Energy and AI 16 (2024) 100360

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.sciencedirect.com/journal/energy-and-ai

# Utilizing Artificial intelligence to identify an Optimal Machine learning model for predicting fuel consumption in Diesel engines

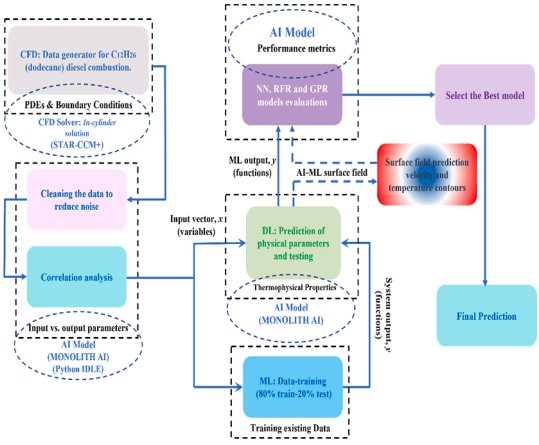
Amirali Shateri, Zhiyin Yang, Jianfei Xie *

*School of Engineering, University of Derby, DE22 3AW, UK*

## HIGHLIGHTS

- An optimal ML model was identified to predict fuel consumptions in diesel engines.
- Various parameters were determined to reflect fuel consumptions using sensitivity analysis.
- Different ML models were performed to consider the complexity of fuel consumption prediction.
- ML models demonstrated good prediction performance for both flow and heat transfer characteristics of fuel combustion.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
AI evaluation
Machine learning
Diesel engine
Fuel consumption
Decarbonization

## ABSTRACT

This paper describes the utilization of artificial intelligence (AI) techniques to identify an optimal machine learning (ML) model for predicting dodecane fuel consumption in diesel combustion. The study incorporates sensitivity analysis to assess the impact levels of various parameters on fuel consumption, thereby highlighting the most influential factors. In addition, this study addresses the impact of noise and implements data cleaning techniques to ensure the reliability of the obtained results. To validate the accuracy of the predictions, the study performs several metrics and validation process, including comparisons with computational fluid dynamics (CFD) results and experimental data. Comprehensive comparisons are made among neural networks (NN), random forest regression (RFR), and Gaussian process regression (GPR) models, taking into account the complexity associated with fuel consumption predictions. The findings demonstrate that the GPR model outperforms the others in terms of accuracy, as evidenced by metrics such as mean absolute error (MAE), mean squared error (MSE), Pearson coefficient (PC), and R-squared ($R^2$). The GPR model exhibits superior predictive ability, accurately detecting and predicting even individual data points that deviate from the overall trend. The significantly lower absolute error values also consistently indicate its higher accuracy compared with the NN and RFR models. Furthermore, the GPR model shows a remarkable speedup, approximately 1.7 times faster than

* Corresponding author
*E-mail address:* j.xie@derby.ac.uk (J. Xie).

https://doi.org/10.1016/j.egyai.2024.100360

Available online 18 March 2024

traditional CFD solvers, and physically captures the momentum and thermal characteristics in a surface field prediction. Finally, the target optimization is assessed using the Euclidean distance as a fitness function, ensuring the reliability of predicted data.

## 1. Introduction

Diesel combustion is a critical process in internal combustion engines, and optimizing its performance is of great importance to achieve high fuel efficiency, less emissions and improved overall thermal performance. Traditional approaches to fuel consumption predictions mainly rely on semi-empirical formulas or physical models. However, these methods often lack accuracy and are limited in their ability to capture complex relations between engine parameters and fuel consumption [1–7]. To address these limitations, ML models have emerged as a promising alternative for fuel consumption predictions. ML models can learn complex patterns and relationships from data, enabling accurate predictions without relying on predefined formulas or models. Several studies have investigated the application of ML models in predicting fuel consumption in diesel combustion [8,24-26]. These models leverage various algorithms and techniques to learn patterns and relationships from input data. A comparative study of these models can provide insights into their performance and suitability for fuel consumption prediction.

A variety of research has been reported to show progressive achievements in predicting the combustion process in terms of different ML models. Yuksel et al. [9] compared the performance of multiple ML models, including support vector machines (SVM), random forests (RF), and artificial neural networks (ANN), to predict the fuel consumption in a marine diesel engine. Their results showed that the SVM model was more capable of capturing complex relationships between engine parameters and fuel consumption. Bappon et al. [10] addressed the growing concern of global warming and the role of vehicle emissions, particularly $CO_2$. By applying eight different ML techniques, a remarkable accuracy of 96 % was achieved using the RF algorithm. Ruan et al. [11] proposed a grey box model (GBM) for predicting fuel consumption in wing-diesel hybrid vessels. The optimal combination, utilizing parallel modelling and the RF algorithm with the inclusion of wing fuel consumption savings, returned a remarkable 41.7 % reduction in root mean square error (RMSE) compared with a white box model (WBM). Badra et al. [12] proposed a methodological approach to optimize engine combustion systems using computational fluid dynamics (CFD) and ML. Their research highlighted the potential of ML in predicting fuel consumption and optimizing engine performance. Mandal et al. [13] developed an artificial neural network (ANN) model to predict the performance and emissions of a compression ignition (CI) engine using biogas flow variation demonstrating a great effectiveness of ML techniques in predicting fuel consumption and emissions in CI engines. Gong et al. [14] conducted a comparative study on fuel consumption prediction methods of heavy-duty diesel trucks by considering 21 influencing factors, suggesting that ML algorithms, such as RF and SVR, can accurately predict fuel consumption in heavy-duty vehicles. Zeng et al. [15] conducted a single-pulse shock tube pyrolysis study of RP-3 jet fuel and developed a kinetic model. Although it focuses on the jet fuel, the developed kinetic model can be utilized in predicting fuel consumption in diesel engines. Kaleli and Akolaş [16] designed an electromechanical EGR cooling system for a diesel engine to reduce emission and fuel consumption in terms of ML and genetic algorithms. Satrio et al. [17] analysed the effect of fuel type selection on the performance and fuel consumption of a steam power plant. Wen et al. [18] investigated the impact of input parameters on the accuracy of an AI model for predicting emissions produced by light diesel vehicles using a gradient boosting regression model. Pereira et al. [19] developed a non-invasive approach for fuel consumption prediction of construction trucks using dedicated sensors and ML. The above studies demonstrate the potential of ML techniques in predicting emissions and fuel consumption under real-world driving conditions.

Furthermore, Wu et al. [20] further developed a deep learning-based framework for accurate long-term prediction of turbulent combustion in engines. They proposed two training techniques, namely unrolled training and injecting noise training, to address the issue of shifted distribution in autoregressive long-term prediction. Tuan et al. [21] compared the performance of ANN and SVM methods in predicting the ignition delay of a diesel engine using diesel and biodiesel fuels. Their results showed that the SVM model outperformed the ANN model in predicting the ignition delay. Park et al. [22] described the development of a lightweight and accurate NOx prediction model for diesel engines using the Explainable Artificial Intelligence (XAI). To select the dominant features, they employed the Shapley Additive Explanations (SHAP) method and the Pearson Correlation Coefficient (PCC) method. Their results showed that the prediction performance in the SHAP method is similar to the base model but only utilizing 30 % of its data size. Pitchaiah et al. [23] optimized the performance of a direct inject CI engine fuelled with diesel-Bael biodiesel blends and dimethyl carbonate (DMC) additive. They concluded that the precision and certainty provided by Response Surface Methodology (RSM) and ANN models could help estimate the engine performance and support the Sustainable Development Goals (SDGs) of the United Nations. Godwin et al. [27] predicted the emission parameters in a dual-fuel spark ignition (SI) engine using ML algorithms. They successfully implemented an Ensemble LS Boost ML framework to efficiently optimize the combustion performance and predict its emission characteristics. Ramachandran et al. [28] investigated the Reactivity Controlled Compression Ignition (RCCI) combustion fuelled with microalgae biodiesel and Compressed Natural Gas (CNG) using two ML models, i.e., Gradient Boosting Regressor (GBR) and LASSO Regression, both of which achieved high accuracy. Sanjeevannavar et al. [29] examined the effect of different biodiesel blends with hydrogen peroxide additive on the performance and emissions in an engine. They found that the XG Boost model provided highly accurate predictions and could help reduce the time and costs associated with traditional engine trials.

Although the findings in previous works help make a better understanding of diesel combustion and fuel consumption predictions using ML and AI tools, there are still gaps between the existing knowledge and future development that need to be fully filled. The potential interests could be focused on applying AI techniques to evaluate and select an optimal ML model for an accurate prediction of fuel consumption, leading to improved predictions and optimized fuel efficiency in diesel combustion. The present study aims to fill the research gap in the evaluation of different ML models for predicting fuel consumption in diesel engines. A novel approach is taken by integrating ML techniques/AI tools with computational fluid dynamics (CFD) to improve the accuracy of predictions. This study also introduces the use of dodecane as a surrogate diesel fuel, allowing for a more comprehensive evaluation of ML models. Moreover, the inclusion of additional influencing factors such as fuel mass fraction, apparent heat release rate, spray penetration, and injection velocity further enriches the predictive capabilities of the ML models. The evaluation of ML models is conducted using various metrics, including Mean Absolute Error, Mean Squared Error, Pearson Coefficient, and R-Squared, ensuring a comprehensive assessment of their performance. Furthermore, the study employs a surface field model to visualize the predictions made by ML models. Ultimately, it hopes to provide a set of recommended designs through target optimization to support the development of more efficient and environment friendly diesel engines.

## 2. CFD simulations

The CFD simulation in this study was conducted using Simcenter STAR-CCM+ software, specifically the *In-Cylinder* Solution add-on. This software enables accurate and efficient in-cylinder CFD simulations of engines, providing a critical analysis of the injection, ignition, and combustion processes within the engine cylinder.

### 2.1. Mesh generation

In CFD simulations, a closed-cycle analysis of a diesel compression-ignition engine is conducted within the interval of 680 to 800° using a 45° sector model. The engine's stroke is measured at 158.54 mm, while the connecting rod length is 270.0 mm. Operating at a constant angular velocity of 1100.0 rpm, displacement of 2.1, and compression ratio of 18:1, the simulation starts at 680° crank angle and runs for 120°. To accurately represent the engine's geometry and flow dynamics, a fully automated approach within Simcenter STAR-CCM+ is employed for mesh generation. Trimmed meshes are utilized, which can adapt to the motion of both piston and valves. This approach allows for capturing the intricate details and complexities of the in-cylinder flow. The meshing process involves different techniques and considerations to maintain fidelity and accuracy. The mesh operation employs trimmed cell mesher, prism layer mesher, and triangle methods. The minimum face quality is set at 0.05 to ensure high-quality mesh elements.

Figs. 1(a) and (b) illustrate the cylinder sector in both side and top views, showing a locally refined mesh at 50 % of the base size around the injection area. This refinement is necessary to capture the detailed physics related to droplet diameter calculations, which play a crucial role in fuel atomization and spray behaviour. Furthermore, the second locally refined mesh at 70 % of the base size is implemented in the entire area around the piston crown. This region experiences significant changes and induced turbulence due to the rotational motion of the piston, necessitating a finer mesh resolution. By employing these meshing techniques and local refinements, the simulation can accurately capture the intricate flow dynamics and combustion processes occurring within the engine cylinder. This comprehensive approach ensures that the CFD simulation provides reliable insights into the engine's performance, fuel-air mixing, and combustion characteristics.

### 2.2. Computational details

A multi-nozzle injector is positioned at the centre of the cylinder head, introducing dodecane ($C_{12}H_{26}$) fuel into the cylinder sector. The specific timing of the injection event occurs between 714.75 and 722.65° of crank angles. The accurate representation of the fuel jet's disintegration process is crucial, and therefore the Huh Atomization model is employed. This model considers the breakup of the fuel jet into smaller droplets, enabling the capture of spray behaviour and subsequent mixing with the surrounding air. To account for the heat transfer and boundary layer effects near the walls, constant temperature conditions ranging from 400 to 450 K are applied at the cylinder walls. The simulations employ the Extended Coherent Flame Model with Combustion Limited by Equilibrium Enthalpy (ECFM-CLEH) to accurately simulate the combustion process within the engine cylinder. This model considers the complex chemistry and equilibrium enthalpy constraints during combustion. Additionally, the ECFM TKI Auto-Ignition model enhances the accuracy of the combustion process representation. To promote an efficient fuel-air mixing, a swirl of 2000 rpm is implemented around the central axis of the cylinder's vertical plane. This swirl is crucial for achieving proper combustion characteristics. The initial pressure and temperature within the engine cylinder are set at constant values of 9.87 bar and 583 K, respectively, providing the necessary initial conditions for the combustion process to initiate and progress within the cylinder.

The *In-Cylinder* solution in STAR-CCM+ offers a powerful tool for optimizing in-cylinder properties and provides various optional models to simulate the combustion process, with an Automatic Composition Initialization feature being particularly useful. This feature determines the initial gas composition in the cylinder based on the equivalence ratio and exhaust gas recirculation (EGR) percentage, ensuring an accurate representation of the combustion process. The simulations consist of three stages, i.e., pre-processing, CFD solver, and post-processing. At the pre-processing stage, the engine geometry is constructed, the physics models are selected, and the boundary conditions are specified. The CFD solver stage involves discretizing the domain and iteratively solving the resulting algebraic systems of governing equations. Finally, at the post-processing stage, the simulation results are analysed and visualized using tabular data, 3D data, vectors (e.g., velocity), scalars (e.g., pressure), and contours.

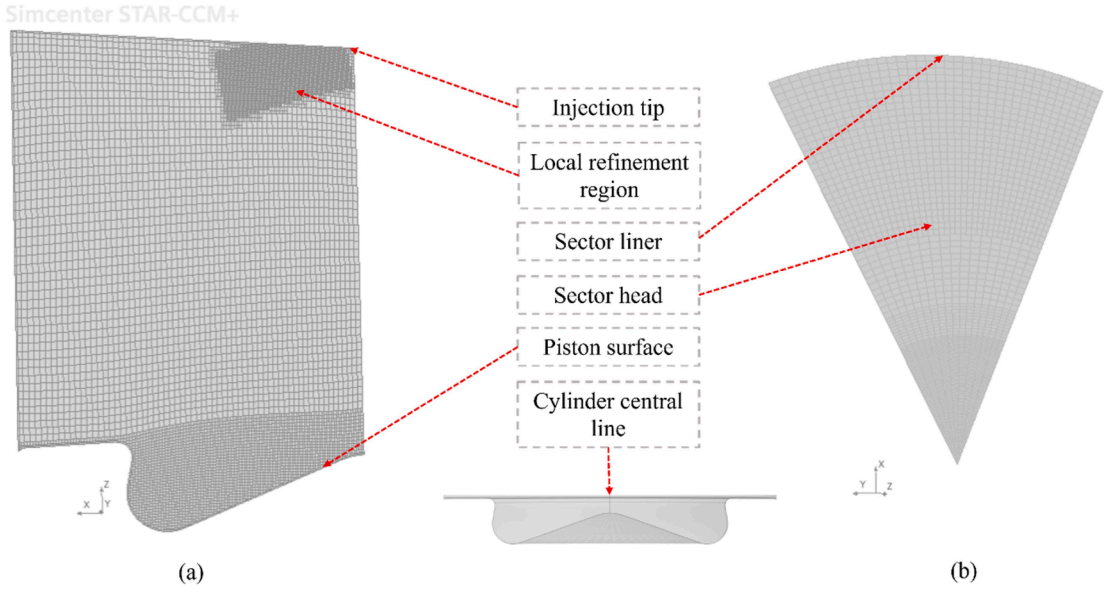



**Fig. 1.** Mesh visualisation of the engine model: (a) side view and (b) top view.

## 3. ML algorithms

The main aim of this study is to utilize ML techniques to predict fuel consumption in diesel combustion. To achieve this, all the necessary data from the CFD simulation, exported from the STAR-CCM+ software, are collected. The Monolith AI platform is then utilized to analyse the data, as well as train and evaluate the ML models. To ensure data quality, the collected data are first categorized and subjected to noise removal processes. Various tools, including sensitivity analysis, can be employed to identify the correlation between inputs and outputs data. This step helps gain insights into the key factors that affect the fuel consumption. The data are then split using the 80–20 % rule, where 80 % is used for training the ML models and the remaining 20 % is reserved for testing and validation. This split ensures that the models are trained on a substantial amount of data while still having a separate dataset for evaluation. The ML models are trained and evaluated using various metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson Coefficient (PC), and R-Squared ($R^2$). These metrics provide a comprehensive assessment of the model's performance, making insightful comparisons between different approaches. In addition to quantitative evaluation, a surface field model is employed to visualize the predictions retuned by deep learning (DL). DL is a subset of machine learning that utilizes neural networks with multiple layers to learn and make predictions from complex data. It is particularly effective when dealing with large amounts of data and complex patterns. The use of raw data (the data extracted from CFD and imported without any training process) and training data in the section of deep learning can be valid if they serve distinct purposes. This visualization technique helps understand the spatial distribution and patterns of fuel consumption within the combustion chamber. Through this iterative process of training, evaluation, and visualization, an optimal ML model is identified, which can accurately predict the fuel consumption in diesel combustion. The selected model is then introduced as the recommended approach for estimating fuel consumption in similar scenarios. The methodology described above ensures a systematic and rigorous approach to harnessing ML techniques for predicting fuel consumption in diesel combustion (see Fig. 2).

### 3.1. Data structure

The prediction of fuel consumption in diesel engines plays a pivotal role in achieving enhanced performance and producing less emissions. In this study, a CFD simulation was conducted starting at 680° crank angle and running for 120°. For each crank angle, a dataset was extracted, resulting in multiple datasets. Each dataset consisted of a varying number of rows, ranging from 9000 to 13,000, and 88 columns including all inputs and their respective subsets of dependencies. It is worth noting that to utilize a parameter as an input, all its dependencies must be taken into consideration. The variation in the number of rows was due to the movement of the piston from top dead centre (TDC) to bottom dead centre (BDC), which created more space in the cylinder. This increased space generated more geometrical data, which in turn led to more operating data. The final size of the dataset, consisting of 120 CSV files (crank angles), each containing 9000–13,000 rows and 88 columns, can be calculated by multiplying these values together. Therefore, the total dataset size is 120 (CSV files) multiplied by an average of 11,000 (rows) multiplied by 88 (columns), resulting in a dataset size of approximately 116,160,000 data points. Regarding the 80–20 % training rule, this typically refers to the practice of splitting the dataset into 80 % for training the machine learning model and 20 % for testing the model's performance. In the context of the given dataset, this would mean using 80 % of the 116,160,000 data points for training, which is approximately 92,928,000 data points, and the remaining 20 % for testing, which is approximately 23,232,000 data points. This approach allows for the model to be trained on a substantial portion of the data while still retaining a separate portion for evaluating its performance.

In this study, several key parameters have been selected based on previous research in the field of fuel consumption and emissions in diesel combustion [3–7]. These parameters include Fuel Ratio (FR), Fuel Mass (FM), Fuel Mass Fraction Averaged (FMFA), Apparent Heat Release Rate (AHRR), Air/Fuel Equivalence Ratio (AFER), Spray Penetration

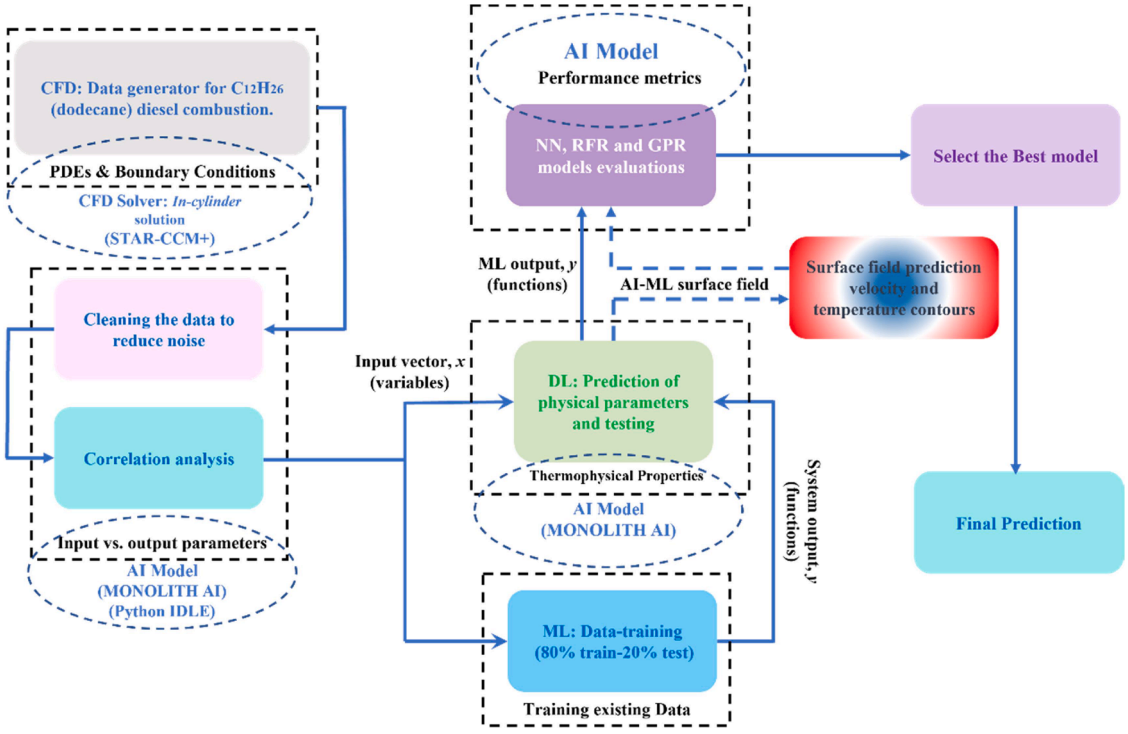



**Fig. 2.** Schematic of the methodology integrating computational fluid dynamics (CFD) tools with artificial intelligence-machine learning (AI-ML) models: an example of predicting internal combustion process in diesel engines.

(LSP-VSP), Injection Velocity (IV), and Cylinder Volume (CV). These parameters have been identified as significant contributors to fuel consumption in diesel engines. In the conducted ML modelling of this study, a set of input variables is utilized to predict specific output parameters. The inputs, which serve as the independent variables, include Liquid-vapour Spray Penetration (LSP-VSP), Injection Velocity (IV), Cylinder Volume (CV), Turbulent Kinetic Energy (TKE), Cylinder Tumble-Y (CT-Y), Cylinder Temperature (CT), Cylinder Swirl (CS), Cylinder Pressure (CP), Air/Fuel Equivalence Ratio (AFER), Air Mass Fraction Averaged (AMFA), Air Ratio (AR), and Air Equivalence Ratio (AER). These inputs capture various aspects of the combustion process, such as spray characteristics, chamber geometry, turbulence, temperature, pressure, and air-fuel mixture properties.

In addition, the ML model aims to predict several output parameters that are crucial for understanding the combustion behaviour. These outputs include Apparent Heat Release Rate (AHRR), Fuel Mass Fraction Averaged (FMFA), Fuel Ratio (FR), and Fuel Mass (FM). AHRR represents the rate at which heat is released during combustion, while FMFA indicates the fraction of fuel mass in the combustion chamber. FR quantifies the ratio of actual fuel mass to the stoichiometric fuel mass, and FM represents the actual mass of fuel present in the combustion chamber. By training the ML model using a dataset that includes the inputs and corresponding output parameters, the model can learn and identify complex relationships and patterns between the input variables and the desired outputs.

The selection of these parameters is crucial as they are interrelated and help understand the intricate relationships between the inputs and outputs of the predictive models. To achieve this, a sensitivity analysis has been conducted using the Sobol method with first-order variable combinations. This advanced analysis technique allows for the examination of both direct effects and interactions between the parameters on the model outputs [24]. This analysis visually demonstrates the effects of each input parameter on the respective model outputs (refer to Fig. 3). The correlation between the inputs and AHRR and FMFA is depicted in Fig. 3. The sensitivity analysis reveals that the chosen inputs have major impacts on the prediction of fuel consumption. Amongst the parameters, AMFA, CT-Y, IV, CV, and AR exhibit the most significant impact on AHRR, while AER, AR, CP, CT, CV, and IV emerge as the most influential parameters affecting FMFA.

The dataset used for analysis was obtained in the CFD simulations, which can introduce various sources of noise. Noise in the data can arise due to computational errors, measurement errors, or other factors that reduce the accuracy and reliability of the collected data [8,11]. Noisy datasets captured from the CFD simulations may contain outliers, duplicate entries, or missing data, which can further impact the quality of the dataset. To improve the quality of the dataset, preprocessing steps were undertaken to remove erroneous or invalid samples. This step helps to ensure that the dataset used for training the models is as clean and reliable as possible. Fig. 4 presents a comparison of learning curves, which illustrate the impact of the amount of training data on the model's error. The metric used in this curve is the Mean Squared Error (MSE), with different percentages indicating the proportion of training data. The blue line on the graph represents the clean data, where noise has been removed or reduced significantly. On the other hand, the red line represents the noisy data, which still contains the effects of noise. It clearly shows that the existence of noise in the dataset can lead to errors or overfitting in the trained models. Specifically, the noisy data exhibits fluctuations in the errors at around 25 % and 70 % of the training data. When using 25 % of the training data, MSE of the noisy data returns to its starting point. Moreover, after utilizing 100 % of the training data, MSE of the noisy data jumps to its maximum value again. This reveals that the presence of noise in the dataset can cause fluctuations and deviations in the errors during training.

After cleaning the dataset, the next step is to apply the train-test split method. This method divides the dataset into two subsets: one is for training the model and the other for testing its performance. The commonly used rule in this split method is the 80–20 % rule, where 80 % of the data is allocated for training and 20 % is reserved for testing [8]. The increase in accuracy using a more extensive training dataset can be attributed to several factors. First, a larger volume of data provides the model with a more comprehensive representation of the underlying patterns and relationships within the dataset. This enables the model to learn more effectively and make more accurate predictions. Second, a more extensive dataset helps to mitigate the impact of potential outliers or noise in the data, leading to improved generalization and robustness of the model. By adhering to the 80–20 % rule, a significant portion of the dataset is dedicated to training the model, allowing it to learn the underlying patterns and relations. The remaining 20 % is then used to evaluate the model's performance on unseen data, providing an estimate of how the model can be effectively generalized to new observations. This division ensures that the model is not overly reliant on the training data and also has the potential to perform well on unseen data.

### 3.2. ML models

This study focuses on regression analysis using supervised learning techniques. Specifically, the aim is to train and compare three reliable ML models to accurately predict the fuel consumption in diesel

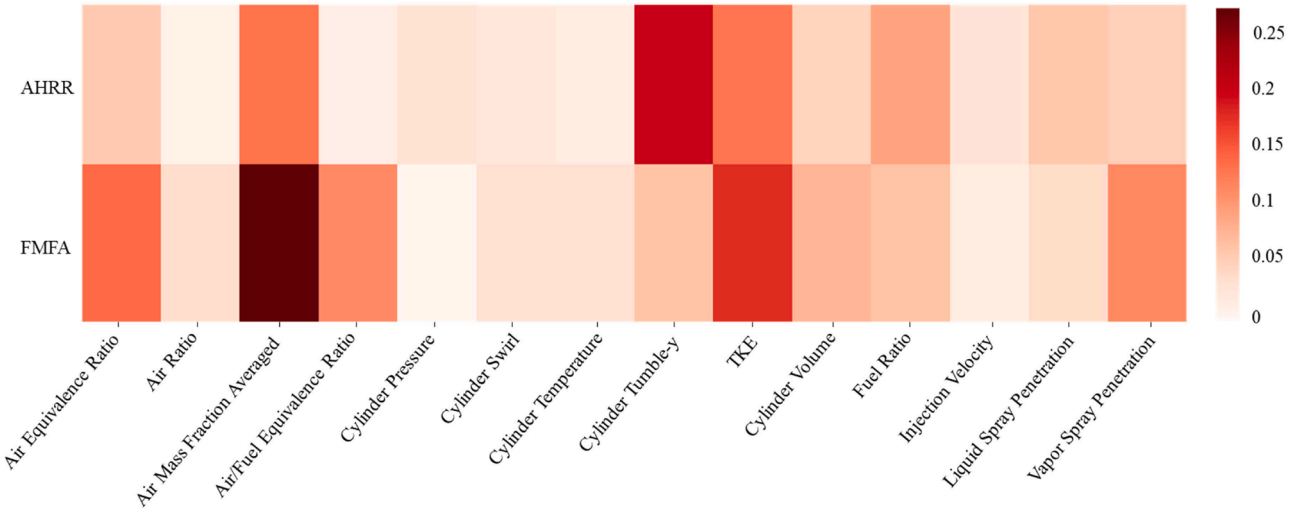



**Fig. 3.** Correlation coefficient heat map of selected parameters with AHRR and FMFA in diesel engines.

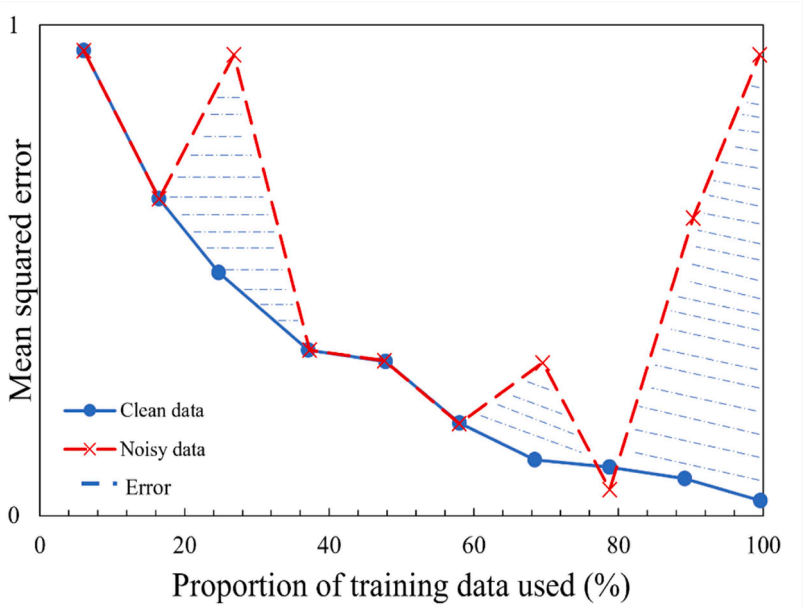



**Fig. 4.** Impact of training data on model's accuracy: a comparison between clean and noisy data.

combustion. The selected ML models for this study include Random Forest Regression (RFR), Neural Network (NN), and Gaussian Process Regression (GPR). Detailed explanations of these models can be found in Refs. [8,30-33], which provide comprehensive insights into their architectures and workings. Given the complexity and non-linearity inherent in diesel combustion, it is crucial to employ robust ML models that can effectively capture the intricate relationships impacting fuel consumption. By utilizing RFR, NN and GPR, researchers can gain valuable insights into the factors affecting fuel consumption in diesel engines. The comparative analysis of these models will provide a better understanding of their performance and help identify the most reliable model for an accurate prediction of fuel consumption.

#### 3.2.1. Model evaluation

To evaluate the generalization capacity of predictive models and mitigate the risk of overfitting, cross-validation is employed. Cross-validation, belonging to the family of Monte Carlo (MC) methods along with the bootstrap, is a statistical technique and can be used to compare and assess learning algorithms [9]. It divides the data into two sections: one is for training the model and the other for model validation. This approach allows for a robust evaluation of the model's performance. A fundamental type of cross-validation is the k-fold cross-validation, which forms the basis for other variations of this technique. In the k-fold cross-validation, the data is initially divided into k segments or folds of roughly equal size. Subsequently, k iterations of training and validation are performed, with each iteration holding out a different fold for validation while utilizing the remaining k-1 folds for model training [31]. This process comprehensively assesses the model's performance across different subsets of the data. The choice of the number of folds (k) typically ranges from two to ten, as this is a commonly applied practice in assessing model performance and avoiding overfitting. By systematically varying the number of folds, it can help improve the stability and robustness of the models under consideration. The performance of the models has been evaluated using various metrics such as R-squared ($R^2$), mean absolute error (MAE), mean squared error (MSE), and Pearson coefficient (PC). $R^2$ measures the proportion of the variance in a dependant variable that can be explained by the independent variables. A value of $R^2$ close to 1 indicates a highly accurate model, which suggests that a large percentage of the variation in the dependant variable is captured by the model. MAE measures the average absolute difference between the predicted values and the actual values. A lower MAE indicates a more accurate model, signifying that the model's predictions are closer to the actual values. MSE calculates the average squared difference between the predicted and actual values. Similar to MAE, a lower MSE indicates a more accurate model, with smaller differences between predicted and actual values. PC measures the linear correlation between two variables. It ranges from −1 to 1, with values closer to 1 indicating a strong positive correlation and values closer to −1 indicating a strong negative correlation. A detailed comparison of these metrics to highlight the model's training process in a more understandable manner is presented in the discussion below.

## 4. Results and discussion

### 4.1. Validation of CFD results

In the context of ML predictions, ensuring the accuracy and reliability of the dataset is of paramount importance. Therefore, it becomes crucial to validate the results obtained in the CFD simulations of diesel combustion. This validation process involves comparing the outcomes of a selected sector with the full cylinder to assess the consistency of the results. Additionally, further validation is achieved through the comparison of in-cylinder pressure with experimental data obtained from previous studies. Fig. 5(a) presents comparison between the predicted pressure using the full cylinder and that using the sector, and also the experimental data [25]. It can be seen clearly that the predicted pressure using the full cylinder agrees extremely well with that using the sector, justifying the use of a sector to save computational costs. It is also evident from Fig. 5(a) that a good agreement is obtained between the predicted pressure and the experimental data [25] and the discrepancies could be mainly due to some difference in fuel type, ignition delay, and operating conditions between the experiment and the CFD simulations. Fig. 5(b) shows that an excellent agreement is obtained between the predicted temperature using the full cylinder and that using the selected sector, confirming the consistency and reliability of the CFD results. Furthermore, comparisons between the CFD predictions and the experimental measurements from Sandia [26] are presented in Figs. 5(c) and (d), which show the liquid and vapour fuel penetration lengths, respectively. The liquid spray penetration initially increases with time and then stabilizes at a quasi-steady value, referred to as the liquid length. Beyond this axial distance, the presence of liquid fuel diminishes. On the other hand, the fuel vapour penetration continues to increase

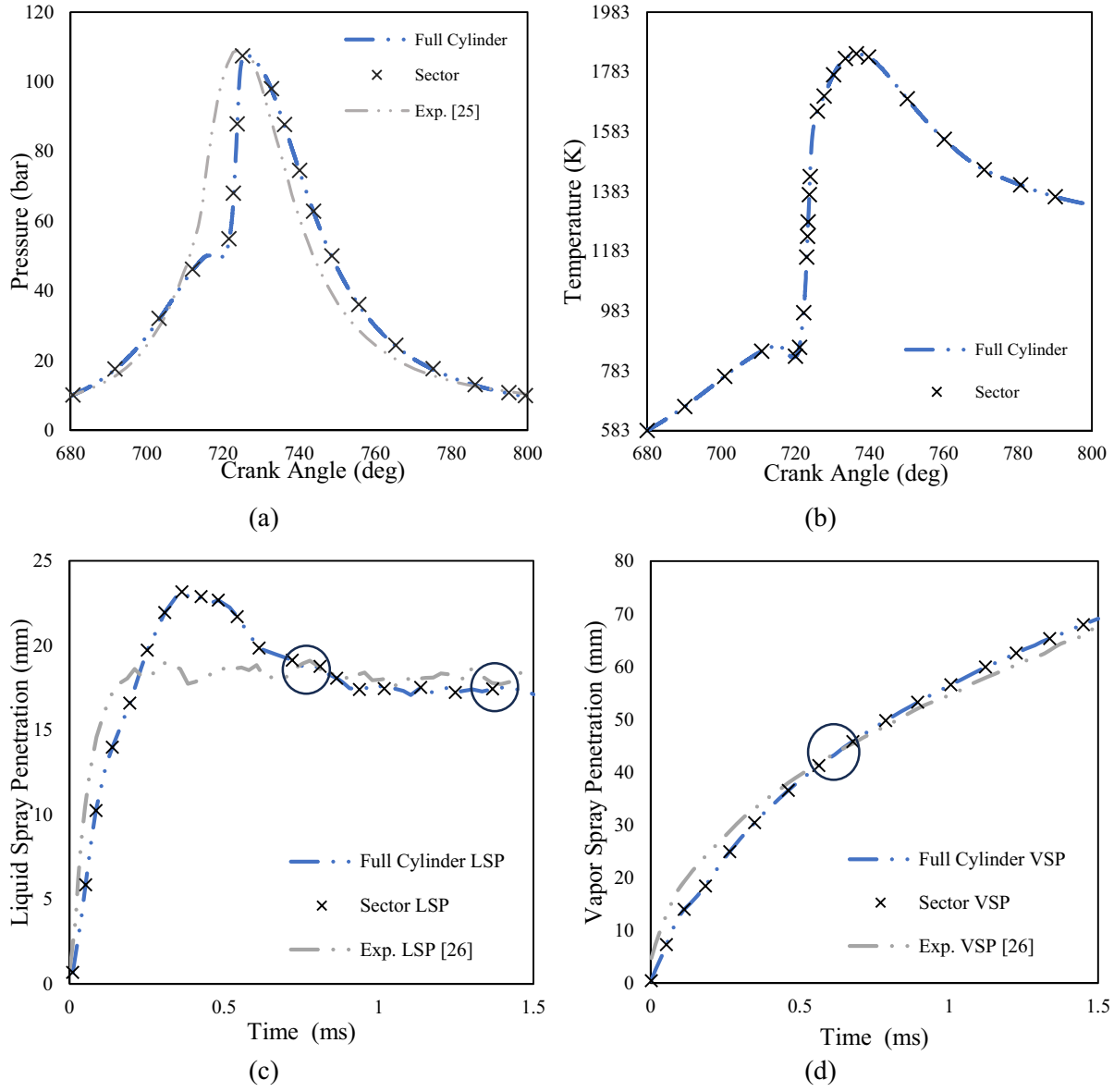



**Fig. 5.** Comparison of pressure (a), temperature (b) and liquid (c) /vapour (d) spray penetration results between the full cylinder and selected sector. Experimental data [25,26] are also included.

with time and plays an important role in fuel-ambient air mixing. A very good agreement between the CFD predictions and the experimental data is clearly observable, suggesting strongly that the characteristics of both the liquid spray and vapour fuel penetration lengths are captured accurately in the simulations.

### 4.2. Comparison of ML models

To evaluate the performance of different ML models, a comparison was conducted amongst Neural Networks (NN), Random Forest Regression (RFR), and Gaussian Process Regression (GPR). In order to facilitate this comparison, an AI tool was employed to track the predicted values against the actual results. Figs. 6(a) and (b) illustrate the comparison between each model's predictions for AHRR and FMFA against actual AHRR and FMFA. This AI technique enables an easily visible comparison of model predictions on a test set, providing insights into their proximity to the true values. While Fig. 6 demonstrates good accuracy and alignment of all models with the actual values at this stage, further comparison with other AI techniques is necessary to determine the best model. To evaluate the models, various metrics are employed, as shown in Table 1. The results indicate that GPR outperforms the others in terms of accuracy, as evidenced by the metrics: MAE, MSE, PC, and $R^2$. For example, in the case of AHRR, GPR achieves an MAE of 0.05124, MSE of 0.06514, PC of 0.99992, and $R^2$ of 0.99982, while RFR and NN exhibit slightly higher values for these metrics. Similar trends are observed for FMFA, FM, and FR. In addition, Fig. 6(c) illustrates the comparison of prediction errors for FM amongst ML models. It clearly demonstrates that the error range for GPR is between 0 and 0.02, while the error range for RFR lies between 0 and 0.08, and for NN it is much larger and exceeds 0.1. These findings emphasize the superior accuracy of the GPR model in comparison with RFR and NN.

Fig. 7 presents the comparison of the models' performance metrics using a Box and Whisker plot. This plot provides insights into the median, minimum and maximum datapoints, as well as the 1st and 2nd quartiles for four key metrics: MAE, MSE, $R^2$, and PC. The results depicted in Fig. 7 clearly demonstrate that the GPR model exhibits highly accurate predictions, outperforming both the RFR and NN models across all the evaluated metrics. The tight clustering, lower median values, and smaller interquartile ranges for the GPR model indicate its superior performance and robustness in accurately predicting fuel

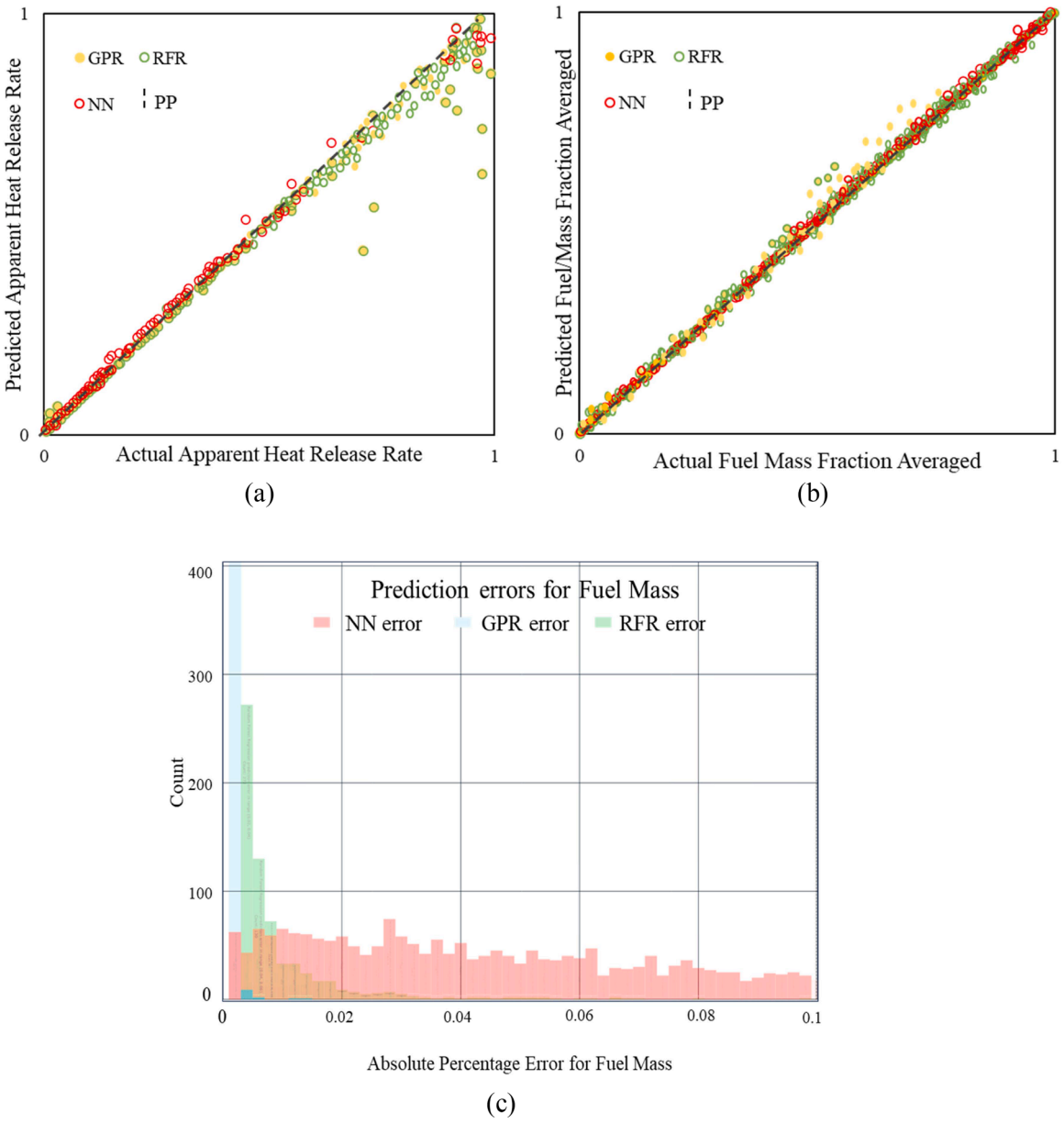



**Fig. 6.** Comparison of predicted and actual values by GPR, NN and RFR models for: (a) AHRR; (b) FMFA; and (c) prediction errors for FM.

**Table 1**
Comparison of model's performance metrics.

| Output | Model | MAE | MSE | PC | $R^2$ |
|---|---|---|---|---|---|
| AHRR | GPR | 0.05124 | 0.06514 | 0.99992 | 0.99982 |
| | RFR | 0.39811 | 0.15328 | 0.99842 | 0.99581 |
| | NN | 0.75616 | 0.99148 | 0.99015 | 0.97292 |
| FMFA | GPR | 0.0 | 0.0 | 1.0 | 1.0 |
| | RFR | 1.00E-05 | 0.0 | 1.0 | 1.0 |
| | NN | 0.00018 | 0.0 | 0.99985 | 0.99972 |
| FR | GPR | 0.0 | 0.0 | 1.0 | 1.0 |
| | RFR | 0.0 | 0.0 | 1.0 | 1.0 |
| | NN | 0.0 | 0.0 | 0.99992 | 0.99981 |
| FM | GPR | 0.0 | 0.0 | 1.0 | 1.0 |
| | RFR | 0.0 | 0.0 | 1.0 | 1.0 |
| | NN | 6.00E-05 | 0.0 | 0.9989 | 0.99977 |

consumption values.

To provide a more comprehensive assessment, Figs. 8(a) and (b) present the comparison of the measured values for FM and AHRR against the predicted values by the proposed models at different crank angles. It is worth emphasizing that although the measured points fall within the uncertainty range of both NN and RFR models, relying solely on the uncertainty range is insufficient when it comes to predicting fuel consumption in a diesel engine. While the uncertainty range provides a measure of the potential variability in the predictions, it cannot guarantee an acceptable accuracy of the model's output. Therefore, it is necessary to go beyond the uncertainty range and thoroughly assess the model's performance and alignment with the measured data. In this regard, the comparison between predicted results by RFR, GPR, and NN models becomes crucial. The significant deviations observed in the NN model's predictions, compared with the measured data, highlight the limitations of relying solely on the uncertainty range as an indicator of model accuracy. To ensure reliable and accurate predictions of fuel consumption, it is imperative to select a model that can demonstrate both good alignment with the measured data and minimal deviations from the true values. This reinforces the importance of evaluating and comparing different ML models, such as RFR and GPR, to identify the most accurate and reliable model to predict the fuel consumption in

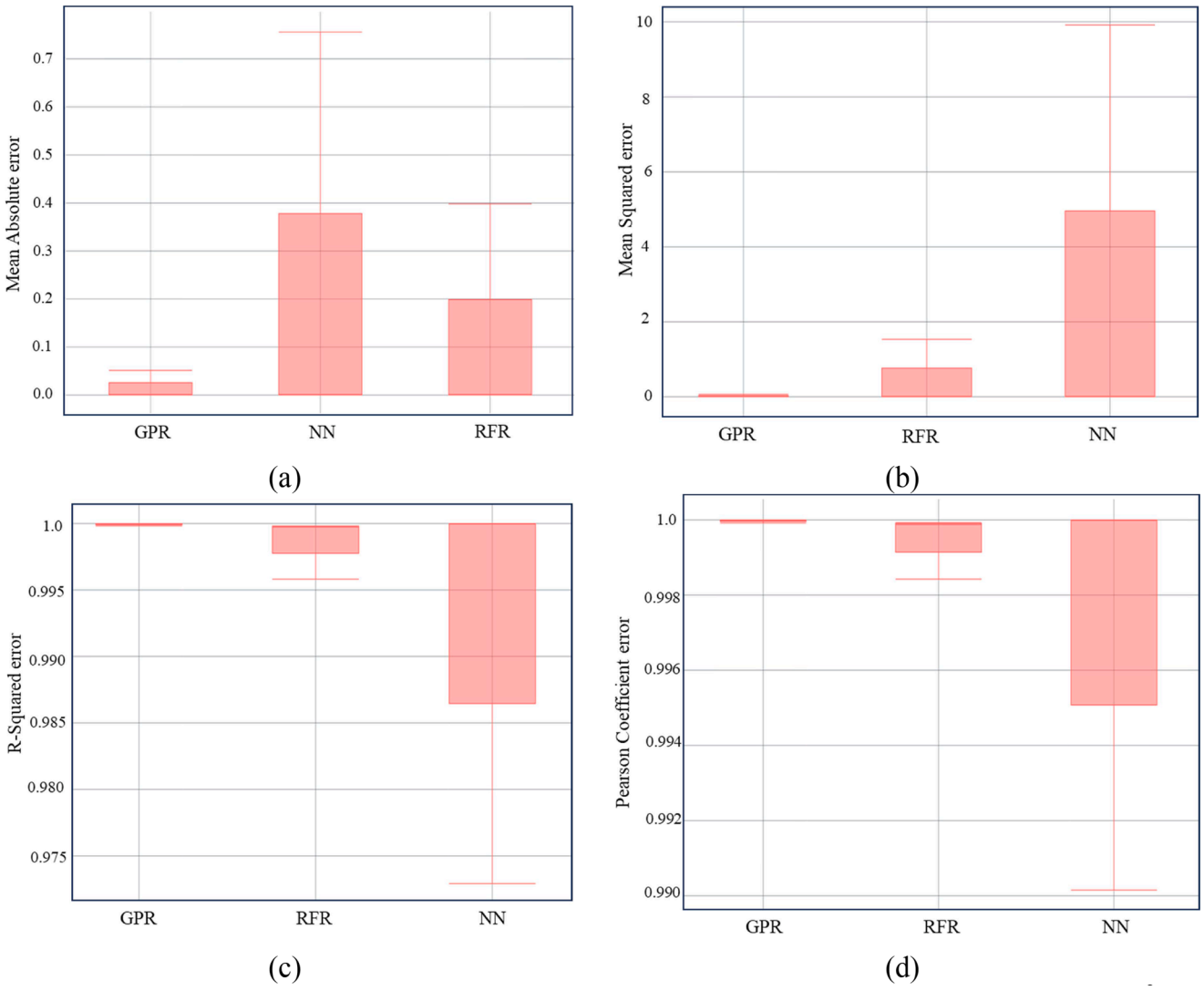



**Fig. 7.** Comparison of performance metrics between different models: (a) MAE; (b) MSE; (c) $R^2$; and (d) PC.

diesel engines. The inset of results in both Figs. 8(a) and (b) highlights the disparity and accuracy of the models around the peak points between 720–730° crank angle. According to Fig. 8, while both RFR and GPR exhibit good alignment with the measured results, there are subtle differences between them when examining the results more closely. For instance, specifically at the peak point and slope within the 730–740° crank angle range in Fig. 8(b), GPR demonstrates the capability to detect and predict even single points that deviate from the overall trend of the graph. This demonstrates the superior predictive ability of the GPR model. The results presented in Table 1 and Figs. 7 and 8 unequivocally indicate that the GPR model outperforms both NN and RFR models.

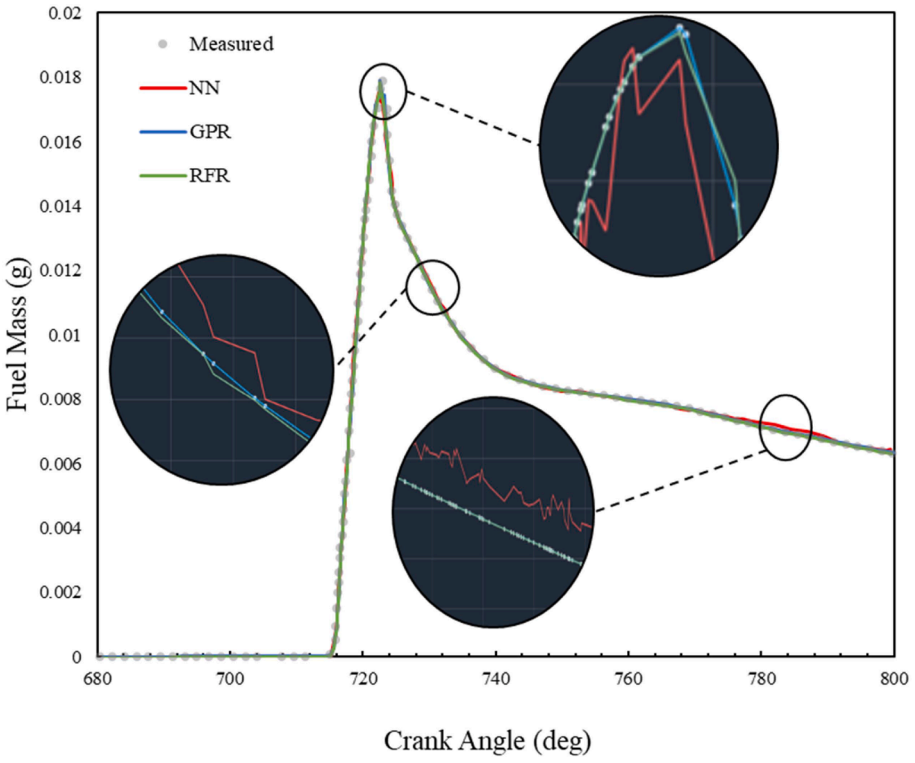
Table 2 presents a comparison of the predictions using NN, RFR, and GPR models against the measured values for FM at random crank angles, including the absolute error in each model. Upon analysis of the results, it is obvious that the GPR model consistently exhibits much lower absolute errors compared with the NN and RFR models. The absolute error values in Table 2 quantify the discrepancy between the predicted values and measured data. Lower absolute error values indicate a good agreement between the model predictions and actual measurements. It can be therefore concluded that the GPR model demonstrates superior accuracy, as evidenced by the considerably lower absolute error values compared with the NN and RFR models. This also indicates that the GPR model can more accurately predict the FM. Notably, the measured value for FM is 0 at a crank angle of 746.05°, and the GPR prediction also aligns perfectly with this value. These findings further support the conclusion that the GPR model outperforms the NN and RFR models in accurately predicting the FM in a diesel engine.

Furthermore, Fig. 9 illustrates comparison of the predicted temperature and pressure by the GPR model against the sector CFD results at different crank angles, and it can be seen that an excellent agreement is reached. Specifically, at a crank angle of 736.5° in Fig. 9(a), the GPR model predicts a maximum temperature of 1841.49 K, while the CFD result is 1835.94 K. Similarly, at a crank angle of 727.3° in Fig. 9(b), the GPR model predicts a maximum pressure of 110.2 bar while the CFD value is 114.8 bar These results highlight the GPR model's ability to precisely predict both the temperature and pressure, further substantiating its selection as the most reliable and accurate model for predicting fuel consumption in a diesel engine.

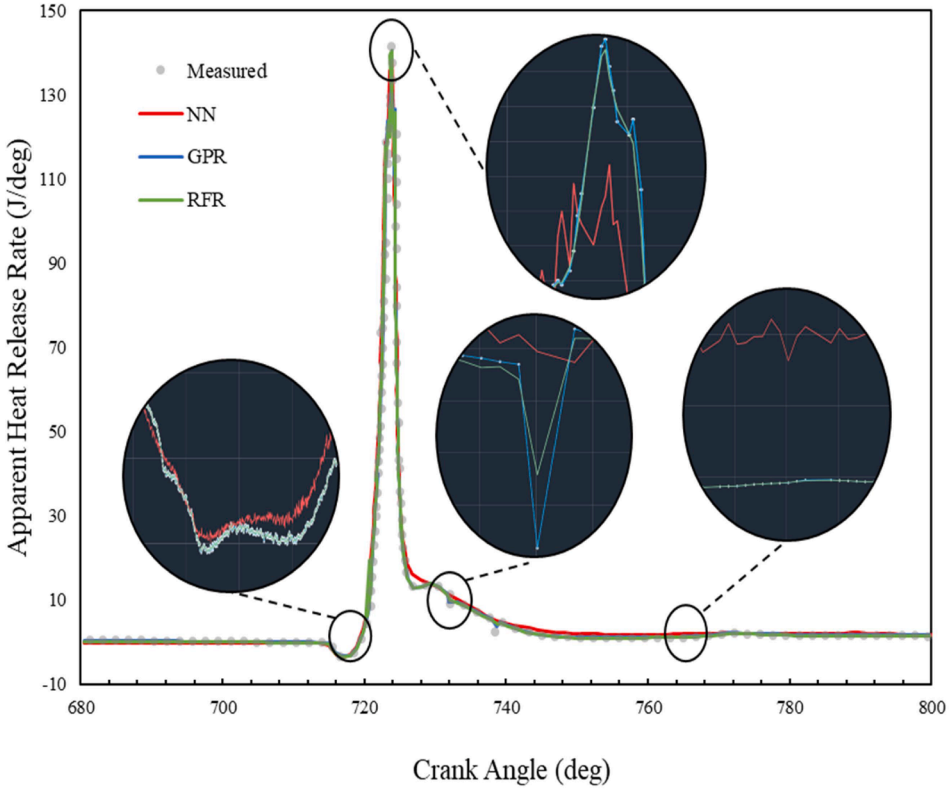
### 4.3. *Surface field prediction*

This section investigates the potential of an innovative AI tool Surface Field in engineering research and development. By leveraging historical 3D simulation data, this approach establishes correlations between 3D geometries and their corresponding surface fields. It explores the prediction capacity of the ML model compared with CFD simulations.

In this study, the surface field model is trained using a DL approach,




(a)




(b)

**Fig. 8.** Comparison of predicted values against measured data for: (a) FM and (b) AHRR using NN, RFF and GPR models. The measured data is represented by white dots, while the predictions from NN model are shown as a red line, GPR as a blue line, and RFR as a green line.

**Table 2**
Comparison of predicted and measured values for FM at random crank angles using NN, RFR, and GPR models.

| Crank Angle | FM Measured | FM in NN | FM in GPR | FM in RFR | NN Ab. Error | GPR Ab. Error | RFR Ab. Error |
|---|---|---|---|---|---|---|---|
| 747.86 | 0.007897548 | 0.007855524 | 0.007897547 | 0.00789714 | 4.20E-05 | 8.00E-10 | 4.08E-07 |
| 773.97 | 0.007019702 | 0.007024102 | 0.007019703 | 0.007020554 | 4.40E-06 | 6.00E-10 | 8.52E-07 |
| 716.44 | 0.004021347 | 0.003914255 | 0.004021309 | 0.0040249 | 0.000107092 | 3.77E-08 | 3.55E-06 |
| 784.49 | 0.006551227 | 0.006672351 | 0.006551228 | 0.006550895 | 0.000121124 | 1.00E-09 | 3.32E-07 |
| 746.05 | 0.007971983 | 0.008029698 | 0.007971983 | 0.007971783 | 5.77E-05 | 0.0 | 2.00E-07 |
| 714.85 | 3.91E-05 | −3.54E-05 | 3.91E-05 | 3.93E-05 | 7.45E-05 | 2.29E-08 | 1.67E-07 |
| 752.25 | 0.007758151 | 0.00780306 | 0.007758151 | 0.007758037 | 4.49E-05 | 4.00E-10 | 1.14E-07 |
| 787.89 | 0.006417309 | 0.006531842 | 0.006417309 | 0.006416862 | 0.000114533 | 2.00E-10 | 4.47E-07 |
| 732.75 | 0.009952324 | 0.009964001 | 0.009952323 | 0.009951039 | 1.17E-05 | 9.00E-10 | 1.29E-06 |
| 789.51 | 0.006354357 | 0.006489141 | 0.006354355 | 0.006354496 | 0.000134784 | 2.40E-09 | 1.39E-07 |

specifically a NN. During training, the model learns the underlying patterns and relationships between the input features and the desired output parameters. This is typically implemented by optimizing the internal parameters of model or its weights based on a chosen loss function and an optimization algorithm. Once the ML model is trained and validated, it can be used to make predictions on new, unseen data. The model takes the input parameters and produces predictions for the surface field parameters of interest. The initial input values for the model are extracted from STAR CCM+ software and are saved in the format of vtk files. The input parameters include point coordinates X, Y, and Z, velocity components i, j and k. The point coordinates X, Y, and Z represent 3D cloud points that specify the geometry of the combustion chamber. In addition, pressure is added as a boundary condition to the model. The outputs of the surface field model are temperature and velocity magnitude, which are predicted based on the given input parameters. Employing NN allows the model to learn complex relationships between the input parameters and the desired surface field outputs, enabling accurate predictions based on the given inputs. Fig. 10 illustrates contours of the velocity magnitudes and Fig. 11 presents the temperature distributions on a XZ plane of the cylinder. Both ML predictions and CFD results are presented at different crank angles. Notably, despite differences in running time and processing resources, a remarkable similarity in the trend of flow patterns and the min/max magnitude of velocity and temperature is observed. These findings highlight the effectiveness of the ML approach in accurately predicting surface fields. The CFD simulations employed parallel computing with 8 cores, resulting in a total CPU time of 31,613 s. In contrast, the ML model was trained for 50,000 steps on a single processor, significantly reducing computational costs. Table 3 provides a comparison of the CPUs utilized in both processes, with the accumulated CPU time highlighted as the key metric. The ML model required only 17,689 s, a remarkable speedup and an approximate 1.7 times faster performance compared with traditional CFD solvers.

One notable advantage of the ML approach is its versatility beyond speed. Once the model is trained and evaluated, it can provide predictions for other desired parameters and their effects on results in real time. This capability underscores the potential of ML models to not only enhance computational efficiency but also offer instant insights into the impact of various parameters on surface field predictions.

Fig. 10 displays the velocity magnitude contours at different crank angles, providing a comparison between the ML prediction and the results obtained in CFD simulations. The findings demonstrate the ML model's exceptional accuracy in detecting and predicting flow fields, particularly in challenging areas like the edges and corners of the cylinder sector. Notably, the contours exhibit a circular shape in some regions, suggesting the presence of vortices. The ML model successfully captures and predicts these vortical structures, showing its ability to discern complex flow phenomena. Moreover, there is a remarkable proximity between the minimum and maximum velocity values obtained from both ML and CFD predictions, underscoring the high level of accuracy achieved by the ML model. Fig. 11 portrays the temperature distribution at different crank angles. Similar to the velocity contours, the ML model accurately predicts the temperature distributions everywhere, encompassing the corners, margins, edges, and walls. The temperature patterns captured by the ML model closely align with those derived from the CFD simulations. In addition, the minimum and maximum temperature values from both the ML model and the CFD predictions at different crank angles agree very well, further substantiating the ML model's precision in predicting temperature distributions. These results exemplify capabilities of the ML model to capture and replicate the momentum and thermal characteristics of diesel combustion process in a diesel engine, also highlighting its potential for advancing surface field predictions.

Notably, increasing the number of training steps is anticipated to yield consistent results having the train loss and validation loss progressively decreased. This underscores the potential of based surface field predictions as a cost-effective and efficient alternative to resource-intensive simulations, emphasizing its relevance and impact on the future development of combustion engineering and science. Fig. 12 presents a comprehensive overview of a model's training process over 50k training steps, where the blue line represents the validation, and the red line indicates the training process on selected data. The inset graphs illustrate the temperature distribution on a XY plane of the cylinder at various training steps, specifically at 20k, 30k, 40k, and 50k, shedding light on the evolution of the accuracy of predictions with an increasing number of training steps. This trend suggests a positive correlation between the number of training steps and the accuracy of predictions in comparison with the CFD results. Thus, the findings imply that more training steps lead to more accurate predictions, as evidenced by the diminishing train loss and validation loss values.

### 4.4. Targeted optimization of fuel consumption

In the context of targeted optimization using GPR as the optimal model, the objective is to identify the best set of inputs that closely match a given list of target outputs. To achieve this, a fitness function that is called the Euclidean distance is used in this approach. The Euclidean distance is a metric method that measures the straight-line distance between two points in a multi-dimensional space [34]. Regarding the targeted optimization, it quantifies the similarity between the predicted outputs obtained from the GPR model and the desired target outputs. By minimizing the Euclidean distance, the GPR model can identify the inputs that result in outputs, which closely match the desired target outputs. The Euclidean distance is calculated by taking the square root of the sum of the squared differences between corresponding elements of the predicted outputs and the target outputs. This fitness function provides a quantitative measure of how close the predicted outputs are to the target outputs [35]. Using the Euclidean distance as the fitness function offers several advantages in the targeted optimization. First, it enables researchers to directly assess the similarity between the predicted and target outputs. By minimizing the Euclidean distance, researchers can identify the set of inputs that produce outputs, which closely meet the desired performance and characteristics. Second, as a widely used and intuitive metric, the Euclidean distance is easy to

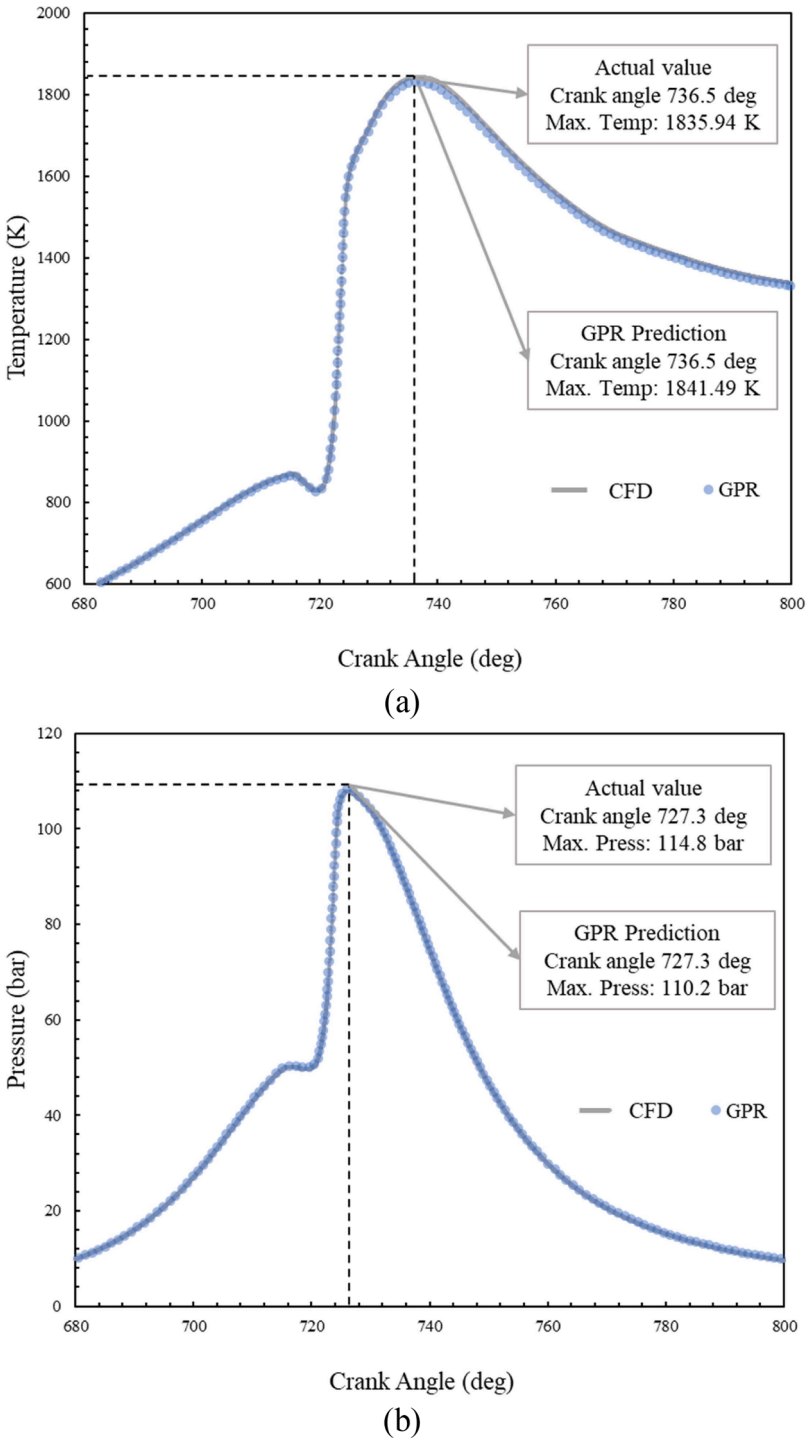



**Fig. 9.** Comparison of predictions using GPR model with CFD results for: (a) temperature and (b) pressure at different crank angles.

interpret. Its calculation is straightforward and can be efficiently implemented into optimization algorithms. This feature helps iteratively refine the inputs and converge towards the optimal solution more efficiently and effectively.

The recommended designs presented in Table 4 are an integral part of the predicted dataset obtained through the application of Gaussian Process Regression (GPR). These designs have undergone rigorous validation as part of this study, ensuring their reliability and accuracy. By utilizing the GPR model, researchers are able to make precise predictions of fuel consumption based on the given parameters and their values. The GPR model, known as a probabilistic approach, offers the robust and reliable means of approximating the unknown target function associated with fuel consumption in a diesel engine. Through an extensive validation, the recommended designs have demonstrated their efficacy in achieving the desired performance and characteristics. By relying on the predictions generated by the GPR model, researchers can confidently select these designs as reliable options to optimize the fuel consumption. The validation process includes a comparison of the predicted outputs from the GPR model with actual measurements or benchmark data. This step ensures that the recommended designs align closely with the expected fuel consumption levels, providing researchers with a reliable basis for decision-making. It is also crucial to note that

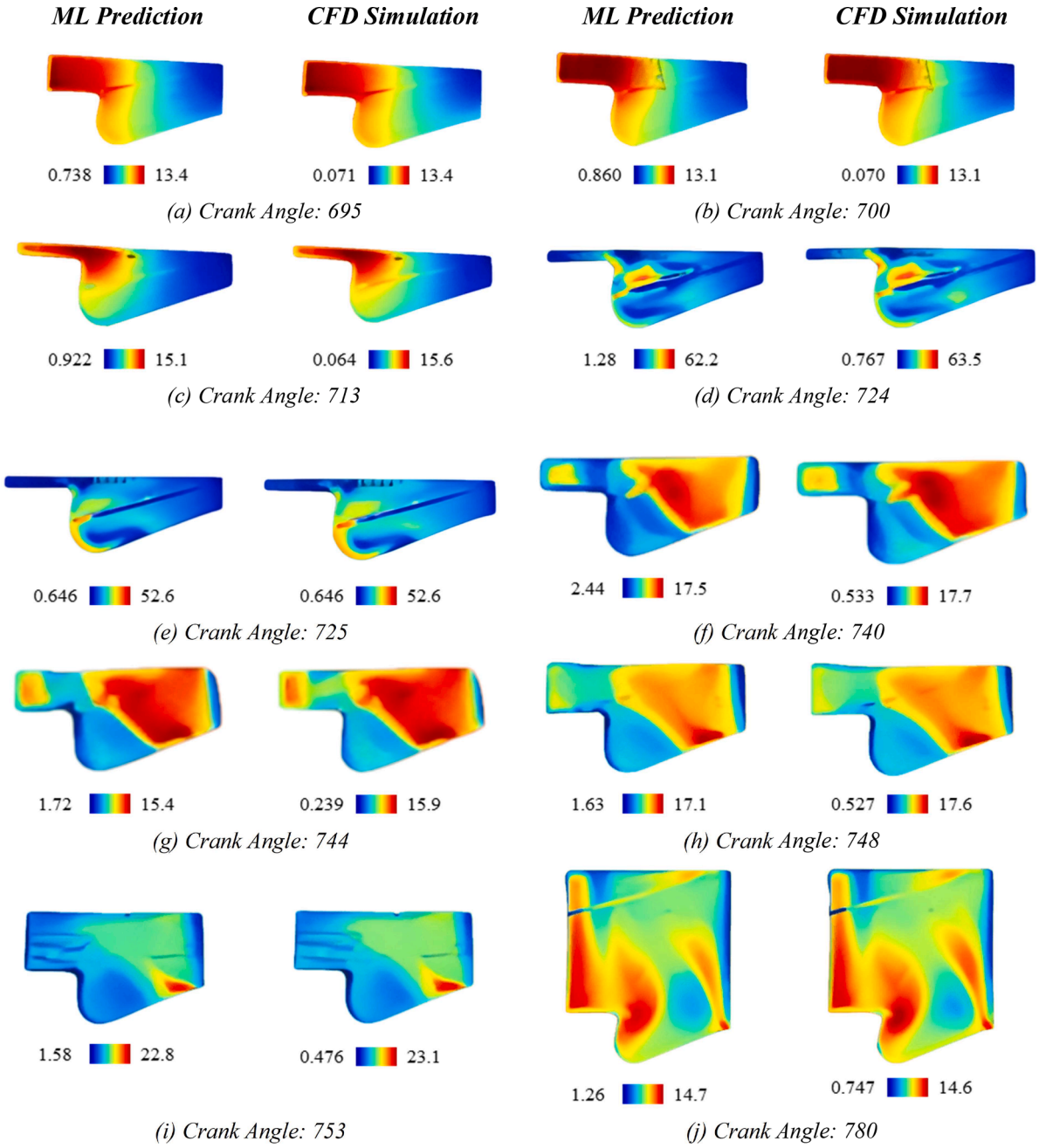



**Fig. 10.** Velocity magnitudes (m/s) of ML predictions and CFD simulations: (a) 695°; (b) 700°; (c) 713°; (d) 724°; (e) 725°; (f) 740°; (g) 744°; (h) 748°; (i) 753°; and (j) 780°.

while the numerical results in Table 4 provide compelling evidence of the efficacy of these design configurations, further experimental validation is necessary to confirm their real-world impact. Rigorous testing and analysis are required to fully evaluate the potential of these optimized designs in practical engine applications.

## 5. Conclusion

This study employs artificial intelligence (AI) techniques to identify an optimal machine learning (ML) model to predict the dodecane fuel consumption in diesel combustion. Through sensitivity analysis, the impact levels of various parameters have been determined, highlighting the most influential ones in fuel consumption. The present work also addresses the impact of data noise and implements data cleaning techniques to ensure the reliability of the obtained results. To validate the accuracy of the predictions, several metrics and validation steps have been conducted. The computational fluid dynamics (CFD) results are compared with experimental data, followed by a comparison of the ML model predictions with the CFD results. Comprehensive comparisons amongst neural networks (NN), random forest regression (RFR), and Gaussian process regression (GPR) models have been performed, taking into account the complexity of fuel consumption prediction. The findings of this study demonstrate that the GPR model outperforms other models in terms of accuracy. Metrics such as mean absolute error (MAE), mean squared error (MSE), Pearson coefficient (PC), and R-squared ($R^2$) consistently indicate the superior performance of the GPR model. The GPR model exhibits superior predictive ability by accurately detecting and predicting some single points that deviate from the overall trend. Furthermore, it consistently demonstrates superior accuracy compared with the NN and RFR models, as evidenced by considerably lower

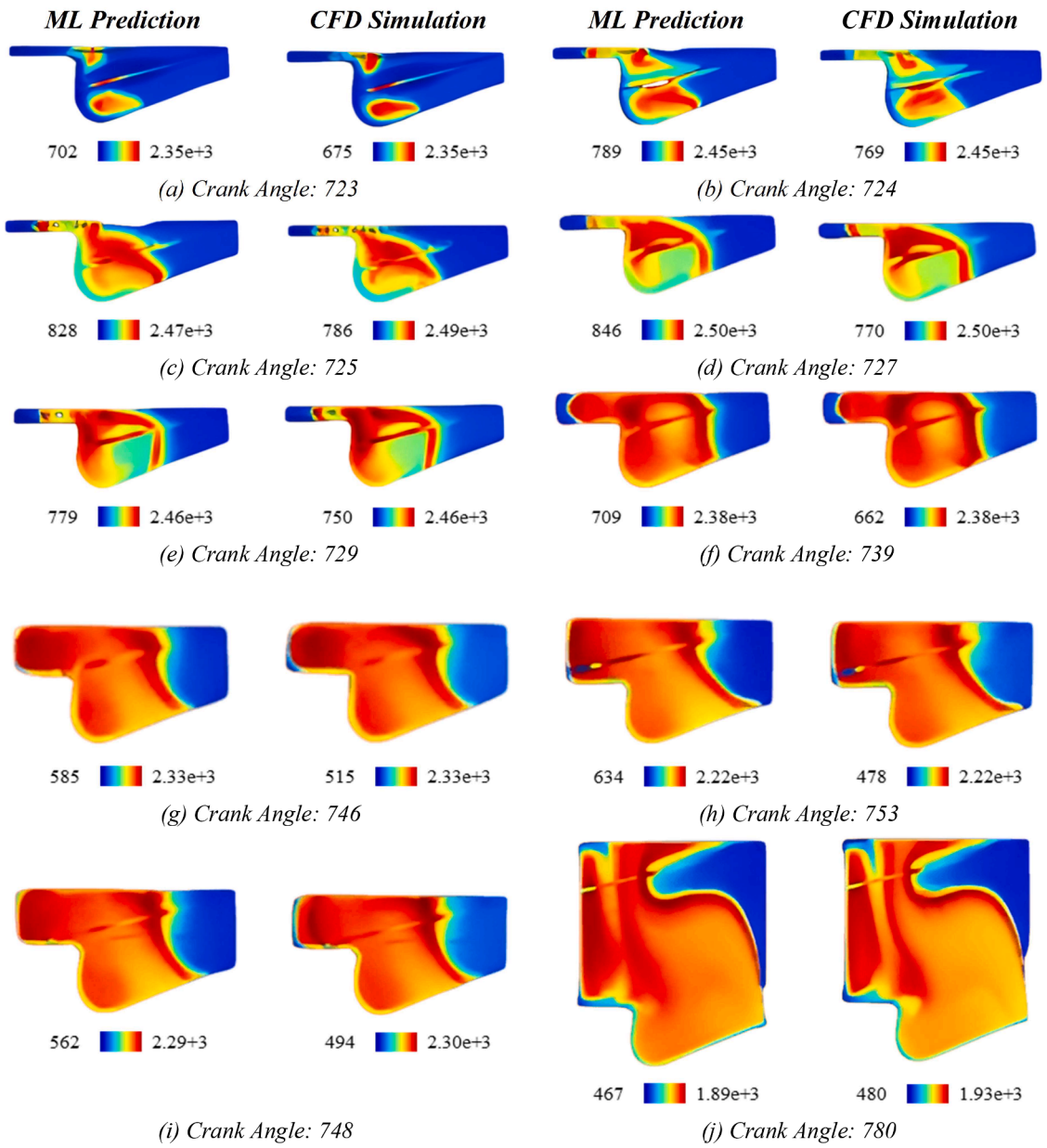



**Fig. 11.** Temperature distributions (K) of ML predictions and CFD simulations: (a) 723°; (b) 724°; (c) 725°; (d) 727°; (e) 729°; (f) 739°; (g) 746°; (h) 753°; (i) 748°; and (j) 780°.

**Table 3**
Comparison of CPU times between CFD simulations and ML model predictions.

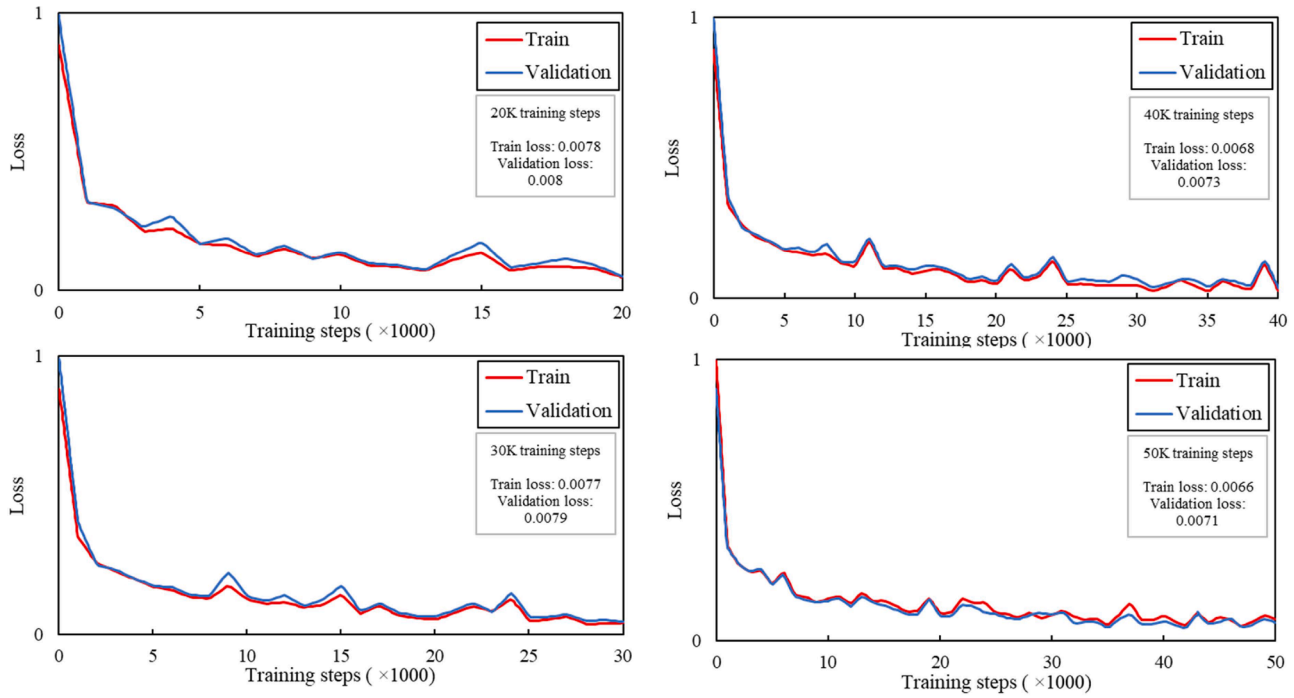
| Module | CFD simulation | ML model | Speed up |
|---|---|---|---|
| Accumulated CPU Time Over All Processes (s) | 31,613 | 17,689 | 1.7 |

absolute error values. Additionally, the GPR model shows a significant speedup, approximately 1.7 times faster than traditional CFD solvers. Regarding surface field prediction, the ML models demonstrate good prediction performance for both flow and heat transfer characteristics. This analysis employing ML techniques provides valuable insights into the underlying physics of complex processes of combustion, effectively showing the impacts of operating parameters on fuel consumption.

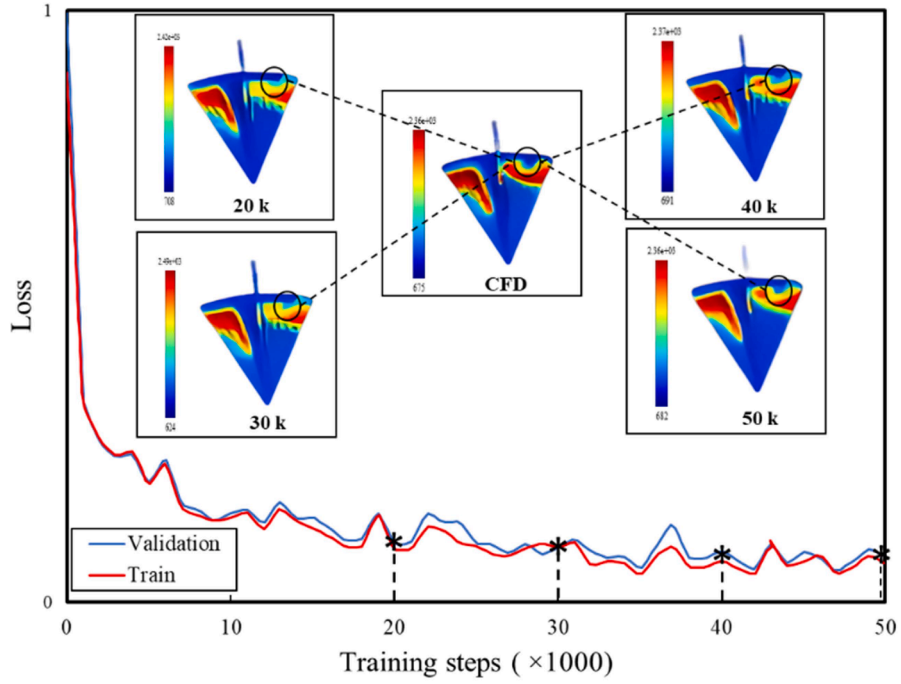
It is also worth pointing out the limitations of the present study. The first limitation is the requirement for data cleaning. Employing ML algorithms introduce an additional step of data cleaning, including noise removal, duplicate data removal, and outlier detection. Running this extra step is time-consuming and requires manual intervention, which may reduce the efficiency of the prediction process. Future studies could focus on developing automated data cleaning techniques or exploring alternative approaches to minimize the preprocessing requirements. Another limitation is the dependency of ML model compatibility on specific data types. The architectures of ML models used in this study may not be universally compatible with all types of data. Different types of data may require specific preprocessing techniques or modifications to the model architecture, limiting the generalizability and applicability of ML models. To overcome these obstacles, more flexible ML models are hoped to handle the diverse data types without modifying the model architecture significantly.

## CRediT authorship contribution statement

**Amirali Shateri:** Writing – original draft, Visualization, Validation,




(a)




(b)

**Fig. 12.** Analysis of the prediction accuracy against training steps: (a) training and validation results for 20–50k training steps and (b) temperature distribution on a XY plane of the cylinder at 20–50k training steps.

Software, Methodology, Investigation, Formal analysis, Data curation. **Zhiyin Yang:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis. **Jianfei Xie:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

**Table 4**
Optimal designs recommended for fuel consumption optimization in diesel engines.

| Recommended Design | Crank Angle (deg) | FMFA | AFER | FR | IV (m/s) | LSP (mm) | VPS (mm) | FM (g) |
|---|---|---|---|---|---|---|---|---|
| # 1 | 745.94 | 0.04710 | 4.2462 | 20.120 | 474.74 | 31.683 | 45.858 | 0.0790 |
| # 2 | 746.17 | 0.04465 | 5.0573 | 19.134 | 477.73 | 30.195 | 47.594 | 0.0774 |
| # 3 | 741.27 | 0.04931 | 4.8295 | 21.143 | 445.60 | 32.709 | 39.998 | 0.0743 |
| # 4 | 742.13 | 0.04323 | 5.5233 | 20.641 | 441.85 | 31.157 | 45.798 | 0.0727 |
| # 5 | 737.67 | 0.04670 | 6.0805 | 20.695 | 464.94 | 32.003 | 35.357 | 0.0708 |
| # 6 | 754.38 | 0.04711 | 4.7135 | 20.794 | 381.64 | 30.630 | 36.968 | 0.0623 |
| # 7 | 755.30 | 0.04307 | 6.7973 | 20.885 | 418.63 | 31.062 | 31.794 | 0.0592 |
| # 8 | 755.69 | 0.04436 | 6.7171 | 17.790 | 467.80 | 34.448 | 30.243 | 0.0569 |
| # 9 | 751.79 | 0.04998 | 4.7135 | 17.428 | 381.64 | 30.630 | 36.968 | 0.0537 |
| # 10 | 724.72 | 0.04924 | 5.7649 | 17.719 | 176.36 | 26.627 | 38.365 | 0.0390 |
| # 11 | 719.77 | 0.04617 | 3.2729 | 16.744 | 43.26 | 25.337 | 45.651 | 0.0248 |

## Declaration of competing interest

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

## Data availability

Data will be made available on request.

## Acknowledgement

The authors would like to express sincere thanks to the AI engineers from Monolith AI for their very useful discussions, and also their permission to use different ML algorithms on the Monolith platform. The Ph.D studentship for Mr. Amirali Shateri and other support provided by the University of Derby are gratefully acknowledged.

## References

[1] Shateri Amirali, jalili Bahram, saffar Saber, Jalili Payam, Ganji Davood Domiri. Numerical study of the effect of ultrasound waves on the turbulent flow with chemical reaction. Energy 2024;289:129707.
[2] Xie Jianfei. Approaches for describing processes of fuel droplet heating and evaporation in combustion engines. Fuel 2024;360:130465.
[3] Papagiannakis RG, Krishnan SR, Rakopoulos DC, Srinivasan KK, Rakopoulos CD. A combined experimental and theoretical study of diesel fuel injection timing and gaseous fuel/diesel mass ratio effects on the performance and emissions of natural gas-diesel HDDI engine operating at various loads. Fuel 2017;202:675–87.
[4] Zheng Jinbao, Wang Jinhua, Zhao Zhibo, Wang Duidui, Huang Zuohua. Effect of equivalence ratio on combustion and emissions of a dual-fuel natural gas engine ignited with diesel. Appl Therm Eng 2019;146:738–51.
[5] Sateesh KA, Yaliwal VS, Soudagar Manzoore Elahi M, Banapurmath NR, Fayaz H, Safaei Mohammad Reza, Elfasakhany Ashraf, EL-Seesy Ahmed I. Utilization of biodiesel/Al2O3 nanoparticles for combustion behavior enhancement of a diesel engine operated on dual fuel mode. J Therm Anal Calorim 2022;147(10):5897–911.
[6] Ayodhya Archit Srinivasacharya, Narayanappa Kumar Gottekere. An overview of after-treatment systems for diesel engines. Environ Sci Pollut Res 2018;25:35034–47.
[7] Shameer PMohamed, Ramesh Kasimani, Sakthivel Rajamohan, Purnachandran Ramakrishman. Effects of fuel injection parameters on emission characteristics of diesel engines operating on various biodiesel: a review. Renew Sustain Energy Rev. 2017;67:1267–81.
[8] Ihme Matthias, Chung Wai Tong, Mishra Aashwin Ananda. Combustion machine learning: principles, progress and prospects. Prog Energy Combust Sci 2022;91:101010.
[9] Yuksel Onur, Bayraktar Murat, Sokukcu Mustafa. Comparative study of machine learning techniques to predict fuel consumption of a marine diesel engine. Ocean Eng 2023;286:115505.
[10] Bappon Suborno Deb, Dey Ashim, Sabuj Shahriar Mahmud, Das Annesha. Toward a machine learning approach to predict the Co2 rating of fuel-consuming vehicles in Canada. In: In 2022 25th International Conference on Computer and Information Technology (ICCIT). IEEE; 2022. p. 384–9.
[11] Ruan Zhang, Huang Lianzhong, Wang Kai, Ma Ranqi, Wang Zhongyi, Zhang Rui, Zhao Haoyang, Wang Cong. A novel prediction method of fuel consumption for wing-diesel hybrid vessels based on feature construction. Energy 2024;286:129516.
[12] Badra Jihad A, Khaled Fethi, Tang Meng, Pei Yuanjiang, Kodavasal Janardhan, Pal Pinaki, Owoyele Opeoluwa, Fuetterer Carsten, Mattia Brenner, Aamir Farooq. Engine combustion system optimization using computational fluid dynamics and machine learning: a methodological approach. J Energy Resour Technol 2021;143(2):022306.
[13] Mandal Adhirath, Cho Haengmuk, Singh Chauhan Bhupendra. ANN prediction of performance and emissions of CI engine using biogas flow variation. Energies (Basel) 2021;14(10):2910.
[14] Gong Jian, Shang Junzhu, Li Lei, Zhang Changjian, He Jie, Ma Jinhang. A comparative study on fuel consumption prediction methods of heavy-duty diesel trucks considering 21 influencing factors. Energies (Basel) 2021;14(23):8106.
[15] Zeng Ping, Wang Bi-Yao, He Ruining, Liang Jinhu, Yang Zhi-Yuan, Xia Zu-Xi, Wang Quan-De. Single-pulse shock tube pyrolysis study of rp-3 jet fuel and kinetic modelling. ACS Omega 2021;6(16):11039–47.
[16] Kaleli Alırıza, Akolaş Halil İbrahim. The design and development of a diesel engine electromechanical EGR cooling system based on machine learning-genetic algorithm prediction models to reduce emission and fuel consumption. In: Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 236; 2022. p. 1888–902.
[17] Satrio Dendy, Trisnayanthi NNAI, Pratilastiarso Joke. Analysis of the effects of fuel type selection on the performance and fuel consumption of a steam power plant. Int. J Adv Sci Eng Inf Technol 2021;11:2046–54.
[18] Wen Hung-Ta, Lu Jau-Huai, Jhang Deng-Siang. Features importance analysis of Diesel vehicles' NOx and CO2 emission predictions in real road driving based on gradient boosting regression model. Int J Environ Res Public Health 2021;18(24):13044.
[19] Pereira Gonçalo, Parente Manuel, Moutinho João, Sampaio Manuel. Fuel consumption prediction for construction trucks: a noninvasive approach using dedicated sensors and machine learning. Infrastructures (Basel) 2021;6(11):157.
[20] Wu Sipei, Wang Haiou, Luo Kai Hong. A robust autoregressive long-term spatiotemporal forecasting framework for surrogate-based turbulent combustion modeling via deep learning. Energy and AI 2023:100333.
[21] Tuan Nguyen Van, Minh Duong Quang, Khoa Nguyen Xuan, Lim Ocktaeck. A study to predict ignition delay of an engine using diesel and biodiesel fuel based on the ANN and SVM machine learning methods. ACS Omega 2023;8(11):9995–10005.
[22] Park Jeong Jun, Lee Sangyul, Shin Seunghyup, Kim Minjae, Park Jihwan. Development of a light and accurate Nox Prediction model for diesel engines using machine learning and Xai methods. Internat. J. Automotive Techn. 2023;24(2):559–71.
[23] Pitchaiah S, Juchelková Dagmar, Sathyamurthy Ravishankar, Atabani AE. Prediction and performance optimisation of a DI CI engine fuelled diesel–Bael biodiesel blends with DMC additive using RSM and ANN: energy and exergy analysis. Energy Convers Manag 2023;292:117386.
[24] Novello Paul, Poëtte Gaël, Lugato David, Congedo Pietro Marco. Goal-oriented sensitivity analysis of hyperparameters in deep learning. J Sci Comput 2023;94(3):45.
[25] Zeuch Thomas, Moréac Gladys, Ahmed Syed Sayeed, Mauss Fabian. A comprehensive skeletal mechanism for the oxidation of n-heptane generated by chemistry-guided reduction. Combust Flame 2008;155(4):651–74.
[26] L.M. Pickett, S. Parrish, S. Kaiser, et al. Engine Combustion Network. 2024. Available at http://www.sandia.gov/ecn/.
[27] Godwin DJesu, Varuvel Edwin Geo, Leenus Jesu Martin M. Prediction of combustion, performance, and emission parameters of ethanol powered spark ignition engine using ensemble Least Squares boosting machine learning algorithms. J Clean Prod 2023;421:138401.
[28] Ramachandran Elumalai, Krishnaiah Ravi, Venkatesan Elumalai Perumal, Parida Satyajeet, Dwarshala Siva Krishna Reddy, Khan Sher Afghan, Asif Mohammad, Linul Emanoil. Prediction of RCCI combustion fueled with CNG and algal biodiesel to sustain efficient diesel engines using machine learning techniques. Case Stud Thermal Eng 2023;51:103630.
[29] Sanjeevannavar Mallesh B, Banapurmath Nagaraj R, Dananjaya Kumar V, Sajjan Ashok M, Badruddin Irfan Anjum, Vadlamudi Chandramouli, Krishnappa Sanjay, Kamangar Sarfaraz, Baig Rahmath Ulla, Yunus Khan TM. Machine learning prediction and optimization of performance and emissions characteristics of IC engine. Sustainability 2023;15(18):13825.
[30] Srinidhi Chetan L, Ciga Ozan, L.Martel Anne. Deep neural network models for computational histopathology: a survey. Med Image Anal 2021;67:101813.

[31] Xue Liang, Liu Yuetian, Xiong Yifei, Liu Yanli, Cui Xuehui, Lei Gang. A data-driven shale gas production forecasting method based on the multi-objective random forest repression. J Petroleum Sci Eng 2021;196:107801.
[32] Cai Haoshu, Jia Xiaodong, Feng Jianshe, Li Wenzhe, Hsu Yuan-Ming, Lee Jay. Gaussian process regression for numerical wind speed prediction enhancement. Renew Energy 2020;146:2112–23.
[33] Hsu Chia-Yu, Lim Sirirat Sae, Yang Chin-Sheng. Data mining for enhanced driving effectiveness: an eco-driving behaviour analysis model for better driving decisions. Int J Prod Res 2017;55(23):7096–109.
[34] Dokmanic Ivan, Parhizkar Reza, Ranieri Juri, Vetterli Martin. Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Process Mag 2015;32(6):12–30.
[35] Liang Jing, Wei Yunpeng, Qu Boyang, Yue Caitong, Song Hui. Ensemble learning based on fitness Euclidean-distance ratio differential evolution for classification. Nat Comput 2021;20:77–87.