2024-01-4306

# Ensemble machine learning techniques for particulate emissions estimation from a highly boosted GDI engine fuelled by different gasoline blends

**Marta Stangierska[1*], Abdullah U. Bajwa[1*], Andrew G.J. Lewis[2], Sam Akehurst[3], James W.G. Turner[4] and Felix C.P. Leach[1]**

1 Department of Engineering Science, University of Oxford, Oxford, UK
2 IAAPS Ltd, Emersons Green, UK
3 University of Bath, Bath, UK
4 King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
*These authors contributed equally to this work



## Abstract

Light-duty vehicle emissions regulations worldwide impose stringent limits on particulate matter (PM) emissions, necessitating accurate modelling and prediction of particulate emissions across a range of sizes (as low as 10 nm). It has been shown that the decision tree-based ensemble machine learning technique known as Random Forest can accurately predict particle size, concentration, and accumulation mode geometric standard deviation (GSD) for particulate emission diameters as low as 23 nm from a highly boosted gasoline direct injection (GDI) engine operating on a single fuel, while also offering insights into the underlying factors of emissions production because of the interpretable nature of decision trees.

This work builds on the prior Random Forest research as its basis and further investigates the relative performance of five decision tree-based machine learning techniques in predicting these particulate emission parameters and extends the work to 10 nm particles. In addition to Random Forest, the selected techniques consist of four gradient boosting models: GBM, XGBoost, LightGBM, and CatBoost. Moreover, the influences of fuel chemistry are assessed by using data from 13 gasoline fuel blends, including blends with ethanol and methanol – common bio- and e-fuels. The results show that the CatBoost model achieves the highest prediction accuracy ($R^2$ between 0.77 and 0.932), even when the feature set is reduced to improve computational efficiency. Random Forest and LightGBM are also shown to be suitable for PM emissions estimation. Permutation feature importance was used to highlight the dependence of PM emissions on both fuel and engine operating parameters – offering new insights into the effect of fuel properties on particulate emissions and their formation in highly boosted engines.

## Introduction

Gasoline Direct Injection (GDI) engine sales in the European Union (EU) exceeded those of diesel engines for the first time in 2017 [1]. Compared to Port Fuel Injection (PFI) engines, GDI engines can offer fuel economy improvements because of precise fuel metering control, superior volumetric efficiency and improved knocking resistance. The knocking resistance is enabled by in-cylinder charge cooling, which has the additional benefit of permitting higher compression ratios [1]. However, direct fuel injection into the combustion chamber reduces the time available for fuel to vaporise and mix with air and also increases the chances for wall wetting; all of which can lead to increased particulate formation [2].

The two primary sources of Particulate Matter (PM) emissions from GDI engines are the fuel (primarily) and lubricant (secondarliy). Partial fuel-rich zones, in particular, and liquid droplet combustion lead to higher levels of particulate matter emissions. In addition, fuel spray impingement on the piston and combustion chamber surfaces (so-called wall-wetting) and any liquid remaining around the injector itself (injector dribble) [3]. These liquid films can lead to incomplete combustion and the development of a diffusion flame, thereby contributing to the formation of PM [4].

Downsized GDI engines are becoming more prevalent because they offer fuel consumption improvements. They operate at higher specific loads and their design typically includes external boosting using either a turbocharger or a supercharger [5]. Market examples of downsized engines include the 21.7 bar BMEP (VW TSI 1.4) and the 22.1 bar BMEP (EB16.4, fitted to a Bugatti Veyron) [6]. Boosted GDI engines have been shown to produce smaller size distributions of particulate matter emissions compared to unboosted engines and diesel engines [1] and such smaller particles are known to be more harmful to human health [7].

PM emissions from vehicles were first regulated in 1992. Both the emission limits and testing methods have been revised multiple times since then [1], following advances in measurement technology, accuracy, and increased scientific understanding of the connection between PM emissions and pulmonary diseases. The current EU regulations (Euro 6) for light-duty vehicles, regulate emissions by mass to 2 mg/km and Particle Number (PN) emissions to $6{\times}10^{11}$ #/km. The regulations specify the size of the particles to be counted at PN23, i.e. 50% count efficiency ($D_{50}$) at 23 nm and a 90% count efficiency ($D_{90}$) at 41 nm [1]. The recently approved Euro 7 regulations that will enter into force from 2026 reduce the size of particles counted to PN10 (i.e. $D_{50}$ of 10 nm and $D_{90}$ of 15 nm). This is because smaller particles are known to be more harmful to human health [7]. Manufacturers are expected to meet the proposed emissions limit by using gasoline particulate filters (GPF) [8].

Particulate emission prediction models will play an important role in the design and control of low particulate emission GDI engines. Both physical and empirical models have been considered [9]. Physical models include CFD-appropriate models both for soot and in-cylinder effects [10, 11]. Empirical models are developed using experimental (engine operation and design, and fuel) data combined with tunable factors. Such models are limited in their predictive strength (i.e. they perform poorly at conditions for which they were not tuned) but are computationally inexpensive, which makes them suitable for engine control. Fuel effects on PM emissions have also been modelled using indices including the Particulate Matter Index (PMI), Particle Number Index (PNI), and Moriya Index, amongst others. These have been described in some detail in a recent review paper [12].

A promising alternative to physical, computationally expensive PM models are data-driven empirical Machine Learning (ML) models. ML models use statistical learning tools to predict outcomes using input features and offer insights on how these inputs affect the target variable [13]. Ensemble algorithms, a subset of ML techniques, combine multiple trained models, often yielding better predictive performance compared to that of a single model.

ML approaches to modelling performance of and emissions from internal combustion engines are challenging due to the complex phenomena involved, such as turbulent air and flow mixing inside the combustion chamber, and in-cylinder temperature and pressure gradients [14]. Artificial neural networks (ANNs) have been frequently used for emissions modelling [15]. Back-propagated ANNs were used by Yuanwang *et al.* [16] as early as 2002 to analyse the effect of cetane number on exhaust emissions from a diesel engine. Good agreement between emissions measurements and predictions have been reported in the literature [17, 18, 19]. Studies have also utilized approaches based on physics-guided artificial neural networks (PGNNs), where ANNs are bounded (or informed) by physical characteristics of the system [20, 21].

ANNs, despite being widely used for estimating engine emissions, have the disadvantage of being opaque in their architecture, which limits the insights that can be drawn about the underlying emissions formation phenomenology. Other ML approaches, like bagging and boosting, exist that offer better interpretability. Bagging ensemble algorithms, such as Random Forest and Extra Trees, train models independently of each other and typically use bootstrapping, which involves resampling the data with replacement [22]. In contrast, boosting algorithms are based on iterative learning, where the weight of the wrongly predicted instances is increased in each iteration. Sahin [23] compared ANN, Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost) methods to predict engine performance and emissions when fuelled with diesel/biodiesel/iso-pentanol blends, concluding that XGBoost exhibited the best performance in predicting emissions parameters. Delgado *et al.* [24] tested 179 classifiers from 17 model families, including neural networks, decision trees, and support vector machines on University of California, Irvine datasets for ML research and concluded that Random Forest (RF) gave the best performance for classification tasks, with boosting models also among the best. Gonzalez *et al.* [22] analysed the performance of 13 bagging and boosting algorithms in 76 different classification tasks on Knowledge Extraction based on Evolutionary Learning (KEEL) software datasets of medium and large size and concluded that while RF gives good performance, it can be outperformed by several boosting algorithms for large datasets.

Papaioannou *et al.* [25] used a Random Forest model to predict particulate emissions from a highly boosted GDI engine, showing good agreement ($R^2_{\text{avg}}$ = 0.94) with experimental data. This was the first time that a multi-output, data-driven model demonstrated accurate results in predicting PM emissions from GDI engines when using a relatively simple and computationally inexpensive method like RF. They also demonstrated the use of the permutation feature importance method to reduce model inputs (from 82 to 9 inputs) and draw insights into parameters that drive PM emissions formation. Some of the data used in the current work comes from the same experimental dataset.

The Random Forest model methodology is schematically depicted in Figure 1. During the training phase, $N$ random samples are selected using bootstrap aggregation with replacement. Next, a regressor tree is fitted to each sample, with the mean squared error used as the splitting criterion for the leaves. Each tree casts a unit vote on their output prediction [26]. RF's benefits over other ensemble methods include robustness to outliers and noise, error convergence, and computational lightness.

The general structure of a gradient boosting model is shown in Figure 2. The model starts with a constant initial guess for the output of all N

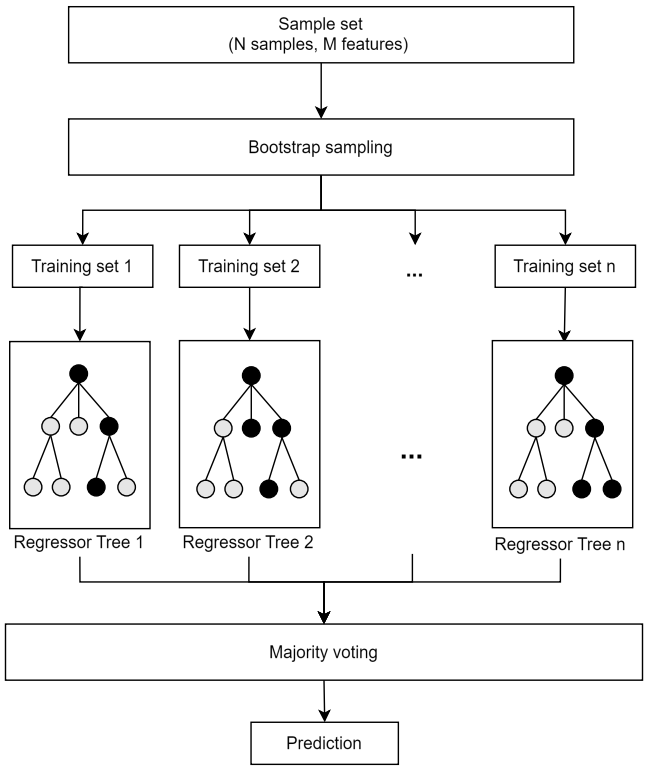


Figure 1: Random forest flow chart for a regression problem. Random Forest is an example of a bagging algorithm, where the final prediction is made by majority voting.

training samples. In each iteration $n$, a new weak learner $F_n(\mathbf{x})$ is added to the ensemble. The weak learner is typically a decision tree. To compute $F_{n+1}(\mathbf{x})$, an estimator $h_m(x)$ is added to the previous weak learner, so that $F_{n+1}(\mathbf{x}_i) = F_n(\mathbf{x}_i) + h_m(\mathbf{x}_i) = y_i$, where $y_i$ is the output value. Hence, the residual $r_i$ is given by $y_i - F_n$. The weight $\alpha_n$ of the weak learner is determined by minimizing the loss function on the training dataset $X$. To ensure the algorithm converges, regularization techniques such as shrinkage are typically applied to the gradient descent step [27]. Gradient boosting models outperform RF in handling sparse and imbalanced datasets, as well as in predicting outliers [22].

Several gradient boosting-based algorithms are available, including XGBoost, GBM, LGBM, and CatBoost. The advantages and disadvantages of each model are shown in Table 1.

eXtreme Gradient Boosting (XGB or XGBoost) uses decision trees as regressors [27]. XGBoost's unique features include automatic ways to remove missing values and approximate split finding. The latter is introduced to increase training speed and reduce overfitting, making XGBoost especially efficient for large datasets [22]. Categorical Boosting (CatBoost) works with oblivious trees, where the splitting criterion is the same for the entire tree level [22]. CatBoost also implements gradients to prevent the prediction shift, defined as a significant difference between the distributions of the training and test samples [27]. CatBoost offers a configurable boosting scheme to prevent overfitting. However, precomputations of the gradients make CatBoost computationally heavy. LightGBM's distinct features include (1) a subsampling technique called Gradient-based One-Side Sampling (GOSS), and (2) Exclusive Feature Bundling (EFB), which bundles sparse features into a single feature. GOSS works by creating multiple training sets for building base trees, composed of a top fraction of instances with higher uncertainty and a random sample fraction of instances with lower uncertainty. The weight of the lower gradient instances is subsequently adjusted to keep the sample distribution the same.

| Model | Advantages | Disadvantages |
|---|---|---|
| RF | Robust, computationally light, errors converge. | Weak for sparse and imbalanced datasets, no categorical variable support. |
| GBM | Good for outliers, adaptable to various problems. | Computationally intensive, prone to overfitting. |
| XGB | Handles missing values, efficient, sparsity-aware, supports regularization techniques. | Complex tuning, resource-intensive. |
| LGBM | Speed-optimized, supports categorical features. | Sensitive to overfitting, not suited for small datasets. |
| CB | Prevents prediction shift, handles categorical features well. | Computationally heavy, not adjustable to custom loss function |

Table 1: Comparison of selected tree-based models. The 5 ML models are: Random Forest (RF), Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (LGBM), eXtreme Gradient Boosting (XGB), and Categorical Boosting (CatBoost).

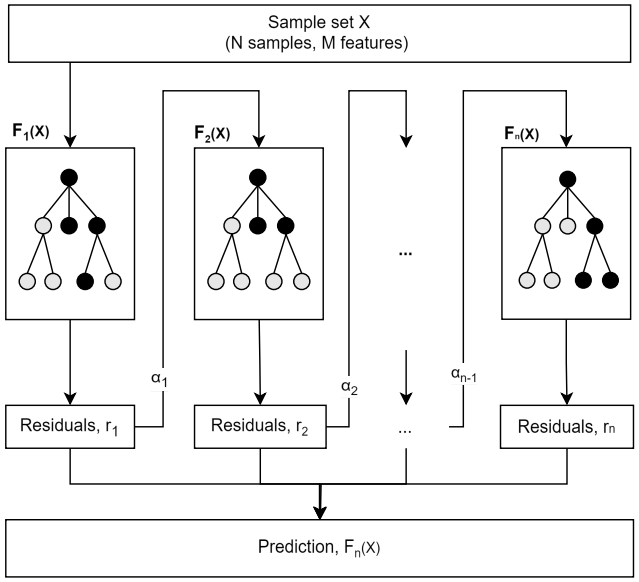


Figure 2: Gradient Boosting flow chart for a regression problem. $F_n(\mathbf{x})$ is a weak learner of nth iteration. Here, $\alpha_n$ is the weight of the $n_{th}$ weak learner.

A number of techniques can be used both to optimise the application of ML but also to gain additional knowledge as a result of the application of ML. Feature engineering facilitates the integration of domain knowledge with ML techniques by selecting and transforming input variables prior to training. A reduction in input dimensionality enhances the interpretability of the model. The permutation feature importance technique enables the identification of relevant input variables by estimating the distribution of measured importance for each variable [13]. Breaking the link with the target variable reveals the significance of the feature in the model, indicated by the resulting drop in the model's score [25]. Further improvements in the ML performance can be achieved by tuning the model's hyperparameters. Hyperparameters are variables whose values control the learning process and determine the values of model parameters that a learning algorithm learns. They cannot be estimated from the data but are adjusted to suit a specific modeling problem.

This work builds upon the RF modeling work done by Papaioannou *et al.* [25] and investigates the relative performance of five bagging and boosting algorithms in predicting GDI PM emissions. The effects of fuel composition and engine operating parameters are also explored, as well as the ability of such machine learning models to estimate emissions when particles down to 23 nm or 10 nm diameter are considered. A permutation feature importance exercise is carried out to highlight the most important parameters for predicting the particulate matter emissions.

## Methodology

Five machine learning models, Random Forest (RF), Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (LGBM), eXtreme Gradient Boosting (XGB), and Categorical Boosting (CatBoost), were trained and validated on a widely-used, previously published experimental dataset of particulate matter emissions from a highly boosted GDI engine [6, 28, 29]. This section will detail the model configuration and briefly overview the experimental dataset.

### *Model*

The form of the model is $\mathbf{y} = f(\mathbf{X})$, where $\mathbf{y}$ is the vector of output features and $\mathbf{X}$ is a vector of input features of length N, $(x_1, x_2, \ldots, x_N)$. 35 input features were used for model development. These include the following 17 engine operating parameters: Air/Fuel Ratio (AFR), angle of max. cylinder pressure, max. cylinder pressure, engine power, engine torque, exhaust CO, exhaust $CO_2$, exhaust $NO_x$, exhaust THC, fuel pressure, injection angle, inlet manifold pressure, intake manifold temperature, intake manifold oxygen, equivalence ratio, exhaust opacity, throttle position; and the following 18 fuel parameters: RON, IBP (IP 123), 90% rec (IP 123), FBP (IP 123), RVP (IP 394/ASTM 5191), and composition (carbon (C), hydrogen (H) and oxygen (O), paraffins, isoparaffins, olefins (including dienes), dienes, naphthenes, aromatics, oxygenates, unknowns, methanol, and ethanol concentration).

Model evaluation was conducted in three parts as summarised in Figure 3. *Experiment 1* considered only a baseline fuel and focused on the prediction of PN23 (the current Euro 6 standard) for different engine operating parameters. *Experiment 2* added the influence of fuel composition (whilst still including the effect of engine operating parameters) by predicting PN23 emissions for 13 different fuels. *Experiment 3* extended the first two experiments by adding prediction of PN10 (the future Euro 7 standard), across all fuels and engine operating points. For all experiments, PN concentration, GSD, and CMD are predicted - the GSD and CMD being the standard deviation and mean diameter (particle size) of a log-normal fit to the particle size distribution [2].

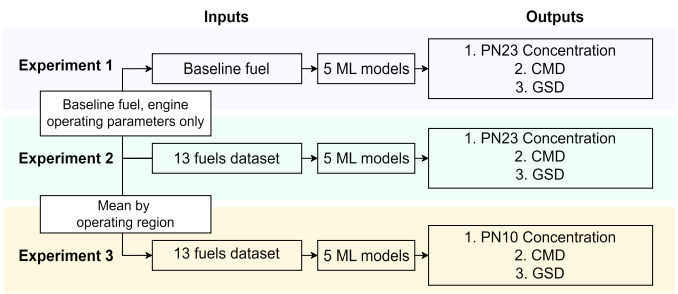


Figure 3: Model inputs and outputs for the 3 experiments described in this work. The 5 ML models used are: Random Forest (RF), Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (LGBM), eXtreme Gradient Boosting (XGB), and Categorical Boosting (CatBoost).

Model development was performed on an *Amazon SageMaker Notebook* (*ml.t3.medium* EC2 instance). RF and GBM were implemented using the Python *scikit-learn* package (version 1.2.0), employing the *multioutput.MultiOutputRegressor* method. Dedicated libraries were used for XGBoost (version 1.7.4), LightGBM (version 3.3.5), and CatBoost (version 1.1.1).

Hyperparameter search was conducted using Python's *GridSearchCV* function, available in the *scikit-optimize* library. In Experiment 3, input features were aggregated per operating region, necessitating a separate hyperparameter search. The chosen hyperparameters are listed in Table 2, and the entire hyperparameter space swept is provided in the appendix.

Table 2: Optimal hyperparameter values for the 5 ML models investigated in this work. The hyperparameters for Experiments 1&2 and Experiment 3 are computed separately.

| Model | Hyperparameter | Exp. 1&2 | Exp. 3 |
|---|---|---|---|
| RF | n_estimators | 34 | 25 |
| | max_depth | 19 | 25 |
| | min_samples_split | 2 | 5 |
| | min_samples_leaf | 4 | 4 |
| | max_features | 1 | 1 |
| GBM | learning_rate | 0.2 | 0.1 |
| | max_depth | 10 | 10 |
| | min_samples_split | 2 | 2 |
| | max_features | sqrt | 0.25 |
| | subsample | 1 | 0.5 |
| XGB | learning_rate | 0.025 | 0.025 |
| | gamma | 0 | 0 |
| | max_depth | 2 | 2 |
| | colsample_bylevel | 0.25 | 0.25 |
| | subsample | 0.5 | 0.5 |
| LGBM | learning_rate | 0.1 | 0.1 |
| | num_leaves | 1024 | 15 |
| | top_rate | 0.4 | 0.7 |
| | other_rate | 0.05 | 0.05 |
| CB | learning_rate | 0.3 | 0.05 |
| | max_depth | 9 | 9 |
| | leaf_estimation_iterations | 1 | 1 |
| | l2_leaf_reg | 3 | 3 |

## *Model Evaluation Metrics*

To assess model accuracy, the coefficient of determination ($R^2$) and mean squared error (MSE) were used. Algorithm efficiency was compared using mean score and fit times, recorded over 50 runs using *GridSearchCV.best_estimator* attributes. Mean score time denotes the average prediction time, while the mean fit time represents average time taken to fit the model to a set of parameters. A 75% to 25% train-test split was selected. The input and target parameters were standardised prior to training using Equation 1 [25], where $x$ is the non-standardised parameter, $\bar{x}$ is the mean of the values of $x$, and $\sigma_x$ is their standard deviation. MSE results are presented on standardised parameters.

$$x_{std} = \frac{(x - \bar{x})}{\sigma_x} \tag{1}$$

All models are trained and evaluated using the same set of 10 random seeds to perform the test-train splits of the database for each experiment.

## *Permutation Feature Importance*

A permutation feature importance study was conducted on a set of uncorrelated features by shuffling each feature randomly several times and recording the drop in model score. This was done to identify parameters that the model considers important for the prediction of PN emissions. Information about the relative importance of different engine/fuel parameters on PN emissions can help develop appropriate control strategies.

Decision-tree-based models' predictions are not significantly affected by correlated input features. Bagging ensembles, such as Random Forests, reduce variance by averaging the predictions of many trees trained on different subsets of the data, hence their effectiveness is not significantly affected by the correlated features. In boosting algorithms, feature selection methods like Gradient-based One-Side Sampling (GOSS) in LightGBM [30], regularization methods in XGBoost and GBM [31, 32], and ordered boosting in CatBoost [33] help mitigate the impact of correlated variables on model performance. However, during the permutation feature importance process, it is necessary to remove highly correlated parameters to avoid misleading importance scores. When features are highly correlated, permuting one feature might not significantly affect the model's performance if another correlated feature can take over its predictive power. To address this, a Pearson correlation test was conducted to identify clusters of highly correlated parameters ($R > |0.8|$). From each identified group, a single representative parameter was selected for model development. Figures 4 and 5 display the correlation heatmaps for the remaining (weakly correlated) engine operating parameters and fuel parameters, respectively. Full heatmaps for all (correlated and uncorrelated) parameters are provided in the Appendix.

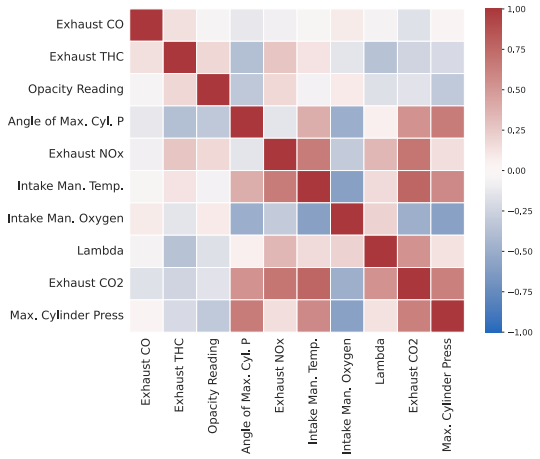


Figure 4: Pearson correlation heatmap for engine operating parameters used in permutation feature importance analysis. Correlated feature clusters were identified, with one representative element retained from each group.

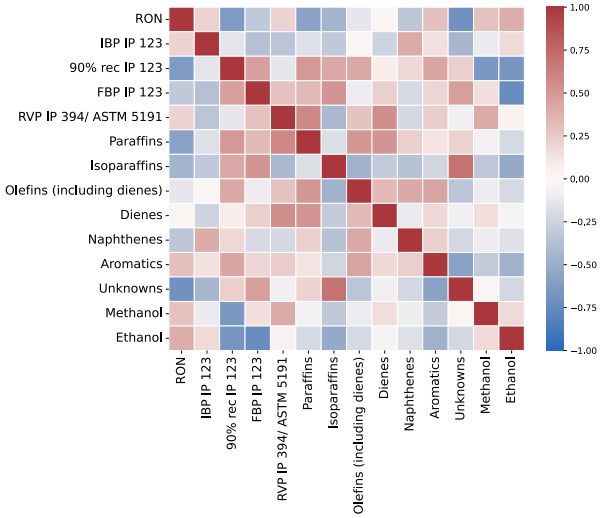


Figure 5: Pearson correlation heatmap for the fuel parameters used in permutation feature importance analysis.

To implement the permutation importance technique, the model is first trained, and then a reference score is computed. Then, each feature is randomly shuffled, and a new score is computed. This

process is iterated several times to ensure a robust result. The importance of each feature is calculated as follows [25]:

$$i_x = s_{ref} - \frac{1}{K}\sum_{k=1}^{K} s_{kj} \tag{2}$$

Here, $s_{ref}$ represents the reference score for the model, $K$ is the number of shuffles, and $s_{kj}$ denotes the new score for the $k^{th}$ shuffle of the feature $j$. The feature is deemed significant if its importance is greater than or equal to 0.05 [34].

The feature effect of $F_x$ is calculated using Equation 3 [35].

$$\begin{aligned}\text{Feature effect of } F_x &= \text{useful information for prediction that } F_x \text{ possesses} \\ &+ \text{interaction power of } F_x \text{ with other features}\end{aligned} \tag{3}$$

## *Engine and Emissions Measurement*

The experimental dataset used in this work contained Particle Number (PN) emissions measured from a 2.0-L 4 cylinder GDI prototype engine (UB100 Ultraboost) using a Cambustion DMS500 analyser, which provided particle concentration, mode diameter (CMD), and GSD. Detailed engine specifications can be found in Turner *et al.* [36]. The engine experiments were performed using a motoring dynamometer with external air, fuel, coolant, and oil conditioning equipment. No after treatment equipment was installed. The dataset and its collection have been published in previous works [6, 28, 29]. The measurement of PN emissions using the DMS500 and subsequent calculation of PN10 and PN23 are detailed in two works by Leach *et al.* [6, 37]. For convenience, the histograms of concentration and size (CMD) values for each tested fuel are shown in Figure 6 and Figure 7.

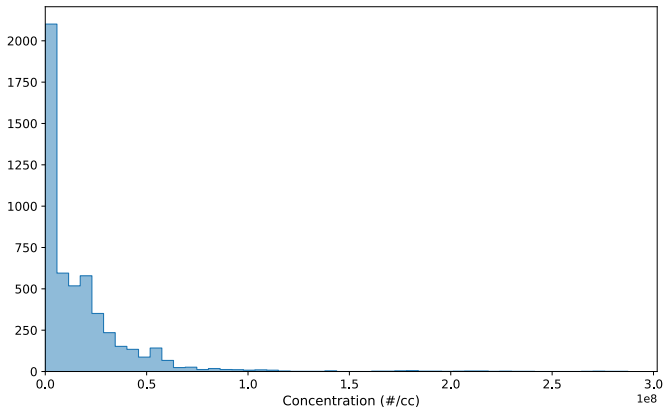


Figure 6: Histogram of PM concentration values for the dataset used in this work.

Important engine specifications are summarised in Table 3.

Table 3: Test engine specifications.

| Engine | UB100 |
|---|---|
| Type | In-line 4 cylinder |
| Displacement | 1991 $cm^3$ |
| Bore x Stroke | 83 mm × 92 mm |
| Valves per cylinder | 2 intake, 2 exhaust |
| Compression ratio | 9:1 |
| Maximum fuel pressure | 200 bar |
| Aspiration | External boost |

## *Experimental Test Conditions*

The first phase of testing was focused on evaluating the effect of engine load and spark/fuel injection timing on particulate emissions.

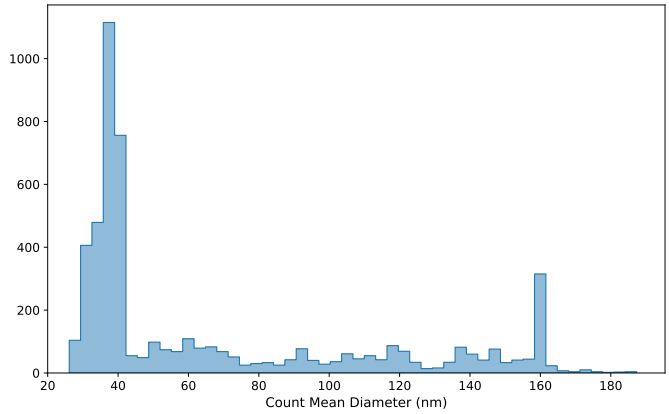


Figure 7: Histogram of count mean diameter (CMD) values for the dataset used in this work.

The engine was operated at a constant speed of 2000 rpm while the load was varied between minimum idle to maximum load (approximately 500 Nm) across nine increments. In the second phase, the fuels were tested across four operating regions - supercharged, naturally-aspirated, supercharged-turbocharged transition, and turbocharged. In addition, in the second phase, the impact of EGR, inlet air temperature, fuel injection timing, exhaust back pressure, and excess air-to-fuel ratio (lambda) were tested. Detailed descriptions of the test conditions were previously given in [6]. Experiment 1 considers all of these data points on the base fuel, Experiments 2 and 3 include all of these points across the fuel matrix.

## *Fuels*

A summary of the properties of the 13 fuels used in experiments 2 and 3 is given in Table 4. Even though these fuels are used herein to assess whether the inclusion of fuel composition indexes can improve ML models' PN emission estimates, they were not selected with PN emissions in mind. They represent a large spread of the fuels likely to be used in highly boosted engines.

Table 4: Test fuel specifications.

| Fuel | RON (-) | MON (-) | DVPE (kPa) | FBP (°C) |
|---|---|---|---|---|
| A | 103.3 | 95 | 26.1 | 177 |
| B | 101.4 | 88.8 | 68 | 176 |
| C | 92.8 | 90.7 | 30.5 | 193 |
| D | 88.6 | 87.3 | 32.9 | 190 |
| E | 95.1 | 82.2 | 28.7 | 138 |
| F | 104.2 | 92.6 | 23.3 | 139 |
| G | 111.6 | 101.2 | 57.4 | 192 |
| H | 95.1 | 85 | 53.1 | 189 |
| I | 98.7 | 86.5 | 97.4 | 173 |
| Base | 97 | 85.3 | 75 | 188 |
| E20 | 99.6 | 85.7 | 57.8 | 184 |
| E85 | 107.4 | 89.5 | 44.4 | 78 |
| GEM | 106 | 88.1 | 84.4 | 74 |

The base fuel (referred to as baseline here) used in these experiments is a gasoline compliant with the EN228 standard. It contains up to 5% by volume ethanol (E5) [6]. Fuel H (again, E5) is also EN228-compliant and contains the minimum research octane number (RON) required by the standard. Fuel I has a higher octane number with a higher Dry Vapour Pressure Equivalent (DVPE) and is a typical winter-grade premium fuel containing no ethanol. Together, Base, H and I are categorised as market-representative fuels. Fuels A through D were blended to demonstrate the effect of RON and motor octane number (MON) on particulate emissions with their RON and MON being broadly independent. Fuels E and F were blended to test the

impact of Laminar Flame Speed on combustion parameters and Fuel G has an artificially boosted RON. Together, Fuels A through G are categorised as research fuels. E20, E85, and GEM are fuels containing higher levels of oxygenates. E20 and E85 containing 20% and 85% by volume ethanol respectively, and GEM is iso-stoichiometric with E85, containing 23% ethanol and 42% methanol. These oxygenate fuels were tested to explore the impact of higher levels of renewable components. Further details of these fuels and their composition can be found in [28, 29].

## Results and Discussion

Results are presented for the validation dataset only. Training dataset results are not included for brevity.

### *Effect of Engine Operating Parameters*

In Experiment 1, the five ML models were trained and tested using the baseline fuel data across the whole range of engine operating parameters. To assess their performance, the results are validated against the reference RF model configured as in Papaioannou *et. al* [25].

Table 5 presents the average score and fit times for all models. The most computationally intensive aspect of each algorithm's execution was the fit time. Notably, Random Forest and Light Gradient Boosting Machine demonstrated the quickest fit times. In contrast, CatBoost required significantly more time, ranging from 200 to 400 times longer due to its ordered boosting scheme. This observation is consistent with findings in Bentejac *et al.* [27]. All algorithms exhibited similar score times, except for CatBoost with default parameters, which performed 10 times faster.

Both Random Forest and CatBoost, in their default configurations, yielded competitive results without requiring computationally expensive hyperparameter tuning. CatBoost, in its default setup, surpassed other algorithms on average and exhibited a performance over eight times faster than its tuned version. Conversely, Random Forest performed slower in its default configuration.

Table 5: Fit and score times for the baseline fuel database.

| Model | Score time (s) | Fit time (s) |
|---|---|---|
| RF | 0.01 | 0.46 |
| GBM | 0.01 | 2.08 |
| XGB | 0.01 | 2.21 |
| LGBM | 0.01 | 0.56 |
| CB | 0.01 | 439.29 |
| RF [a] | 0.02 | 3.10 |
| RF [b] | 0.02 | 3.10 |
| CB [b] | 0.002 | 50.6 |

[a] Hyperparameters recommended in [25]
[b] Default multi-output regressor version

The average performance of the algorithms in predicting each of the particulates' characteristics across 10 runs is depicted in Figure 8 ($R^2$) and Figure 9 (MSE), with both figures showing consistent performance trends. Note that $R^2$ is a measure of accuracy, whereas MSE is a measure of error. Generally, all algorithms demonstrated good $R^2_{avg}$ performance, with scores ranging between 0.75 and 0.90. However, they exhibited significantly degraded performance in predicting the GSD - the worst performing algorithm for this measure was XGB ($R^2_{GSD}$ = 0.44, $MSE_{GSD}$ = 0.56) and the best performing algorithm was CatBoost ($R^2_{GSD}$ = 0.78, $MSE_{GSD}$ = 0.22). Overall, CatBoost exhibiting the best performance, could be attributed to it being specifically designed to handle, among other things, correlated inputs.

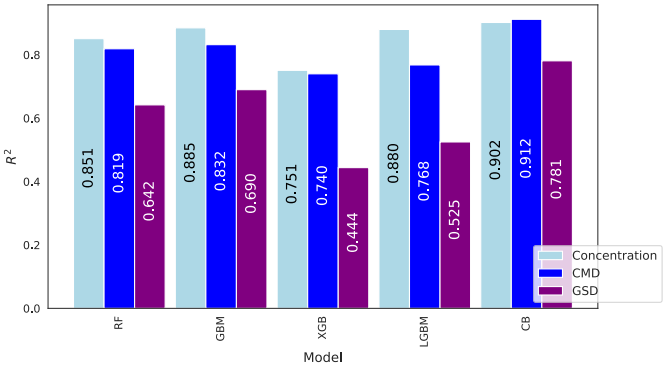


Figure 8: The $R^2$ performance of the multi-output regression problem by the five chosen algorithms on the baseline fuel dataset. The algorithms used were Random Forest (RF), Gradient Boosting Machine (GBM), eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and CatBoost (CB). The model outputs are PN23 Concentration, Count Mean Diameter (CMD), and Geometric Standard Deviation (GSD).

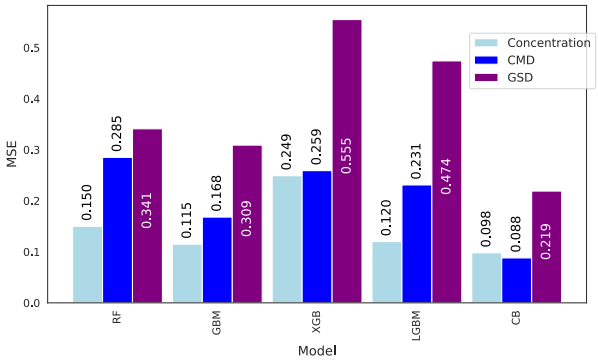


Figure 9: The MSE performance of the multi-output regression problem by the five chosen algorithms on the baseline fuel dataset.

### Permutation feature importance results for baseline fuel

Figure 10 shows the relative importance of the 10 features considered in the models. Each column represents a different ML algorithm and each row represents a parameter. The colours indicate the permutation feature importance score of these parameters within the algorithm. Note that the presence of a feature in this figure implies either a positive or a negative correlation. There is clear consistency between the five different ML models used. Opacity reading, max. cylinder pressure, angle of max. cylinder pressure, intake manifold oxygen, lambda and exhaust CO, $NO_x$, $CO_2$, and THC are all identified as significant input features across all of the top-performing models. It can also be seen that the XGB model relies on the least parameters to predict PN and attributes comparable importance to each of selected features.

It is worth considering the physical underpinning of the importance of these parameters. The opacity reading is a coarse, smoke-based, exhaust emissions measurement and therefore it ought to correlate closely with PN emissions formation. CO and THC exhaust emissions indicate incomplete combustion, which is linked to increased PM emissions [25]. Max. cylinder pressure, angle of max. cylinder pressure and $CO_2$ emissions correlate with engine load, which is known to strongly affect PM formation [4, 25]. Intake manifold oxygen levels and intake manifold temperature are indicative of Exhaust Gas Recirculation (EGR). EGR reduces in-cylinder temperature, impacting PN in two competing ways: (1) by decreasing pyrolysis, thus limiting smaller particulate formation, and (2) by reducing soot oxidation rates, promoting the formation of smaller

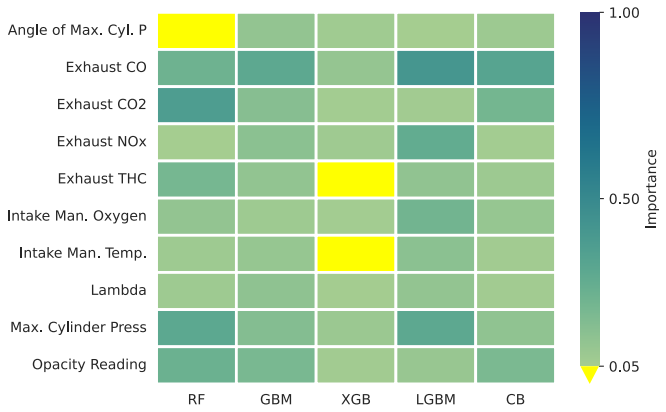


Figure 10: Permutation feature importance results for the baseline dataset. Features with importance greater than or equal to 0.05 are considered significant.

particles [38]. Exhaust lambda correlates closely with PM emissions—PM emissions are higher for rich mixtures compared to stoichiometric ones due to reduced soot oxidation [9].

## Reduced models

It is desirable to limit the number of input parameters needed by a model to reduce experimental effort required to collect the data. The permutation feature importance results are used to guide model reduction by removing parameters that do not have a significant influence on the model results. The resulting reduced models use between 47% and 59% of the original feature space, as has been summarised in Table 6 for the various algorithms.

Table 6: Permutation feature importance results for the baseline dataset for full and reduced models.

| Model | No. of features | | Conc. | CMD | GSD |
|---|---|---|---|---|---|
| | Orig. | Red. | $\Delta$ $R^2$ (Orig. vs Red.) (%) | | |
| RF | 17 | 9 | 2.8 | 16.0 | 29.0 |
| GBM | 17 | 10 | -1.9 | 13.3 | 15.5 |
| XGB | 17 | 8 | -7.7 | 8.1 | -3.6 |
| LGBM | 17 | 10 | 1.4 | 18.9 | -57.1 |
| CB | 17 | 10 | -4.4 | 4.8 | 7.8 |

Figures 11 and 12 respectively illustrate the $R^2$ and MSE results for the reduced models and Table 6 compares the reduced models' performance to that of full models. The reduced models predict CMD with higher accuracy than the original models, likely due to the removal of correlated features in reduced models. For XGBoost and LGBM, the test $R^2_{avg}$ score for the reduced models decreases by 1 and 13 %, respectively. The other algorithms show improvements of 2% to 16%. For XGBoost, reducing features of 52% (9 features) results in a 1.1% decrease in prediction accuracy. Furthermore, for the two algorithms that experience a drop in the $R^2$ score, the decrease is primarily attributed to lower accuracy in predicting the GSD, which holds substantially lower importance for characterizing emissions. XGBoost and LightGBM demonstrated the lowest $R^2_{avg}$ among the models. Overall, the models' performance is affected in only relatively minor way despite substantial reductions in the number of parameters. The best performance is exhibited by CB ($R^2_{avg}$ = 0.89, $MSE_{avg}$ = 0.12), followed by RF ($R^2_{avg}$ = 0.88, $MSE_{avg}$ = 0.17) and GBM ($R^2_{avg}$ = 0.87, $MSE_{avg}$ = 0.13).

## *Predicting PN emissions from different fuels*

Experiment 2 was performed on a dataset containing information on engine operation and fuel characteristics for the full dataset (see Figure 3) of 13 fuels. For reference, the best physical model

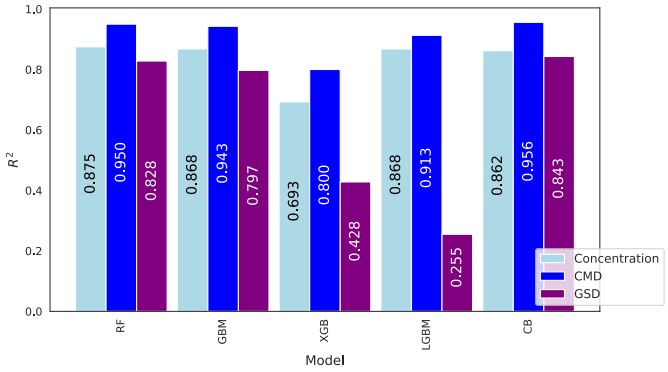


Figure 11: The $R^2$ performance of the reduced models on the baseline fuel dataset.

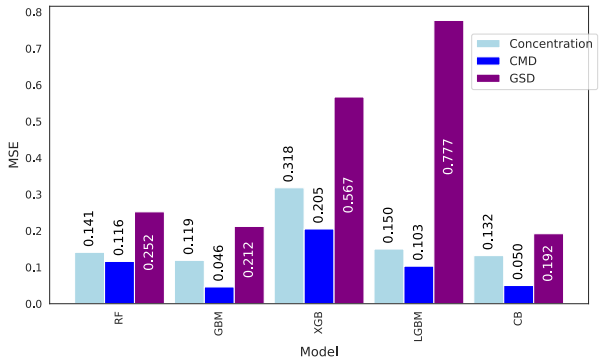


Figure 12: The MSE performance of the reduced models on the baseline fuel dataset.

performance (various PM indices) achieved $R^2$ values of 0.28 to 0.82 for predicting PN concentration across all fuels [28]. The $R^2$ and MSE of the ML algorithms are shown in Figures 13 and 14, respectively. Here, as this is the first time that PN emissions from a GDI engine have been predicted using ML models (to the authors' knowledge), there is no benchmark model to compare to. A substantial improvement in performance for all models, compared to both the baseline dataset and previous index-based approaches is evident - $R^2$ values above 0.9 are seen for both PN23 concentration and CMD for almost all models. This good performance can be attributed to both the inclusion of additional features (the addition of fuel parameters has increased the number of features to 35 from 17) and the increased volume of data in this database (adding the different fuels has roughly quadrupled the size of the dataset).

The algorithms predict CMD with the highest accuracy ($R^2_{avg}$ = 0.93, $MSE_{avg}$ = 0.07) and GSD with the lowest accuracy ($R^2_{avg}$ = 0.78, $MSE_{avg}$ = 0.23). Similarly to Experiment 1, the highest average performance is achieved by the CB algorithm ($R^2_{avg}$ = 0.93, $MSE^2_{avg}$ = 0.07), closely followed by RF and GBM algorithms (both achieve $R^2_{avg}$ = 0.92 and $MSE^2_{avg}$ = 0.08)

The target vs. predicted variables for the model performances for concentration, CMD, and GSD, trained and tested on the same sample dataset, are shown in Figures 15, 16, and 17, respectively. CB, RF, and GB perform the best among the algorithms and for this particular seed; RF performs better than CB.

## Permutation feature importance results for the full 13-fuel dataset

The results of the permutation feature importance study for Experiment 2 are presented in Figure 18, which displays the features

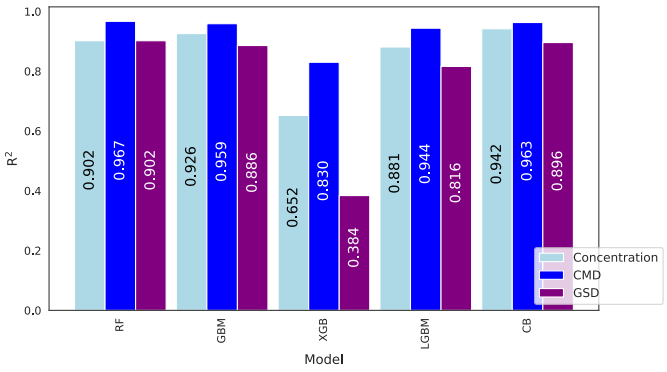


Figure 13: Multi-output regression $R^2$ scores for predicting PN emissions from 13 different fuels.

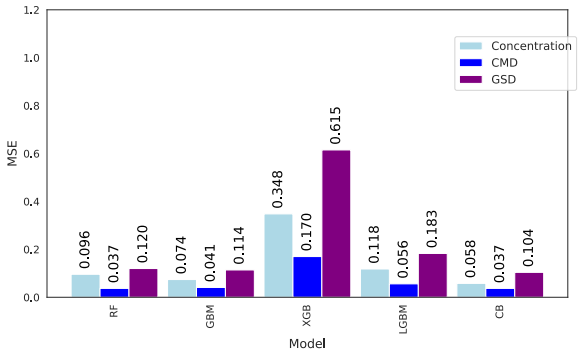


Figure 14: Multi-output regression MSE results for predicting PN emissions from 13 different fuels.

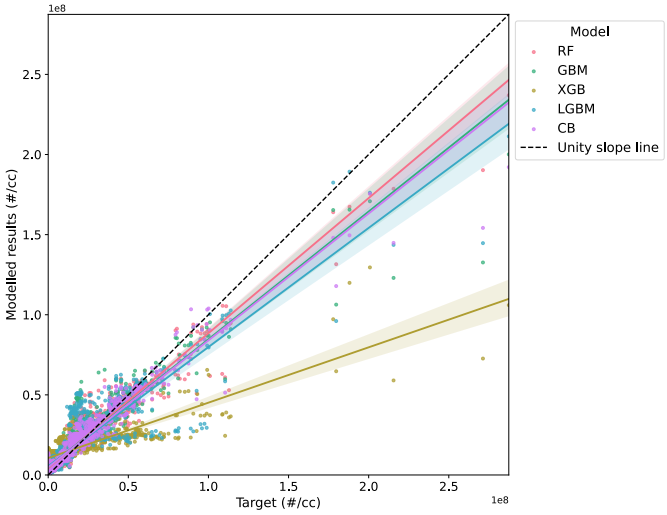


Figure 15: Concentration prediction vs. target for the 13-fuel dataset of the tested models.

which were found significant in at least one model. Apart from XGBoost (which had only three significant parameters), the algorithms identified a combination of fuel parameters and engine operating parameters as significant features. The significant features include specific engine operating parameters such as intake manifold oxygen, exhaust $CO_2$, max. cylinder pressure all of which align with previous findings for baseline fuel. Additionally, angle of max. cylinder pressure, $NO_x$, THC, intake manifold temperature, lambda, and opacity readings were all identified as significant by at least one algorithm. These parameters are known to be related to PM emissions

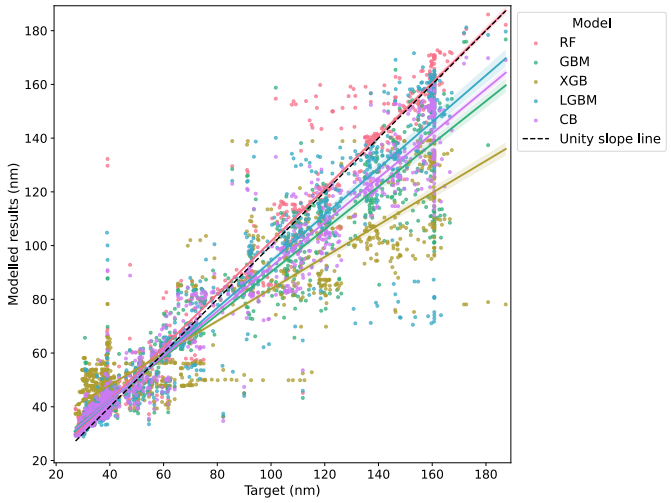


Figure 16: CMD prediction vs. target for the 13-fuel dataset of the tested models.

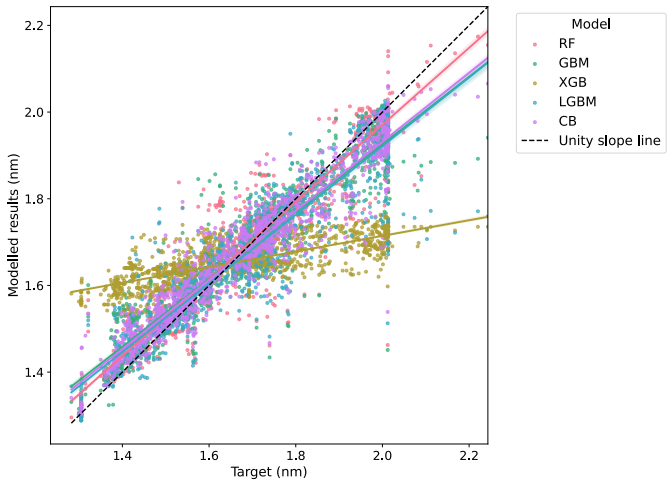


Figure 17: GSD prediction vs. target for the 13-fuel dataset of the tested models.

as discussed in the previous section [9, 25].

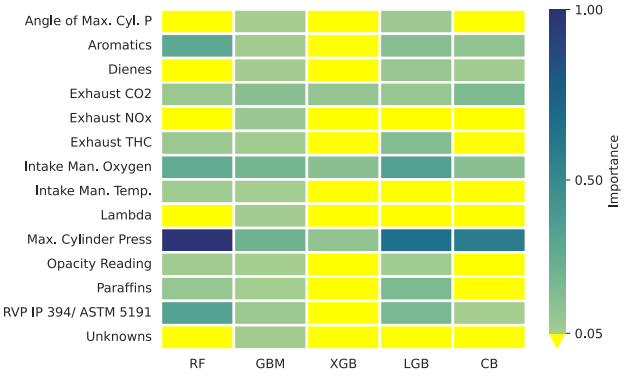


Figure 18: Permutation feature importance results for all fuels. Only features considered statistically significant in at least one model are included.

Among the fuel parameters, all four best-performing models identify RVP (Reid Vapor Pressure) and aromatics as significant factors. Vapour Pressure (whether RVP or DVPE), a measure of the volatility

of gasoline, is a key component of a number of the indices that have been used to predict PM emissions in the past and is known to have a substantial influence on in-cylinder fuel evaporation and hence PM emissions formation [12]. Aromatics with high boiling points and high double bond equivalent (DBE) values tend to produce higher PN emissions [39]. Other parameters highlighted by at least one model include fuel components such as paraffins, dienes, and 'unknowns' content. Paraffins have been reported to contribute to a sooting tendency in a fuel blend hence correlating to PM formation [40]. Conjugated dienes (a less reactive form of dienes that also have high DBEs) are produced during the petroleum refining process and are known to contribute to the instability of petroleum products [29].

## Reduced models

The performance of the reduced models is summarised in Table 7. Again, some substantial reductions in the number of features required to run the models are seen, with the reduced models only utilizing between 9% and 40% of the original feature space. It is observed that the performance of GBM, CatBoost and XGBoost decrease, with XGBoost showing a significant drop for each target variable. Conversely, the performance of RF and LGBM improves by a small amount. Overall, the models' performance is affected in only relatively minor way despite substantial reductions in the number of parameters.

Table 7: Permutation feature importance results for the 13-fuels dataset for full and reduced models.

| Model | No. of features | | Conc. | CMD | GSD |
|---|---|---|---|---|---|
| | Orig. | Red. | $\Delta$ $R^2$ (Orig. vs Red.) (%) | | |
| RF | 35 | 9 | 3.5 | -0.7 | -0.2 |
| GBM | 35 | 14 | -1.3 | -1.3 | -1.9 |
| XGB | 35 | 3 | -29.3 | -10.1 | -43.2 |
| LGBM | 35 | 10 | 2.4 | 1.5 | 6.4 |
| CB | 35 | 6 | -3.1 | -1.0 | -9.4 |

The results of the Reduced Models are shown in Figure 19 ($R^2$) and Figure 20 (MSE). XGBoost performed the worst among the algorithms; therefore, features selected by this algorithm will not be further analysed. The best performance is exhibited by RF ($R^2_{avg}$ = 0.93, $MSE_{avg}$ = 0.07), GB and LightGBM (both performing at $R^2_{avg}$ = 0.91, $MSE_{avg}$ = 0.09). The performance of RF and LightGBM improved in comparison to the full models, while requiring 9, 10 and 14 features respectively. Out of these features, more than a third are fuel characteristic parameters known beforehand. The performance of CatBoost ($R^2_{avg}$ = 0.89, $MSE_{avg}$ = 0.11) degraded in comparison to the full model; however, it remains competitive with other models.

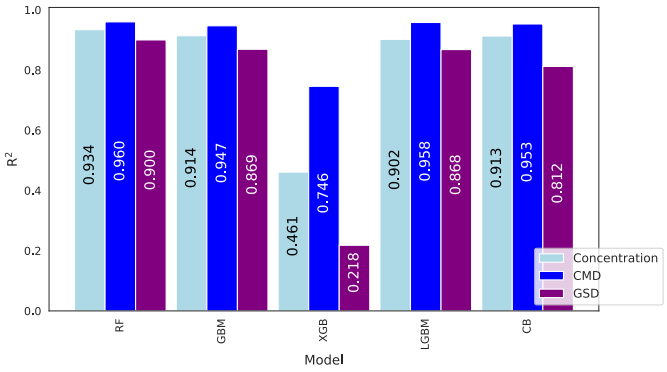


Figure 19: Multi-output score results for predicting PN emissions from 13 different fuels with reduced models.

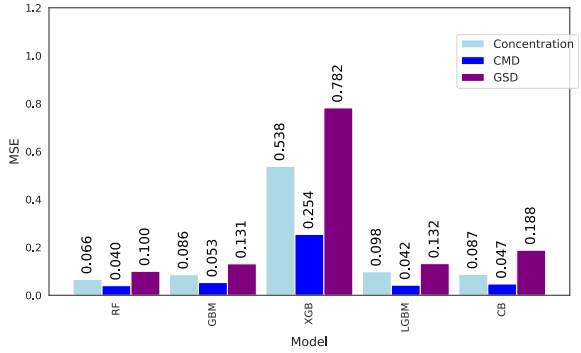


Figure 20: Multi-output MSE results for predicting PN emissions from 13 different fuels with reduced models.

## *PN10 results*

The results of Experiment 3, predicting the PN10 (PN down to 10 nm diameter) for all fuels, are depicted in Figures 21 ($R^2$) and 22 (MSE). In contrast to Figures 13 and 14, which show the equivalent plot for the PN23 across all fuels, overall the $R^2$ values are lower, perhaps indicating a higher variability in the measurement of PN10 compared to PN23. CatBoost ($R^2_{avg}$ = 0.84, $MSE_{avg}$ = 0.16), Random Forest ($R^2_{avg}$ = 0.83, $MSE_{avg}$ = 0.17) and GBM ($R^2_{avg}$ = 0.81, $MSE_{avg}$ = 0.19) emerge as the top-performing algorithms. Across all five models, GSD remains the variable with the lowest prediction score, as observed previously.

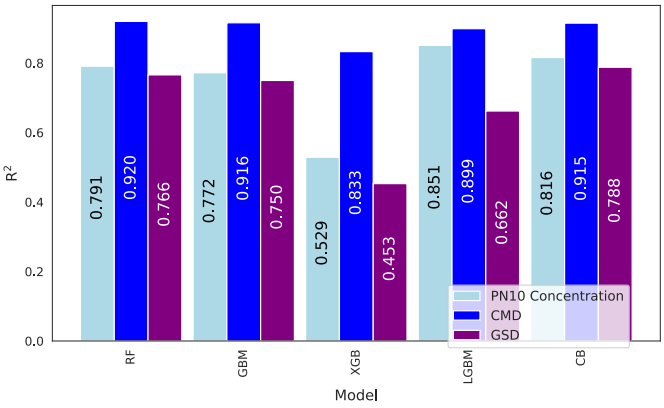


Figure 21: Multi-output regression $R^2$ results for PN10 data.

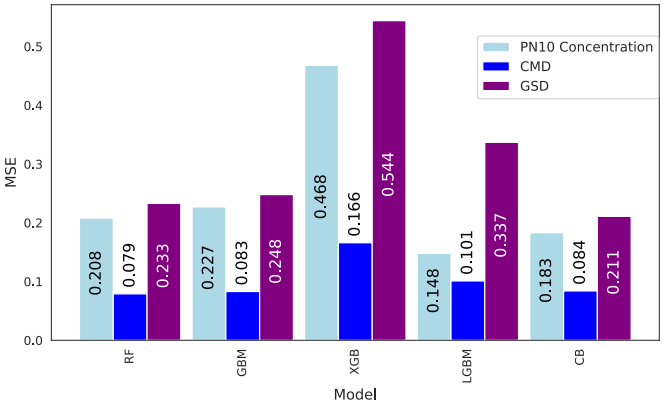


Figure 22: Multi-output regression MSE results for PN10 data.

The target vs. predicted variable for the model performances for

PN10 concentration, CMD and GSD, trained and tested on the same sample dataset, are shown in Figures 23, 24 and 25 respectively.

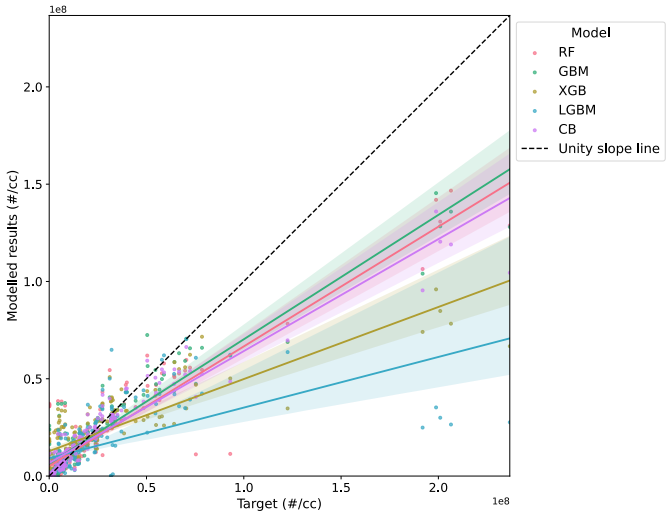


Figure 23: Concentration predictions vs. targets for the PN10 dataset of the tested models.

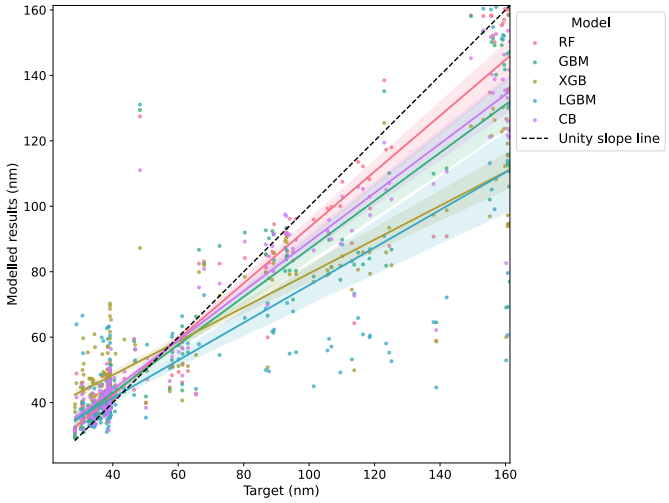


Figure 24: CMD predictions vs. targets for the PN10 dataset of the tested models.

## Permutation Feature Importance for PN10 dataset

The statistically significant features identified by each model are depicted in Figure 26. The features identified align with those discussed in Experiments 1 and 2. Additionally, the olefins (including dienes) parameter, napthenes and RON emerged as significant. Olefins content is known to impact PM emissions [41, 42]. Olefins facilitate larger soot particle formation at medium and high speeds. Reduction in olefin content decreases particulate mass but not particulate number, indicating that it favours the smallest particles, as shown in Wei *et al.* [42]. This may explain why reduced CatBoost, the sole model to identify olefin content as a significant feature, achieved a higher $R^2$ in this Experiment, which is looking at PN10. Naphthenes are known to promote PN formation, although their influence is lower than that of aromatics [3]. Higher Research Octane Numbers are often achieved with higher levels of aromatic components (which have a very high RON) that promote PM emissions formation as discussed elsewhere in this paper. [43].

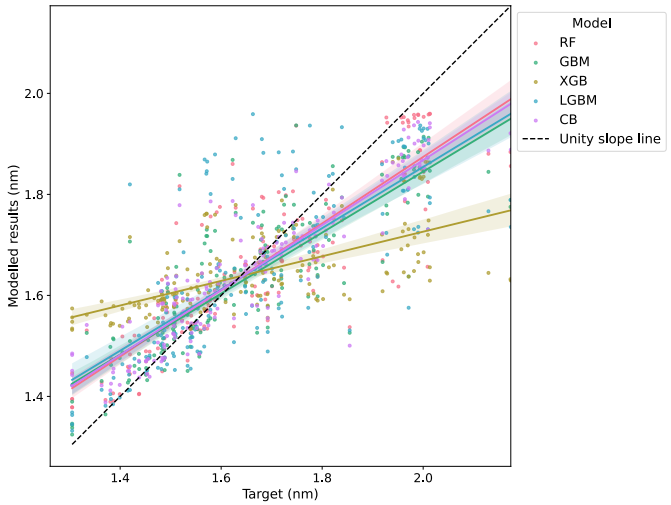


Figure 25: GSD predictions vs. targets for the PN10 dataset of the tested models.

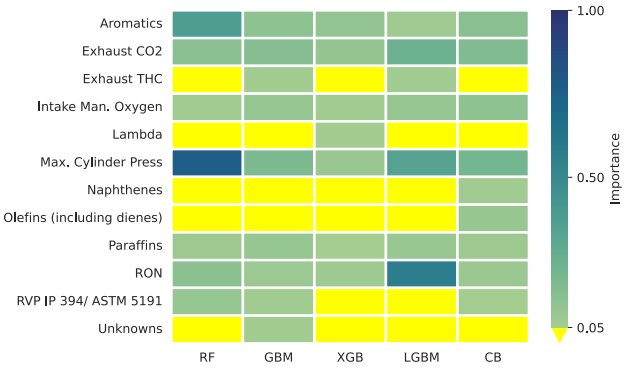


Figure 26: Permutation feature importance results for sub-10 nm dataset.

## Reduced Model

Table 8 shows a summary of the results for the reduced dataset. Overall, a decrease in scores is observed, with slight improvements of up to 1.3% seen in CatBoost and GBM. Of note is that GBM, LightGBM and RF require fewer features for the PN10 dataset, compared to the PN23 dataset (Table 7) and the feature space is between 20% and 26% of its original size. The reduction in scores between original and reduced models is most significant for GSD, as observed previously, but the differences are lower than before, perhaps due to accuracy being generally lower.

Table 8: Permutation feature importance results for the sub-10 nm dataset for full and reduced models.

| Model | No. of features | | PN10 Conc. | CMD | GSD |
|---|---|---|---|---|---|
| | Orig. | Red. | $\Delta$ $R^2$ (Orig. vs Red.) (%) | | |
| RF | 35 | 7 | -1.6 | -0.4 | -7.0 |
| GBM | 35 | 9 | 5.7 | 0.3 | -2.1 |
| XGB | 35 | 7 | -1.5 | -6.6 | -15.7 |
| LGBM | 35 | 7 | 1.1 | -3.8 | -12.5 |
| CB | 35 | 9 | 0.9 | -0.3 | -0.5 |

The results of the Reduced Models are shown in Figure 27 ($R^2$) and Figure 28 (MSE). The best performing algorithm is CatBoost ($R^2_{avg}$ = 0.84, $MSE_{avg}$ = 0.16), as before, followed by GBM ($R^2_{avg}$ = 0.82,

$MSE_{avg}$ = 0.18) and RF ($R^2_{avg}$ = 0.80, $MSE_{avg}$ = 0.20). Note, however, that CatBoost is significantly more computationally expensive than other algorithms. These findings suggests that, if compute power is particularly expensive, exploring algorithms with good performance other than CatBoost, such as RF, GBM, LightGBM and Random Forest, could be beneficial for optimal speed and accuracy.

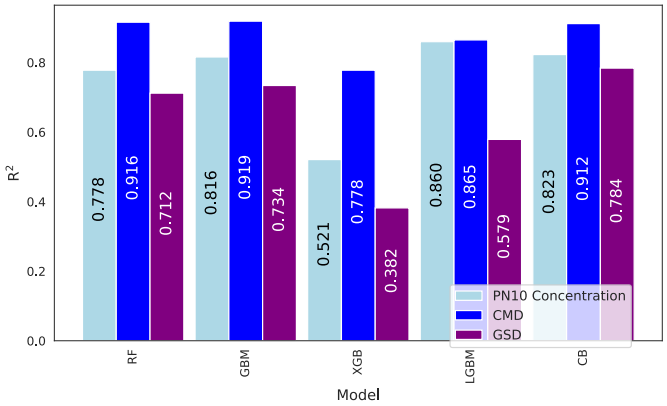


Figure 27: $R^2$ results for PN10 data for reduced models.

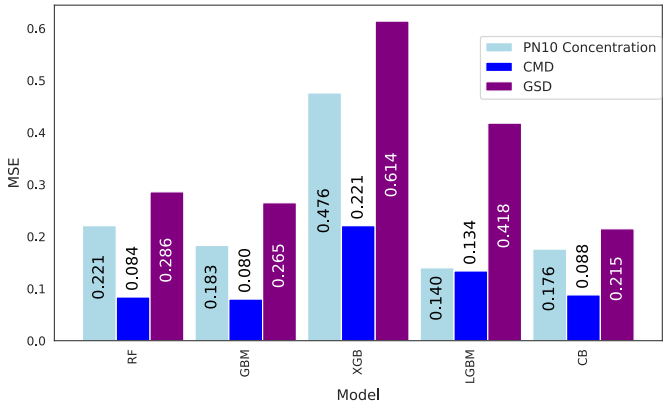


Figure 28: MSE results for PN10 data for reduced models.

## Summary / Conclusions

Five machine learning algorithms were tested for a multi-output regression problem of predicting size, concentrations, and geometric standard deviation (GSD) of particles emitted from a downsized 4-cylinder GDI engine operating at up to 35 bar BMEP. The algorithms tested included Random Forest (a bagging method), and boosting methods: LightGBM, XGBoost, GBM, and CatBoost. 13 different fuels were examined, comprising market-representative fuels, research fuels, and oxygenate fuels.

Ordered CatBoost demonstrated the highest accuracy for the two PN23 datasets among the algorithms tested, and LightGBM demonstrated the highest accuracy for the PN10 dataset. To reduce dimensionality, feature engineering and a Pearson correlation test were implemented, reducing input features from 35 to 9, encompassing both fuel and engine operating parameters.

Permutation feature importance tests were used to determine significant engine operation parameters and fuel characteristics, further reducing dimensionality. For the 13-fuel dataset, the best-performing models were CatBoost and Random Forest, yielding $R^2_{avg}$ values of 0.934 and 0.924, respectively. Reduced Random Forest performed at an $R^2_{avg}$ value of 0.931 using only 6 out of 35 features. Fuel parameters identified as significant and correlated with emissions by the best-performing models included Reid Vapor Pressure and aromatics content, which are both parameters that have fed into previously developed PM indices.

The ML algorithms in this work have demonstrated, for the first time, better prediction of the effect of fuel composition on PN emissions from this engine and dataset compared to three PM indices, which have previously been evaluated on this dataset. This demonstrates the utility of such ML models and their potential for superior performance compared to existing approaches.

A notable aspect of this work is the testing on both 10 nm and 23 nm diameter databases, which is useful due to the high engine-out PN of GDI engines and the more stringent Euro 7 regulation to be introduced from 2026 for light vehicles which will include 10 nm particles for the first time. Top-performing models demonstrated the capability to predict particulate emissions, with both full and reduced CatBoost models achieving $R^2$ values of 0.84 and MSE of 0.16. Olefins were also identified as contributors to small particulate formation, identified as significant by CatBoost.

By complementing the ML algorithms with a permutation importance method, accurate predictions of particulate emissions characteristics from GDI engines and a better understanding of their dependence on fuel and engine operating parameters can be gained.

To extend this research, a physics-guided neural network model (PGNN) could be developed to incorporate principles of physics in emissions modelling. Other models, such as GPFlow and Artificial Neural Networks (ANNs), could also be explored. In addition, larger, more diverse, datasets would aid generalisability and further work on PN10, perhaps using a dataset taken with a regulatory-grade instrument, would be of interest.

## References


1. B. Giechaskiel, A. Joshi, L. Ntziachristos, and P. Dilara, "European regulatory framework and particulate matter emissions of gasoline light-duty vehicles: A review," *Catalysts*, vol. 9, no. 7, p. 586, 2019.

2. M. Raza, L. Chen, F. Leach, and S. Ding, "A review of particulate number (pn) emissions from gasoline direct injection (gdi) engines and their control techniques," *Energies*, vol. 11, no. 6, 2018.

3. Leach, Felix, Knorsch, Tobias, Laidig, Christoph, and Wiese, Wolfram, "A review of the requirements for injection systems and the effects of fuel quality on particulate emissions from gdi engines," *SAE Technical Paper 2018-01-1710*, 2018.

4. P. Eastwood, *Particulate emissions from vehicles*. Chichester, England: John Wiley  Sons, 2008.

5. B. Lecointe and G. Monnier, "Downsizing a gasoline engine using turbocharging with direct injection," *SAE Technical Paper 2003-01-0542*, 2003.

6. F. Leach, R. Stone, D. Richardson, A. Lewis, S. Akehurst, J. Turner, S. Remmert, S. Campbell, and R. Cracknell, "Particulate emissions from a highly boosted gasoline direct injection engine," *International Journal of Engine Research*, vol. 19, 2017.

7. B. Alfoldy, B. Giechaskiel, W. Hofmann, and Y. Drossinos, "Size-distribution dependent lung deposition of diesel exhaust particles," *Journal of Aerosol Science*, vol. 40, no. 8, pp. 652–663, 2009.

8. Official Journal of the European Union, "Regulation (EU) 2024/1257 of the European Parliament and of the Council."

https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401257, 2024. [Accessed 01-06-2024].

9. F. Leach, R. Stone, D. Fennell, D. Hayden, D. Richardson, and N. Wicks, "Predicting the particulate matter emissions from spray-guided gasoline direct-injection spark ignition engines," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 231, no. 6, pp. 717–730, 2017.

10. Del Pecchia, Marco, Sparacino, Simone, Pessina, Valentina, Fontanesi, Stefano, Breda, Sebastiano, Irimescu, Adrian, and Di Iorio, Silvana, "Development of a sectional soot model based methodology for the prediction of soot engine-out emissions in gdi units," *SAE Technical Paper 2020-01-0239*, 2020.

11. A. Bajwa, G. Zou, F. Zhong, X. Fang, F. C. Leach, and M. H. Davy, "Development of a semi-empirical physical model for transient nox emissions prediction from a high-speed diesel engine," *International Journal of Engine Research*, pp. 1–14, 2024.

12. Leach, Felix, Chapman, Elana, Jetter, Jeff J., Rubino, Lauretta, Christensen, Earl D., St. John, Peter C., Fioroni, Gina M., and McCormick, Robert L., "A review and perspective on particulate matter indices linking fuel composition to particulate emissions from gasoline engines," *SAE International Journal of Fuels and Lubricants*, vol. 15, no. 1, pp. 3–28, 2021.

13. A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, pp. 1340–1347, 04 2010.

14. M. Aliramezani, C. R. Koch, and M. Shahbakhti, "Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: A review and future directions," *Progress in Energy and Combustion Science*, vol. 88, p. 100967, 2022.

15. K. Karunamurthy, A. A. Janvekar, P. L. Palaniappan, V. Adhitya, T. T. K. Lokeswar, and J. Harish, "Prediction of ic engine performance and emission parameters using machine learning: A review," *Journal of Thermal Analysis and Calorimetry*, vol. 148, no. 9, pp. "3155–3177", 2023.

16. D. Yuanwang, Z. Meilin, X. Dong, and C. Xiaobei, "An analysis for effect of cetane number on exhaust emissions from engine with the neural network," *Fuel*, vol. 81, no. 15, pp. 1963–1970, 2002.

17. H. Wang, C. Ji, C. Shi, Y. Ge, H. Meng, J. Yang, K. Chang, and S. Wang, "Comparison and evaluation of advanced machine learning methods for performance and emissions prediction of a gasoline wankel rotary engine," *Energy*, vol. 248, p. 123611, 2022.

18. S. Uslu and M. B. Celik, "Performance and exhaust emission prediction of a si engine fueled with i-amyl alcohol-gasoline blends: An ann coupled rsm based optimization," *Fuel*, vol. 265, p. 116922, 2020.

19. X. Fang, F. Zhong, N. Papaioannou, M. H. Davy, and F. C. Leach, "Artificial neural network (ann) assisted prediction of transient nox emissions from a high-speed direct injection (hsdi) diesel engine," *International Journal of Engine Research*, vol. 23, no. 7, pp. 1201–1212, 2022.

20. K. F. Lee, N. Eaves, S. Mosbach, D. Ooi, J. Lai, A. Bhave, A. Manz, J. N. Geiler, J. A. Noble, D. Duca, and C. Focsa, "Model guided application for investigating particle number (pn) emissions in gdi spark ignition engines," *SAE International Journal of Advances and Current Practices in Mobility*, vol. 1, no. 1, pp. 76–88, 2019.

21. W. Park, J. Lee, K. Min, J. Yu, S. Park, and S. Cho, "Prediction of real-time no based on the in-cylinder pressure in diesel engines," *Proceedings of the Combustion Institute*, vol. 34, no. 2, pp. 3075–3082, 2013.

22. S. Gonzalez, S. Garcia, J. D. Ser, L. Rokach, and F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Information Fusion*, vol. 64, pp. 205–237, 2020.

23. S. Şahin, "Comparison of machine learning algorithms for predicting diesel/biodiesel/iso-pentanol blend engine performance and emissions," *Heliyon*, vol. 9, no. 11, p. e21365, 2023.

24. M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, no. 90, pp. 3133–3181, 2014.

25. N. Papaioannou, X. Fang, F. Leach, A. Lewis, S. Akehurst, and J. Turner, "A random forest algorithmic approach to predicting particulate emissions from a highly boosted gdi engine," *SAE Technical Paper 2021-24-0076*, 2021.

26. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

27. C. Bentejac, A. Csorgo, and G. Martinez-Munoz, "A comparative analysis of gradient boosting algorithms," *The Artificial intelligence review*, vol. 54, no. 3, pp. 1937–1967, 2021.

28. F. Leach, R. Stone, D. Richardson, A. Lewis, S. Akehurst, J. Turner, V. Shankar, J. Chahal, R. Cracknell, and A. Aradi, "The effect of fuel composition on particulate emissions from a highly boosted gdi engine - an evaluation of three particulate indices," *Fuel*, vol. 252, pp. 598–611, 2019.

29. F. Leach, R. Stone, D. Richardson, J. Turner, A. Lewis, S. Akehurst, S. Remmert, S. Campbell, and R. Cracknell, "The effect of oxygenate fuels on pn emissions from a highly boosted gdi engine," *Fuel*, vol. 225, pp. 277–286, 08 2018.

30. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

31. J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, 11 2000.

32. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.

33. J. M. Chin, "Catboost: Unbiased boosting with categorical features."

34. A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, pp. 1340–1347, 04 2010.

35. S. Oh, "Predictive case-based feature importance and interaction," *Information Sciences*, vol. 593, pp. 155–176, 2022.

36. Lewis, A.G.J., Akehurst, S., Brace, C.J., Copeland, C., Bredda, S.W., Fernandes, J.X., Turner, J.W.G., Popplewell, A., Patel, R., Johnson, T.R., Darnton, N.J., Richardson, S., Martinez-Botas, R., Romagnoli, A., Tudor, R.J., Bithell, C.I., Jackson, R.,

Burluka, A.A., Remmert, S.M., and Cracknell, R.F., "Ultra boost for economy: Extending the limits of extreme engine downsizing," *SAE International Journal of Engines*, vol. 7, no. 1, pp. 387–417, 2014.

37. F. Leach, A. Lewis, S. Akehurst, J. Turner, and D. Richardson, "Sub-23 nm particulate emissions from a highly boosted gdi engine," *SAE Technical Paper 2019-24-0153*, 2019.

38. C. Hergueta, M. Bogarra, A. Tsolakis, K. Essa, and J. Herreros, "Butanol-gasoline blend and exhaust gas recirculation, impact on gdi engine emissions," *Fuel*, vol. 208, pp. 662–672, 2017.

39. K. Aikawa, T. Sakurai, and J. J. Jetter, "Development of a predictive model for gasoline vehicle particulate matter emissions," *SAE International Journal of Fuels and Lubricants*, vol. 3, no. 2, pp. 610–622, 2010.

40. M. Edney, J. Barker, J. Reid, D. Scurr, and C. Snape, "Recent advances in the analysis of gdi and diesel fuel injector deposits," *Fuel*, vol. 272, p. 117682, 07 2020.

41. Z. He, L. Zhang, L. Guibin, Y. Qian, and X. Lu, "Evaluating the effects of olefin components in gasoline on gdi engine combustion and emissions," *Fuel*, vol. 291, p. 120131, 2021.

42. J. Wei, Z. Yin, Y. Qian, C. Wang, and B. Chen, "Comparative effects of olefin content on the performance and emissions of a modern gdi engine," *Energy & Fuels*, vol. 33, no. 11, pp. 10499–10507, 2019.

43. Remmert, Sarah, Campbell, Steven, Cracknell, Roger, Schuetze, Andrea, Lewis, Andrew, Giles, Karl, Akehurst, Sam, Turner, James, Popplewell, Andrew, and Patel, Rishin, "Octane appetite: The relevance of a lower limit to the mon specification in a downsized, highly boosted disi engine," *SAE International Journal of Fuels and Lubricants*, vol. 7, no. 3, pp. 743–755, 2014.

## Contact Information

Felix Leach,
Department of Engineering Science,
University of Oxford, Oxford, OX1 3PJ, UK
felix.leach@eng.ox.ac.uk

## Acknowledgments

The authors acknowledge the Technology Strategy Board (now Innovate UK), the UK's innovation agency, for the partial funding of this work. Consortium members GE Precision Engineering, Jaguar Land Rover, Shell, Lotus Engineering, CD-Adapco, Imperial College London, University of Bath and the University of Leeds have all made various portions of this work possible. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

## Definitions, Acronyms, Abbreviations

| | |
|---|---|
| **AFR** | Air/Fuel Ratio |
| **ANN** | Artificial Neural Network |
| **BMEP** | Brake Mean Effective Pressure |
| **CatBoost** | Categorical Boosting |
| **CB** | Categorical Boosting |
| **CFD** | Computational Fluid Dynamics |
| **CMD** | Count Mean Diameter |
| **DBE** | Double Bond Equivalent |
| **DVPE** | Dry Vapour Pressure Equivalent |
| **EFB** | Exclusive Feature Bundling |
| **EGR** | Exhaust Gas Recirculation |
| **EU** | European Union |
| **FBP** | Final Boiling Point |
| **GBM** | Gradient Boosting Machine |
| **GDI** | Gasoline Direct Injection |
| **GOSS** | Gradient-based One-Side Sampling |
| **GPF** | Gasoline Particulate Filter |
| **GSD** | Geometric Standard Deviation |
| **IBP** | Initial Boiling Point |
| **KEEL** | Knowledge Extraction based on Evolutionary Learning |
| **LGBM** | Light Gradient Boosting Machine |
| **ML** | Machine Learning |
| **MON** | Motor Octane Number |
| **MSE** | Mean Squared Error |
| **PFI** | Port Fuel Injection |
| **PGNN** | Physics-Guided Neural Network |
| **PM** | Particulate Matter |
| **PMI** | Particulate Matter Index |
| **PN** | Particle Number |
| **PNI** | Particle Number Index |
| **RF** | Random Forest |
| **RVP** | Reid Vapor Pressure |
| **RON** | Research Octane Number |
| **SVM** | Support Vector Machine |
| **THC** | Total Hydrocarbon |
| **XGB** | eXtreme Gradient Boosting |

# APPENDIX

## *Hyperparameters*

Table 9: Grid search values for hyperparameters.

| Model | Hyperparameter | Grid search values |
|---|---|---|
| RF | n_estimators | 2, 12, 23, 34, 45, 56, 67, 78, 89, 100 |
| | max_depth | 1, 3, 6, 9, 11, 14, 17, 19, 22, 25 |
| | min_samples_split | 2, 10, 19 |
| | min_samples_leaf | 1, 4, 6, 8, 10 |
| | max_features | sqrt, 1 |
| GBM | learning_rate | 0.025, 0.05, 0.1, 0.2, 0.3 |
| | max_depth | 2, 3, 5, 7, 10, unlimited |
| | min_samples_split | 2, 5, 10, 20 |
| | max_features | log2, sqrt, 0.25, 1 |
| | subsample | 0.15, 0.5, 0.75, 1.0 |
| XGB | learning_rate | 0.025, 0.05, 0.1, 0.2, 0.3 |
| | gamma | 2, 3, 5, 7, 10, unlimited |
| | n_estimators | 2, 5, 10, 20 |
| | colsample_bylevel | log2, sqrt, 0.25, 1 |
| | subsample | 0.15, 0.5, 0.75, 1.0 |
| LGBM | learning_rate | 0.025, 0.05, 0.1, 0.2, 0.3 |
| | num_leaves | 3, 7, 15, 31, 127, 1024 |
| | top_rate | 0.2, 0.4, 0.6, 0.7 |
| | other_rate | 0.05, 0.1, 0.3 |
| CB | learning_rate | 0.025, 0.05, 0.1, 0.2, 0.3 |
| | max_depth | 3, 6, 9 |
| | leaf_estimation_iterations | 1, 10 |
| | l2_leaf_reg | 1, 3, 6, 9 |

***Input features list***

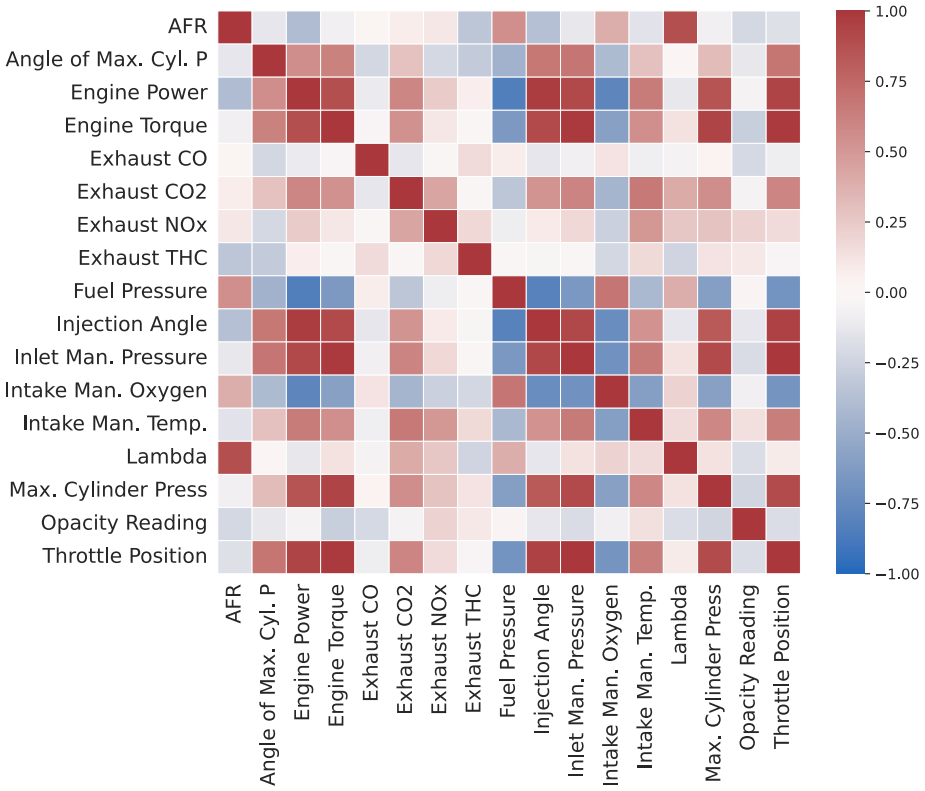


Figure 29: Full Pearson correlation heatmap for engine operating parameters.

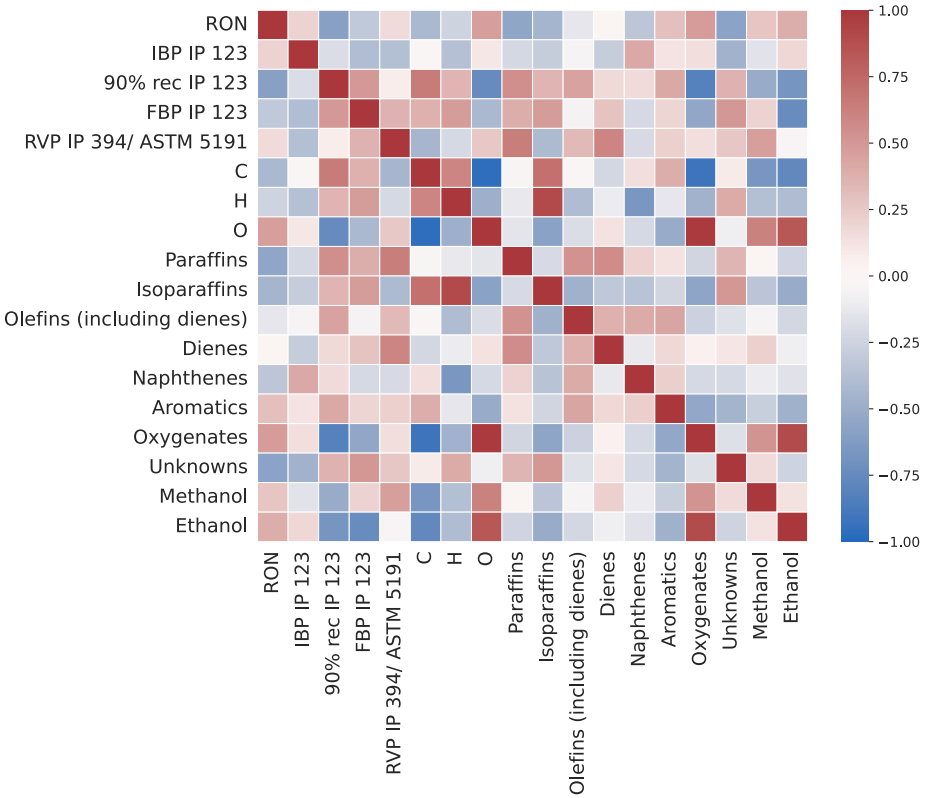


Figure 30: Full Pearson correlation heatmap for fuel parameters.