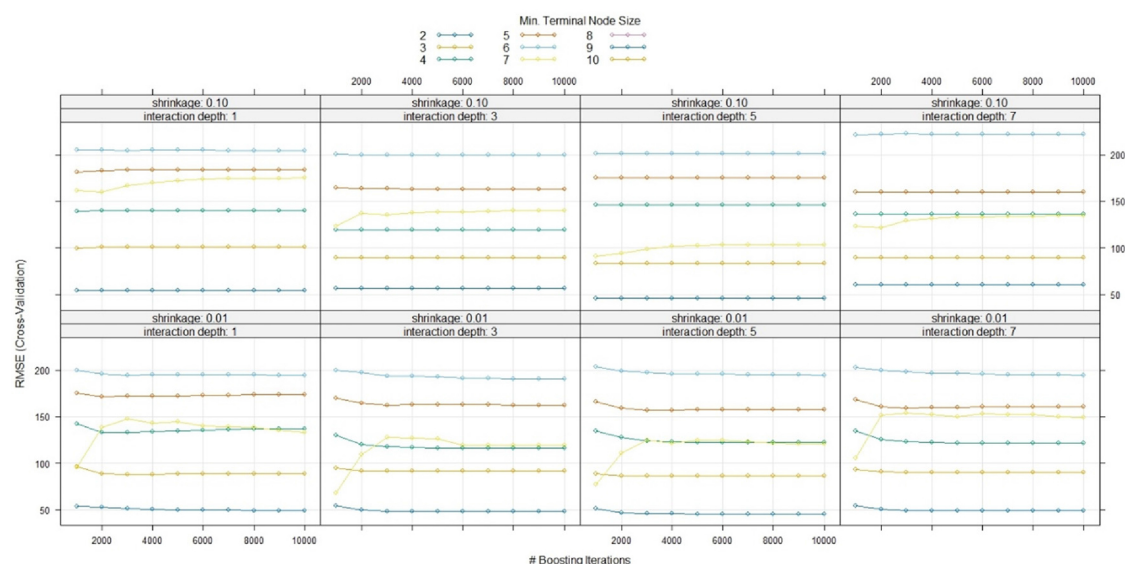


Fig. 4. Grid-search results for the physically-informed NOx model.

Overall, the selected configuration successfully captured the nonlinear combustion emission dynamics associated with NOx formation while maintaining robust predictive behavior across validation folds.

Fig. 4 shows the grid search results for the Physically-Informed NOx model. For the Enhanced NOx feature set, gradient boosting hyperparameters were tuned over the same predefined grid applied to all other models, ensuring methodological consistency. The search encompassed variation in learning rate, tree depth, minimum terminal node size, and the number of boosting iterations. The objective was to identify the configuration yielding the lowest cross-

validation RMSE within the defined parameter space. Based on this procedure, the optimal hyperparameter combination for the Enhanced NOx model was determined as 10,000 boosting iterations, an interaction depth of 5, a shrinkage value of 0.01, and a minimum terminal node size of 2. These settings were subsequently adopted for model training and evaluation.

The grid search results for the Basic Soot model are shown in Fig. 5. For the Standard soot input configuration, the gradient boosting model was tuned using the same hyperparameter search grid employed for the other cases, ensuring a consistent optimization framework. The tuning procedure evaluated

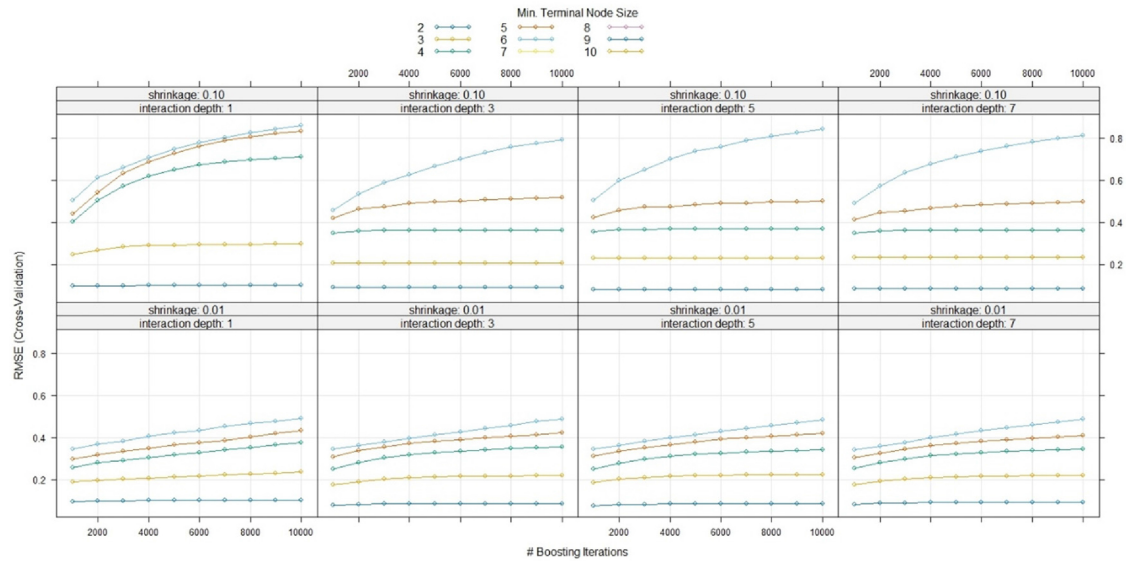


Fig. 5. Grid search results for the basic soot model.

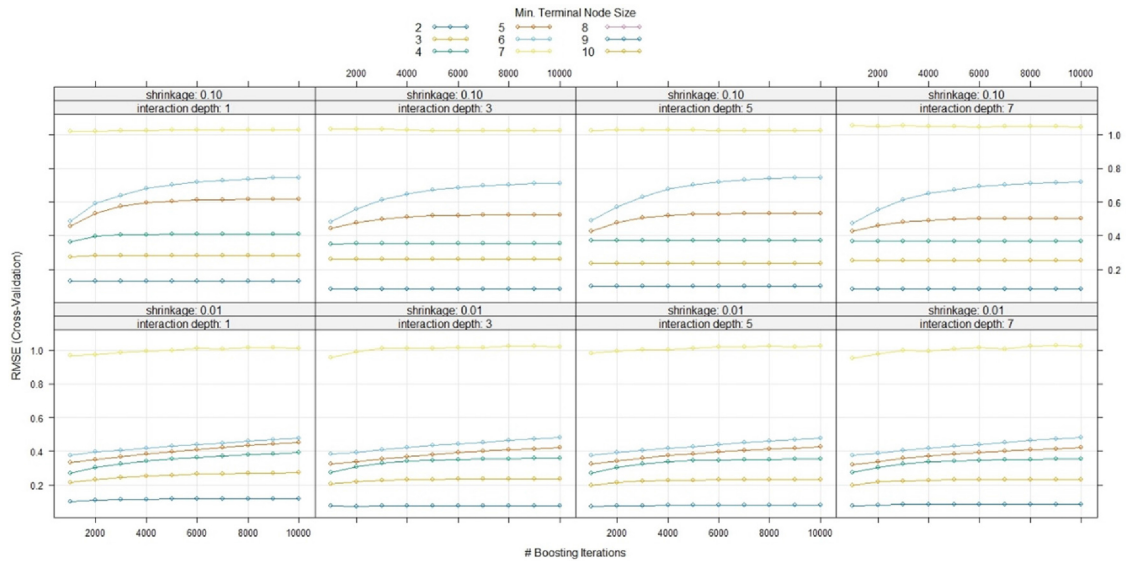


Fig. 6. Grid search results for the physically-informed soot model.

combinations of learning rate, interaction depth, minimum terminal node size, and number of boosting iterations to identify the setting that minimized cross-validation RMSE. Following this systematic search, the optimal hyperparameters for the Standard soot model were determined as 1,000 boosting iterations, an interaction depth of 5, a shrinkage value of 0.01, and a minimum terminal node size of 2. These parameters were subsequently selected for the final model training stage.

The grid search results for the Physically-Informed Soot model are shown in Fig. 6. For the Enhanced soot feature set, the gradient boosting model was tuned

using the same hyperparameter grid utilized across all remaining configurations to ensure methodological consistency. The grid search systematically explored multiple values of learning rate, tree depth, minimum terminal node size, and boosting iterations, with model performance evaluated via cross-validation RMSE. Based on this procedure, the optimal hyperparameter set for the Enhanced soot model was identified as 2,000 boosting iterations, an interaction depth of 3, a shrinkage value of 0.01, and a minimum terminal node size of 2. These parameters were subsequently adopted for the final model training process.

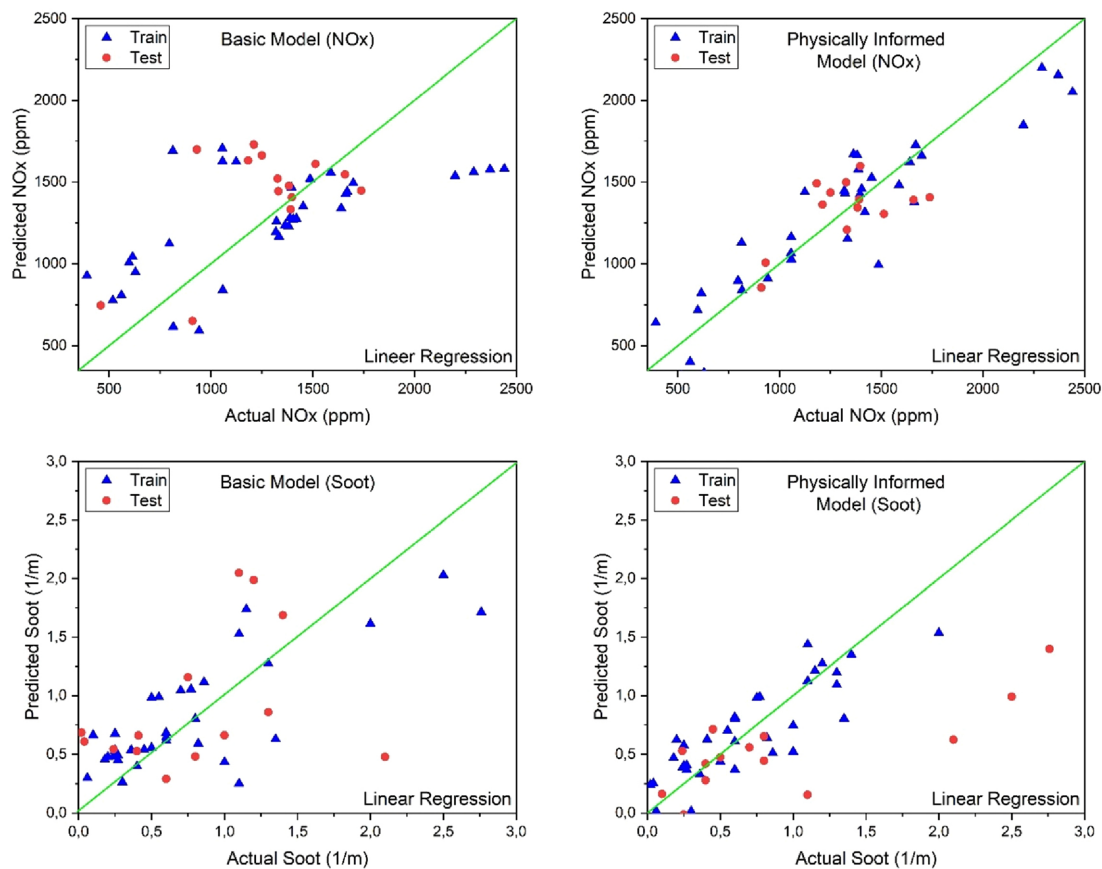


Fig. 7. Comparison of basic and physically informed models for NOx and soot prediction using linear regression.

3. Results and discussions

In this study, two different modelling approaches were evaluated for the prediction of NOx and soot emissions. The first approach, the Basic model, is a fundamental machine learning framework that utilizes only fuel types (diesel, biodiesel, and ammonia) as input variables and does not incorporate any physical parameters related to the combustion process. This model aims to statistically capture the general tendency of emission formation based on fuel composition and represents the simplest reference level for comparison, as it contains no information about the underlying physical processes. In contrast, the Physically Informed model includes not only fuel types but also combustion related parameters derived from experimental in-cylinder pressure data, such as IMEP, MPRR, Pmax, Pmax_CA, CA10–CA90, ignition delay, and combustion duration. By integrating these variables, the model reflects the chemical, thermodynamic, and temporal dynamics governing emission formation, thereby providing a physically coherent prediction framework. The development of both models involved data processing, derivation of

combustion indicators, RFE-based feature selection, evaluation of model complexity using BIC, implementation of linear regression, Random Forest, and Gradient Boosting algorithms, hyperparameter optimization, and final assessment using R^2 and RMSE performance metrics. In this context, the primary objective of the Results and Discussions section is not merely to compare the overall predictive accuracy of the models, but rather to highlight the performance distinction between the Basic and Physically Informed models and thereby assess the effect of incorporating physical knowledge into the prediction framework.

When evaluating the results in Fig. 7, it is observed that the Basic model exhibits significant deviations from the reference line in both NOx and soot predictions. Using only fuel type as an input variable, this model systematically underestimated predictions, particularly at high NOx and medium-to-high soot levels. The wide dispersion of data points indicates that the model fails to represent the physical dynamics of the combustion process. In contrast, the Physically Informed model, thanks to the use of parameters that directly reflect combustion physics,

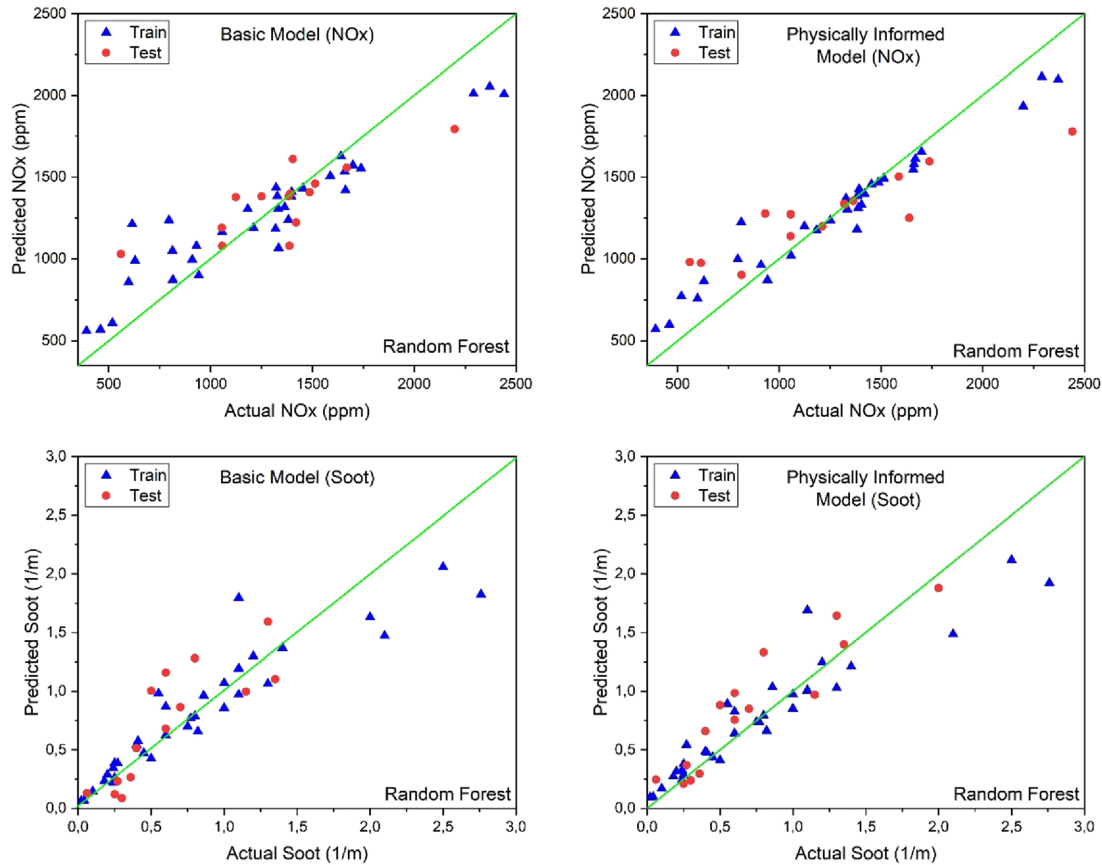


Fig. 8. Comparison of basic and physically informed models for NOx and soot prediction using linear regression.

such as IMEP, MPRR, Pmax, Pmax_CA, CA10–CA90, and ignition delay, provided a distribution much closer to the reference line in both NOx and soot predictions, with a noticeable reduction in deviations. In particular, the improvement in NOx predictions at high values and the tight clustering at low levels in soot predictions demonstrate that incorporating physical information into the model significantly increases prediction success. These findings are consistent with the fact that emission formation mechanisms are highly sensitive to physical processes such as excess combustion, pressure rise rate, combustion time, and mixture formation, independent of fuel type, confirming that the Physically Informed model offers a much more reliable and physically grounded prediction structure compared to the Basic model.

The Random Forest results presented in Fig. 8 show that the performance difference between the Basic and Physically Informed models is consistent with the trends observed in the Linear Regression model, but is more pronounced. In the Basic model, both NOx and soot predictions show significant deviations from the reference line and exhibit a systematic underestimation tendency, particularly at high NOx

levels. This situation, as in the Linear Regression results, demonstrates that a model based solely on fuel type is insufficient to represent the complex physical dynamics of the combustion process. In contrast, the Physically Informed model shows data points clustered much closer to the reference line, with a marked improvement in prediction accuracy for both training and test data. This improvement parallels the performance increase seen in the Linear Regression model, but achieves greater success due to the Random Forest's ability to capture non-linear relationships. Particularly in soot predictions, the significant reduction in the high dispersion observed in the Basic model demonstrates that the effect of including physical combustion parameters in the model is more pronounced in the Random Forest structure. Overall, the results in Fig. 3 clearly demonstrate that the Physically Informed model offers a much superior prediction performance compared to the Basic model in both linear and non-linear approaches.

The Gradient Boosting results presented in Fig. 9 demonstrate that the performance difference between the Basic and Physically Informed models is most pronounced among the three methods used in this study.

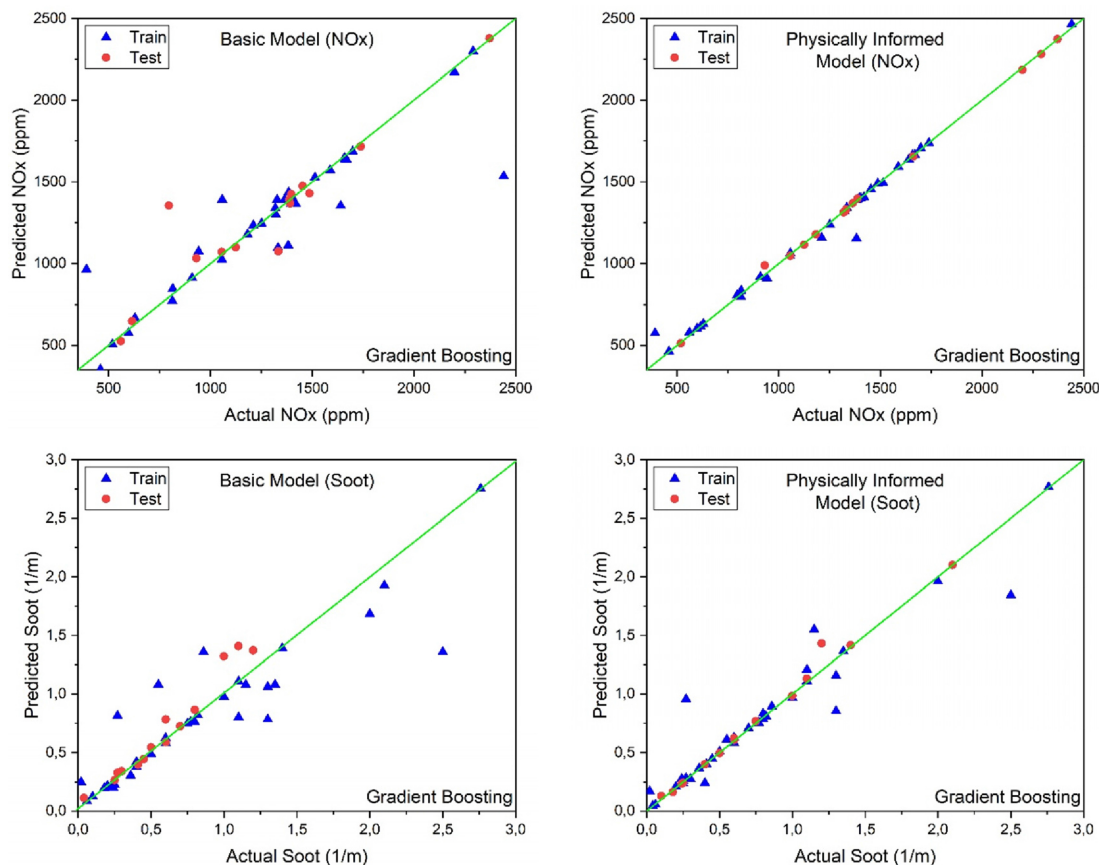


Fig. 9. Comparison of basic and physically informed models for NOx and soot prediction using GBM.

Although the Basic model shows better fit in both NOx and soot predictions compared to Linear Regression and Random Forest results, it is noteworthy that the data points still exhibit significant deviations from the reference line and a pronounced tendency to underestimate at high soot levels. This situation once again confirms that models that do not incorporate combustion physics cannot fully represent complex emission processes. In contrast, the Physically Informed model, combined with the Gradient Boosting algorithm, provided the highest accuracy among all models; it was observed that the data points showed an almost ideal alignment on the reference line in both the training and test sets. This result demonstrates that the data structure enriched with physical parameters, when combined with the sequential learning mechanism of Gradient Boosting, significantly improves emission prediction performance. In particular, the attainment of a near-error-free distribution in NOx predictions and the much lower deviations in soot predictions compared to previous models demonstrate that the strongest effect of incorporating physical information into the model emerges under the GBM structure. Overall,

Fig. 4 clearly shows that the Physically Informed model, in conjunction with the Gradient Boosting algorithm, delivered the most successful and reliable prediction performance in this study.

In the study, residuals are expressed as percentage deviations rather than absolute differences and are presented in violin plot format in the Figure. The percentage residual shows how much and in which direction the model estimate deviates from the actual measurement for each data point. This approach aimed to evaluate the relative magnitude of error in emission components with different size ranges, such as NOx and soot, on a comparable scale. The violin plots in the figures now show the density structure of the value distribution, with black boxes (25th–75th percentile range), black line extensions ($1.5 \times \text{IQR}$ range) and a white dot (median) indicating the position and spread of the distribution.

Fig. 10 shows the percentage residual distributions obtained for NOx and soot predictions using the Basic and Physically Informed models. In the Basic model, the residual distributions for both NOx and soot are spread over a wider range, with median values below zero; this indicates that the model has a systematic

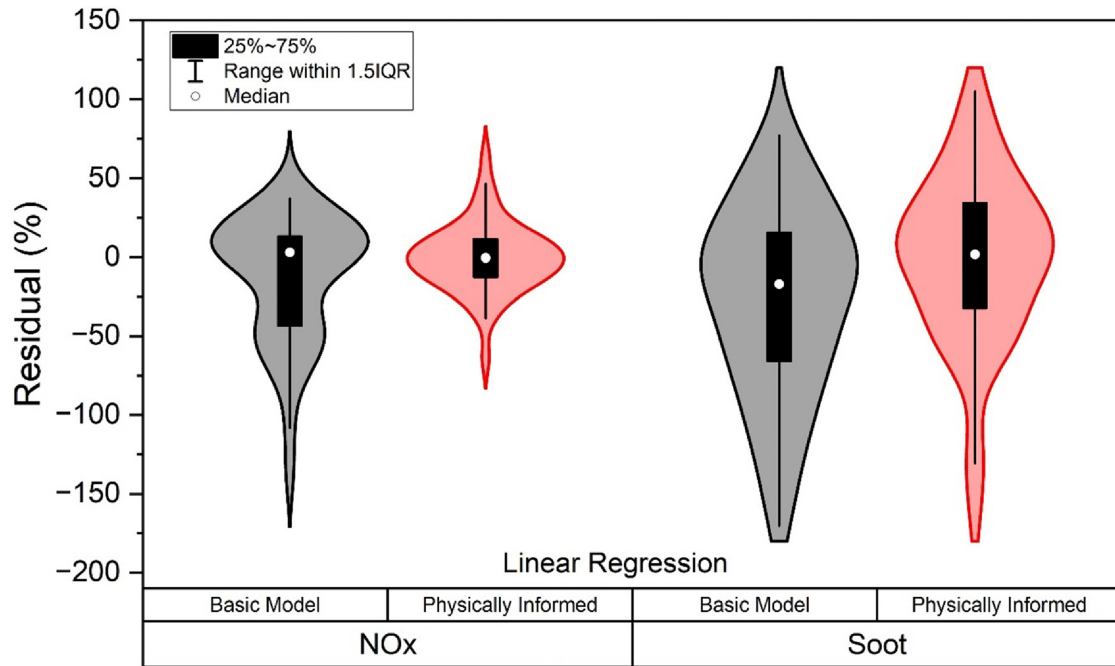


Fig. 10. Percentage residual distributions for NOx and soot predictions using linear regression.

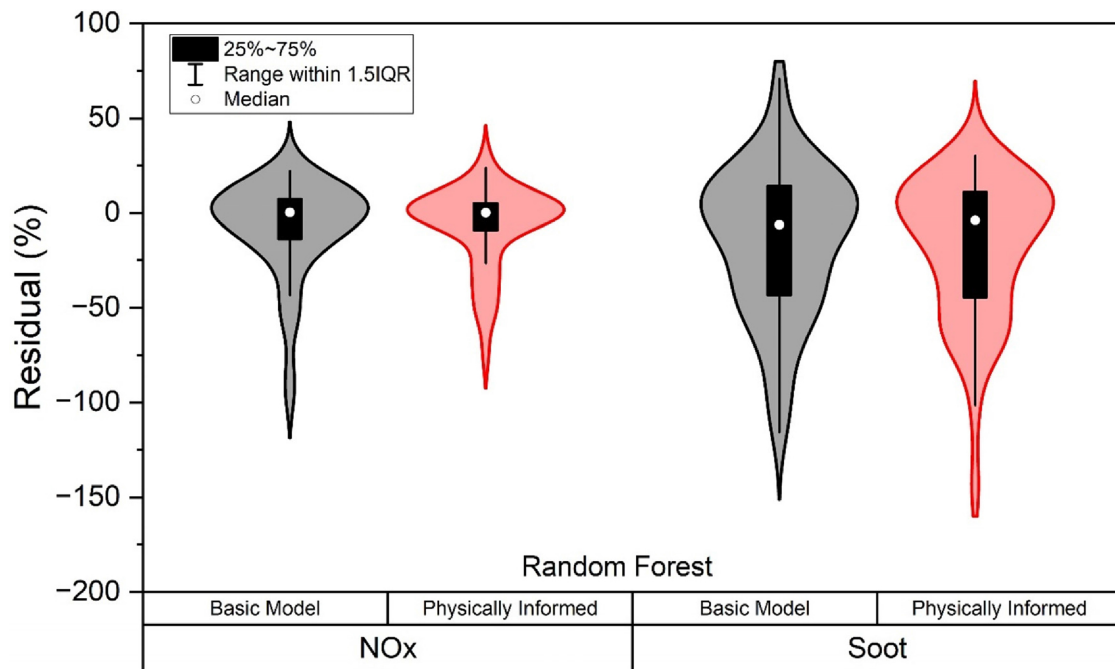


Fig. 11. Percentage residual distributions for NOx and soot predictions using random forest.

tendency to underestimate. In contrast, the Physically Informed model shows a significantly narrower distribution, fewer extreme values, and a median approaching zero. This improvement demonstrates that incorporating physical combustion parameters into the model not only reduces error but also significantly eliminates the model's prediction bias.

Fig. 11 shows the percentage residual distributions for NOx and soot estimates obtained using the Random Forest model. In the basic model, it is noteworthy that the residual distributions for both NOx and soot are spread over a fairly wide range, with extreme deviations of up to $\pm 150\%$ observed, particularly in the soot component. This

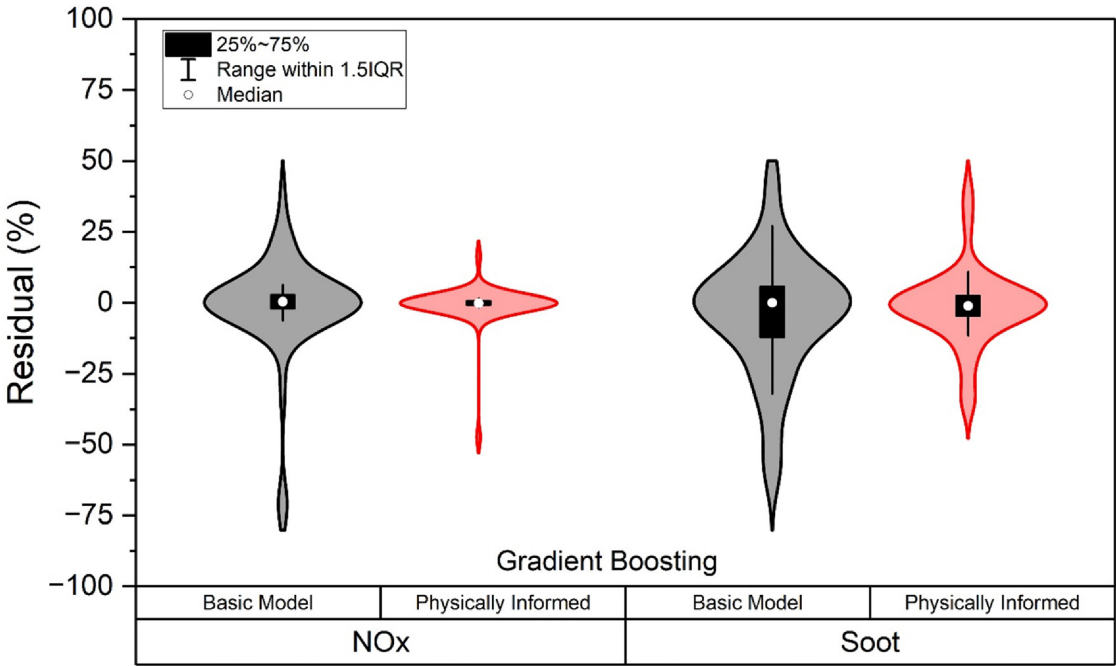


Fig. 12. Percentage residual distributions for NOx and soot predictions using GBM.

situation demonstrates that approaches lacking physical parameters are limited in representing the complex character of emission formation, despite the non-linear learning capacity of Random Forest. In contrast, the Physically Informed model shows a significantly narrowed distribution, with extreme values substantially reduced and median values approaching zero. This improvement demonstrates that adding combustion physics variables such as IMEP, MPRR, Pmax, and CA10–CA90 to the model creates a strong interaction under the Random Forest structure and significantly reduces prediction errors in terms of both magnitude and direction. In particular, the residual distribution becoming more compact for NOx and the convergence of the median on the soot side becoming more pronounced clearly demonstrate that the Physically Informed model, combined with the Random Forest algorithm, has significantly improved emission prediction performance.

Fig. 12 shows the percentage residual distributions for NOx and soot estimates obtained using the Gradient Boosting model. In the basic model, it can be seen that the residual distributions for both emission components are still spread over a relatively wide range and that there are significant deviations, particularly in the negative direction, in the soot predictions. However, the fact that the distribution is more compact compared to the Linear Regression and Random Forest models confirms that the sequential learning structure of Gradient Boosting naturally reduces the

error. In the Physically Informed model, the residual distributions reach their narrowest form, extreme values almost completely disappear, and median values approach zero. This demonstrates that physical combustion parameters, when combined with Gradient Boosting’s structured learning process, maximise emission prediction accuracy. Particularly in NOx predictions, the residual distribution concentrating within a very low range and the elimination of systematic errors on the soot side clearly demonstrate that the Physically Informed model exhibits the most stable and lowest bias prediction performance among the three methods used in this study under Gradient Boosting.

Table 2 comprehensively presents the NOx and soot prediction performance of the Basic and Physically Informed models under three different regression algorithms—Linear Regression, Random Forest, and Gradient Boosting Machine. When the results are evaluated overall, it is seen that adding physical combustion parameters to the model provides a consistent performance increase in all methods. In the Linear Regression results, it is noteworthy that the R^2 values for the Basic model remained at 0.37 for both the training and test sets in NOx prediction, whereas these values increased to 0.84 for training and 0.81 for testing in the Physically Informed model. The same trend is observed in the RMSE and MAE values, with the RMSE in the Basic model ranging from 408 to 322 ppm, while in the Physically

Table 2. Performance metrics (R^2 , RMSE, MAE) of basic and physically informed models for NOx and soot prediction across all regression algorithms.

Algorithm	Emission	Model	Set	R^2	RMSE	MAE
Linear Regression	NOx	Basic	Train	0.37	408.00	305.45
			Test	0.37	322.03	255.13
		Physically Informed	Train	0.84	204.78	146.12
			Test	0.81	242.53	181.22
	Soot	Basic	Train	0.48	0.47	0.23
			Test	0.46	0.51	0.28
		Physically Informed	Train	0.71	0.25	0.12
			Test	0.58	0.79	0.34
Random Forest	NOx	Basic	Train	0.74	226.17	142.10
			Test	0.78	229.79	157.22
		Physically Informed	Train	0.84	190.05	87.82
			Test	0.82	220.14	117.59
	Soot	Basic	Train	0.89	0.33	0.12
			Test	0.92	0.21	0.17
		Physically Informed	Train	0.90	0.20	0.10
			Test	0.99	0.06	0.03
Gradient Boosting Machine	NOx	Basic	Train	0.91	139.00	94.21
			Test	0.88	187.64	130.82
		Physically Informed	Train	0.99	29.45	15.57
			Test	0.99	43.13	28.17
	Soot	Basic	Train	0.90	0.31	0.13
			Test	0.94	0.19	0.17
		Physically Informed	Train	0.99	0.04	0.02
			Test	0.99	0.05	0.02

Informed model it is reduced by approximately half (204–243 ppm). and the MAE decreasing from 305–255 ppm to 146–181 ppm, indicating that physical inputs provide a significant improvement for linear regression. A similar trend is observed in soot predictions, where the test R^2 value of the Basic model is only 0.46, while in the Physically Informed model this value increases to 0.58, RMSE decreases from 0.47 to 0.25, and MAE decreases from 0.23 to 0.12. Random Forest results offer significantly higher accuracy compared to Linear Regression, and performance has improved further with the inclusion of physical parameters. For NOx, the R^2 values of the Basic model ranged from 0.74 to 0.78, while the Physically Informed model brought these values to 0.84–0.82 levels. MAE values decreased from 142–157 ppm to 87–117 ppm, and RMSE decreased from 226–229 ppm to 190–220 ppm. For soot prediction, RF was one of the most successful methods, with the Basic model's test R^2 value at 0.92, while the Physically Informed model increased this value to 0.99. Furthermore, the decrease in the test RMSE for soot from 0.21 to 0.06 and the MAE from 0.17 to 0.03 demonstrates that the physical combustion parameters provided a very strong and clear improvement in particle formation estimation.

The trends observed in Table 2 are consistent with findings reported in previous studies on physics-informed and combustion-feature-based ma-

chine learning models applied to emission prediction. Several studies have shown that incorporating physically meaningful combustion parameters, such as IMEP, pressure rise rate, and combustion phasing, significantly improves prediction accuracy for NOx and soot compared to models relying solely on fuel-type or operating-condition inputs (Alessandro et al., 2025). Similar reductions in prediction error and improvements in model robustness have been reported in recent works focusing on hybrid data-driven–physics-based frameworks for internal combustion engines (Z. Chen et al., 2025). Therefore, the results presented in Table 1 are in good agreement with the existing literature and further confirm the effectiveness of incorporating combustion-derived features into emission prediction models.

Despite the strong prediction performance achieved by the physically informed models, several limitations of the present study should be acknowledged. First, the dataset was obtained from a single-cylinder diesel engine operated at a fixed engine speed and limited load conditions. Therefore, the trained models may not directly generalize to engines with different geometries, combustion systems, or transient operating conditions without retraining or recalibration. Also, we acknowledge that the physically informed features used in this study rely on high-resolution in-cylinder pressure measurements, which historically represented a limitation for practical

implementation, particularly in applications where continuous pressure monitoring was not available. In large marine main engines, however, in-cylinder pressure has long been obtained routinely through indicator valves for combustion analysis and performance assessment, albeit typically on a periodic rather than continuous basis. With the progressive adoption of modern electronically controlled marine engines and advanced condition monitoring systems, access to combustion-related parameters has been steadily improving. Consequently, key combustion indicators such as IMEP, pressure rise rate, combustion phasing, and combustion duration are no longer purely experimental quantities but are increasingly becoming practically accessible parameters in marine applications. While the absolute model coefficients remain engine-specific, the proposed modelling strategy is therefore transferable and can be progressively extended to different marine engine types, fuel combinations, and operating conditions, provided that appropriate training data are available.

The improved prediction capability obtained by incorporating combustion-related parameters can be directly linked to well-established NO_x and soot formation mechanisms in diesel and dual-fuel combustion. Nitrogen oxide formation in compression ignition engines is predominantly governed by the thermal (Zeldovich) mechanism, which exhibits an exponential dependence on local flame temperature and residence time at high temperatures (Marinović & Mitrović, 2025). Parameters such as IMEP, maximum pressure, maximum pressure rise rate, and ignition delay provide indirect but physically meaningful information about in-cylinder temperature levels, combustion intensity, and heat release phasing. Ignition delay and higher MPRR generally indicate more intense premixed combustion, leading to elevated peak temperatures and, consequently, increased NO_x formation, as widely reported in the literature (Qin et al., 2026; Shen et al., 2025).

Similarly, soot formation is strongly influenced by local equivalence ratio, mixing quality, and combustion duration (Koob et al., 2026). The CA10–CA90 combustion interval and total combustion duration reflect the temporal evolution of fuel–air mixing and oxidation processes within the cylinder. Extended combustion durations and delayed heat release phases tend to promote fuel-rich regions and incomplete oxidation, favouring soot formation, whereas improved mixing and more homogeneous combustion reduce soot levels. In addition, variations in heat release characteristics observed under different fuel strategies, particularly in dual-fuel and high-ammonia conditions, alter flame structure and oxidation pathways, further affecting soot emissions (C. Chen & Liu, 2023). Therefore, the inclusion of

these combustion indicators allows the model to implicitly capture the underlying thermochemical processes governing NO_x and soot formation, explaining the observed reduction in prediction error beyond purely statistical improvements.

4. Conclusions

In this study, a fuel type based fundamental model for estimating NO_x and soot emissions in diesel engines was comparatively evaluated against a physics-based model incorporating combustion characteristics derived from in-cylinder pressure data. The impact of parameters such as IMEP, maximum pressure rise rate, maximum pressure, P_{max}-CA, CA10–CA90, ignition delay, and total combustion time, integrated into the physical model, on emission prediction performance has been quantitatively analysed.

The results obtained show that the physics-based model provides a significant increase in accuracy compared to the basic model in all machine learning algorithms. In the linear regression model, the test R² value of the basic NO_x model was 0.37, while in the physics-based model, this value increased to 0.81; RMSE decreased from 322 ppm to 243 ppm, and MAE decreased from 255 ppm to 181 ppm.

The contribution of physical information in the Random Forest algorithm has become more pronounced; while the test R² value of the basic NO_x model was 0.78, it reached 0.82 in the physics-based model. The NO_x RMSE value decreased from 229 ppm to 220 ppm, and the MAE decreased from 157 ppm to 117 ppm. In temperature estimation, although the basic model had a test R² value of 0.92, the physics-based model increased this value to 0.99; the test RMSE decreased significantly from 0.21 to 0.06, and the MAE decreased from 0.17 to 0.03.

The highest accuracy was achieved with the Gradient Boosting Machine (GBM) method. While the test R² value for NO_x was 0.88 in the basic GBM model, the physics-based model reached 0.99. The test RMSE decreased significantly from 187.6 ppm to 43.1 ppm, and the MAE decreased significantly from 130.8 ppm to 28.1 ppm. For the Is component, the test R² value of the basic model was 0.94, while the physics-based model reached a value of 0.99; the RMSE decreased from 0.19 to 0.05, and the MAE decreased from 0.17 to 0.02.

These numerical findings demonstrate that integrating cylinder pressure-based combustion indicators into the model provides high accuracy in predicting both NO_x and soot emissions. The physics-based approach, particularly with the Gradient Boosting method reaching an R² of 0.99, has provided the most reliable prediction framework for representing the emission behaviour of alternative fuel strategies.

Symbol / Abbreviation	Definition
IMEP	Indicated Mean Effective Pressure (bar)
MPRR	Maximum Pressure Rise Rate (bar/°CA)
Pmax	Maximum in-cylinder pressure (bar)
Pmax_CA	Crank angle corresponding to maximum in-cylinder pressure (°CA)
CA10–CA90	Crank angle interval between 10% and 90% cumulative heat release, representing combustion duration (°CA)
CHR	Cumulative Heat Release — total energy released as a function of crank angle
HRR	Heat Release Rate
ID	Ignition Delay
BIC	Bayesian Information Criterion
RFE	Recursive Feature Elimination
RF	Random Forest
GBM	Gradient Boosting Machine
OOB	Out-of-Bag
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
R ²	Coefficient of Determination
nTree	Number of trees in Random Forest
n.trees	Number of boosting iterations in Gradient Boosting Machine
interaction.depth	Maximum depth of individual trees in Gradient Boosting model
shrinkage	Learning rate controlling contribution of each weak learner in GBM
n.minobsinnode	Minimum number of observations per terminal node (tree leaf)
NH ₃	Ammonia — alternative fuel used in dual-fuel operation
B0, B20, B40, B100	Diesel–biodiesel blend ratios by volume (%)
CO, CO ₂ , HC, NOx, SOx, PM	Carbon monoxide, carbon dioxide, hydrocarbons, nitrogen oxides, sulfur oxides, particulate matter
Basic Model	Fuel type based machine learning model without combustion parameters
Physically-Informed Model	Data-driven model incorporating combustion-derived physical indicators
Enhanced Feature Set	Input configuration including physical combustion parameters (e.g., IMEP, MPRR, CA10–CA90)
Standard Input Configuration	Baseline input set including only fuel type variables

Nomenclature

Declaration of interests

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contribution

The entire study, “Bridging Physics and Data: Hybrid Modelling of NOx and Soot Emissions in Alternative-Fuel Diesel Engine” was conducted by Fatih OKUMUŞ.

Data availability

The datasets generated and analyzed during the current study are not publicly available due to their use in ongoing research but are available from the corresponding author on reasonable request.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Afape, J. O., Willoughby, A. A., Sanyaolu, M. E., Obiyemi, O. O., Moloi, K., Jooda, J. O., & Dairo, O. F. (2024). Improving millimetre-wave path loss estimation using automated hyperparameter-tuned stacking ensemble regression machine learning. *Results in Engineering*, 22, 102289. <https://doi.org/10.1016/j.rineng.2024.102289>.
- Ait tchakoucht, T., Elkari, B., Chaibi, Y., & Kousksou, T. (2024). Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems. *Energy Reports*, 12, 988–999. <https://doi.org/10.1016/j.egy.2024.07.002>.
- Alessandro, R., Giacomo, S., Vittorio, R., & Enrico, C. (2025). Machine learning assisted modeling of ignition delay in a light-duty gasoline compression ignition engine. *Transportation Engineering*, 20, 100316. <https://doi.org/10.1016/j.treng.2025.100316>.
- Bariş, O. (2024). Prediction of NOx emissions for hydrogen combustion engines using thermodynamical model in steady and transient conditions. *International Journal of Hydrogen Energy*, 110, 138–143. <https://doi.org/10.1016/j.ijhydene.2025.02.113>.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Chen, C., Liang, J., Sun, W., Yang, G., & Meng, X. (2025). An automatically recursive feature elimination method based on threshold decision in random forest classification. *Geo-Spatial Information Science*, 28(4), 1494–1519. <https://doi.org/10.1080/10095020.2024.2387457>.
- Chen, C., & Liu, D. (2023). Review of effects of zero-carbon fuel ammonia addition on soot formation in combustion. *Renewable*

- and Sustainable Energy Reviews, 185, 113640. <https://doi.org/10.1016/j.rser.2023.113640>.
- Chen, Z., Ju, P., Gong, M., Shi, X., Qin, C., & Shi, L. (2025). Study on control-oriented emission predictions of ammonia-diesel dual-fuel engine with combustion identification. *Energy*, 333, 137390. <https://doi.org/10.1016/j.energy.2025.137390>.
- Curinao, J., Cepeda, F., Escudero, F., Dworkin, S. B., & Demarco, R. (2024). Understanding soot formation: A comprehensive analysis using reactive models in Inverse Non-Premixed Flames. *Combustion and Flame*, 267, 113569. <https://doi.org/10.1016/j.combustflame.2024.113569>.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5). <https://doi.org/10.1214/aos/1013203451>.
- Göksel, B. F. (2022). A literature survey on exergy analyses of marine diesel engine and power systems. *Seatific Engineering Research Journal*. <https://doi.org/10.14744/seatific.2022.0008>.
- Heywood, J. B. (2018). *Internal combustion engine fundamentals* (Second edition). McGraw-Hill Education.
- Jia, Y. (2024). Research on Factor Interaction Effects and Nonlinear Relationships in Quantitative Models. 2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 106–109. <https://doi.org/10.1109/MLBDBI63974.2024.10823702>.
- Kanberoğlu, B., Okumuş, F., Sönmez, H. I., Gonca, G., Kökkülünk, G., Kaya, C., & Aydın, Z. (2025). Performance investigation and emission analysis of a diesel engine operated on hydrogen and ammonia. *Applied Thermal Engineering*, 281, 128702. <https://doi.org/10.1016/j.applthermaleng.2025.128702>.
- Karatuş, Ç. (2025). Evaluating the impact of inefficient maintenance practices of diesel engines on the ship's operational efficiency. *Seatific Journal*, 5(1). <https://doi.org/10.29187/2792-0771.1036>.
- Koch, T., Tiemann, C., Hamm, T., Ecker, H. J., & Bick, W. (2008). Single cylinder test engines. *ATZautotechnology*, 8(11), 38–42. <https://doi.org/10.1007/BF03247098>.
- Koob, P., Ferraro, F., Magens, E., Heinze, J., Soworka, T., Behrendt, T., Eggels, R. L. G. M., Hasse, C., & Nicolai, H. (2026). Exploring Soot Pathways: High-Fidelity Large Eddy Simulation Investigation of Soot Formation and Oxidation in Rich–Quench–Lean Combustion Systems Under Real Conditions. *Journal of Engineering for Gas Turbines and Power*, 148(1), 011017. <https://doi.org/10.1115/1.4069469>.
- Manafifard, M. (2024). A new hyperparameter to random forest: Application of remote sensing in yield prediction. *Earth Science Informatics*, 17(1), 63–73. <https://doi.org/10.1007/s12145-023-01156-8>.
- Marinović, L. M., & Mitrović, D. M. (2025). Analysis of NOx emission reduction and changes in exergy by flue gas re-circulation during natural gas combustion. *Thermal Science*, 29(5 Part A), 3429–3439.
- Mashruk, S., Shi, H., Mazzotta, L., Ustun, C. E., Aravind, B., Meloni, R., Alnasif, A., Boulet, E., Jankowski, R., Yu, C., Alnajideen, M., Paykani, A., Maas, U., Slefarski, R., Borello, D., & Valera-Medina, A. (2024). Perspectives on NOx Emissions and Impacts from Ammonia Combustion Processes. *Energy & Fuels*, 38(20), 19253–19292. <https://doi.org/10.1021/acs.energyfuels.4c03381>.
- Mohammad, A., Rezaei, R., Hayduk, C., Delebinski, T., Shahpour, S., & Shahbakhti, M. (2023). Physical-oriented and machine learning-based emission modeling in a diesel compression ignition engine: Dimensionality reduction and regression. *International Journal of Engine Research*, 24(3), 904–918. <https://doi.org/10.1177/14680874211070736>.
- Mwita, M., Mbelwa, J., Agbinya, J., & Sam, A. E. (2023). The Effect of Hyperparameter Optimization on the Estimation of Performance Metrics in Network Traffic Prediction using the Gradient Boosting Machine Model. *Engineering, Technology & Applied Science Research*. <https://doi.org/10.48084/etasr.5548>.
- OKUMUŞ, F. (2024). *Experimental and numerical analysis of the effects of using ammonia and biodiesel as triple fuel on performance and emissions in diesel engines*. Yıldız Technical University.
- Okumuş, F., & EkmekçiOğlu, A. (2021). Modeling of general cargo ship's main engine powers with regression based machine learning algorithms: comparative research. *Mersin University Journal of Maritime Faculty*. <https://doi.org/10.47512/meujmaf.923874>.
- Okumuş, F., Kanberoğlu, B., Gonca, G., Kökkülünk, G., Aydın, Z., & Kaya, C. (2024). The effects of ammonia addition on the emission and performance characteristics of a diesel engine with variable compression ratio and injection timing. *International Journal of Hydrogen Energy*, 64, 186–195. <https://doi.org/10.1016/j.ijhydene.2024.03.206>.
- Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 154–168). Springer. https://doi.org/10.1007/978-3-642-31537-4_13.
- Qin, H., Zhou, S., Xu, T., Zhao, Y., Yu, J., Guan, Y., & Bi, W. (2026). Insights into the co-combustion properties and NOx emissions of ammonia/dimethoxymethane via ReaxFF molecular dynamics and kinetic numerical simulations. *Fuel*, 406, 137088. <https://doi.org/10.1016/j.fuel.2025.137088>.
- Qiu, T., Liu, Z., Lei, Y., Ma, X., Chen, Z., Li, N., & Fu, J. (2024). Research on input parameter optimization for NOx deep learning prediction model. *International Journal of Engine Research*, 25(12), 2111–2124. <https://doi.org/10.1177/14680874241272818>.
- Radha Srikakolapu, S., Honganur Raju, M., Nath Thatoi, D., M B, S., Rajgor, M., Kumari, A., & K, K. P. (2025). Clean Energy Pathways for Marine Engines: Technological and Environmental Insights into Biodiesel-Alcohol Blends. *Clean Energy*, zkaf059. <https://doi.org/10.1093/ce/zkaf059>.
- Sangeetha, J. M., & Alfia, K. J. (2024). Financial stock market forecast using evaluated linear regression based machine learning technique. *Measurement: Sensors*, 31, 100950. <https://doi.org/10.1016/j.measen.2023.100950>.
- Seo, J., & Park, S. (2023). Optimizing model parameters of artificial neural networks to predict vehicle emissions. *Atmospheric Environment*, 294, 119508. <https://doi.org/10.1016/j.atmosenv.2022.119508>.
- Shaparia, N., Pelay, U., Bougeard, D., Levasseur, A., François, N., & Russeil, S. (2024). Investigation of Wall Boiling Closure, Momentum Closure and Population Balance Models for Refrigerant Gas–Liquid Subcooled Boiling Flow in a Vertical Pipe Using a Two-Fluid Eulerian CFD Model. *Energies*, 17(17), 4225. <https://doi.org/10.3390/en17174225>.
- Shen, Z., Pan, Y., Zhou, H., Zhang, D., & Zhu, T. (2025). Optimization research to minimize fuel consumption/NOx emission of a two-stroke low speed marine diesel engine. *International Journal of Engine Research*, 26(11), 1726–1741. <https://doi.org/10.1177/14680874251331307>.
- Sofos, F., Drikakis, D., & Kokkinakis, I. W. (2025). Enhancing indoor temperature mapping: High-resolution insights through deep learning and computational fluid dynamics. *Physics of Fluids*, 37(1), 015206. <https://doi.org/10.1063/5.0250478>.
- Sun, W., Wang, X., Guo, L., Zhang, H., Zhu, G., Jiang, M., You, C., Ma, X., & Ling, Q. (2025). Study on combustion process and boundary condition optimization of ammonia/biodiesel dual-fuel engine. *Process Safety and Environmental Protection*, 201, 107605. <https://doi.org/10.1016/j.psep.2025.107605>.
- Testa, L., Chiamonti, D., Prussi, M., & Bensaid, S. (2024). Challenges and opportunities of process modelling renewable

- advanced fuels. *Biomass Conversion and Biorefinery*, 14(7), 8153–8188. <https://doi.org/10.1007/s13399-022-03057-0>.
- Tütüncü, K., & Allahverdi, N. (2009). *Modeling the performance and emission characteristics of diesel engine and petrol-driven engine by ANN*. <https://doi.org/10.1145/1731740.1731803>.
- Wang, M., Zhang, J., Zhang, S., & Ma, Q. (2010). Predication Emission of an Marine Two Stroke Diesel Engine Based on Modeling of Radial Basis Function Neural Networks. *2010 Second WRI Global Congress on Intelligent Systems*, 184–188. <https://doi.org/10.1109/GCIS.2010.230>.
- Wang, Z., & Yang, X. (2024). NO_x Formation Mechanism and Emission Prediction in Turbulent Combustion: A Review. *Applied Sciences*, 14(14), 6104. <https://doi.org/10.3390/app14146104>.
- Wu, X., Feng, Y., Gao, Y., Xia, C., Zhu, Y., Shreka, M., & Ming, P. (2023). Numerical simulation of lean premixed combustion characteristics and emissions of natural gas-ammonia dual-fuel marine engine with the pre-chamber ignition system. *Fuel*, 343, 127990. <https://doi.org/10.1016/j.fuel.2023.127990>.
- Yang, Y., Ma, M., Zhou, L., Wang, W., & Li, F. (2025). Study on the effect of soot generation from metal oxide/biodiesel nanofluid fuel combustion. *Renewable Energy*, 243, 122498. <https://doi.org/10.1016/j.renene.2025.122498>.
- Zettervall, N., & Nilsson, E. J. K. (2024). Semi-global Chemical Kinetic Mechanism for FAME Combustion Modeling. *Combustion Science and Technology*, 196(7), 997–1014. <https://doi.org/10.1080/00102202.2022.2108321>.