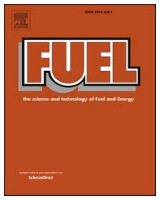
Fuel 389 (2025) 134590

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel



Full Length Article

Data-driven optimization of turbulent kinetic energy and tumble-y in combustion engines: A comparative study of machine learning models

Amirali Shateri [a], Zhiyin Yang [a], Yun Liu [b], Jianfei Xie [a,*]

[a] School of Engineering, University of Derby, DE22 3AW, UK
[b] Hebei Key Laboratory of Low Carbon and High Efficiency Power Generation Technology, School of Energy, Power and Mechanical Engineering, North China Electric Power University, Baoding 071003 China

ARTICLE INFO

Keywords:
Combustion process
Machine learning
Computational fluid dynamics
Target optimization

ABSTRACT

This paper presents an innovative approach to optimising the cold flow dynamics in combustion engines by integrating machine learning (ML) techniques with computational fluid dynamics (CFD). The research focuses on predicting and optimising critical pre-combustion parameters, such as turbulence kinetic energy (TKE) and tumble-y, which are pivotal for enhancing the air–fuel mixing during the intake and compression phases. Three ML models, Random Forest Regression (RFR), Gaussian Process Regression (GPR), and Neural Networks (NN), are evaluated for their predictive capabilities. The GPR model outperforms the others, demonstrating superior accuracy and reduced uncertainty, as highlighted by metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson Coefficient (PC), and R-squared (R^2). Additionally, the ML-based approach achieves a remarkable 21.6x speedup compared with traditional CFD solvers, significantly reducing the computational costs while maintaining high fidelity in capturing momentum and thermal characteristics. The optimization results underscore the critical role of TKE and tumble-y in creating favourable conditions for efficient combustion. For instance, as demonstrated in Design #1 (TKE = 396.56 J/kg, Tumble-y = -0.1535, Temp. = 846.42 K, Pres. = 1.52 bar) and Design #2 (TKE = 366.77 J/kg, Tumble-y = -0.1535, Temp. = 549.59 K, Pres. = 2.81 bar), higher TKE and optimized tumble-y values enhance air motion dynamics, promoting better fuel–air mixing and thermal performance. The rigorous assessment of optimization results using the Euclidean distance as a fitness function validates the reliability of the predictions and highlights the potential of ML models for efficient, scalable and cost-effective design exploration. Therefore, the present work provides a beneficial relationship between CFD simulation and experimental findings on cold flow dynamics and how these might play a leading role in pre-combustion process. Results provide a frame-shifting pathway toward optimization of engine design for the improvement of thermal efficiency, and meeting sustainability targets.

1. Introduction

Internal combustion engines (ICEs) have been central to modern transportation, powering diverse vehicles, including heavy-duty trucks, road vehicles, and off-road machinery. As global demands for efficiency and environmental sustainability grow, optimizing combustion processes in ICEs has become a priority. Heavy vehicles used for transporting goods require robust engines capable of withstanding long-haul journeys and heavy loads. Similarly, road vehicles such as cars and buses demand engines that balance power and fuel efficiency, particularly in urban and suburban environments. Off-road vehicles, including construction and agricultural machinery, face unique operational challenges, requiring specialized engine designs to meet their performance needs. The ongoing development of ICE technology focuses on improving thermal efficiency, reducing emissions, and minimizing environmental impacts. A wide range of strategies has been explored, including low-temperature combustion engines [1], advanced combustion strategies [2], and the use of alternative fuels such as synthesis gas [3]. Researchers have also examined the effects of ultrasound waves on combustion [4], alternative fuels for emission reduction [5,6], and air filter pressure drops on emissions [7]. Dual-fuel combustion approaches in compression-ignition engines [8], reactivity-controlled compression ignition (RCCI) modes in biodiesel/natural gas engines [9], and RCCI combustion strategies [10] have also demonstrated significant potential

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

https://doi.org/10.1016/j.fuel.2025.134590
Received 3 June 2024; Received in revised form 14 January 2025; Accepted 3 February 2025
Available online 6 February 2025

for improving ICE performance. Other innovations include asymmetric fuel injection in hydrogen engines [11], pilot fuel injection optimization in dual-fuel engines [12], and advanced injection strategies for premixed charge compression ignition (PCCI) combustion [13]. Research on optimizing spray angle and combustion chamber geometry has further contributed to improving engine efficiency and reducing emissions [14]. Recent studies have also investigated fundamental processes such as fuel droplet heating and evaporation. Xie [15], for instance, provided critical insights into how these processes influence combustion in ICEs. Alternative fuels like ammonia have gained significant attention due to their potential to mitigate greenhouse gas emissions. Studies on ammonia jets have revealed key mixing characteristics and flammability limits, demonstrating their feasibility for ICE applications [16]. Additionally, performance analyses of ammonia-diesel dual-fuel engines have shown that optimizing ammonia energy ratios enhances thermal efficiency and reduces emissions, offering a promising pathway toward low-carbon transportation [18].

Parallel to advancements in combustion strategies and alternative fuels, machine learning (ML) has emerged as a transformative tool in ICE optimization. ML algorithms excel at handling the complex, multidimensional nature of combustion, emissions, and heat transfer processes. By leveraging large datasets generated through experiments and simulations, ML models can predict engine performance, optimize design parameters, and reduce the need for costly physical testing. Wu et al. [19] developed high-fidelity turbulent combustion surrogate models using deep learning, enabling real-time simulations of combustion processes. Ihme et al. [17] reviewed ML's role in combustion science, emphasizing its potential to uncover actionable insights from large-scale datasets. These advancements highlight the critical role of ML in driving innovations in ICE technology. The application of ML in ICE research has advanced significantly, demonstrating its effectiveness across a variety of domains, including combustion phasing, emissions modelling, heat transfer analysis, and power output prediction. In combustion phasing, artificial neural networks (ANNs) have been used to model nonlinear processes in natural gas spark-ignition engines, accurately predicting parameters such as peak cylinder pressure, ignition lag, and burn duration. These models reduce computational costs while maintaining high predictive accuracy [20]. Similarly, ML algorithms have been employed to predict engine emissions. Comparative studies between random forest (RF) and ANN models have shown that while RF provides quick initial predictions, ANN achieves higher precision, particularly for emissions forecasts [24,36].

In addition to emissions modelling, ML has proven valuable in analysing heat transfer dynamics in ICEs. ANN models have been used to study the heat transfer characteristics of spark-ignition natural gas engines, revealing how combustion chamber geometries and fuel–air ratios influence thermal performance [35]. Furthermore, ML models have been applied to predict engine power output. RF models, for example, have accurately estimated indicated mean effective pressure (IMEP), a key performance metric, with minimal errors. These models offer cost-efficient alternatives for experimental design and engine calibration [36]. The integration of ML with CFD simulations has further enhanced ICE optimization. ML-driven approaches have improved mesh generation, refined turbulence and combustion models, and automated calibration processes, reducing computational costs and human effort [24–31]. Studies by Yang et al. [32] and Zhou et al. [33] have demonstrated ML's ability to predict engine power output and emissions while aiding in combustion modelling and fuel optimization. For example, ML models have been used to calibrate RCCI strategies, improving combustion efficiency and reducing emissions in biodiesel/natural gas engines [10]. The versatility of ML has been further demonstrated in studies aimed at improving operational efficiency and reducing fuel consumption. For instance, Karunamurthy et al. [34] reviewed ML applications in predicting engine performance metrics such as brake power, brake-specific fuel consumption, and brake thermal efficiency under various operating conditions. These studies collectively illustrate the transformative impact of ML on ICE research and its potential to accelerate the development of environmentally friendly engine technologies.

While previous studies have focused on ML-driven enhancements in mesh generation, turbulence modelling and emissions prediction, this study takes a novel approach by leveraging CFD-validated data to train ML models for targeted optimization of in-cylinder flow dynamics. Recent works, such as Liu and Wang [52], have explored ML-assisted modelling of mixing timescales for large-eddy simulations (LES) in turbulent premixed combustion, demonstrating the potential of ML to refine physical models under extreme conditions. Similarly, studies like Liu et al. [53] have showcased ML-assisted predictions of exhaust gas temperature in natural gas engines, providing insights into the relationship between ML algorithms and traditional physical models in ICE research. Building on these advancements, this study integrates ML and CFD simulations to create a reduced-order modelling framework [58–60] tailored to ICE optimization. By combining the predictive power of ML with detailed CFD simulations, this methodology provides a robust framework for uncovering the complex relationships between turbulence kinetic energy (TKE), tumble, and other critical flow parameters. Unlike traditional iterative simulations, which are computationally intensive, the trained ML models allow for rapid exploration of different operating conditions, enabling adaptable initial and boundary conditions (IC/BCs) tailored to specific engine requirements. This integration not only accelerates the optimization process but also provides engineers with insights into hidden layers of data, such as the effects of varying TKE and tumble intensities on air–fuel mixing during the cold flow phase [38]. These advancements demonstrate the potential for ML to enhance both the efficiency and accuracy of ICE development processes, offering scalable and cost-effective solutions for modern engine optimization challenges. This study addresses this gap by introducing an innovative methodology that leverages ML to enhance CFD simulations and optimize key combustion parameters in ICEs. The research critically evaluates multiple ML models, including Random Forest Regression (RFR), Gaussian Process Regression (GPR) and Neural Networks (NN), using extensive metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson Coefficient (PC) and R-squared (R^2) to identify the most effective model for predicting complex in-cylinder phenomena. The selected ML model is then applied to optimize the engine designs by targeting crucial parameters like turbulence kinetic energy (TKE) and Tumble-y, demonstrating the capability to significantly enhance the engine performance while reducing computational costs. Furthermore, the research explores the application of a surface field model to predict flow velocities, achieving a considerable speedup in the design process compared with traditional methods. This approach represents a novel integration of ML and CFD, enabling rapid and accurate predictions of engine behaviour and optimization of design parameters.

These contributions advance the state-of-the-art by providing a unified, efficient framework for ICE optimization. This study not only reduces the reliance on computationally expensive simulations but also introduces a robust ML-driven process for achieving optimal engine configurations. By addressing the dual objectives of enhancing performance and promoting environmental sustainability, this research demonstrates a vital step toward the development of next-generation ICEs.

2. 3D modelling of engine

2.1. Engine configuration

The present study adopted the same engine configuration as that used in a comprehensive experimental study by Schiffmann et al. [21]. The commercial CFD software, STAR CCM+, was employed to conduct the simulations. The geometry consists of a four-valve cylinder (two exhaust valves and two intake valves) with the piston positioned at the top dead centre. To reduce the computational costs, the computational

domain includes only half of the engine with the symmetry boundary condition implemented along the central plane. The use of symmetry conditions is a widely accepted practice in CFD simulations when the geometry and flow characteristics are symmetric, as it allows for significant computational savings without compromising the accuracy of results. For instance, Shateri et al. [4] employed symmetry conditions in a combustion chamber to investigate turbulent flow with chemical reactions, demonstrating the reliability of this approach. Similarly, Klayborworn and Pakdee [45] considered half of the computational domain for a round combustion jet burner, explicitly stating that this simplification saves computational resources while preserving all essential information. Additionally, the study by Blunier et al. [46] applied symmetry assumptions in the analysis of annular combustors with symmetric designs, further validating this approach in symmetric configurations. Fig. 1 presents a three-dimensional (3D) model of the engine. Table 1 provides the engine specifications, including the bore diameter of 92 mm, stroke length of 86 mm, crank radius of 40 mm, connecting rod length of 145 mm, and start angle of 300 degrees.

Table 1
Geometric parameters of the proposed engine.

Engine specification	Value	Unit
Bore	92	mm
Stroke	86	mm
Crank radius	40	mm
Connecting Rod length	145	mm
Start angle	300	degree
Compression ratio	10:1	–

Table 2
Mesh characteristics of the engine model.

Mesh specification	Value	Unit
Mesh min size	0.5	mm
Prism layer total thickness	0.25	mm
Number of prism layers	2	–
Sharp edge angle	5	degree
Mesh type	Trimmed	
Mesh motion	Morph/ remesh/ map solution	

2.2. Mesh generation

Mesh resolution and quality are critical for accurately representing the engine geometry and capturing in-cylinder fluid dynamics. In this study, a fully automated approach in Simcenter STAR-CCM + was used to generate trimmed meshes that adapt to the motion of the piston and valves. Automatic remeshing was employed whenever cell distortion necessitated adjustments, ensuring fidelity in flow dynamics. The baseline mesh specifications are detailed in Table 2, and the mesh configuration from top and side views is shown in Fig. 2. The mesh resolution was set to 0.5 mm, corresponding to the fine resolution described in the *Simcenter STAR-CCM + In-cylinder Solution White Paper* [57]. This resolution was selected based on findings in the white paper, which demonstrated that while coarse meshes (1 mm and 0.75 mm mesh sizes) show mesh dependency for flow dynamics parameters such as TKE and tumble ratio, the fine resolution of 0.5 mm yields mesh-independent results. Additionally, the finer mesh better captures key flow features, including intake jet dynamics and tumble evolution, and provides good agreement with the experimental particle image velocimetry (PIV) data [57]. By adopting this validated mesh resolution, this study ensures high accuracy in capturing the in-cylinder fluid dynamics and turbulence characteristics.

2.3. CFD setting for the engine

The simulation was initiated at a 300-degree crank angle and concluded at a 720-degree crank angle. The intake port inlet serves as the entry point for air while the exhaust port outlet acts as the exit point with both boundaries modelled as pressure outlets. To model the property of air accurately, a multi-component gas mixture, which was composed of 23.31 % oxygen and 76.69 % nitrogen by mass volume rate, was used. To ensure the reliability and validation of the results, the initial conditions within the cylinder and ports were based on the data obtained from the experiment [21]. The initial absolute pressure was set to 1.47 bar, while the initial temperature was 1052.57 K. At a 300-degree crank angle, the intake port pressure value was specified as 0.98 bar, and the temperature was set at 316.77 K. For the exhaust port, the initial absolute pressure was determined as 1.35 bar, and the initial temperature was established as 986.68 K.

To accurately capture the turbulent flow characteristics inside the engine, the Realizable k-ε Two-Layer turbulence model with All y + Wall Treatment under the Reynolds-Averaged Navier-Stokes (RANS) approach was employed for all simulations in this study. This model was selected for its robust performance in handling complex internal flows, such as those encountered in engine intake and exhaust systems, where the flow separation, recirculation and anisotropy are critical phenomena [22,23]. The Realizable k-ε model improves upon the standard k-ε model by satisfying the realizability constraints on Reynolds stresses and providing a more accurate representation of turbulent kinetic energy dissipation, particularly in regions of strong curvature or rotation. This enhancement enables better predictions in flows with significant anisotropy or separation, such as in-cylinder flows in combustion engines [48,49]. The model relies on two transport equations: one for turbulent kinetic energy (k) (i.e., Eq. (1)) and the other for the dissipation rate (ε) (i.e., Eq. (2)) [48]:

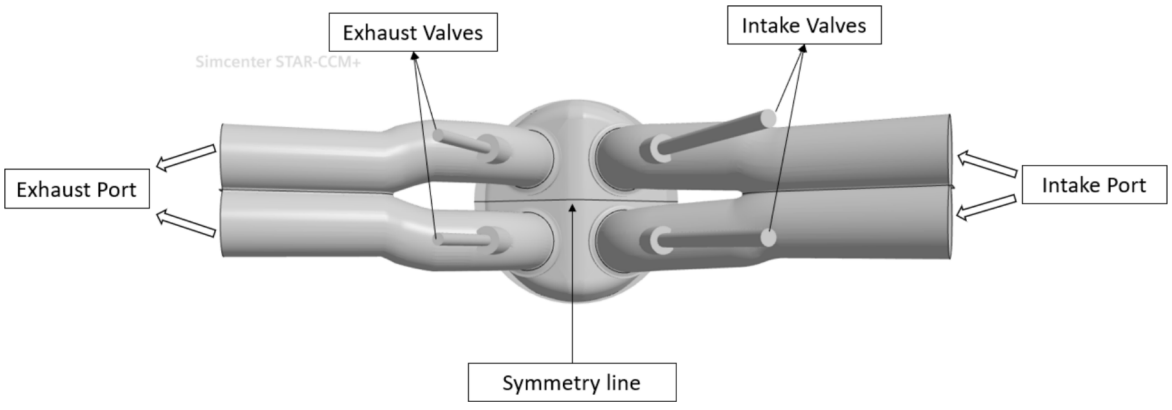


Fig. 1. Three-dimensional (3D) model of the proposed engine.

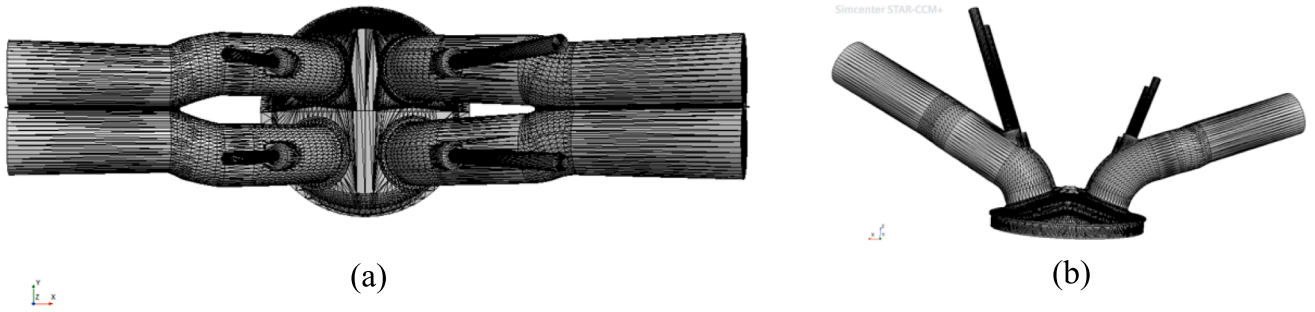


Fig. 2. Mesh visualisation of the engine model: (a) top view and (b) side view.

$$\frac{\partial(\rho k)}{\partial t}+\frac{\partial(\rho k u_i)}{\partial x_i}=\frac{\partial}{\partial x_j}\left[\Gamma_k\frac{\partial k}{\partial x_j}\right]+G_k-\rho\varepsilon \tag{1}$$

$$\frac{\partial(\rho \varepsilon)}{\partial t}+\frac{\partial(\rho \varepsilon u_i)}{\partial x_i}=\frac{\partial}{\partial x_j}\left[\Gamma_\varepsilon\frac{\partial \varepsilon}{\partial x_j}\right]+\rho C_\varepsilon\left(\frac{G_k\varepsilon}{k}+C_{\varepsilon 2}\frac{\varepsilon^2}{k}\right) \tag{2}$$

Here, ρ is the fluid density, and k is the turbulent kinetic energy, and ε denotes the turbulent kinetic energy dissipation rate. u_i and x_i are the velocity components and coordinates along the i_{th} direction. The effective diffusivities for k and ε (Γ_k and Γ_ε), are defined as $\Gamma_k=\frac{\mu_t}{\sigma_k}$ and $\Gamma_\varepsilon=\frac{\mu_t}{\sigma_\varepsilon}$, where μ_t is the turbulent viscosity and σ_k, σ_ε, are the turbulent Prandtl numbers for k and ε, respectively. Additionally, G_k is the production term for k, and C_ε and $C_{\varepsilon 2}$ are model constants. The Realizable k-ε model introduces a variable damping function (f_μ) that adjusts the critical coefficient C_μ dynamically to ensure the model adheres to physical constraints on turbulent stresses. This enhancement improves the model's ability to predict boundary layers, flow separation, and recirculating flows, making it suitable for engine simulations.

The All y^+ Wall Treatment integrates wall functions for high y^+ regions and a two-layer approach for resolving low y^+ regions. This ensures an accurate representation of near-wall turbulence, a critical factor in capturing boundary layer phenomena. The velocity distribution near walls is modelled using the law of the wall:

$$u^+=\frac{1}{\kappa}\ln(E'y^+) \tag{3}$$

Here, u^+ is non-dimensional velocity, y^+ is non-dimensional wall distance, κ is Von Kármán constant, and E' is defined as $E'=\frac{E}{f}$, with E is the log law offset, and f is the roughness function [48–50]. This model is part of the in-cylinder module provided by Star-CCM + and is identified as a high-accuracy option under specific conditions, including the mesh resolution and flow parameters used in this study. According to Star-CCM + documentation [57], the Realizable k-ε model with All y + Wall Treatment is particularly well-suited for in-cylinder simulations, where high accuracy is required to capture the transient flow structures, turbulent mixing and boundary layer effects.

3. ML framework for combustion process

In recent years, there has been a growing interest in CFD and heat

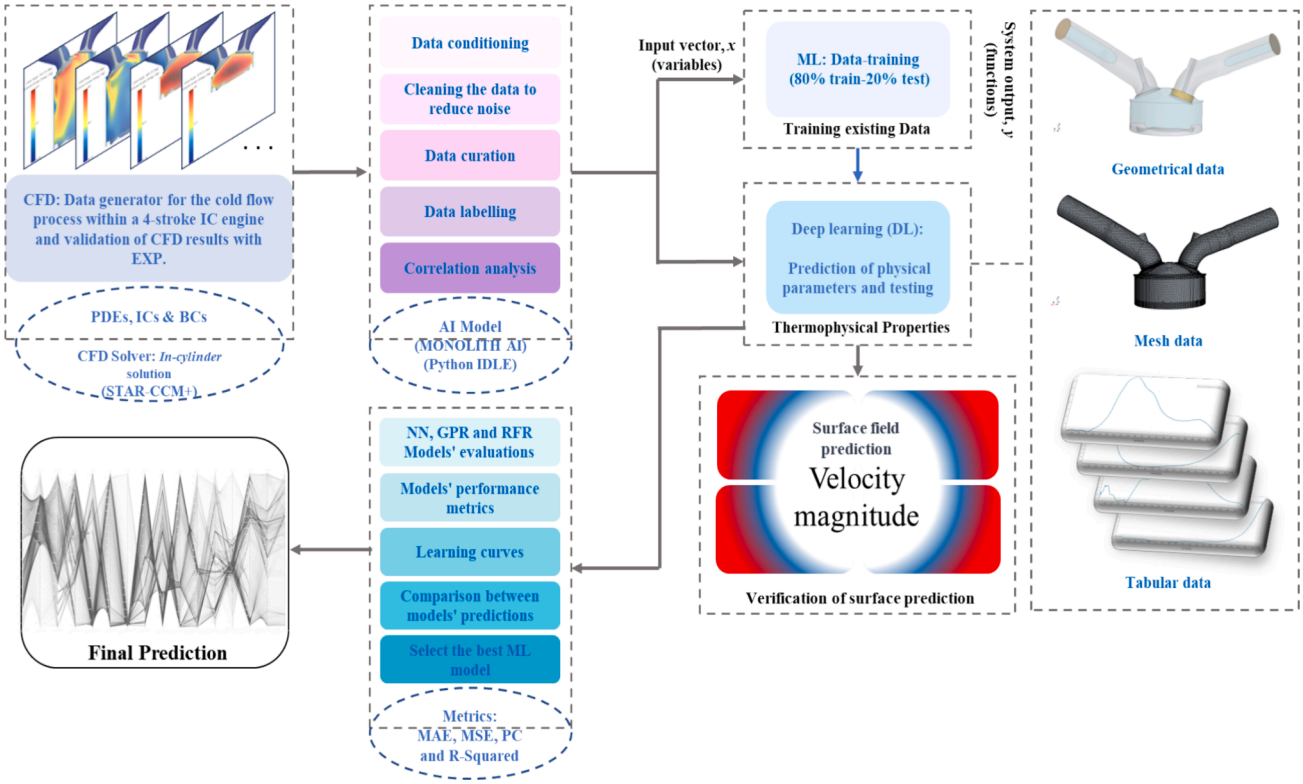


Fig. 3. Flowchart of CFD simulation, data extraction, model training, and evaluation.

transfer communities using ML models to expedite their simulations. One platform that has gained prominence in this regard is Monolith AI [37]. A methodology incorporating ML techniques has been employed to enhance the precision of CFD simulations, particularly in the context of ICE simulations (see Fig. 3). The flowchart in Fig. 3 illustrates the comprehensive process starting with CFD simulation, data extraction, and the train-test split procedure. Following this, the flow moves on to DL model training, surface model predictions, and the application of three distinct ML models: Random Forest Regression (RFR), Neural Network (NN), and Gaussian Process Regression (GPR). Subsequently, the evaluation of these models is conducted using various metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Prediction Correlation (PC), and R-squared (R^2). Finally, the flow culminates in the generation of the final prediction based on the preceding stages. During the training process, the ML models learn the underlying correlations between the input variables, such as initial conditions, boundary conditions, and the output variables, including pressure, swirl, TKE, and tumble-y. The trained ML models are subsequently validated using a separate test set to assess its generalization ability and accuracy in predicting engine performance.

The integration of ML with CFD is a promising approach for accelerating engine optimization and extracting hidden insights from complex flow simulations. However, the implementation of this hybrid methodology presents several challenges that must be addressed to ensure its effectiveness. Generating high-quality data for training ML models requires careful curation of CFD simulation results. This includes ensuring that the simulation data is complete, consistent, and representative of the full parameter space. In the present study, significant efforts were devoted to maintaining consistency across crank angles, operating conditions and geometrical configurations. Additionally, data conditioning was performed to remove the outliers and correct inconsistencies arising from numerical instabilities in CFD simulations. Labelling CFD-generated data with appropriate metadata, such as crank angles, TKE, tumble-y, and geometrical parameters, is essential for an effective ML training. This process was particularly challenging because the multidimensional nature of CFD data required precise segmentation and annotation. Automated scripts were developed to extract and label key flow parameters across multiple simulation cases. CFD simulations inherently contain noise due to discretization errors, numerical approximations and convergence tolerances. This noise can degrade the performance of ML models if not properly addressed. In this study, noise-reduction techniques, including statistical filtering and smoothing algorithms, were applied to minimize the noise while preserving important physical trends. Representing the geometrical features of the in-cylinder flow domain in a format suitable for ML was another significant challenge. CFD simulation data was converted into the VTK (Visualization Toolkit) format, which allowed for the spatial representation of velocity fields, pressure distributions, and turbulence parameters. This process required considerable preprocessing, including meshing consistency checks and data interpolation, to ensure compatibility with ML workflows.

3.1. Data structure

In this study, a CFD simulation was conducted starting at a 300-degree crank angle and running for 420 degrees, covering the complete engine cycle. For each crank angle, a dataset was extracted, resulting in 420 datasets. Each dataset consisted of a varying number of rows, ranging from 9500 to 10000, and 18 columns. The variation in the number of rows was due to the movement of the piston from top dead centre to bottom dead centre, which created more space in the cylinder and generated additional geometrical and operational data. Aggregating the data across all 420 crank angles resulted in a total of approximately 4 million rows of data. Following best practices in machine learning (ML) [17,35], 80 % of this data (approximately 3.2 million rows) was allocated for training the ML models, while 20 % (approximately 800k rows) was reserved for testing. This extensive dataset, collected over a wide range of crank angles, enabled the ML models to capture intricate relationships between the input parameters (e.g., cylinder temperature, pressure, and mass flow rates) and the output parameters (e.g., cylinder pressure, TKE, and tumble-y). Furthermore, retraining experiments using subsets of the training data confirmed that this allocation ensured a balance between the computational efficiency and model prediction accuracy.

In this study, a carefully curated set of parameters were selected to train the machine learning (ML) model for predicting the turbulence kinetic energy (TKE) and tumble-y, which are pivotal for understanding the in-cylinder air motion dynamics during the cold flow phase of an internal combustion engine. These parameters include Cylinder Temperature (CT), Cylinder Tumble-x (CT-X), Cylinder Mass (CM), Exhaust Port Pressure (EPP), Exhaust Port Temperature (EPT), Exhaust Port Mass Flow Rate (EPMFR), Intake Port Mass Flow Rate (IPMFR), Intake Port Pressure (IPP), and Intake Port Temperature (IPT). Each parameter was chosen for its direct or indirect influence on TKE or tumble-y, ensuring the ML model captures the intricate physical relationships governing these phenomena. CT and CM are fundamental thermodynamic variables that influence the turbulence generation within the cylinder. CT impacts air density and velocity gradients, which are primary contributors to turbulence formation, while CM governs bulk flow characteristics, affecting the momentum that drives swirl and tumble motion. The inclusion of CT-x reflects the rotational motion of in-cylinder air, which is critical for evaluating the intensity and distribution of the rotational flow. Tumble motion plays a crucial role in enhancing the air–fuel mixing and turbulence, particularly during the intake and compression phases. The exhaust and intake port parameters provide critical boundary conditions that shape in-cylinder flow behaviour. For example, IPMFR determines the volume of fresh air entering the cylinder, which directly influences the initial turbulence levels. IPP and IPT set thermodynamic conditions at the intake, affecting the shear layer dynamics that contribute to both TKE and tumble formation. Similarly, EPMFR, EPP and EPT characterize the expulsion of exhaust gases, which indirectly influence the dissipation or maintenance of tumble motion. By integrating these parameters, the ML model can establish a comprehensive understanding of the thermodynamic state, mass distribution and flow dynamics during the engine cycle. This systematic selection enables the ML model to predict TKE and tumble-y with high accuracy, uncovering the underlying physical relationships and providing valuable insights for optimizing in-cylinder flow processes under varying engine operating conditions.

Furthermore, the ML model aims to predict several critical output parameters essential for comprehending combustion process, namely Cylinder Pressure (CP), Cylinder Swirl (CS), Cylinder tumble-y (CT-Y), and TKE. To comprehensively address this, a sensitivity analysis employing the Sobol method with the first-order variable combinations was carried out. This sophisticated analytical approach facilitates the exploration of both direct effects and interactions among the parameters concerning the model outputs [40–42]. The outcomes of this analysis provide a visible representation of the impact of each input parameter on the corresponding model outputs. As shown in Fig. 4, the correlation between the inputs and outputs is depicted, revealing notable findings. Specifically, the maximum correlation coefficients are observed for temperature as an input on TKE, registering at 0.3148; while for CS, it stands at 0.3195, and for CP, it is 0.2070. Additionally, the correlation coefficient for CM as an input on CT-Y is reported at 0.2019. This analysis provides valuable insights into the intricate connections between the input parameters and the model outputs.

Fig. 5 presents the learning curves, which illustrate the relationship between the MSE and the proportion of used data for training. The curve reveals that as the proportion of training data increases, the accuracy of the ML model is improved. Specifically, the model trained by 80% of the available data exhibits twice the accuracy of the model trained by only 20%. This observation suggests that a larger volume of training data

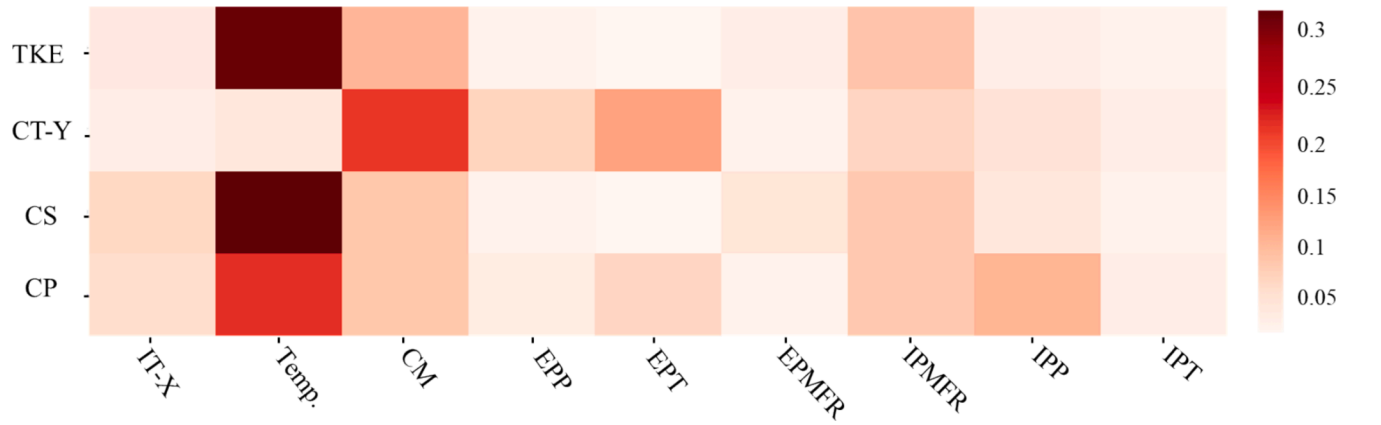


Fig. 4. Correlation coefficient heat map of selected parameters with Cylinder Pressure (CP), Cylinder Swirl (CS), Cylinder Tumble-y (CT-Y) and TKE in IC engine.

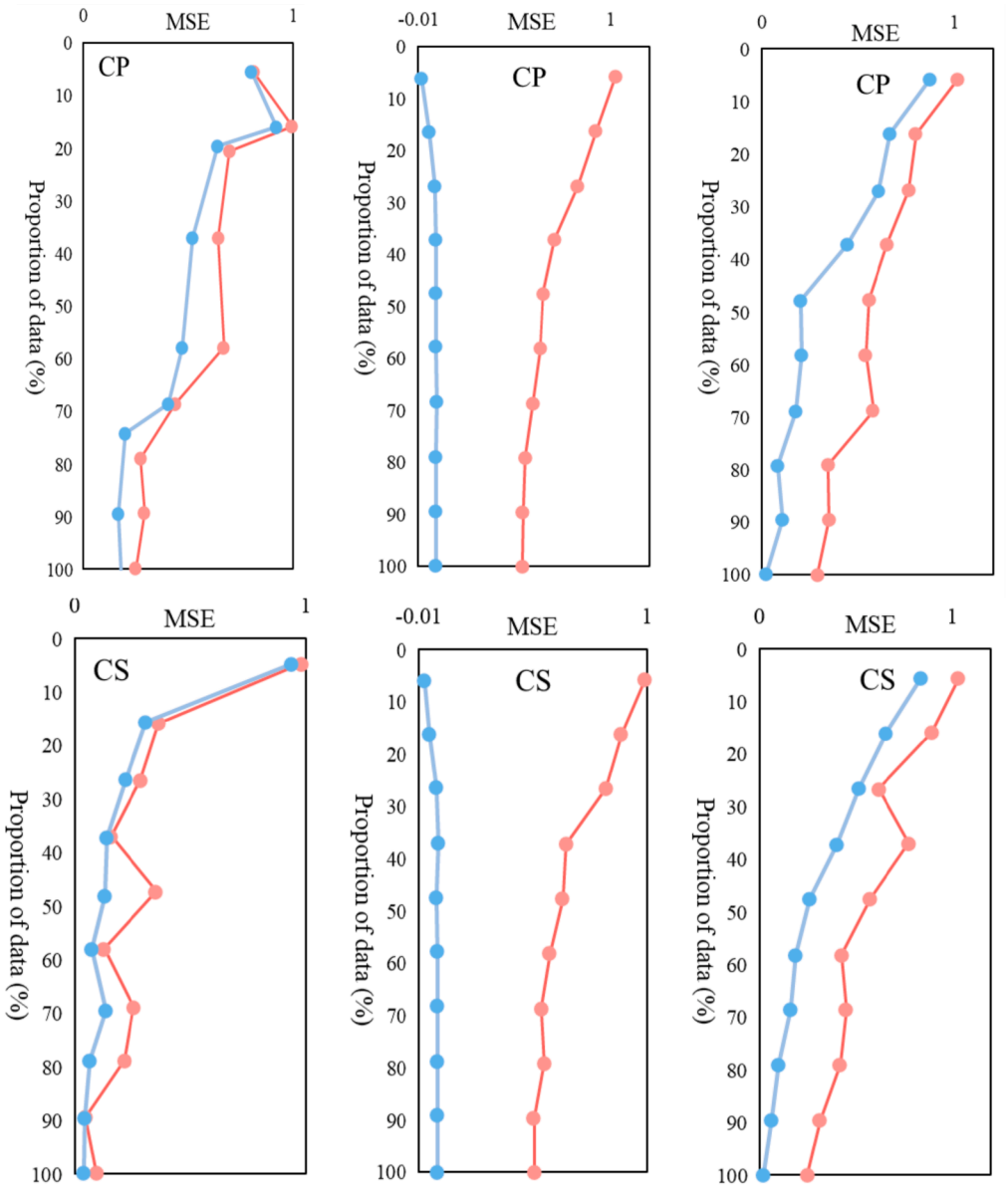


Fig. 5. Train split learning curves of RFR, NN and GPR models for CP, CS, CT-Y and TKE.

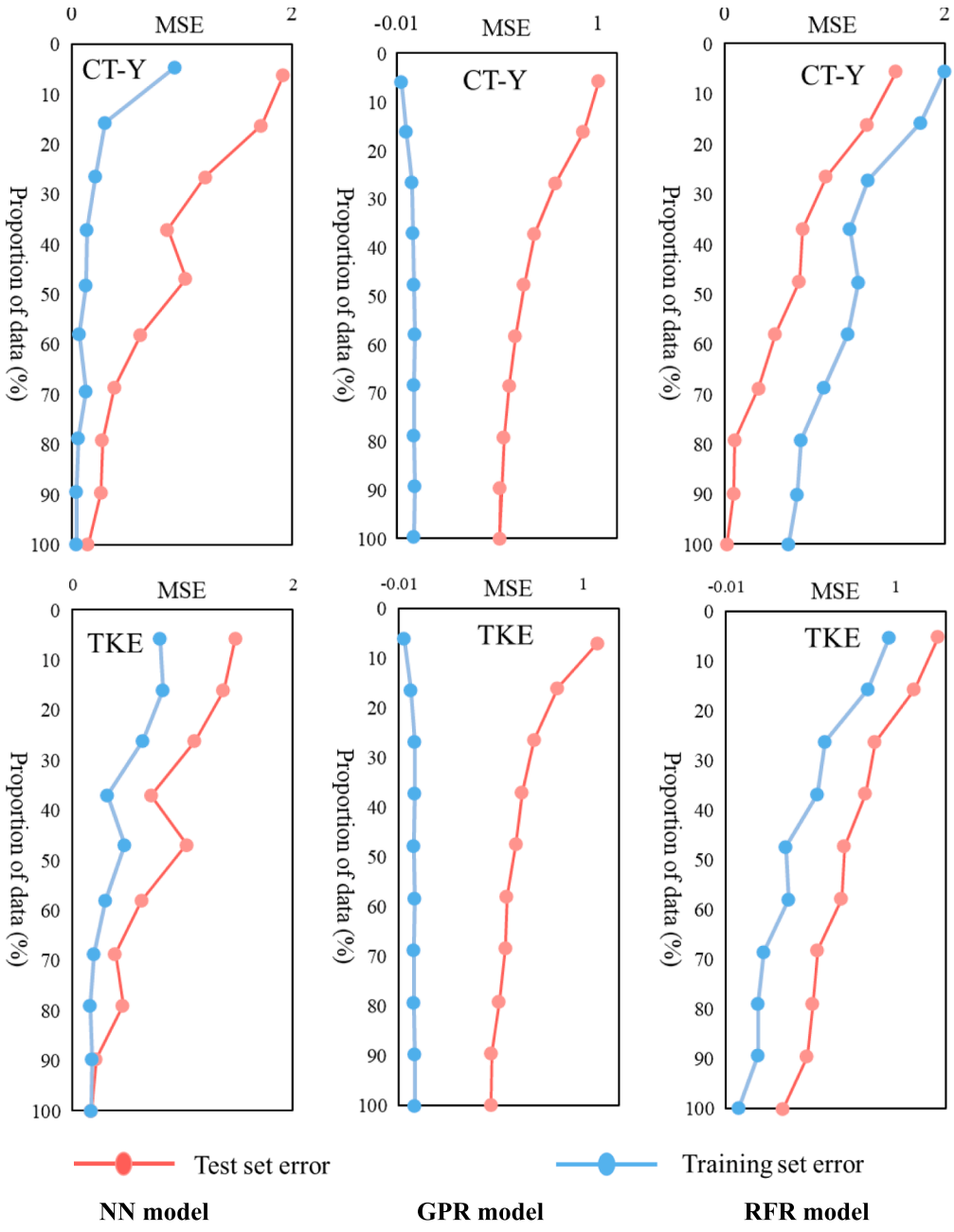


Fig. 5. (*continued*).

facilitates the development of a more accurate and reliable ML model. 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 [17,39].

3.2. ML models

The Monolith AI platform provides a suite of ML models, including NN, RFR, GPR, Support Vector Machines (SVM), Polynomial Regressions (PR), and Decision Trees (DT). The selection of RFR, GPR and NN for this study was guided by their adaptability to the specific challenges of modelling cold flow phenomena in combustion engines. These phenomena involve highly nonlinear interactions between turbulence kinetic energy (TKE), tumble, pressure, and temperature. The chosen models were selected based on their ability to capture the complex relationships, scale effectively with large datasets and quantify uncertainty in predictions. RFR was chosen for its ensemble-based approach, which reduces overfitting and provides feature importance metrics, offering insights into the influence of various parameters. Its ability to handle the high-dimensional data and resilience to noise make it an effective choice for regression tasks involving intricate patterns [17,39]. GPR was selected for its probabilistic framework, which models uncertainties in predictions and provides confidence intervals, a crucial feature in engineering applications. GPR is particularly suited for moderate-sized datasets, where its ability to provide both predictions and uncertainty bounds adds significant value. Compared with Gradient Boosting (GB), which is not available in the Monolith AI platform, GPR offers a more accessible and interpretable framework for this study. NN models were included for their ability to learn complex, high-

dimensional patterns from large datasets. They are particularly effective in analysing relationships between TKE, tumble and other cold flow parameters, enabling deeper insights into nonlinear interactions. While GB models are well-regarded for their performance in regression tasks, their exclusion in this study is due to their unavailability within Monolith AI and their sensitivity to hyperparameter tuning, which adds complexity to their implementation [17,39]. While SVM was considered, its use was limited in this study due to its computational requirements, particularly for kernel-based methods in regression tasks. As noted in [52], SVM had the smallest error, but it required the highest computational resources. This trade-off made SVM less practical for large-scale, nonlinear problems, despite its strong performance in smaller datasets.

3.3. Evaluation of ML models

In the realm of predictive modelling, the assessment of model generalization and the mitigation of overfitting should be considered critically. Cross-validation, which is a statistical technique in Monte Carlo (MC) method family, plays a pivotal role in achieving these objectives. By dividing the data into distinct sections for model training and validation, cross-validation enables a comprehensive evaluation of a model's performance, thus enhancing its generalization capacity and reducing the risk of overfitting [43]. One of the fundamental types of cross-validation is the *k*-fold cross-validation, which has been widely adopted in engineering. This technique involves dividing the dataset into *k* segments or folds, followed by *k* iterations of training and validation. During each iteration, a different fold is held out for validation while the remaining (*k*-1) folds are used for model training. The use of *k*-fold cross-validation allows for the assessment of a model's performance across various subsets of the data, contributing to a more robust evaluation process. The selection of the number of folds, *k*, is a critical consideration, with the typical range being from two to ten, to ensure robust model assessment and guard against overfitting [44]. For this study, a 5-fold approach was employed to achieve these objectives. In the evaluation of predictive models, several key performance metrics come into play. R^2, for instance, measures the proportion of the variance in a dependent variable that can be explained by the independent variables. A higher R^2 value signifies a more accurate model, indicating that a significant portion of the variation in the dependent variable is captured by the model. MAE and MSE are used to quantify the average differences between the predicted and actual values, with lower values indicating higher model accuracy. In addition, the PC is employed to assess the linear correlation between two variables, providing insights into the strength and direction of the relationship [41]. This approach contributes to the advancement of predictive modelling techniques and the development of more reliable and accurate models for various applications. To evaluate the models, various metrics are employed and presented in Table 3. 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 CP, GPR achieves an MAE of 0.0556, MSE of 0.06923, PC of 1, and R^2 of 1, while RFR and NN exhibit slightly higher values for these metrics. Similar trends are observed for CS, CT-Y and TKE. Fig. 6 illustrates a Box and Whisker plot that compares the performance metrics of the models. This visible representation offers insights into the median, minimum and maximum data points, as well as the 1st and 2nd quartiles for four key metrics: R^2 and PC. The findings presented in Fig. 6 unequivocally indicate that the GPR model demonstrates remarkably precise predictions, surpassing both the RFR and NN models across all the assessed metrics. The lower median values, and narrower interquartile ranges for the GPR model signify its superior performance and resilience in accurately forecasting the engine performance values. The primary reason for GPR's outperformance lies in its probabilistic nature, which allows it to provide not only point predictions but also uncertainty estimates. This capability is particularly advantageous for modelling complex in-cylinder flow phenomena characterized by nonlinear and multidimensional interactions. Unlike RFR, which may produce inconsistent predictions in regions of sparse data, and NN, which requires extensive tuning and larger datasets for optimal performance, GPR excels in adapting to sparse data regions through its kernel-based approach [54–56]. The choice of kernel, such as the radial basis function (RBF) kernel used in this study, allows GPR to capture smooth and nonlinear variations in TKE, tumble-y, and other key parameters. This makes GPR well-suited for capturing spatial and temporal dependencies inherent in the CFD-generated data. Moreover, GPR's ability to quantify predictive uncertainty provides engineers with actionable insights for making informed decisions during optimization.

Table 3
Comparison of performance metrics for the cylinder pressure, swirl, tumble-y and TKE between NN, RFR and GPR models.

Output	ML model	MAE	MSE	PC	R^2
CP	GPR	0.0556	0.0923	1	1
CP	NN	1.2633	3.3478	0.99965	0.99909
CP	RFR	0.6658	1.0726	1	1
CS	GPR	0	0	1	1
CS	NN	0.00059	5.00E-05	0.99981	0.99911
CS	RFR	0.00012	0	1	1
CT-Y	GPR	0	0	1	1
CT-Y	NN	0.00728	9.00E-05	0.99978	0.99949
CT-Y	RFR	0.00024	0	1	1
TKE	GPR	0.00039	0	1	1
TKE	NN	2.07397	3.34937	0.99988	0.99951
TKE	RFR	0.04213	0.00728	1	1

4. Results and Discussion

An essential aspect of ensuring the reliability of the simulation results involves comparing them with the data derived from a previous experimental study [21], which is considered as a benchmark case. Detailed comparison has been presented in Sec. 4.1 to establish the credibility and validity of the simulation approach. Sec. 4.2 presents the in-depth predictions generated in ML models concerning a variety of engine parameters. The prediction for the parameters includes the pressure, swirl, TKE, and tumble-y.

4.1. Validation of CFD results against experimental data

The present CFD results are validated by comparing with the experimental findings of Transparent Combustion Chamber-III (TCC-III) Optical Engine obtained at the University of Michigan [21]. The engine features a spark-ignition, 2-valve, 4-stroke design with a pancake-shaped combustion chamber, a geometrical compression ratio of 10:1, and bore × stroke dimensions of 92 × 86 mm. Extensive optical access is provided by a full quartz cylinder and a 70 mm-diameter flat quartz piston window. High-frequency piezo-resistive absolute-pressure transducers were installed in the intake and exhaust systems to record intra-cycle pressure at 0.5 CAD intervals. PIV measurements were conducted in four planes using a high-speed camera and a high-repetition-rate dual-cavity frequency doubled Nd:YLF laser. This comprehensive experimental setup yielded detailed data on pressure, temperature, and velocity fields, which served as a benchmark for the CFD simulations in this study. As discussed in Sec. 2, the engine configuration of the simulation is the same as in the experiment. Nevertheless, to minimize the computational costs, the simulation focuses on the crank angle ranging from 300 to 720 degrees, whereas the experimental study encompasses the full crank angle range of 0–720 degrees. In the context of combustion process, the crank angle of 0 degrees typically corresponds to the top dead centre of the piston at the beginning of the intake stroke, and the 180-degree mark represents the bottom dead centre at the end of the exhaust stroke. Therefore, the simulation focusing from 300 to 720 degrees captures the latter part of the power stroke and the entire exhaust stroke, aligning with the relevant operational characteristics

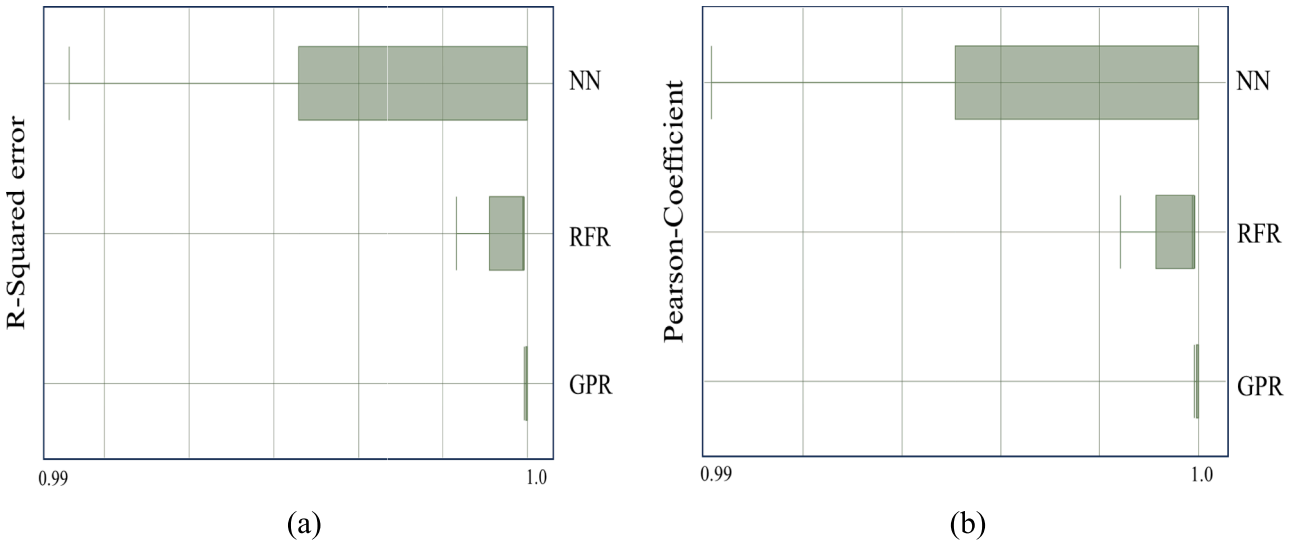


Fig. 6. Comparison of the performance metrics between NN, RFR and GPR models: (a) R-squared (R^2) and (b) Pearson Coefficient (PC).

while optimizing computational resources.

Fig. 7 illustrates the comparison of key in-cylinder flow metrics derived from CFD simulations and experimental data for the TCC-III engine. Fig. 7(a) highlights the intake and exhaust valve lift profiles as a function of crank angle. The results demonstrate a strong agreement between the CFD and experimental valve lifts, as evidenced by the overlapping curves for both intake and exhaust valves. This alignment verifies the accurate representation of valve dynamics in the simulation, which is essential for capturing the air-flow characteristics during the intake and exhaust phases. Fig. 7(b) presents the in-cylinder trapped mass variation with crank angle. The trapped mass, critical for determining the mass of air available for combustion, shows excellent agreement between CFD and experimental results. Both profiles closely follow the same trend, with minimal deviation observed across the crank angles. Fig. 8 (a-b) show the comparison between the predicted cylinder pressure/temperature at different crank angles and the corresponding experimental data. It can be seen from Figs. 7 and 8 that an excellent agreement between the CFD results and the experimental data is reached. This provides compelling evidence of the robust accuracy achieved in the present CFD simulations, demonstrating its capability to accurately capture the key characteristics of the combustion process.

The statistical comparison between the CFD results and experimental data for 10 random crank angles is depicted in Fig. 9, comprising: (a) Temperature, (b) Pressure, and (c) Trapped Mass. The analysis reveals a generally good alignment between the two datasets across all three parameters, supported by the respective RMSE values. For temperature in Fig. 9(a), the Bland-Altman plot indicates a mean difference of 1.48 K, suggesting a minor bias in the CFD predictions compared against the experimental data. The limits of agreement, spanning from −2.41 K to 5.37 K, demonstrate a satisfactory level of precision, particularly considering the RMSE value of 1.29 K, which falls within acceptable ranges for thermal simulations. For pressure in Fig. 9(b), the agreement is exceptional, with a mean difference of −0.02 bar, reflecting negligible bias. The narrow limits of agreement (−0.05 bar to 0.02 bar) and the corresponding RMSE value of 0.11 bar confirm that the CFD predictions closely align with the experimental data, capturing the pressure trends with high fidelity. For trapped mass in Fig. 9(c), the mean difference of 2.87 mg highlights a minor positive bias, where the CFD model tends to slightly overestimate the trapped mass relative to experimental measurements. However, the limits of agreement (0.38 mg to 5.37 mg) and the RMSE of 1.92 mg indicate that the CFD predictions remain within acceptable engineering tolerances.

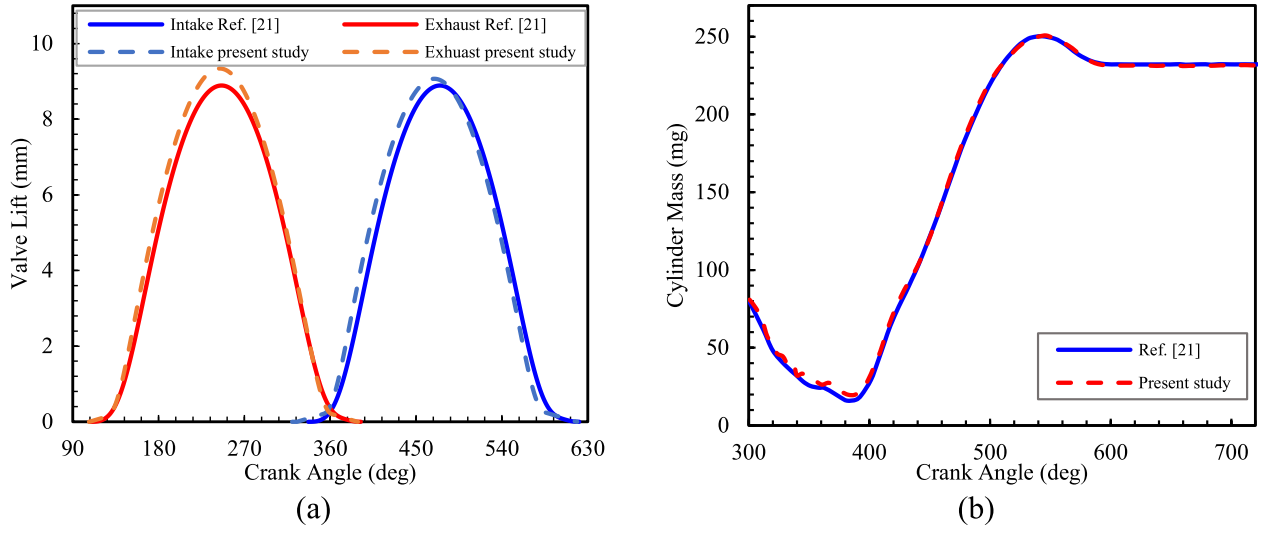


Fig. 7. Comparison of (a) intake/exhaust valve lift and (b) in-cylinder trapped mass between CFD results and experimental data.

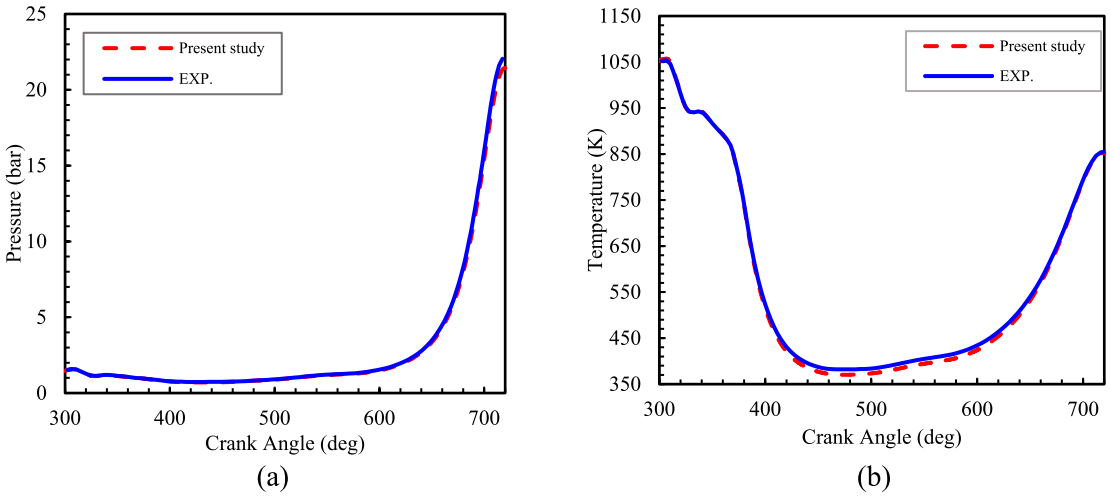


Fig. 8. Comparison of (a) cylinder pressure and (b) cylinder temperature between CFD results and experimental data.

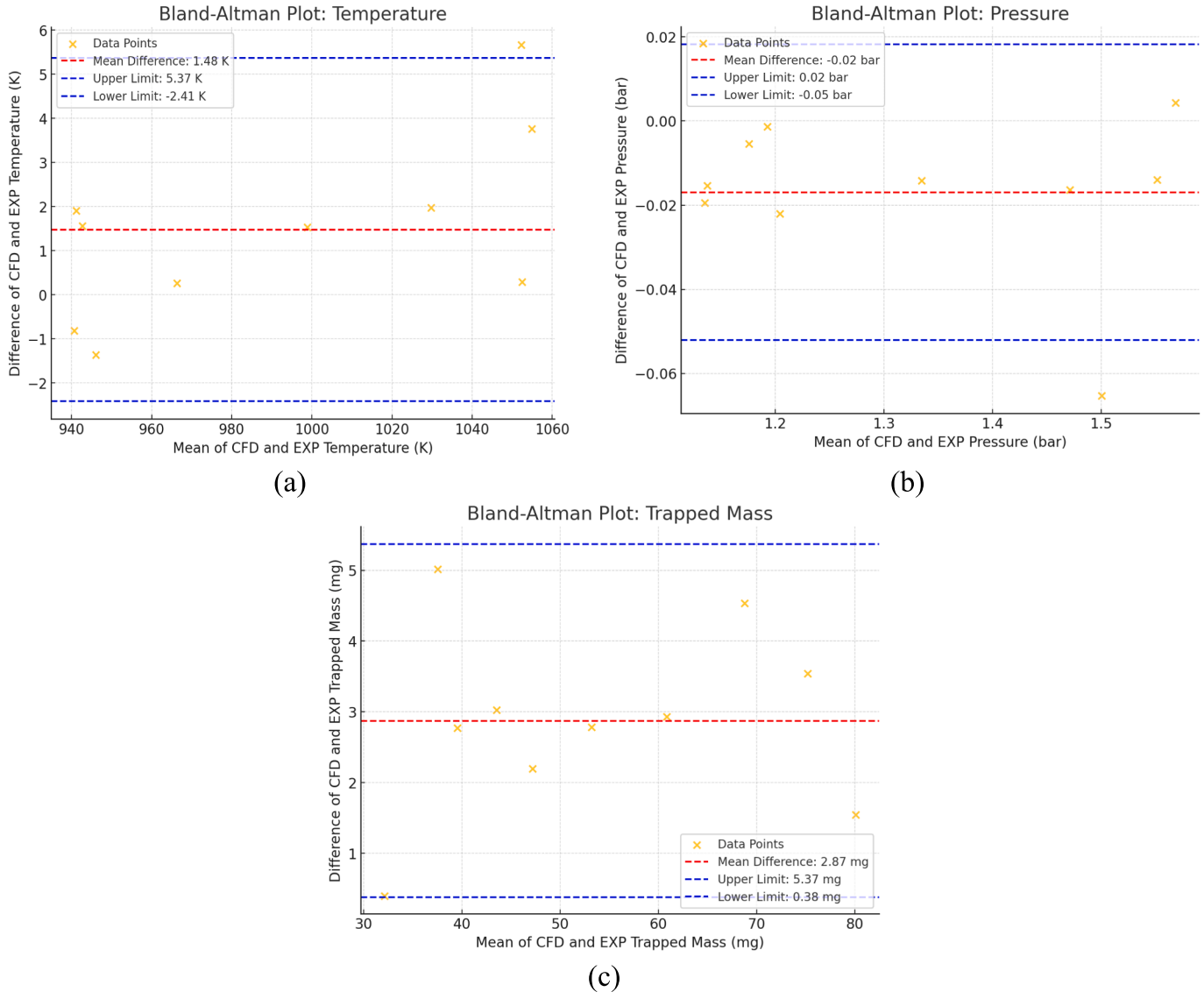


Fig. 9. Bland-Altman analysis of CFD and experimental data for random crank angles: (a) temperature; (b) pressure; and (c) in-cylinder trapped mass.

4.2. In-Cylinder visualization

Fig. 10 shows the velocity magnitude of a cold flow at different crank angles. As mentioned in Sec. 2, this study commences the simulation from a crank angle of 300 degrees, progressing toward the cylinder's top dead centre position. Observations reveal that the exhaust valve remains open during the crank angle range of 300–360 degrees, followed by the opening of the intake valve, which is subsequently closed after reaching a crank angle of 600 degrees. This valve motion is also reflected in the absolute pressure contour, where the cylinder chamber experiences the highest pressure when the valves are closed (See Fig. 11). Absolute pressure provides a comprehensive representation of the overall

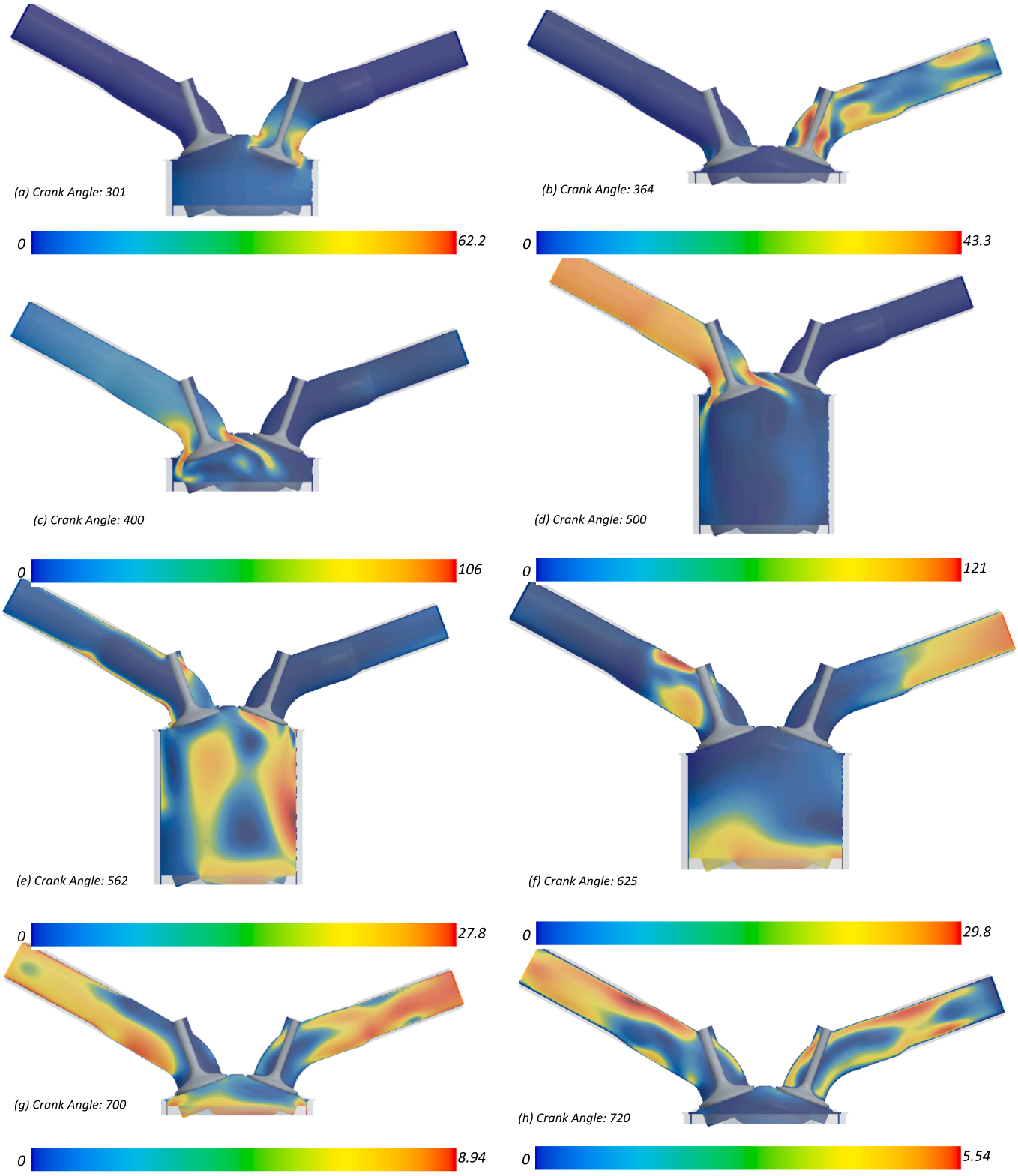


Fig. 10. Velocity magnitude (m/s) of a cold flow at different crank angles: (a) 301 degrees; (b) 364 degrees; (c) 400 degrees; (d) 500 degrees; (e) 562 degrees; (f) 625 degrees; (g) 700 degrees; and (h) 720 degrees.

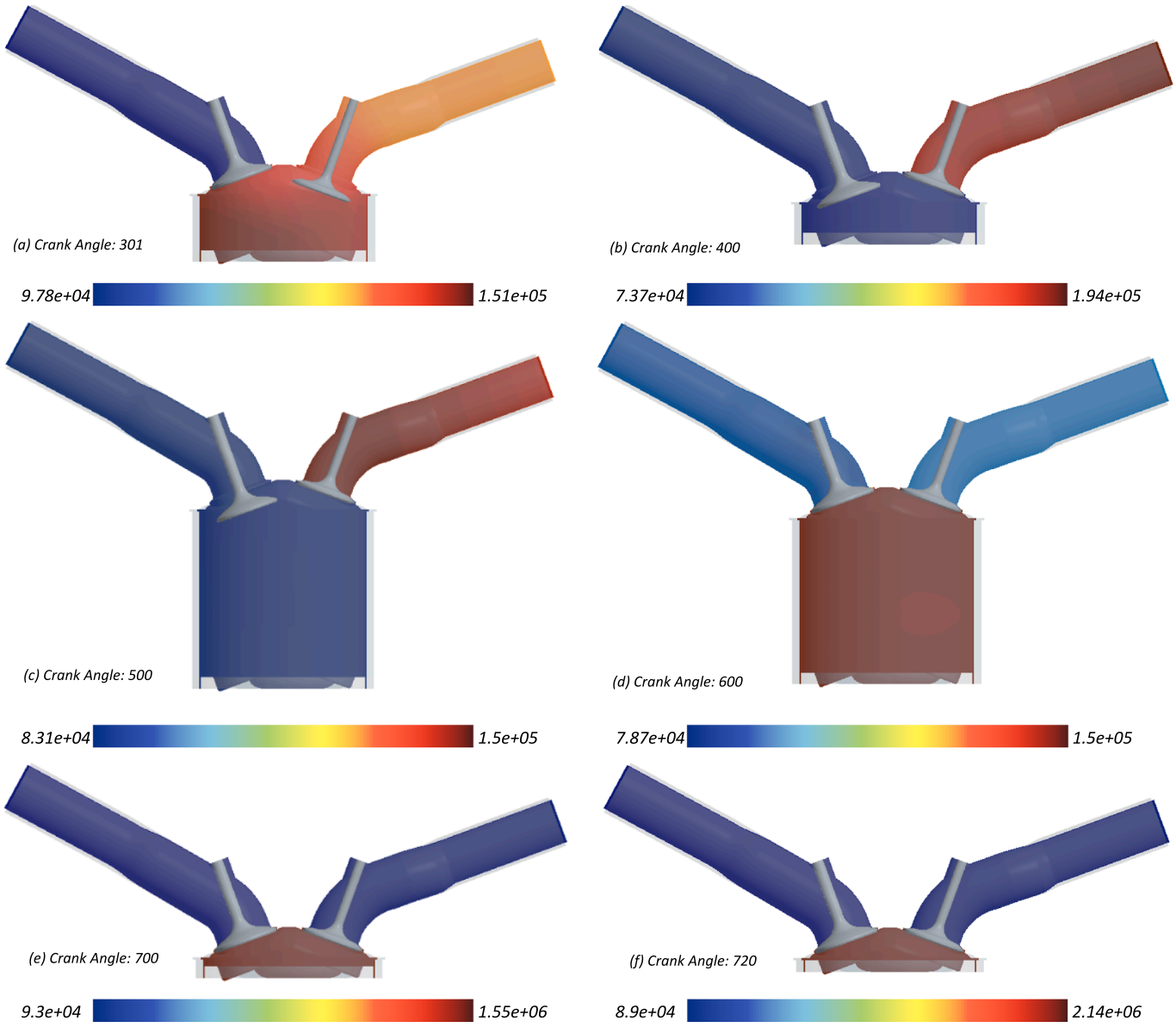


Fig. 11. Absolute pressure (Pa) of a cold flow at different crank angles: (a) 301 degrees; (b) 400 degrees; (c) 500 degrees; (d) 600 degrees; (e) 700 degrees; and (f) 720 degrees.

pressure inside the cylinder, incorporating both effects of the combustion process and the valve dynamics. In addition, absolute pressure allows a direct comparison of pressures at different levels across different crank angles, enabling a clear understanding of the variations in pressure distribution throughout the engine cycle.

4.3. Prediction of ML models

After training the ML models, assessing their quality and accuracy using the test dataset is essential. This evaluation process is crucial in determining the reliability and performance of the model. Figs. 12(a-d) present the data of pressure, swirl, TKE and tumble-y predicted with ML models, specifically focusing on the instances where the predictions align perfectly with the actual values. Upon the examination of these findings in Fig. 12, it becomes evident that all the models perform exceptionally well. However, the GPR and RFR show better alignment compared with the NN model. The predictions from GPR closely align with the actual values across the entire crank angle range, demonstrating its ability to capture the complex relationships and quantify uncertainties. RFR, while effective, shows slightly lower accuracy in regions with rapidly changing phenomena, such as TKE during intake and compression phases. NN predictions, though promising, exhibit higher uncertainties in regions requiring more extensive data or optimized architectures.

The comparison of NN, RFR, and GPR predictions, as shown in Figs. 13(a-d), highlights the varying performance of these models across different crank angle ranges for predicting pressure, swirl, TKE, and tumble-y. The GPR model demonstrates superior accuracy across the entire crank angle range, particularly in regions with high gradients (e.g., Fig. 13(a) at 620–700 crank angles for pressure and Fig. 13(c) for TKE peaks at 500–550 crank angles). This superior performance is attributed to the ability of GPR to capture non-linear trends with higher fidelity due to its Bayesian framework, which inherently models uncertainty and adapts well to sparse or complex data distributions. The RFR model, with its robust averaging mechanism, performs well in steady-state regions. For instance, it accurately captures the plateau observed in Fig. 13 (c) for TKE between 350–400 crank angles. However, its discrete, tree-based approach shows limitations in regions with rapid transitions,

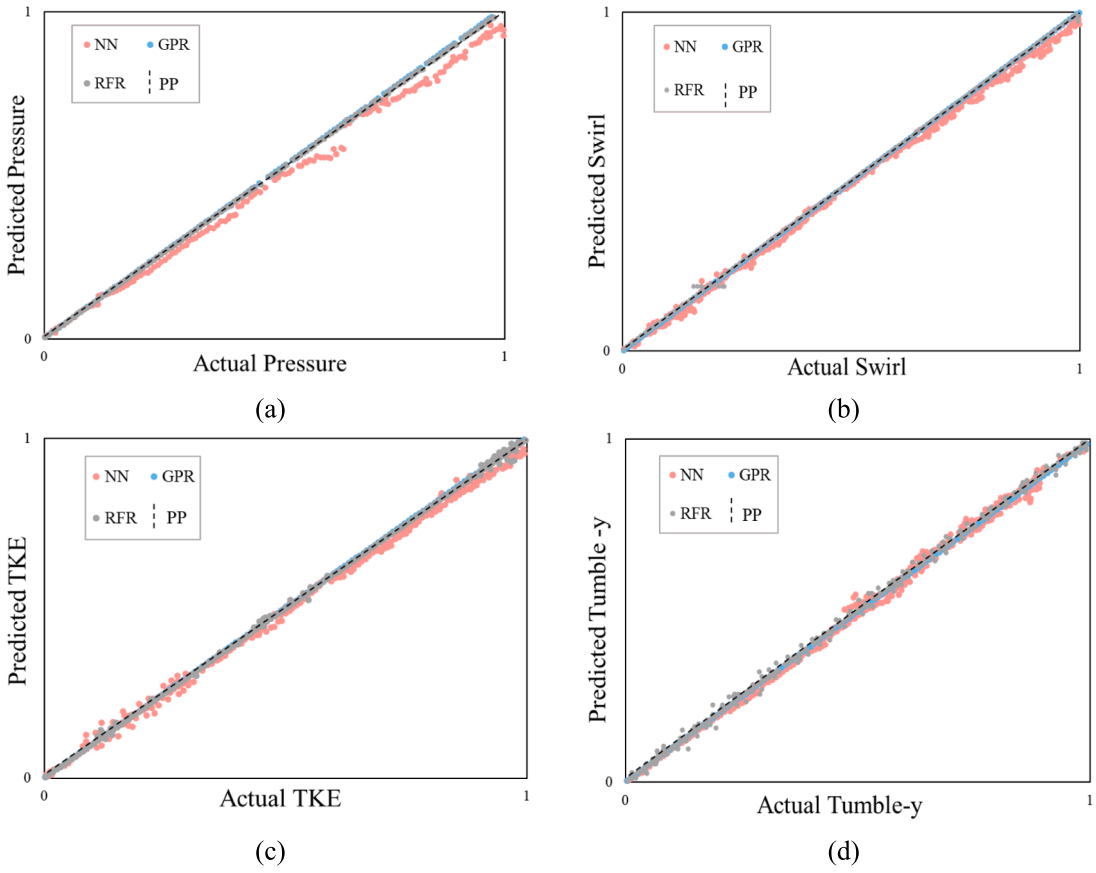


Fig. 12. Comparison of predicted and actual values by RFR, GPR and NN models for: (a) CP; (b) CS; (c) TKE; and (d) CT-Y.

such as the TKE peak at 500–550 crank angles, where it exhibits slightly higher errors compared with GPR. Similarly, in Fig. 13(d), RFR performs reliably for tumble-y predictions in stable regions but shows minor deviations in high-gradient regions. The NN model, while performing reasonably well overall, demonstrates higher uncertainties, particularly in regions with rapid transitions or high variability, as evidenced in Figs. 13(a-d). For example, Fig. 13(b) shows the swirl predictions, where the NN model struggles to maintain consistency in regions with dynamic changes in flow characteristics. Fig. 13(d) for tumble-y further emphasizes this trend, where the wide uncertainty bands indicate challenges in capturing the complex non-linear relationships between input parameters. This behaviour can be attributed to the sensitivity of NN to hyperparameter tuning and the challenges of training on highly non-linear phenomena. The wide uncertainty bands in the NN model predictions suggest potential underfitting in certain regions or limitations in learning the intricate interactions of the dataset.

These differences underline the importance of selecting appropriate ML models for different crank angle ranges and phenomena. For example, GPR is well-suited for scenarios requiring high precision and detailed uncertainty quantification, while RFR provides robust predictions in more stable regions. NN models, although requiring further optimization, offer a versatile framework capable of handling a variety of input–output relationships with appropriate tuning. The predictions generated by the NN model demonstrate promising results having uncertainty ranges observed between 31.27 and 68.20 J/kg for TKE. The predicted value for TKE in the model is 47.06 J/kg. Likewise, the model accurately predicted tumble-y values within an uncertainty range of −0.111 to 0.012 along with a model prediction of −0.048. The predicted pressure was about 2 MPa, and the swirl value was 0.215. While the NN model predictions align closely with the measured data, it is essential to acknowledge some minor differences within the uncertainty range. Several recommendations can be considered to optimize the model and reduce the uncertainty range:

1) Increasing the size of the training dataset may enhance the model's capability of capturing a broader range of scenarios and improving its generalization stability.
2) Fine-tuning the model's hyperparameters, such as adjusting the learning rate, number of layers, or activation functions, can help optimize its performance.
3) Exploring different NN architectures or considering ensemble methods, such as combining multiple models, could enhance the prediction accuracy and reduce the uncertainty.
4) Continuous evaluation and iteration, and further refining the model based on new data, can contribute to its ongoing improvement and uncertainty reduction.

The NN model can be further refined by implementing these optimization strategies, allowing for even more accurate predictions and a narrower uncertainty range. This iterative process of model improvement is a vital step to ensure the robustness and reliability of the predictions, continually aligning them with the measured data. Tables 4-6 present a detailed comparison of predicted and measured values for CP, CS and CT-Y at random crank angles, utilizing NN, RFR and GPR models. Upon analysing the results, it becomes obvious that the GPR model exhibits an exceptional performance, with instances where the GPR absolute error is zero (dash lines), signifying precise predictions that align closely with the measured values. This remarkable accuracy underscores the GPR model's ability to effectively capture the underlying patterns and relationships within the dataset, resulting in highly reliable predictions for CP, CS, and CT-Y. Furthermore, the RFR model also demonstrates notable performance, consistently yielding low absolute error

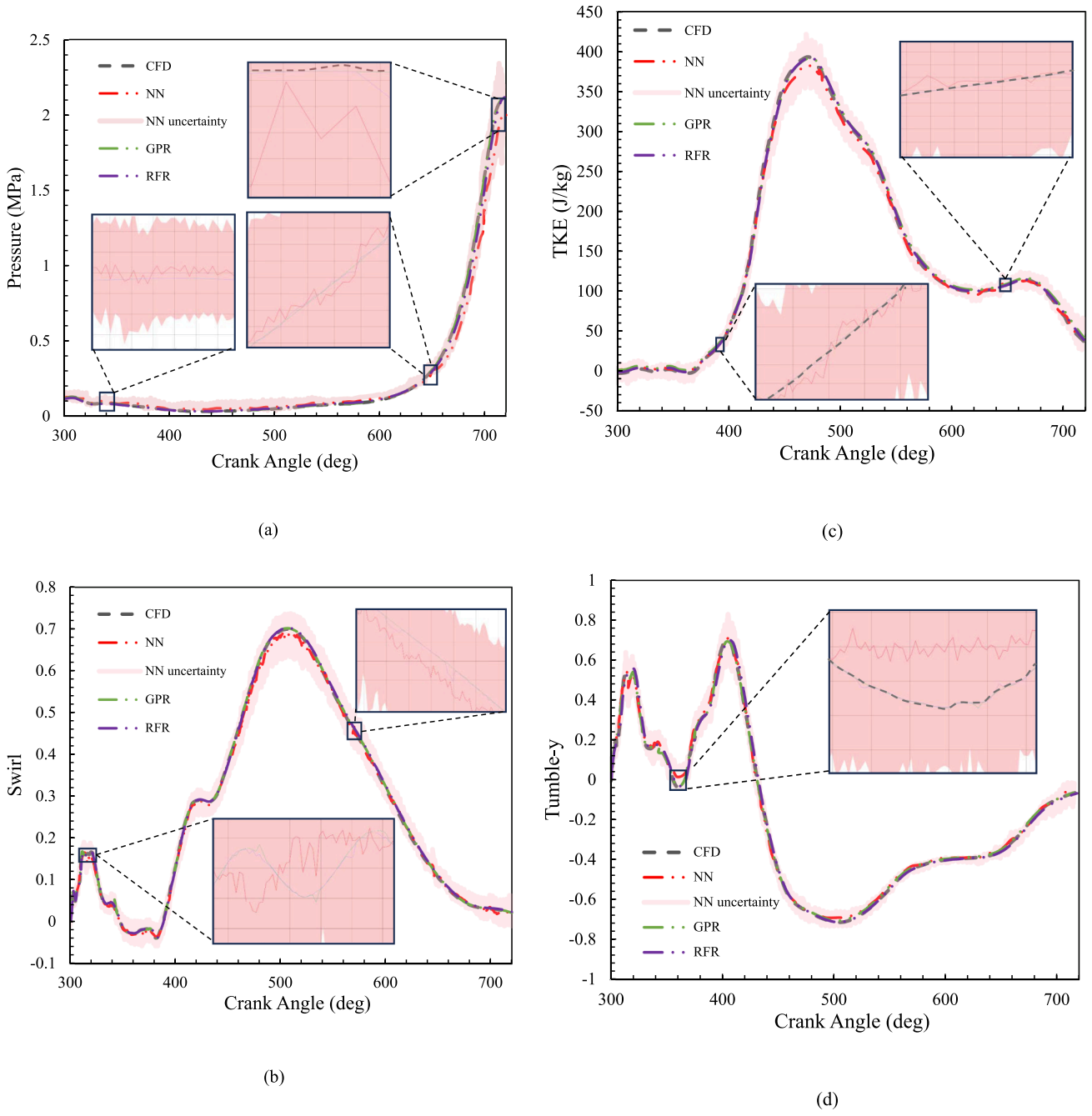


Fig. 13. (a) Pressure; (b) swirl; (c) TKE; and (d) tumble-y predicted by NN, RFR and GPR models at different crank angles.

Table 4
Comparison of predicted and measured values for the cylinder pressure (CP) at random crank angles using NN, RFR, and GPR models.

Crank Angle	Measured	NN	GPR	RFR	NN Ab. Error	GPR Ab. Error	RFR Ab. Error
368.19	0.06577	0.08494	0.06411	0.06339	0.01917	0.00166	0.00238
403.65	0.03817	0.03873	0.03798	0.03711	0.00056	0.00019	0.00106
487.87	0.04483	0.05912	0.04483	0.04563	0.01429	0.00000	0.0008
536.63	0.06953	0.08971	0.06947	0.07413	0.02018	6E-05	0.0046
580.96	0.08613	0.10691	0.08832	0.09255	0.02078	0.00219	0.00642
611.98	0.12848	0.13863	0.12881	0.13637	0.01015	0.00033	0.00789
651.88	0.32434	0.31454	0.32434	0.33402	0.0098	0.00000	0.00968
688.55	1.10828	1.07552	1.11509	1.11926	0.03276	0.00681	0.01098
711.31	2.04319	2.01317	2.03559	2.09463	0.03002	0.0076	0.05144

Table 5
Comparison of predicted and measured values for the cylinder swirl (CS) at random crank angles using NN, RFR, and GPR models.

Crank Angle	Measured	NN	GPR	RFR	NN Ab. Error	GPR Ab. Error	RFR Ab. Error
305.83316	0.07038	0.05925	0.07182	0.05193	0.011126	0.001440	0.018444
351.17310	0.01795	0.01814	0.01795	0.01706	0.000186	0.000001	0.000887
400.71176	0.13628	0.13543	0.13634	0.13642	0.000845	0.000064	0.000141
441.74462	0.31695	0.31367	0.31888	0.31414	0.003277	0.001925	0.002812
512.37777	0.69635	0.68474	0.69761	0.69987	0.011619	0.001260	0.003511
600.46568	0.31666	0.31599	0.31666	0.32345	0.000668	0.000000	0.006792
691.63949	0.03338	0.03032	0.03382	0.03576	0.003062	0.000435	0.002378
707.86847	0.02897	0.02107	0.02965	0.03130	0.007897	0.000682	0.002333
305.83316	0.07038	0.05925	0.07182	0.05193	0.011126	0.001440	0.018444

Table 6
Comparison of predicted and measured values for the cylinder tumble-y (CT-Y) at random crank angles using NN, RFR, and GPR models.

Crank Angle	Measured	NN	GPR	RFR	NN Ab. Error	GPR Ab. Error	RFR Ab. Error
303.29	0.15502	0.16188	0.15372	0.14849	0.006866	0.001293	0.006525
350.55	0.09873	0.09271	0.09687	0.09360	0.006024	0.001861	0.005134
400.21	0.66571	0.63483	0.66195	0.65613	0.030874	0.003758	0.009574
458.59	−0.56986	−0.56280	−0.56396	−0.56266	0.007057	0.005904	0.007197
505.55	−0.71356	−0.68961	−0.71356	−0.71369	0.023956	0.000000	0.000129
549.79	−0.54930	−0.52759	−0.54874	−0.54937	0.021709	0.000557	0.000066
620.32	−0.39215	−0.38874	−0.39693	−0.39336	0.003413	0.004774	0.001205
695.33	−0.11130	−0.11267	−0.11179	−0.11430	0.001365	0.000492	0.003000
303.29	0.15502	0.16188	0.15372	0.14849	0.006866	0.001293	0.006525

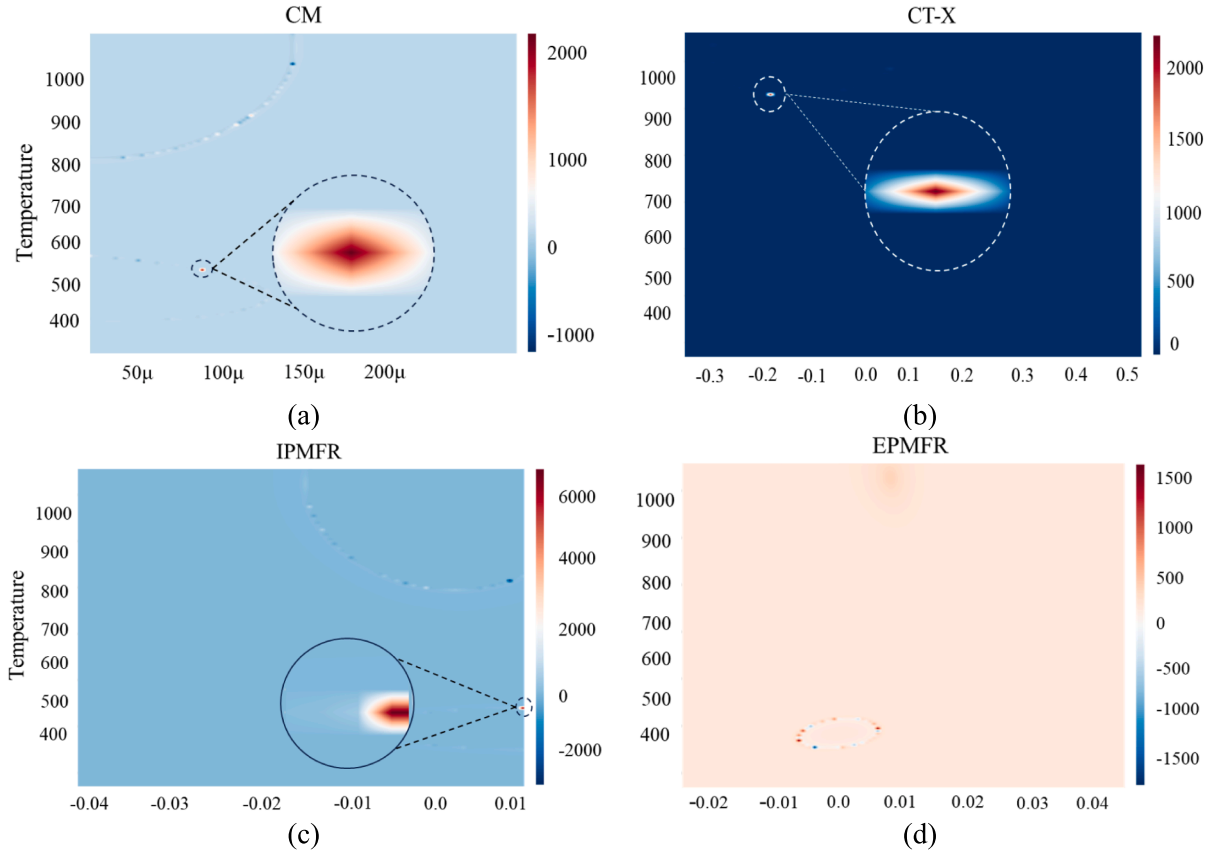


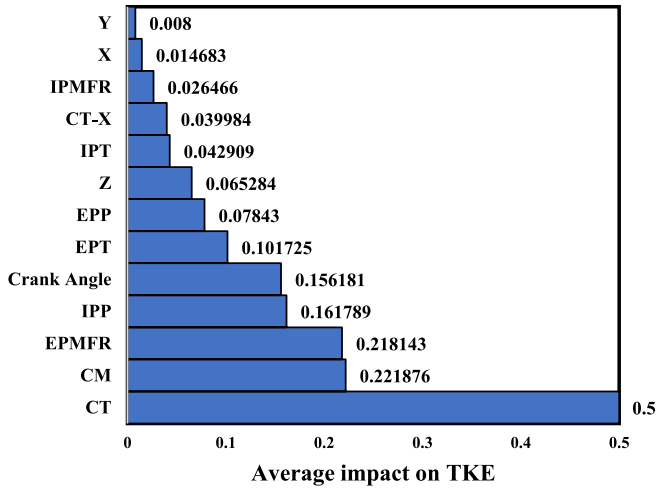
Fig. 14. Mean uncertainty maps for Gaussian Process Regression (GPR) model in prediction of engine performance: (a) CM; (b) CT-X; (c) IPMFR; and (d) EPMFR.

values and indicating its capacity to provide accurate predictions comparing with the measured values.

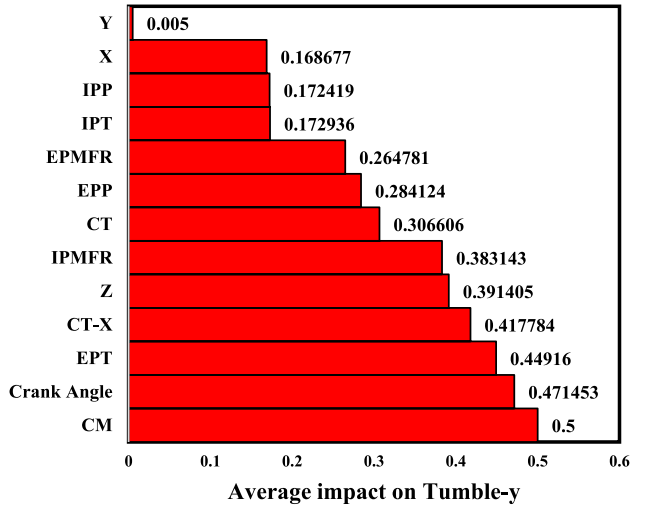
The analysis of the uncertainty maps in Figs. 14(a-d) is essential to evaluate the reliability and robustness of GPR model in predicting engine combustion performance. These uncertainty contours, which represent the mean uncertainty for outputs such as CM, CS, CT-Y, and TKE, provide valuable insights into the model's performance under different conditions. The fixed input of temperature across all figures, along with the second inputs such as Cylinder Mass (CM), Cylinder tumble-x (CT-X), Intake port mass flow rate (IPMFR), and Exhaust port mass flow rate (EPMFR), offer a comprehensive view of the model's predictive capabilities. The low uncertainty exhibited in all four outputs

underscores the high reliability of the GPR model in predicting engine performance. The small areas of high uncertainty, which is delineated by dash lines in figures, serve as valuable indicators for further data collection, model refinement, and the inclusion of more complex phenomena to enhance the prediction accuracy of ML models.

The sensitivity analysis of the GPR model was conducted to evaluate the influence of various input parameters on two critical outputs: TKE and tumble-y. The results are depicted in Figs. 15(a) and (b), which illustrate the average impact of each input parameter on TKE and tumble-y, respectively. For TKE in Fig. 15(a), Cylinder Temperature (CT) exhibited the highest normalized impact (0.5), followed by Cylinder Mass (CM) (0.221876) and Exhaust Port Mass Flow Rate (EPMFR) (0.218143). These findings highlight the importance of thermal and mass-related parameters in influencing turbulence generation within the cylinder. Conversely, geometric parameters such as X, Y, and Z directions showed minimal influence, emphasizing the predominance of operational conditions over geometry in shaping turbulence dynamics. For tumble-y in Fig. 15(b), Cylinder Mass was identified as the most influential factor (0.5), followed by Crank Angle (0.471453) and Exhaust Port Temperature (EPT) (0.44916). This underscores the importance of mass distribution and timing in controlling the rotational air motion. Notably, Cylinder Tumble-x (CT-X) and Geometrical Z-direction (Z) also played significant roles, indicating the interplay between tumble-y and spatial parameters. Similar to TKE, geometric features such as X and Y directions demonstrated lower impacts, further emphasizing the dominance of thermal and mass-related inputs. This dual analysis provides critical insights into the relative importance of input parameters for both TKE and tumble-y, enabling a deeper understanding of the engine's in-cylinder dynamics.



(a)



(b)

Fig. 15. Sensitivity analysis of input parameters on: (a) TKE and (b) tumble-y predictions using GPR model.

While the ML models demonstrate strong agreement with the CFD results across most crank angle ranges, a systematic bias can be observed in certain regions, particularly in NN and RFR predictions. For instance, as shown in Fig. 13(c), the NN model slightly underpredicts TKE in the crank angle range of 400–500 while showing a tendency to overpredict in the range of 500–600. Similarly, Fig. 13(a) reveals that the RFR model exhibits minor overprediction of pressure at the high crank angle region of 620–700, where sharp gradients are observed. These biases, although within acceptable error margins, indicate limitations in capturing the specific flow dynamics under certain conditions. The systematic underprediction in the NN model can be attributed to the sensitivity of the model to local minimums during training, particularly in regions with sparse or imbalanced data. This may cause the NN to be generalized less effectively in areas of rapid changes, leading to underestimation or overestimation of key parameters. Additionally, the overprediction in the RFR model can be linked to its discrete tree-based structure, which struggles to interpolate effectively in regions with high gradients or nonlinearities. These behaviours highlight the need for refined data curation, such as ensuring balanced representation of all operating conditions during training, and optimization of hyperparameters to improve the model's generalization capabilities.

Interestingly, GPR does not exhibit significant bias in its predictions, as evident from its close alignment with the CFD results across all crank angles as listed in Tables 4-6. This can be attributed to its probabilistic nature, which allows it to effectively capture the uncertainty and variability in data. However, the computational cost of GPR models increases with dataset size, which could be a limitation for scaling the approach to more complex datasets or simulations. Identifying and addressing these systematic biases is critical for further improving the accuracy and reliability of the ML models. Possible mitigation strategies include increasing the size and diversity of the training dataset, employing data augmentation techniques to balance the representation of complex scenarios, and incorporating additional physical constraints or physics-informed features into the ML training process.

4.4. *Surface field prediction*

The innovative AI tool Surface Field has demonstrated the significant promise of surface field prediction in engineering research. This approach empowers engineers to leverage historical 3D simulation data to establish correlations between 3D geometries and their corresponding surface fields. To show the predictive capability of the model, Fig. 16 illustrates the velocity magnitude at the cylinder centre on a Y plane for both ML prediction and CFD simulation. Additionally, the error distribution contour further highlights the similarity between the ML predictions and CFD results, demonstrating a consistent trend in flow patterns and comparable min/max velocity magnitudes. This similarity is particularly noteworthy given the significant difference in running time and processing resources between the two approaches. In the case of CFD simulations, 8 cores were employed for parallel computing, resulting in an accumulated CPU time of 966,185 s. The ML approach, on the other hand, involved training the model for 80,000 steps on a single processor, significantly reducing the computational resource requirements. Table 7 presents a detailed comparison of the CPU time and wall-clock time utilized in the computing process for both CFD simulations and the ML model. For accumulated CPU time, the CFD solver utilized 8 cores of Intel Xeon Gold 5215 (13.75 M Cache, 2.50 GHz) and required 966,185 s, while the ML model required only 44,748 s. This results in a remarkable speedup of approximately 21.6 times faster in

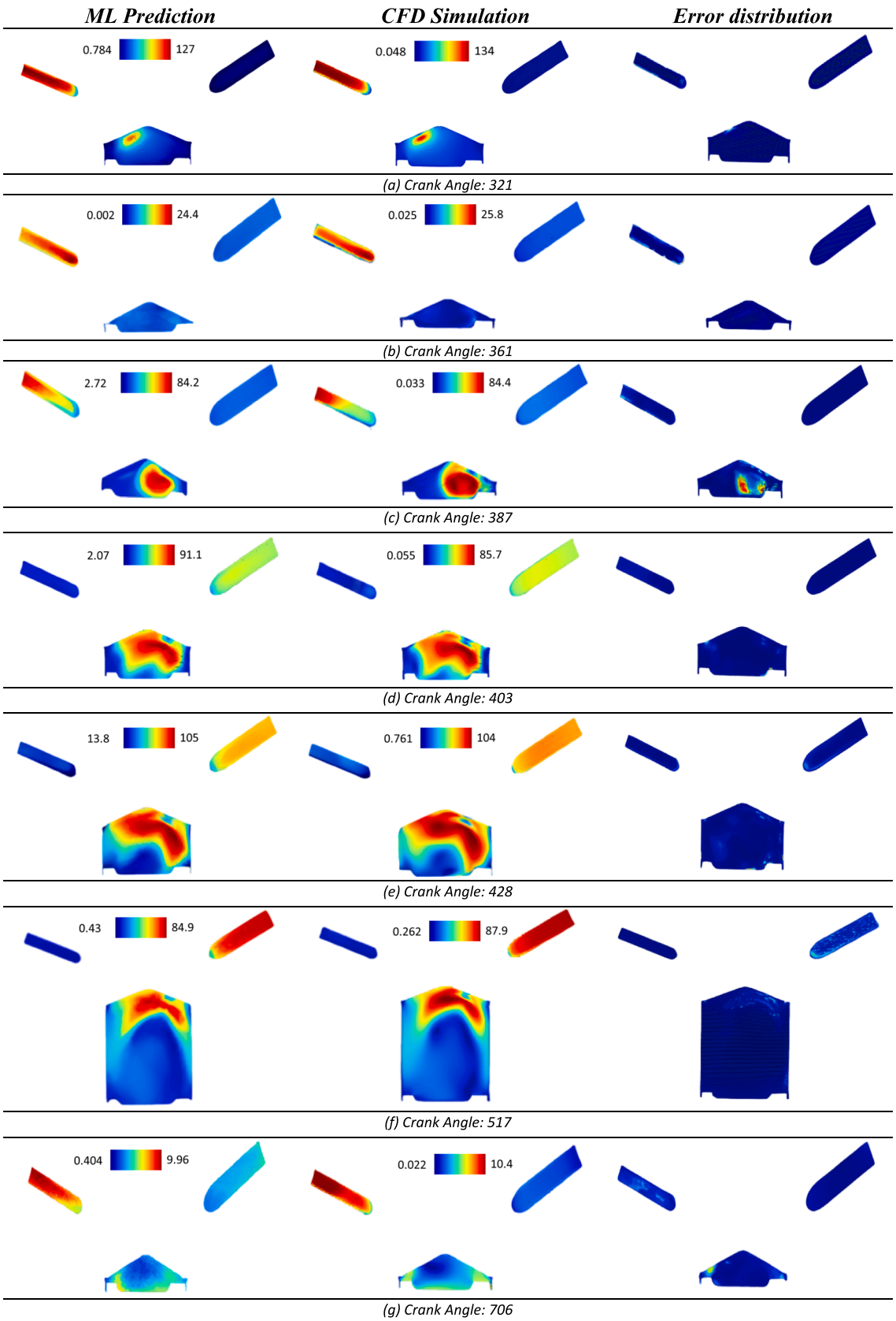


Fig. 16. Velocity magnitude (m/s) of ML prediction (left panel) and CFD simulation (middle panel), and error distribution (right panel): (a) 321 degrees; (b) 361 degrees; (c) 387 degrees; (d) 403 degrees; (e) 428 degrees; (f) 517 degrees; and (g) 706 degrees.

Table 7
Comparison of CPU time between CFD simulation and ML model.

Module	CFD simulation	ML model	Speedup
Accumulated CPU time over all processes (s)	966,185	44,748	21.6
Wall-clock time (hours)	33.5	12.4	2.7

terms of accumulated CPU time. However, when considering wall-clock time, the CFD simulation required approximately 33.5 h for a single run, while the ML model completed its training and inference in 12.4 h, yielding a real-world speedup of approximately 2.7 times faster. These findings highlight the dual advantages of ML models in significantly reducing the computational costs and achieving faster real-world runtimes. This efficiency positions ML as a highly advantageous alternative for optimizing engineering processes compared with traditional CFD solvers.

Significantly, the increase in the number of training steps is expected to produce consistent outcomes, with both the training loss and validation loss decreasing progressively. This highlights the potential of AI-driven surface field prediction as a cost-efficient and effective substitute for resource-intensive simulations, underlining its importance and influence on the future advancement of combustion engineering and science. Fig. 17 provides a comprehensive depiction of a model's training process across 100 k training steps, with the blue line denoting the validation and the red line representing the training process on selected data. The smaller graphs within the figure demonstrate the velocity magnitude at specific training steps (20 k, 30 k, 40 k, 60 k, and 80 k), offering insight into the evolution of prediction accuracy as the number of training steps increases. Particularly, the minimum and maximum values of velocity magnitude are recorded at 4.61 m/s and 64.6 m/s at 20k training steps, respectively, corresponding to the train loss and validation loss of 0.11 and 0.12. As the training progresses, the velocity magnitude demonstrates a discernible improvement and reaches the minimum and maximum values of 2.71 m/s and 84.2 m/s at 80k training steps with the training loss and validation loss being decreased to 0.059 and 0.075. This trend suggests a positive correlation between the number of training steps and the accuracy of prediction results while being in comparison with CFD simulation results.

4.5. Targeted optimization of TKE and tumble-y

In the context of targeted optimization using the GPR model, the aim is to identify the best sets of inputs that can closely match a given list of targeted outputs. Compared with traditional iterative methods, this approach offers researchers a more efficient and expedited option of achieving a product with the desired performance and characteristics in ICEs. Specifically, in the case of optimizing TKE and tumble-y, the focus is on increasing TKE to enhance the air motion dynamics during the cold flow phase of the engine cycle, emphasizing the pre-combustion features. TKE has been shown to enhance the swirl and circulation of air, promoting better mixing of air and fuel, as highlighted in the literature [47,51]. Improved swirl and turbulence during the intake and compression stages can create more favourable conditions for flame initiation and propagation once combustion begins. For example, studies indicate that increased TKE enhances swirl effects, reducing burn duration and improving flame stretching, which can contribute to better

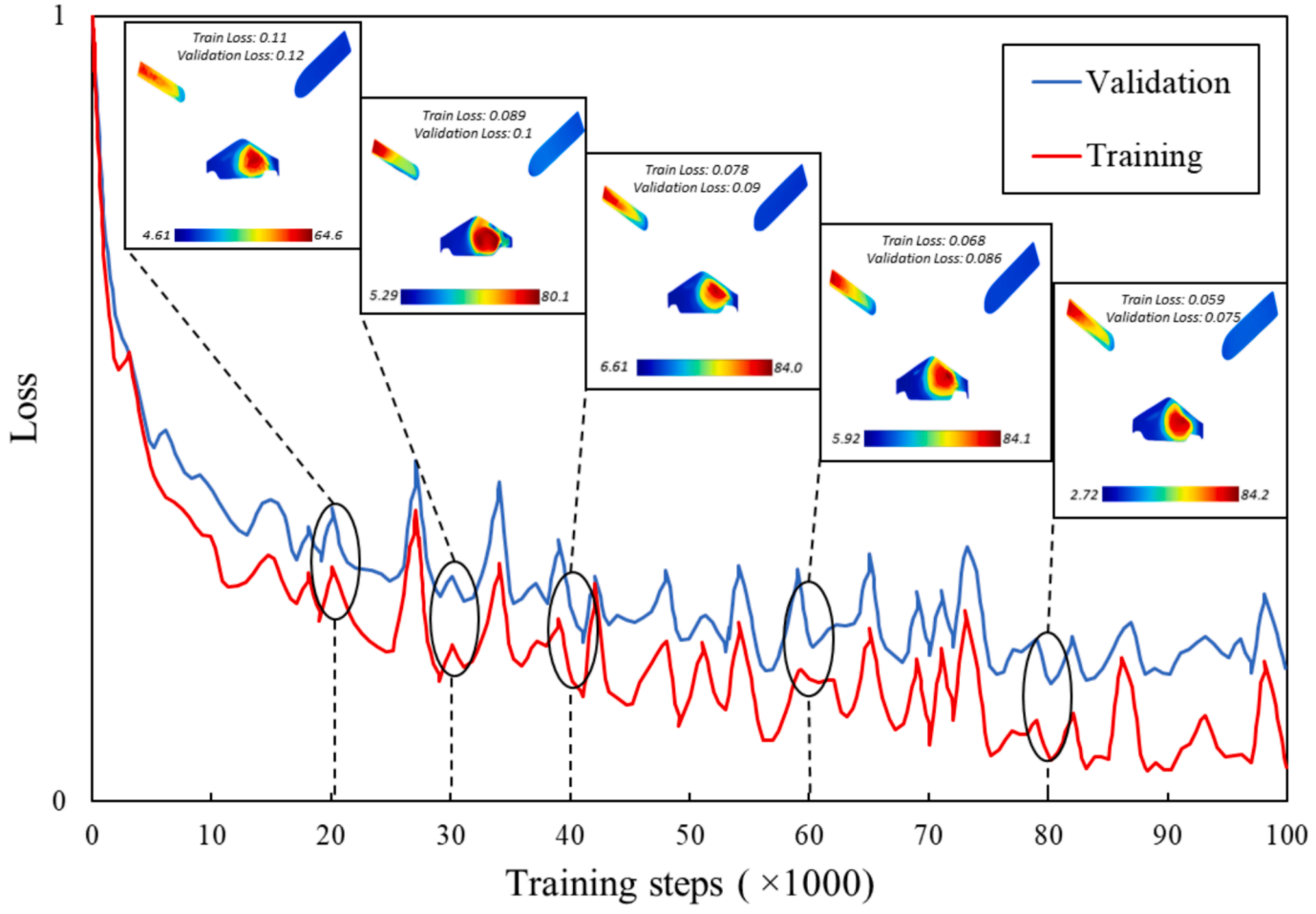


Fig. 17. Analysis of the prediction accuracy against training steps.

air–fuel mixing and heat release rates [51]. Additionally, the presence of swirl at optimal levels minimizes the combustion duration and optimizes heat transfer [47]. These findings align with our focus on optimizing TKE in the pre-combustion phase to improve mixing and in-cylinder flow dynamics. However, it is worth noting that this study focuses solely on the cold flow simulations and does not include direct combustion or emissions analysis. As a result, while improved air–fuel mixing has been shown in the literature to potentially influence the combustion efficiency and emissions, validating these effects would require dedicated combustion modelling, which falls outside the scope of this work.

Conversely, the optimization process also considers minimizing tumble during the cold flow phase. Tumble refers to the rotational motion of the air–fuel mixture inside the cylinder, which, when excessive, can lead to increased heat loss and incomplete mixing. By reducing tumble, greater control over the air–fuel mixture dynamics can be achieved, creating more favourable conditions for combustion initiation. Using the GPR model, engineers can systematically analyse the effects of input parameters on both TKE and tumble-y, identifying configurations that enhance in-cylinder flow characteristics during the intake and compression phases. This targeted optimization minimizes the computational costs and provides deeper insights into the complex interactions between TKE, tumble and pre-combustion air motion dynamics. Regarding the targeted optimization using the GPR model, the evaluation of optimization progress is crucial to ensure the attainment of desired performance and characteristics. One commonly used approach to assess the effectiveness of optimization is to utilise a fitness function, such as the Euclidean distance [39,40]. The choice of the Euclidean distance as the fitness function in this study is guided by its intuitive interpretation, its compatibility with the optimization problem, and its widespread applicability in multidimensional regression tasks. The Euclidean distance is a well-known metric that measures the straight-line distance between two points in a multidimensional space. Its continuous and differentiable nature makes it particularly suited for optimization problems involving gradient-based approaches, as it facilitates efficient exploration of the input space. In the context of targeted optimization, the Euclidean distance can be employed as a fitness function to evaluate how closely the outputs generated in the GPR model match the desired target outputs. By calculating the Euclidean distance between the predicted and target outputs, one can quantify the discrepancy and assess the optimization progress. A smaller Euclidean distance indicates a closer match between the predicted outputs and the desired target outputs, signifying higher optimization achievement.

The Monolith AI platform provides several fitness functions, including Euclidean distance, Manhattan distance, Collinearity coefficient, and Amplitude correlation coefficient. Among these, the Euclidean distance was chosen in this study for several reasons. The Euclidean distance offers a clear and straightforward interpretation of optimization quality. A smaller Euclidean distance directly indicates a closer match between predicted outputs and target outputs, providing an easy-to-understand measure of success. The nature of the optimization problem in this study involves continuous, multidimensional data. The Euclidean distance is well-suited for capturing relationships and dependencies within such spaces, making it effective for evaluating optimization progress. Due to its differentiable nature, the Euclidean distance integrates seamlessly with gradient-based optimization algorithms, enabling efficient and effective exploration of the input parameter space. While other metrics like the Manhattan distance measure the sum of absolute differences between points, the Euclidean distance considers the square root of the sum of squared differences. This makes the Euclidean distance more sensitive to large discrepancies, which is beneficial when precise matching of outputs to targets is critical. The Manhattan distance could be advantageous in scenarios where deviations along individual dimensions are of equal importance, as it weights all deviations linearly. However, in this study the squared weighting in the Euclidean distance aligns better with the goal of reducing large discrepancies, which are often more impactful in engine optimization.

The recommended designs in Table 8 and Fig. 18 demonstrate the remarkable capabilities of utilizing the GPR model combined with the targeted optimization to enhance the engine performance. These designs have been carefully selected to increase TKE and decrease tumble-y, aiming to improve the fuel efficiency and achieve a more efficient engine performance. At a crank angle of 342 degrees, it exhibits a TKE value of 396.56 J/kg, indicating a high level of turbulence. This robust turbulence promotes an optimal fuel–air mixing, resulting in improved fuel efficiency. Moreover, the negative tumble-y value of −0.15352237 signifies a reduction in rotational motion, leading to enhanced combustion completeness. These combined factors along with a temperature of 846.42 K and a pressure of 1.52 bar create an optimal combustion environment that can significantly enhance the engine performance. Similarly, the second recommended design shows a TKE value of 366.77 J/kg at a crank angle of 314 degrees. With a negative tumble-y value of −0.15348984, this design further reduces the rotational motion and promotes a more efficient fuel–air mixing. The temperature of 549.59 K and pressure of 2.81 bar support the overall combustion process, leading to improved engine performance. These recommended designs consistently demonstrate the ability of the GPR model and targeted optimization to identify design configurations that increase TKE and decrease tumble-y. The optimized parameters, including temperature, pressure, and tumble-y, work synergistically to create an ideal combustion environment. By maximizing the turbulence and minimizing the rotational motion, these designs effectively enhance the fuel–air mixing and combustion efficiency, hoping a return of decreased fuel consumption and lower emissions. It is also important to highlight that although the numerical findings in Table 8 offer convincing indications of the optimal design configurations, additional experimental validation is essential to verify their real-world significance. Thorough testing and analysis are crucial to fully assess the potential of these optimized designs in practical engine applications such as different engine configurations.

Testing physical engines under varying conditions incurs significant cost, time, and resources. In such scenarios, the presented research demonstrates how ML models coupled with CFD simulations can bridge this gap to grant access to the engineers by unveiling hidden patterns or unseen data. The use of trained ML models enables the derivation of the inter-relating of the most important parameters, such as TKE, tumble,

Table 8
Comparison of performance metrics for recommended engine designs using GPR model.

Design	Crank Angle (deg)	CT-X	CM (kg)	EPT (K)	IPT (K)	Temp. (K)	CP (bar)	CS	TKE (J/kg)	CT-Y
# 1	342.37	0.27090	2.08E-05	905.19	328.47	846.42	1.52	0.2440	396.56	−0.15352237
# 2	314.93	0.32768	3.76E-05	921.30	314.11	549.59	2.81	0.2769	366.77	−0.15348984
# 3	384.50	0.14556	0.000131	910.69	314.78	467.96	1.83	0.2785	357.92	−0.15348984
# 4	327.69	0.05168	9.89E-05	944.48	323.23	627.34	2.17	0.2659	355.00	−0.15348993
# 5	322.98	0.47194	9.41E-05	793.61	312.24	413.63	2.90	0.2782	350.09	−0.15348984
# 6	501.18	−0.17476	0.000106	833.36	306.95	407.49	0.34	0.2806	347.11	−0.15348984
# 7	360.95	0.03740	0.000126	864.46	305.23	480.32	1.51	0.2938	331.78	−0.15349085
# 8	579.61	−0.23877	0.000215	823.58	298.64	398.81	0.66	0.2981	325.89	−0.15349335
# 9	367.63	0.13691	6.60E-05	920.50	331.82	449.82	1.94	0.2790	322.15	−0.15348984

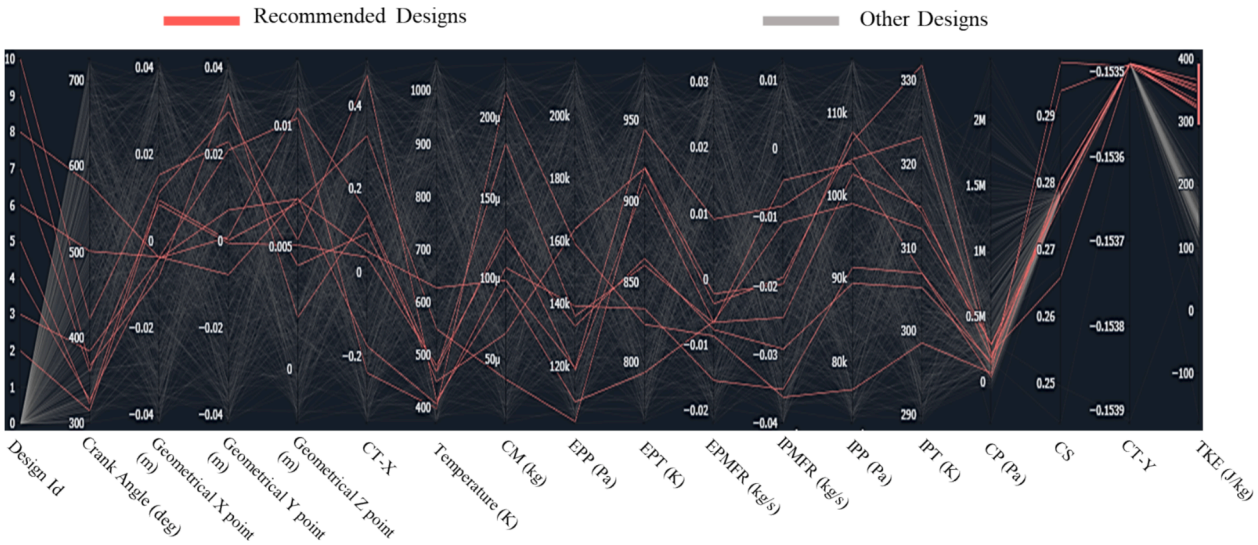


Fig. 18. Parallel coordinates graph for recommended designs (red lines) and other designs (grey lines).

pressure, and temperature, for gaining detailed insight into in-cylinder flow dynamics during an engine cycle at cold flow. The insights allow the user to create adaptable initial and boundary conditions (IC/BC) while avoiding iterative simulations and time-consuming real-life testing.The ability of ML models to predict outcomes under various input configurations allows engineers to optimize the in-cylinder flow processes for different operating conditions without running multiple simulations. For example, TKE varied along the intake and compression phases can have its swirl or circulation modified by changing the dynamics. Improved TKE results in considerable improvement in air and fuel mixing-sketching favourable scenarios for initiating the combustion process by improving the conditions at the time of spark propagation or ignition. Similarly, minimizing excessive tumble intensity through ML-guided optimization can provide better control over air–fuel motion, reducing undesirable effects such as heat losses and incomplete mixing. Moreover, the ML-based methodology developed in this study offers practical benefits beyond optimizing in-cylinder air motion. It enables:

1) ML reduces the computational burden associated with iterative simulations, allowing engineers to efficiently explore a wide range of design configurations and operational conditions.
2) By accessing the hidden layers of trained ML models, engineers can predict the outcomes of varying conditions, such as different IC/BC setups, without the need for physical testing.
3) The integration of CFD and ML methodologies is easily scalable to other engine geometries and operating conditions, making it a versatile tool for the automotive and energy industries.

The implications of these advancements extend to both research and industrial settings. Researchers can use ML to simulate challenging scenarios, such as extreme pressure and temperature conditions, which are difficult to replicate experimentally. The manufacturers will accelerate their work of making cleaner and more efficient engines with the help of designs guided by ML. Furthermore, the approach presented in this study aligns with ongoing efforts to meet increasingly stringent regulatory standards for emissions and fuel efficiency, providing a cost-effective alternative to traditional experimental testing. This, therefore, provides an efficient, scalable, and cost-effective solution to some of the optimization challenges in engines and points toward the transformative potential that could be obtained with the integration of ML and CFD in developing high-performance engine systems.

5. Conclusion

This study has demonstrated the transformative potential of integrating machine learning (ML) with computational fluid dynamics (CFD) to optimize the internal combustion engines. By leveraging ML models such as Random Forest Regression (RFR), Gaussian Process Regression (GPR) and Neural Networks (NN), the research provides significant insights into predicting critical engine parameters. The GPR model emerged as the optimal choice due to its superior accuracy and reduced uncertainty, as indicated by metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson Coefficient (PC), and R-squared (R^2). These results underscore the reliability of GPR in predicting engine dynamics, such as turbulence kinetic energy (TKE) and tumble-y, which play pivotal roles in enhancing fuel–air mixing and pre-combustion air motion. The research highlights the remarkable computational efficiency achieved through ML integration. Compared with traditional CFD solvers, the ML approach resulted in a 21.6x speedup, reducing the computational burden significantly while maintaining accuracy. The results also illustrate the ability of ML models to predict the surface field phenomena, such as velocity distributions, with comparable fidelity to CFD simulations. This capability is particularly valuable in reducing the need for costly and time-intensive physical testing, providing engineers access to hidden patterns and insights into in-cylinder flow dynamics. By predicting outcomes under various input configurations, the trained ML models enable the optimization of initial and boundary conditions (IC/BC) without iterative simulations, significantly accelerating the design process. The study further emphasizes the practical implications of optimizing TKE and tumble-y during the cold flow phase. For example, increasing TKE improves swirl and circulation, leading to better mixing of air and fuel, thereby creating favourable conditions for efficient combustion initiation. Similarly, minimizing excessive tumble intensity through ML-guided optimization allows better control over air–fuel motion, reducing undesirable effects such as heat loss and incomplete mixing. These findings demonstrate the capability of ML-driven optimization to improve thermal efficiency and lay the groundwork for sustainable and efficient engine designs.

While the research demonstrates clear advantages, limitations such as the reliance on specific input parameters and simulation data are acknowledged. Future studies could overcome these obstacles by incorporating more diverse datasets, exploring the impact of varying operating conditions, and integrating hybrid models that combine the physics-based approaches with data-driven ML techniques. Such advancements would further enhance the robustness and scalability of ML-

CFD frameworks for broader applications in the automotive and energy industries. This study provides a scalable, cost-effective and efficient solution for engine optimization challenges. By bridging the gap between traditional CFD simulations and experimental testing, it offers a promising pathway for achieving cleaner and more efficient engines, aligning with global sustainability goals and regulatory standards.

CRediT authorship contribution statement

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

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.

Acknowledgement

The authors express their gratitude to Monolith AI for granting permission to utilize their Monolith platform. Amirali Shateri likes to acknowledge the University of Derby for the Ph. D studentship (contract no. S&E_Engineering_0722) and the support provided.

Data availability

Data will be made available on request.

References

[1] Krishnamoorthi M, Malayalamurthi R, He Z, Kandasamy S. A review on low temperature combustion engines: Performance, combustion and emission characteristics. Renew Sustain Energy Rev 2019;116:109404.
[2] Leach F, Kalghatgi G, Stone R, Miles P. The scope for improving the efficiency and environmental impact of internal combustion engines. Transp Eng 2020;1:100005.
[3] Paykani A, Chehrmonavari H, Tsolakis A, Alger T, Northrop WF, Reitz RD. Synthesis gas as a fuel for internal combustion engines in transportation. Prog Energy Combust Sci 2022;90:100995.
[4] Shateri A, Bahram J, Saber S, Payam J, Domiri Ganji Davood. Numerical study of the effect of ultrasound waves on the turbulent flow with chemical reaction. Energy 2024;289:129707.
[5] Bae C, Kim J. Alternative fuels for internal combustion engines. Proc Combust Inst 2017;36(3):3389–413.
[6] Kalghatgi GT. Developments in internal combustion engines and implications for combustion science and future transport fuels. Proc Combust Inst 2015;35(1):101–15.
[7] Dziubak T, Karczewski M. Experimental Studies of the Effect of Air Filter Pressure Drop on the Composition and Emission Changes of a Compression Ignition Internal Combustion Engine. Energies 2022;15(13):4815.
[8] Pham Q, Park S, Agarwal AK, Park S. Review of dual-fuel combustion in the compression-ignition engine: Spray, combustion, and emission. Energy 2022;250:123778.
[9] Madihi R, Pourfallah M, Gholinia M, Armin M, Ghadi AZ. Thermofluids analysis of combustion, emissions, and energy in a biodiesel (C11H22O2)/natural gas heavy-duty engine with RCCI mode (Part II: Fuel injection time/Fuel injection rate). International Journal of Thermofluids 2022;16:100200.
[10] Agarwal AK, Singh AP, García A, Monsalve-Serrano J. Challenges and opportunities for application of reactivity-controlled compression ignition combustion in commercially viable transport engines. Prog Energy Combust Sci 2022;93:101028.
[11] Huang J, Gao J, Wang Y, Yang Ce, Ma C, Tian G. Effect of asymmetric fuel injection on combustion characteristics and NOx emissions of a hydrogen opposed rotary piston engine. Energy 2023;262:125544.
[12] Kumar M, Bhowmik S, Paul A. Effect of pilot fuel injection pressure and injection timing on combustion, performance and emission of hydrogen-biodiesel dual fuel engine. Int J Hydrogen Energy 2022;47(68):29554–67.
[13] Lu Y, Fan C, Chen Y, Liu Y, Pei Y. Effect of injection strategy optimization on PCCI combustion and emissions under engine speed extension in a heavy-duty diesel engine. Fuel 2023;332:126053.
[14] Hao C, Zhang Z, Wang Z, Bai H, Li Y, Li Y, et al. Investigation of spray angle and combustion chamber geometry to improve combustion performance at full load on a heavy-duty diesel engine using genetic algorithm. Energ Conver Manage 2022;267:115862.
[15] Xie J. Approaches for describing processes of fuel droplet heating and evaporation in combustion engines. Fuel 2024;360:130465.
[16] Zhu G, Wang Y, Zuo Q, Chen W, Shen Z, Yang X, et al. Numerical investigation gaseous ammonia basic jet and mixing characteristics in the constant volume vessel. Int J Hydrogen Energy 2024;80:68–81.
[17] Ihme, Matthias, Wai Tong Chung, and Aashwin Ananda Mishra. Combustion machine learning: Principles, progress and prospects. Progr Energy Combust Sci 91 (2022): 101010.
[18] Zuo Q, Chen L, Chen W, Zhu G, Wang Z, Ma Y, et al. Performance analysis of ammonia energy ratio on an ammonia-diesel engine in different fuel supply modes. Fuel 2025;384:134038.
[19] Sipei Wu, Wang H, Luo KH. A robust autoregressive long-term spatiotemporal forecasting framework for surrogate-based turbulent combustion modeling via deep learning. Energy AI 2023:100333.
[20] Huang Q, Liu J, Ulishney C, Dumitrescu CE. On the use of artificial neural networks to model the performance and emissions of a heavy-duty natural gas spark ignition engine. Int J Engine Res 2022;23(11):1879–98.
[21] Schiffmann P, Gupta S, Reuss D, Sick V, Yang X, Kuo T-W. TCC-III engine benchmark for large-eddy simulation of IC engine flows. Oil, Gas Sci Technol 2016;71(1):3.
[22] Kang KY, Baek JH. Tumble flow and turbulence characteristics in a small four-valve engine. SAE Technical Paper 1996;No. 960265.
[23] Shuliang, Liu, Li Yufeng, and Lu Ming. Prediction of tumble speed in the cylinder of the 4-valve spark ignition engines. No. 2000-01-0247. SAE Technical Paper, 2000.
[24] Liu J, Huang Q, Ulishney C, Dumitrescu CE. Comparison of random forest and neural network in modeling the performance and emissions of a natural gas spark ignition engine. J Energy Res Technol 2022;144(3):032310.
[25] Bhattacharya A, Majumdar P. Artificial Intelligence-Machine Learning Algorithms for the Simulation of Combustion Thermal Analysis. Heat Transfer Eng 2023:1–18.
[26] Brunton SL, Noack BR, Koumoutsakos P. Machine learning for fluid mechanics. Annu Rev Fluid Mech 2020;52:477–508.
[27] Baker, Nathan, Frank Alexander, Timo Bremer, Aric Hagberg, Yannis Kevrekidis, Habib Najm, et al. Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. USDOE Office of Science (SC), Washington, DC (United States), 2019.
[28] Xu, Yongjun, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Liu et al. Artificial intelligence: A powerful paradigm for scientific research. The Innovation 2, no. 4 (2021).
[29] Nassehi A, Zhong RY, Li X, Epureanu BI. Review of machine learning technologies and artificial intelligence in modern manufacturing systems. In: In Design and Operation of Production Networks for Mass Personalization in the Era of Cloud Technology. Elsevier; 2022. p. 317–48.
[30] Raschka S, Patterson J, Nolet C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information 2020;11(4):193.
[31] Meng, Chuizheng, Sungyong Seo, Defu Cao, Sam Griesemer, and Yan Liu. When physics meets machine learning: A survey of physics-informed machine learning. arXiv preprint arXiv:2203.16797 (2022).
[32] Yang R, Xie T, Liu Z. The application of machine learning methods to predict the power output of internal combustion engines. Energies 2022;15(9):3242.
[33] Zhou L, Song Y, Ji W, Wei H. Machine learning for combustion. Energy AI 2022;7:100128.
[34] Karunamurthy K, Janvekar AA, Palaniappan PL, Adhitya V, Lokeswar TTK, Harish J. Prediction of IC engine performance and emission parameters using machine learning: A review. J Therm Anal Calorim 2023:1–23.
[35] Liu, Jinlong, Cosmin Dumitrescu, and Christopher Ulishney. Investigation of Heat Transfer Characteristics of Heavy-Duty Spark Ignition Natural Gas Engines Using Machine Learning. No. 2022-01-0473. SAE Technical Paper, 2022.
[36] Liu J, Ulishney C, Dumitrescu CE. Random forest machine learning model for predicting combustion feedback information of a natural gas spark ignition engine. J Energy Res Technol 2021;143(1):012301.
[37] Zhang N, Wood O, Yang Z, Xie J. AI-Guided Computing Insights into a Thermostat Monitoring Neonatal Intensive Care Unit (NICU). Sensors 2023;23(9):4492.
[38] Jamil A, Baharom MB, Rashid A, Aziz A. IC engine in-cylinder cold-flow analysis–A critical review. Alex Eng J 2021;60(3):2921–45.
[39] Dokmanic I, Parhizkar R, Ranieri J, Vetterli M. Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Process Mag 2015;32:12–30.
[40] Liang J, Wei Y, Boyang Qu, Yue C, Song H. Ensemble learning based on fitness Euclidean-distance ratio differential evolution for classification. Nat Comput 2021;20:77–87.
[41] Shateri A, Yang Z, Xie J. Utilizing artificial intelligence to identify an optimal machine learning model for predicting fuel consumption in diesel engines. Energy AI 2024;16:100360.
[42] Novello P, Poëtte G, Lugato D, Congedo PM. Goal-oriented sensitivity analysis of hyperparameters in deep learning. J Sci Comput 2023;94:45.
[43] Yuksel O, Bayraktar M, Sokukcu M. Comparative study of machine learning techniques to predict fuel consumption of a marine diesel engine. Ocean Eng 2023;286:115505.

[44] Xue L, Liu Y, Xiong Y, Liu Y, Cui X, Lei G. A data-driven shale gas production forecasting method based on the multi-objective random forest repression. J Pet Sci Eng 2021;196:107801.
[45] Klayborworn S, Pakdee W. Effects of porous insertion in a round-jet burner on flame characteristics of turbulent non-premixed syngas combustion. Case Stud Therm Eng 2019;14:100451.
[46] Indlekofer T, Faure-Beaulieu A, Noiray N, Dawson J. The effect of dynamic operating conditions on the thermoacoustic response of hydrogen rich flames in an annular combustor. Combust Flame 2021;223:284–94.
[47] Zhang D, Hill PG. Effect of swirl on combustion in a short cylindrical chamber. Combust Flame 1996;106(3):318–32.
[48] Shih T-H, Liou WW, Shabbir A, Yang Z, Zhu J. A new k-ϵ eddy viscosity model for high reynolds number turbulent flows. Comput Fluids 1995;24(3):227–38.
[49] Wilcox DC. One-equation and two-equation Models. Turbul Model CFD 2006;1:192–209.
[50] Versteeg, Henk Kaarle. An introduction to computational fluid dynamics the finite volume method, 2/E. Pearson Education India, 2007.
[51] Davis GC, Davis CC, Mikulec A, Kent JC, Tabaczynski RJ. Modeling the effect of swirl on turbulence intensity and burn rate in SI engines and comparison with experiment. SAE Trans 1986:427–34.
[52] Liu J, Wang H. Machine learning assisted modeling of mixing timescale for LES/PDF of high-Karlovitz turbulent premixed combustion. Combust Flame 2022;238:111895.
[53] Liu J, Huang Q, Ulishney C, Dumitrescu CE. Machine learning assisted prediction of exhaust gas temperature of a heavy-duty natural gas spark ignition engine. Appl Energy 2021;300:117413.
[54] Shateri A, Yang Z, Xie J. Machine learning-based molecular dynamics studies on predicting thermophysical properties of ethanol–octane blends. Energy Fuel 2025.
[55] Rasmussen, Carl Edward. Gaussian processes in machine learning. In: Summer school on machine learning, pp. 63-71. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.
[56] Breiman L. Classification and regression trees. Routledge; 2017.
[57] Siemens Digital Industries Software. Validating Simcenter STAR-CCM+ In-cylinder Solution. White paper. https://www.plm.automation.siemens.com/media/global/it/SiemensSW-Validating-Simcenter-STAR-CCM-In%20Cylinder-Solution_%20WhitePaper_tcm56-72011.pdf.
[58] Xiao D, Heaney CE, Mottet L, Fang F, Lin W, Navon IM, et al. A reduced order model for turbulent flows in the urban environment using machine learning. Build Environ 2019;148:323–37.
[59] Zhao F, Hung DLS. Applications of machine learning to the analysis of engine in-cylinder flow and thermal process: A review and outlook. Appl Therm Eng 2023;220:119633.
[60] Gupta R, Jaiman R. Three-dimensional deep learning-based reduced order model for unsteady flow dynamics with variable Reynolds number. Phys Fluids 2022;34(3).