Soot Emission Modeling of a Compression Ignition Engine Using Machine Learning

Saeid Shahpouri [*] **Armin Norouzi** [*] **Christopher Hayduk** [**]
Reza Rezaei [**] **Mahdi Shahbakhti** [*] **Charles Robert Koch** [*]

[*] *Mechanical Engineering Department, University of Alberta, Edmonton, Canada*
[**] *IAV GmbH, Gifhorn, Germany*

Abstract: Control of real driving soot emissions in diesel vehicles requires accurate predictive models for engine-out soot emissions. This paper presents an innovative modeling approach that combines a physics-based model and a black-box model to predict soot from a 4.5-liter compression ignition engine under varying load and speed conditions. The physical model is based on an experimentally validated 1D engine model in GT-power. In contrast, the black-box model is designed by investigating different machine learning approaches, including a Bayesian neural network (BNN), support vector machine (SVM), regression tree, and an ensemble of regression tree. The experimental data from running the engine at 219 load and speed conditions are collected and used for training and testing the soot model. The least absolute shrinkage and selection operator (LASSO) feature selection method is used on the GT model outputs to find the most critical parameters in soot prediction. The grey-box modeling results are compared with those from the black-box as well as the physical model. The results show that the grey-box SVM and black-box single hidden layer BNN method provide the best performance with a coefficient of determination (R^2) of 0.95. For most cases, grey-box models outperform the black-box models with the same Machine Learning (ML) algorithm by comparing R^2 of the test data, but this difference becomes negligible when a single hidden layer neural network is used.

Keywords: Diesel engines, Soot emissions, Machine learning, grey-box modeling, Physical model, data-driven modeling

1. INTRODUCTION

Heavy-duty and medium-duty diesel engines are commonly used in the transportation sector, and they are a significant source of soot emission generation which is highly harmful to human health (Omidvarborna et al., 2015). New legislation regulate soot emission under real driving emission (RDE) (EuroVI, 2016). Despite substantial engine calibration effort, robust emission control solutions can not be guaranteed. Developing predictive soot models and model-based soot emission control and calibration are a promising the solution to address RDE soot emissions in vehicles. However, predicting soot emissions is challenging and often more complex than predicting other engine-out emissions (Omidvarborna et al., 2015; Rezaei et al., 2020). State-of-the-art physical soot models still do not predict engine-out soot emissions for broad engine speed and load conditions. Furthermore, physical models for emissions often require high computational efforts, especially for soot emissions (Omidvarborna et al., 2015; Gordon et al., 2020; Norouzi et al., 2019; Shahpouri and Houshfar, 2019; Amani et al., 2018) that can not be accommodated in ECU for real-time emission control RDE conditions. This paper investigates Machine Learning (ML) methods in a black-box and grey-box framework to assess their accuracy for use in engine emission controls.

NOMENCLATURE

ANN	Artificial Neural Network
BNN	Bayesian neural network
ECU	Engine Control Unit
ERT	Ensemble of Regression Trees
GA	Genetic Algorithm
HL	Hidden Layer
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Machine Learning
MSE	Mean Square of Error
MSL	Min Samples Leaf
RBF	Radial Basis Function
RDE	Real Driving Emission
RMSE	Root Mean Square of Error
RT	Regression Tree
SNN	Shallow Neural Network
SVM	support vector machine

Data-driven or black-box emission models use the experimental engine data for emission prediction. These models use the measured inputs and outputs of the engine. Artificial Neural Network (ANN) and Support Vector Machine (SVM) are the most common data-driven models used for emission modeling of diesel engines (Khurana et al.,

[⋆] Corresponding Author: Saeid Shahpouri-(e-mail: shahpour@ualberta.ca).

2021; Norouzi et al., 2020b,a; Aliramezani et al., 2020a). Although data-driven models have low real time computation, they do not contain any physical relations of the system dynamics. They are also limited in responding to the changes in the underlying physics and, for the conditions outside of the training data, where extrapolation results in poor accuracy. In transient engine conditions this could occure. To overcome this, grey-box modeling approaches have been used to combine the advantages of the physical and supervised data-driven methods. Grey-box models usually use a simple physical model (0D or 1D) to limit computation. Detailed physical 3D models and unsupervised data-driven methods have been combined for clustering purposes, e.g., dividing the combustion chamber into different regions based on soot production (Zhou et al., 2015); however, the resulting model is still far too complex for real-time engine control. In this study, a computationally efficient 1D physical model is developed and used in a grey-box modeling platform for soot emission prediction.

For grey-box modeling, the physical model is first parameterized using the experimental data. The grey-box model then uses the experimental data (inputs and outputs) and internal states of the physical model. Grey-box emission modeling has been used for NOx, CO, and HC emission modeling in literature (Bidarvatan et al., 2014). Soot emission modeling using grey-box techniques has been studied where the grey-box model consisting of a 1D GT-Power physical model and a 3-layer ANN was used for soot emission modeling (Rezaei et al., 2020). Physical knowledge about the soot formation was used to select the ANN inputs in the data-driven portion. In a similar study (Mohammad et al., 2021), ANN and SVM methods were trained and deployed for grey box and black box emission modelling. Both studies show that grey-box emission modeling could improve the soot prediction accuracy compared to black-box data-driven methods.

The current study expands our previous works (Rezaei et al., 2020; Mohammad et al., 2021) which used only ANN and SVM for grey-box and black-box soot modeling. In this study, a new grey-box modeling platform is designed, and results are assessed for a different engine. The new platform uses a new combination of inputs that are optimally selected using search and Bayesian algorithms. Five different ML methods for soot emission prediction are evaluated. To do this, first, the engine data including in cylinder pressure was collected to parameterize the 1D physical model. Then, steady-state experimental data for wide ranges of speeds and loads were collected using the experimental engine setup. The physics-based 1D model was developed using the experimental data and calibrated by a Genetic Algorithm (GA) based on experimental in-cylinder pressure. Next, the black-box and grey-box models were parameterized and run using the data-driven ML methods. The LASSO feature selection algorithm was then employed to select the most significant features for both the black-box and grey-box approaches. Using the selected features for the black-box and grey-box methods, ML models were implemented, and their hyperparameters were optimized. Consequently, the main contributions of this work can be summarized here:

- Engine instrumentation and collection of soot emission data for a diesel engine's broad operating condition.
- Created a grey-box soot emission model by designing a physical 1D model (GT-power) and optimum ML model selection.
- Applied ML methods of Regression tree, ensemble learning, and Bayesian Neural Network compared to commonly used SVM and Shallow Neural network.

2. EXPERIMENTAL SETUP

A 4-cylinder medium-duty diesel engine (Cummins QSB4.5 160 - Tier 3/Stage IIIA) is operated, and engine-out soot is collected over a wide range of engine speeds and loads. The engine characteristics are listed in Table 1.

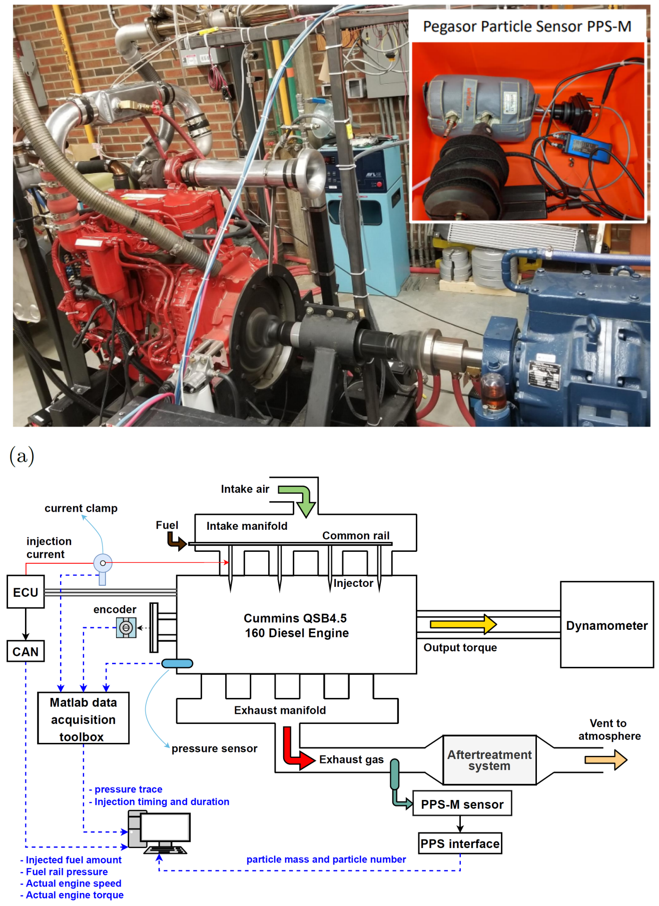
Table 1. Engine specifications

Parameter	**Value**
Engine type	In-Line, 4-Cylinder
Displacement	4.5 L
Bore × Stroke	102 mm × 120 mm
Peak torque	624 N.m @ 1500 rpm
Peak power	123 kW @ 2000 rpm
Aspiration	Turbocharged and Charge Air Cooled
Fuel Injection	3 Pulses
Certification Level	Tier 3 / Stage IIIA

The experimental engine setup is shown in Fig. 1. A Kistler piezoelectric pressure sensor and Pico current clamp are used to measure engine in-cylinder pressure and injector command signal, respectively. Additionally, fueling information, including fuel amount and rail pressure along with intake air pressure, engine speed, and load, are recorded from Cummins ECU through INSITE Pro Cummins software through the INLINE6 interface. A Pegasor Particle Sensor (PPS-M) is used to measure particulate matter (soot) emission. PPS-M is able to detect particle sizes from 1 $\mu g/m^3$ to 290 mg/m^3 with a sampling rate of 100 Hz and sensor-to-noise-ratio (SNR) equal to 100 dB which is suitable for engine exhaust.The main PPS-M sensor's characteristics are listed in Table 2. The sensor is connected to the engine-out exhaust through a heated line to measure soot and to a computer to record particle mass and particle number. The color map soot data for 219 engine stationary operating conditions from a wide range of engine speed (x-axis) and load (y-axis) are shown in Fig. 2. The black dots represent experimental points.

3. GREY-BOX AND BLACK-BOX MODELS

The physical model, black-box, and grey-box are described briefly in this section. The first step for the physical and grey-box models is to develop and parameterize the GT-Power physics-based model. The GT suite software, which contains several physical and chemical sub-models that can simulate complex combustion processes, is used to develop the diesel engine's physical model. DIPulse model is employed as the combustion model because it can deal with multi-injection combustion engines. Approximately 15% of the raw experimental data is used to calibrate



(a)

(b)

Fig. 1. Engine experimental setup

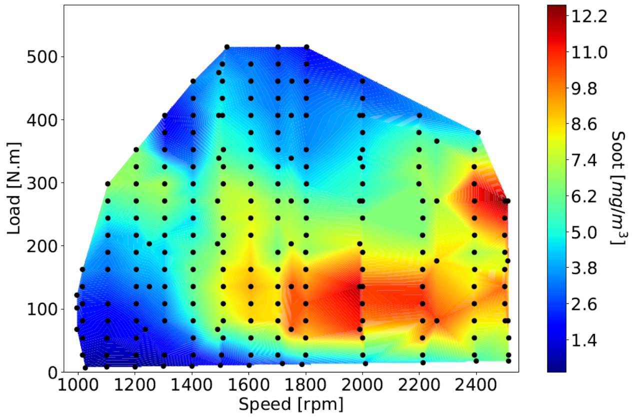


Fig. 2. Engine-out soot measurements over speed and load in mg/m^3

the combustion model using the GA algorithm in GT-suit software.

The Hiroyasu model (Hiroyasu et al., 1983) is used for physical soot prediction. This model is calibrated using 8% of the raw experimental data utilizing the GA algorithm in GT-suit software. The model calibration process is shown schematically in Fig. 3 where GA-based algorithm for optimally calibration of combustion parameters and soot model multi-players are highlighted in green and blue blocks. The GA algorithm gets the experimental data of the soot emissions and in cylinder pressure trace data for some optimization points. Then, the GA algorithms try different values for the combustion model and soot model multipliers. These multipliers for combustion model are: Entertainment Rate Multiplier, Ignition Delay Multiplier, Premixed Combustion Rate Multiplier and Diffusion Combustion Rate Multiplier. For the soot model, the multipliers are Soot Formation Multiplier and Soot Burn up Multiplier. These GA algorithms find the optimized values of these multipliers to minimize the deviation between experimental and simulation diagram of in cylinder pressure trace and value of soot emissions. It must be noted that the calibration process for soot emissions and in cylinder pressure trace are separate (2 different GA algorithm were used).

Table 2. The main PPS-M sensor specifications

Parameter	**Value**
Sensor temperature	200 °C
Extracted sample temperature	-40 up to 850 °C
Dilution	No need
Time response	0.2 s
Measured particle size range	10 nm and up
Particle number range	300 up to 10^9 1/cm^3
Particle mass range	10^{-3} up to 300 mg/m^3
Sample pressure	–20 kPa to +100 kPa
Length/Weight	40 cm/ 3.3 kg
Clean air/Nitrogen supply	10 LPM @ 0.15 MPa
Operating voltage	24 V
Power consumption	6 W

Fig. 4 shows the in-cylinder pressure trace for different load and speed conditions. Case I (136 $[N.m]$ in 1200 $[rpm]$) and case IV (353 $[N.m]$ in 2400 $[rpm]$) are among optimization points for model calibration while case II (Case II: 271 $[N.m]$ in 1600 $[rpm]$) and case III (271 $[N.m]$ in 2000 $[rpm]$) are not among calibration points. As a result, the physical model has acceptable accuracy at both calibrated and not calibrated points for various load and speed conditions.

The structure of both black-box and grey-box soot modeling are shown schematically in Fig.5. As shown in Fig.5, the grey-box and black-box model inputs are identical, including the injection properties (total mass of injected fuel, Start of Injection (SOI) and fuel rail pressure), intake manifold pressure, Brake Mean Effective Pressure (BMEP), and engine speed. These experimental features are directly used in the feature selection algorithm in the black-box model. However, these features and additional features, based on the physical-based model are used in the grey-box model. Finally, as shown in Fig. 3, the selected experimental engine data for grey-box and black-box models are used to train a data-driven model using different ML techniques.

The LASSO feature selection algorithm is used for both black-box and grey-box models to find the most effective soot prediction parameters. LASSO is a regression analysis method that performs both variable selection and regularization to enhance the model's prediction accuracy. For a given pair of training raw data $((x_{i,raw}, y_i))$ where $x_{i,raw}$ is all available raw inputs (array of 125 for grey-box

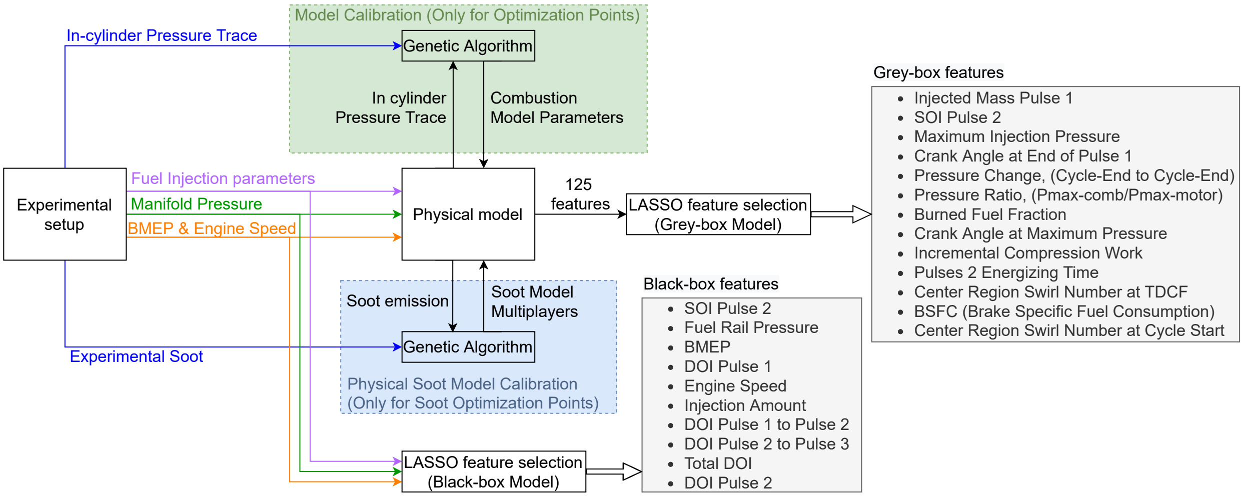

Model Calibration (Only for Optimization Points)
In-cylinder Pressure Trace
Genetic Algorithm
In cylinder Pressure Trace
Combustion Model Parameters
Grey-box features
Injected Mass Pulse 1
SOI Pulse 2
Maximum Injection Pressure
Crank Angle at End of Pulse 1
Pressure Change, (Cycle-End to Cycle-End)
Pressure Ratio, (Pmax-comb/Pmax-motor)
Burned Fuel Fraction
Crank Angle at Maximum Pressure
Incremental Compression Work
Pulses 2 Energizing Time
Center Region Swirl Number at TDCF
BSFC (Brake Specific Fuel Consumption)
Center Region Swirl Number at Cycle Start
Experimental setup
Fuel Injection parameters
Manifold Pressure
BMEP & Engine Speed
Physical model
125 features
LASSO feature selection (Grey-box Model)
Soot emission
Soot Model Multiplayers
Genetic Algorithm
Experimental Soot
Physical Soot Model Calibration (Only for Soot Optimization Points)
Black-box features
SOI Pulse 2
Fuel Rail Pressure
BMEP
DOI Pulse 1
Engine Speed
Injection Amount
DOI Pulse 1 to Pulse 2
DOI Pulse 2 to Pulse 3
Total DOI
DOI Pulse 2
LASSO feature selection (Black-box Model)

Fig. 3. Physical model calibration and feature selection process

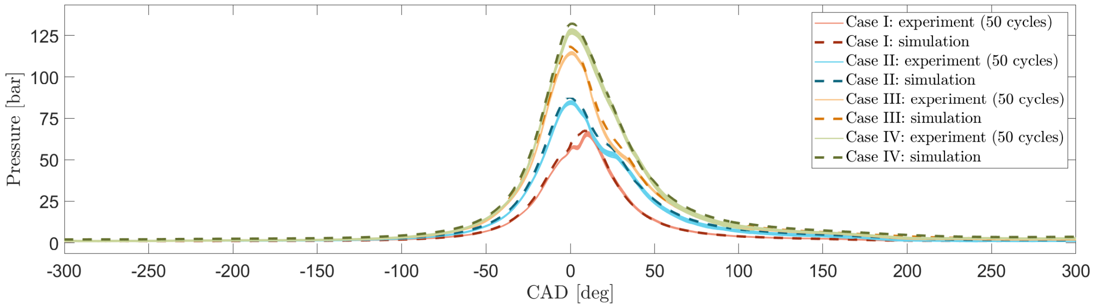


Fig. 4. Physical-based model validation for four operating points. For experimental cases, in cylinder pressure trace for 50 cycles are plotted. (Case I: 136 $[N.m]$ in 1200 $[rpm]$, Case II: 271 $[N.m]$ in 1600 $[rpm]$, Case III: 271 $[N.m]$ in 2000 $[rpm]$, and Case IV: 353 $[N.m]$ in 2400 $[rpm]$]

and array of 21 for black-box), the Lasso regression cost function is defined as

$$J(\theta) = \mathrm{MSE}(\theta) + \lambda \sum_{i=1}^{m} |\theta_i| \tag{1}$$

Where MSE is Mean Squared Error which is defined as

$$\mathrm{MSE}(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{2}$$

where $\hat{y}_i$ is a predicted output equals to $\theta^T x_i$ and m is number of training data points. In the LASSO regression, a penalty variable (λ) is used in the cost function to penalizes the l_1 norm. This tends to push the weights down to precisely zero (induces sparsity in the solution) resulting in performing an automatic feature selection (Géron, 2019). This sparsity depends on λ, which is tuned based on the cross-validation method in the current study.

The feature selection process is schematically shown in Fig. 3. Using LASSO 10 out of 21 features for the black-box model and 13 out of 125 features for the grey-box model are selected. The LASSO algorithm's penalty parameter, which affects the number of selected features, is chosen by the cross-validation method for both black-box and grey-box models.

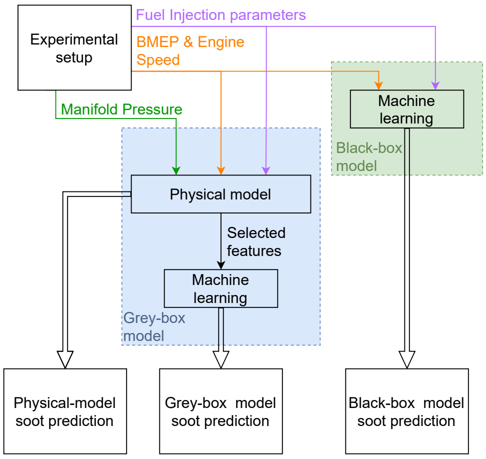


Fig. 5. Overview of the grey-box and black-box soot emission model

4. MACHINE LEARNING METHODS

Four supervised learning algorithms for regression purposes that are used are: Regression Trees (RT), Ensemble of the Regression Trees (ERT), Support Vector Machine (SVM), and Neural Network (NN). These are used to train both black and grey-box soot model.

The K-fold cross-validation algorithm with five folds is used in the models' training procedure for avoiding overfitting. K-fold cross-validation first shuffles the dataset randomly and then splits the data into k groups. In each iteration, K-fold algorithms select one group as a fold, fits a model on the rest of the groups (out of the fold), and evaluates it on the fold set (Rodriguez et al., 2009).

In general, a data-driven ML method is an optimization problem to find the best fit for a given data set subject to system constraints. For a given data set, $\mathcal{D}_{train} = (x_i, y_i)$, where x_i is input features and y_i is the measured output, ML method is an optimization problem of fitting a parameterized model, $\hat{y} = h_\theta(x_i)$, where θ is the parameters set. This optimization problem is defined as

$$\begin{aligned} \min_{\theta} \quad & J(\theta) \\ \text{s.t.} \quad & \phi(\theta) \end{aligned} \tag{3}$$

where $\phi(\theta)$ is a function of θ that represent constraints of optimization, and $J(\Theta)$ is a cost function which defined as

$$J(\Theta) = \bar{J}(\Theta) + \lambda L_2(\Theta) \tag{4}$$

where first term, $\bar{J}(\Theta)$, is defined based on error $e_i(\Theta) = \hat{y}_i - y_i$ to minimize prediction error and second term, $L_2(\Theta)$, is L_2 loss function, which is added to regulate parameters, Θ. These regulatory parameters or penalized variable, λ, provide a trade-off between the flatness of the model and minimizing the training error tolerance (Bishop, 2006).

Each ML method has hyperparameters to control the learning process in its optimization problem (Eq. 3), such as tolerated error (inside constraint function $\phi(\theta)$), regularization parameter (λ), optimization iteration stop criteria. Finding an optimum hyperparameter that yields an optimum model helps allow an automatic and efficient modeling process. In general, an ML algorithm such as A_Λ has N hyperparameters defined as $\Lambda = \lambda_1, \lambda_2, ..., \lambda_N$, the optimum hyperparameters are achieved by solving following optimization problem (Hutter et al., 2019)

$$\Lambda^* = \arg\min_{\Lambda} V(h_\theta(x_i), \mathcal{D}_{train}, \mathcal{D}_{valid}) \tag{5}$$

where $V(h_\theta)$ measures performance of a model generated by algorithm A_Λ that is calculated based on cross-validation data set $\mathcal{D}_{valid}$. In this study, Bayesian optimization (Snoek et al., 2012) is used to tune hyperparameters of RT, SVM, and Ensemble of the Regression Trees (ERT) models. Here, the MSE is used as an evaluation function in Eq. 5 as

$$V(\lambda) = MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{m}(\hat{y_i} - y_i)^2 \tag{6}$$

where $\mathrm{A}_\Lambda \in \{\mathrm{RT}, \mathrm{ERT}, \mathrm{SVM}\}$ and m is size of training set. Inside this optimization, the model is trained based on training set $\mathcal{D}_{train}$ and evaluated using cross-validation set $\mathcal{D}_{valid}$ based on K-fold algorithm. Then, $MSE(\lambda)$ is calculated based on the whole training set including both $\mathcal{D}_{train}$ and $\mathcal{D}_{valid}$. Thus, the K-fold algorithm is run inside hyperparamter optimization problem.

The grid search method (Géron, 2019) is used to tune the NN-based method (ANN and BNN). The main difference is that in grid search, all the possible hyperparameters combinations within a given range are tried, while for a Bayesian method a search along the space of hyperparameters learning with high probability is performed. The preference of a grid search for ANN and BNN over Bayesian optimization is to evaluate all combinations of layer and neuron to allow comparison to our previous work (Mohammad et al., 2021). In the grid search all combination of layer $L \in \{1, 2\}$ and neurons $s_l \in (1, 40)$ are considered where L is number of layers and s_l is number of neurons in l^{th} layer. To avoid a deeper network (due to the limited number of training data points) the layers and neuron's upper limit set to 2 layers and 40 neurons, respectively.

The summary of developed models, hyperparameter optimization method, optimized parameters, and each model configuration are listed in Table 3.

Regression Tree (RT) is a predictive modeling approach where the main algorithm to train RT is Classification and Regression Trees (CART) (Breiman et al., 1984). The cost function of RT based on notation in Eq. 3 is

$$J(\theta) = \frac{m_{left}}{m}\mathrm{MSE}_{left} + \frac{m_{right}}{m}\mathrm{MSE}_{right} \tag{7}$$

where θ has two component including k (class of instances) and t_k (threshold to split), $\hat{y} = \frac{1}{m_{node}}\sum_{i\in\text{node}} y(i)$, and the objective is to first split the training set into two subset m_{left} and m_{right} using k and t_k. One of the essential regularization parameters of the CART algorithm is that the minimum number of samples required at a leaf node (min samples leaf (MSL)) to avoid overfitting when dealing with regression tasks (Géron, 2019). In this optimization, the tree's maximum depth is five (in $\phi\theta$), which means the splitting optimization runs in five iterations. According to Table 3 the optimized value for MSL for the black-box and grey-box model are one and five, respectively.

Several decision trees are constructed in ERT training, where aggregation methods, number of learners, and MSL are three primary hyperparameters to tune. Two main aggregation methods are Bagging (short for Bootstrap aggregation) and Boosting (originally called hypothesis boosting). Bagging uses the same algorithm for every predictor but trains them on a different random subset of the training set. Boosting refers to combining several weak learners and using a sequential architecture to increase total model accuracy to make a strong learner (Géron, 2019). Different regression models such as RT or SVM can be used inside ensemble learning, but the ensemble of regression trees was used in this study. Bayesian optimization in each iteration is used for both boosting and bagging methods and it results in utilizing boosting architecture for both black-box and grey-box. As the number of grey-box features is higher than the black-box features, the ERT requires a higher number of learners for the grey-box model.

Support Vector Machine (SVM) is a powerful ML method that capable of performing both classification and regression tasks. In SVM, convex quadratic programming is solved to find a correlation between input-output. The cost function of SVM, based on notation of Eq. 3, is

$$J(\theta) = \frac{1}{2}||\theta||_2^2 + \lambda\sum_{i=1}^{m}(\zeta_i^+ + \zeta_i^-) \tag{8}$$

where $\zeta_{\mathbf{i}}^-$ and $\zeta_{\mathbf{i}}^+$ are slack variables that perform as penalty variables to overcome a possible infeasibility of an

Table 3. Training and optimization of ML-based model hyperparameters

Method	Opt. method	Opt. hyperparameters	Model type	opt. Model conf.

RT	Bayesian	min samples leaf (MSL)	black-box
			grey-box
				MSL = 1
				MSL = 5

ERT	Bayesian	Ensemble method, min samples leaf, and number of learners	black-box grey-box	Boosting, 28 Learners, and MSL = 4 Boosting, 35 Learners, and MSL = 5
SVM	Bayesian	kernel function, γ, and (λ), and ϵ	black-box grey-box	kernel function: Quadratic kernel function: Gaussian
1-HL ANN	Grid search	Number of neurons in each layer	black-box grey-box	network conf.: [19] network conf.: [4]
2-HL ANN	Grid search	Number of neurons in each layer	black-box grey-box	network conf.: [25,31] network conf.: [4, 13]
1-HL BNN	Grid search	Number of neurons in each layer	black-box grey-box	network conf.: [31] network conf.: [31]
2-HL BNN	Grid search	Number of neurons in each layer	black-box grey-box	network conf.: [16, 13] network conf.: [10, 22]

optimization problem where SVM cannot find any value to fit a function for the defined tolerance. The constraint function, $\phi(\theta)$ is defined as

$$\phi(\theta) = \begin{cases} y_i - h_\theta(x_i) \leq \epsilon + \zeta_i^+ \\ h_\theta(x_i) - y_i \leq \epsilon + \zeta_i^- \\ \zeta_i^-, \zeta_i^+ \geq 0 \end{cases} \tag{9}$$

where ϵ is maximum tolerable deviation for all training data – see (Norouzi et al., 2020b; Aliramezani et al., 2020a; Norouzi et al., 2020a) for more detail. The SVM Kernel Trick can also be used to solve optimization in high dimensional feature space instead of the input space. In this study, linear, Polynomial, and Gaussian RBF kernels are considered in optimization, which are defined as

$$K(x_i, x_j) = \begin{cases} {x_i}^T x_j & \text{Linear} \\ ({x_i}^T x_j + c)^n & \text{Polynomial} \\ exp(-\gamma \|x_i - x_j\|_2^2) & \text{Gaussian RBF} \end{cases} \tag{10}$$

where n is degree of polynomial and γ is scale of RBF kernel – see (Aliramezani et al., 2020b) for more detail. Bayesian optimization results in a Quadratic kernel for the black-box and Gaussian RBF kernel for the grey-box model. In these models, λ and ϵ are 0.78 and 0.32 for the black-box and 9.6 and 0.004 for the grey-box model.

Another standard method for soot modeling is a Neural Network (NN), a set of algorithms that distinguish the correlation between a set of data using rules containing three main layers: input layer, Hidden Layer (HL), and output layer. Shallow neural networks (Called ANN in this study) consist of only 1 or 2 HL while adding more layers creates a deep network (Hassoun et al., 1995). The cost function of NN-based modeling is defined as

$$J(\theta) = \sum_{i=1}^{m} (h_\theta(x_i) - y_i) + \frac{\lambda}{2} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\theta_{j,i}^{(l)})^2 \tag{11}$$

where L is number of total layers (including input, output, and hidden layer), s_l is number of neurons in l^{th} layer, and m is size of training set. For single layer and two-layer NN, L equals 3 and 4, respectively. The main tuning parameters of ANN are the number of the HL ($L_{HL} = L - 2$) and the number of neurons (s_2 and s_3) in the HL, where a grid search method is used to find the optimum number of neurons in each hidden layer (Mohammad et al., 2021). To add probability in ANN, Bayesian-based Shallow NN, denoted BNN, is also developed for both black and grey-box soot. In BNN, the weight and bias values are calculated using similar method to ANN while minimizing a combination of squared errors and weights to determine the correct combination to produce a network that generalizes well (Foresee and Hagan, 1997). In this optimization method, all ANN and BNN configuration combinations are evaluated, and the best model based on cross-validation data evaluation is obtained. Table 3 show single layer (3-layer network) and two hidden layers (4-layer network) configuration for ANN and BNN networks for both black-box and grey-box soot models.

All these models are evaluated using the test data set in the next part, and results will be discussed.

5. RESULTS AND DISCUSSIONS

To develop the data-driven part of both the black-box and grey-box model, the collected data is divided into the training and test data as shown in Fig. 6. For this data set, 80% is used as training data, $\mathcal{D}_{train}$, while $\mathcal{D}_{valid}$ is also included in training for K-fold validation with 5 fold. 20% (44 operating points) of all 219 operating points are randomly selected to test the developed models accuracy ($\mathcal{D}_{test}$). The test data points are only used for evaluation of the finalized model.

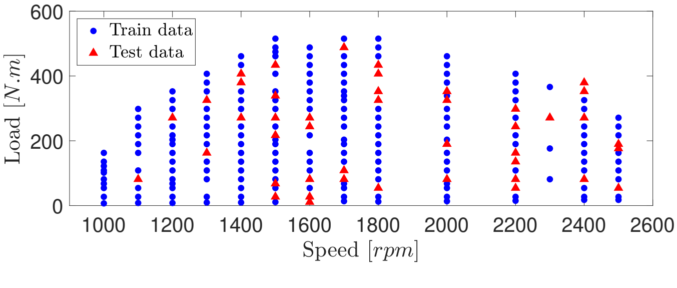


Fig. 6. Training and test data for ML approaches- 175 data points as train dataset (80%) and 44 data points as test dataset (20%)

How the methods and models perform are summarized in Table 4. The coefficient of determination R^2, Root Mean Square of Error (RMSE), and maximum of absolute prediction error $|E_{max}|$ from both training and test data are used to evaluate the different models.

The physical soot model is inaccurate in predicting soot emission as shown in Figs 7 (a) This is attributed to

Table 4. ML-based data-driven soot model comparison

Methods	Criteria	RT	ERT	SVM	1-HL NN	2-HL NN	1-HL BNN	2-HL BNN
black-box	R^2_{train}	0.98	0.99	0.97	0.97	0.98	0.99	0.99
	R^2_{test}	0.87	0.91	0.93	0.90	0.92	0.95	0.94
	$RMSE_{train}$ $[mg/m^3]$	0.48	0.52	0.66	0.66	0.63	0.22	0.20
	$RMSE_{test}$ $[mg/m^3]$	1.33	1.07	0.98	1.19	1.10	0.83	0.93
	$\|E_{train,max}\|$ $[mg/m^3]$	1.6	2.0	2.5	3.2	3.2	1.1	1.1
	$\|E_{test,max}\|$ $[mg/m^3]$	5.0	3.1	4.8	4.4	4.5	2.9	4.3
grey-box	R^2_{train}	0.97	0.99	0.98	0.96	0.96	0.99	0.99
	R^2_{test}	0.92	0.93	0.95	0.90	0.92	0.95	0.95
	$RMSE_{train}$ $[mg/m^3]$	0.62	0.06	0.48	0.73	0.72	0.34	0.09
	$RMSE_{test}$ $[mg/m^3]$	1.09	1.00	0.81	1.2	0.88	0.88	0.97
	$\|E_{train,max}\|$ $[mg/m^3]$	1.9	0.2	1.6	2.8	3.1	1.5	0.4
	$\|E_{test,max}\|$ $[mg/m^3]$	2.9	3.7	1.9	3.6	2.3	2.3	2.6

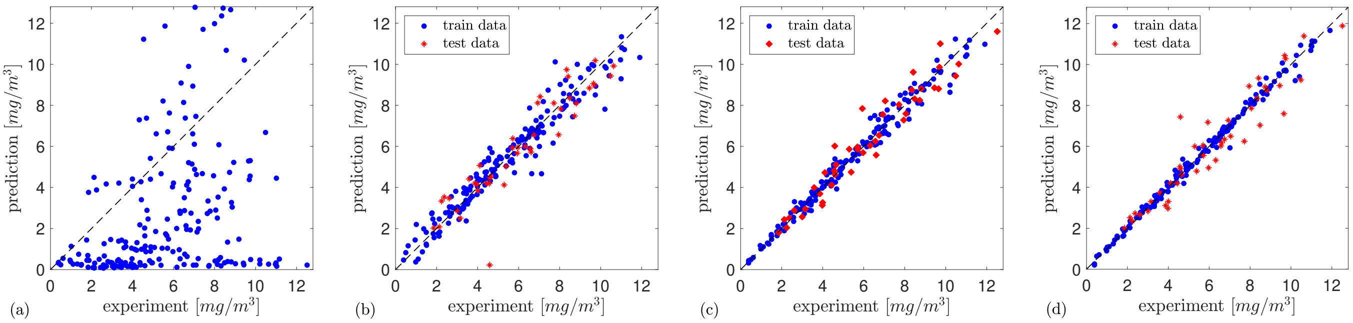


Fig. 7. Prediction vs experiment: (a) Physics-based model, (b) black-box SVM, (c) grey-box SVM, (d) black-box BNN

the complex soot formation and oxidation process that depends on many factors that can not be captured by a 1D physical model (Omidvarborna et al., 2015). This trend was observed in the previous study that used 1D physical GT power model for soot emissions prediction (Rezaei et al., 2020). This motivates the use of data-driven methods for soot emission prediction. The results of black-box and grey-box methods are analyzed from two perspectives. Although using optimization, cross-validation, and feature selection methods can improve ML techniques for different ML methods such as RT, ERT, and SVM, this improvement is limited for neural network-based modeling. Adding physical-based features can improve these prediction models for soot prediction significantly. The results are summarized in Table 4. For instance, in the RT method, using grey-box techniques improves R^2 by 5.4% while decreases RMSE and maximum absolute error by 42.0% for the the test data compared to black-box method. For ERT and SVM methods, using grey-box techniques improves R^2 by 2% and decreases RMSE about 7% and 17% for ERT and SVM respectively. The grey-box SVM model significantly reduced absolute error about 50% compare to the black-box SVM. Although the R^2 and RMSE error of grey-box and black-box neural network methods are similiar, a significant improvement in maximum prediction error is achieved by using grey-box model. Using grey-box techniques decreases maximum absolute error by 18%, 49%, 20%, and 40% for single and two layer ANN and BNN, respectively.

According to the maximum R^2 and minimum RMSE, the grey-box support vector machine has the most accurate prediction. Figs 7 (b) and (c) show experimental versus prediction for both black-box and grey-box methods. BNN methods (1-layer and 2-layers) and SVM method show a similar performance, but the BNN required more training time than SVM (Norouzi et al., 2020a), making SVM preferable. However, the R^2 and $RMSE$ in Table 4 indicate that the black-box BNN model attains almost the same accuracy as the grey-box model. Single-layer BNN in the black-box method model has the best accuracy among other ML methods (high R^2 and low $RMSE$ in test). The test R^2 for 1-HL BNN black-box model is identical to the relevant grey-box model. The grey-box 1-HL BNN has higher RMSE, but it has a lower maximum prediction error (0.6 mg/m^3 higher). Fig 7 (d) shows experimental versus prediction for 1-HL BNN with 31 neurons where maximum prediction error is 2.9 mg/m^3. Although the same grey-box approach method shows better accuracy, it requires significant effort to develop a physical-based model in GT-power. In general, using grey-box techniques provides a modest improvement of steady-state soot emission modeling accuracy which matches with literature (Mohammad et al., 2021; Rezaei et al., 2020). One main advantages of using grey-box modeling is the engine parameters can be varied due to the physical basis of the engine model for scenario testing. Whereas for black-box model, new experimental data is needed to do scenario testing. This is particularly relevant to transient emission testing, since measurements are difficult to obtain but could be modeled in the grey box model. The grey box modelling which uses GT Power model is suitable for real time control due to computational requirements. So, black box models are suggested for real time control. For other purposes like calibration, grey box modelling is a better option.

6. CONCLUSIONS

Grey-box, black-box and physical based emission modeling techniques were used to develop soot emission models for a diesel engine. Using only the physical model results in inaccurate soot emission prediction, which justify using data-driven methods. Experimental engine data was used for black-box model training and to parameterize a physical-based model which was used in a grey-box model. LASSO and K-fold algorithms were used for feature selection and hyperparameter tuning of black-box and grey-box methods to evaluate five different ML models; RT, ERT, SVM, SNN, and BNN. These methods were applied to the black-box and grey-box models. Grey-box SVM and 1 layer BNN black-box methods showed the best performance with test R^2 errors equal to 0.95. Although the test R^2 is the same for both methods, the maximum test error and RSME are lower for SVM grey-box method, so the grey-box SVM is determined to be the best model for soot emission prediction in this study. This conclusion is dependent on data-set collected. Adding the physical model is advantageous on modeling in reducing maximum test data variance. Grey-box transient emission modeling is planned for the future work.

REFERENCES

Aliramezani, M., Norouzi, A., and Koch, C.R. (2020a). Support vector machine for a diesel engine performance and NOx emission control-oriented model. *IFAC-PapersOnLine*, 53(2), 13976–13981.

Aliramezani, M., Norouzi, A., and Koch, C. (2020b). A grey-box machine learning based model of an electrochemical gas sensor. *Sensors and Actuators B: Chemical*, 321, 128414.

Amani, E., Akbari, M., and Shahpouri, S. (2018). Multi-objective cfd optimizations of water spray injection in gas-turbine combustors. *Fuel*, 227, 267–278.

Bidarvatan, M., Thakkar, V., Shahbakhti, M., Bahri, B., and Aziz, A.A. (2014). Grey-box modeling of HCCI engines. *Applied Thermal Engineering*, 70(1), 397–409.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. springer.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). *Classification and regression trees*. CRC press.

EuroVI (2016). commission regulation (EU) 2016/646 of 20 april 2016 amending regulation (EC) NO692/2008 as regards emissions from light passenger and commercial vehicles (Euro 6). In *Euro 6 regulation*.

Foresee, F.D. and Hagan, M.T. (1997). Gauss-newton approximation to bayesian learning. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 3, 1930–1935. IEEE.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gordon, D., Wouters, C., Wick, M., Xia, F., Lehrheuer, B., Andert, J., Koch, C.R., and Pischinger, S. (2020). Development and experimental validation of a real-time capable field programmable gate array–based gas exchange model for negative valve overlap. *International Journal of Engine Research*, 21(3), 421–436.

Hassoun, M.H. et al. (1995). *Fundamentals of artificial neural networks*. MIT press.

Hiroyasu, H., Kadota, T., and Arai, M. (1983). Development and use of a spray combustion modeling to predict diesel engine efficiency and pollutant emissions: Part 1 combustion modeling. *Bulletin of JSME*, 26(214), 569–575.

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

Khurana, S., Saxena, S., Jain, S., and Dixit, A. (2021). Predictive modeling of engine emissions using machine learning: A review. *Materials Today: Proceedings*, 38, 280–284.

Mohammad, A., Rezaei, R., Hayduk, C., Delebinski, T.O., Shahpouri, S., and Shahbakhti, M. (2021). Hybrid physical and machine learning-oriented modeling approach to predict emissions in a diesel compression ignition engine. In *2021 SAE World Congress, 2021-01-0496*. SAE International.

Norouzi, A., Aliramezani, M., and Koch, C.R. (2020a). A correlation based model order reduction approach for a diesel engine NOx and bmep dynamic model using machine learning. *International Journal of Engine Research*, First Published July 10, 2020.

Norouzi, A., Ebrahimi, K., and Koch, C.R. (2019). Integral discrete-time sliding mode control of homogeneous charge compression ignition (hcci) engine load and combustion timing. *IFAC-PapersOnLine*, 52(5), 153–158.

Norouzi, A., Gordon, D., Aliramezani, M., and Koch, C.R. (2020b). Machine Learning-based Diesel Engine-Out NOx Reduction Using a plug-in PD-type Iterative Learning Control. In *2020 IEEE Conference on Control Technology and Applications (CCTA)*, 450–455. IEEE.

Omidvarborna, H., Kumar, A., and Kim, D.S. (2015). Recent studies on soot modeling for diesel combustion. *Renewable and Sustainable Energy Reviews*, 48, 635–647.

Rezaei, R., Hayduk, C., Alkan, E., Kemski, T., Delebinski, T., and Bertram, C. (2020). Hybrid phenomenological and mathematical-based modeling approach for diesel emission prediction. In *WCX SAE World Congress Experience, 2020-01-0660*. SAE International.

Rodriguez, J.D., Perez, A., and Lozano, J.A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3), 569–575.

Shahpouri, S. and Houshfar, E. (2019). Nitrogen oxides reduction and performance enhancement of combustor with direct water injection and humidification of inlet air. *Clean Technologies and Environmental Policy*, 21(3), 667–683.

Snoek, J., Larochelle, H., and Adams, R.P. (2012). Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*.

Zhou, H., Soh, Y.C., and Wu, X. (2015). Integrated analysis of CFD data with k-means clustering algorithm and extreme learning machine for localized hvac control. *Applied Thermal Engineering*, 76, 98–104.