Robust key parameter identification of dedicated hybrid engine performance indicators via K-fold filter collaborated feature selection

He, Xu; Li, Ji; Zhou, Quan; Lu, Guoxiang; Xu, Hongming

DOI:
10.1016/j.engappai.2023.107114

Engineering Applications of Artificial Intelligence 126 (2023) 107114

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Robust key parameter identification of dedicated hybrid engine performance indicators via K-fold filter collaborated feature selection

Xu He [a], Ji Li [a,*], Quan Zhou [a], Guoxiang Lu [b], Hongming Xu [a]

[a] Department of Mechanical Engineering, School of Engineering, University of Birmingham, Birmingham, B15 2TT, UK
[b] BYD Auto Co Ltd, Guangzhou City, China

ARTICLE INFO

Keywords:
Data-driven modelling
Feature selection
Parameter identification
K-fold cross validation
Internal combustion engine

ABSTRACT

Dedicated hybrid engine technology using auxiliary electronic components has been proven as an energy-saving solution to public concerns about energy consumption and carbon emissions. This paper proposes a generic approach of K-fold filter-collaborated feature selection (KFFC-FS) to robustly identify the key parameters of three engine performance indicators, i.e., volumetric efficiency, thermal efficiency, and fuel consumption. By using this approach, five filters are collaborated to provide a robust rank of feature importance and avoid the feature overestimation caused by the single filter. Meanwhile, the K-fold cross validation method is introduced to avoid random precision issues and overfitting, further enhancing the robustness of key parameter identification for the independent engine performance indicators. In this research, the modelling data is collected from an experimental test bench with a BYD 1.5L gasoline engine. Under the basics of the studied three engine performance indicators by using a multiple-layer perceptron network, the proposed approach further reduces by at least 10.3% root-mean-square error (RMSE) and at least 30% reduction of the model inputs.

1. Introduction

Dedicated hybrid engines have been developed, which keep over 40% thermal efficiency while meeting the emission regulations (Lu et al., 2021) to improve fuel economy and reduce vehicle emissions (Ehsani et al., 2021). The success of dedicated hybrid engines depends on the usage of advanced electronic modules such as automobile exhaust gas recirculation (EGR) systems (Koch et al., 2023), variable valve timing with intelligence (Demir et al., 2022), and other electronic systems (Li et al., 2018). These auxiliary systems provide additional degrees of freedom in engine development, but those variables would increase system complexity and further development cost (Li et al., 2021). Therefore, the fast and low-cost development of dedicated hybrid engines is an urgent need to help accelerate their commercialisation. Various modelling methods on the multiple engine performance indicators are applied to save many experimental costs and have proven to be an efficient approach for the rapid development of dedicated hybrid engines (Cervantes-Bobadilla et al., 2023; Liu, 2022; Zhao et al., 2023).

By summarising the previous research on engine modelling, the modelling methods could be divided into the three main groups, i.e., the physics-based white box modelling, the data-driven black box modelling, and fusion grey box modelling. Considering the physics phenomenon in the engines, the white box modelling applies the specific physical principles of the engines as the modelling limitations and further constructs the engine models. Jose et al. applied the computational fluid dynamics (CFD) simulation to build the twin-entry radial turbines of the engines. The physics-based one-dimensional model minimises the fitting parameters and keeps the prediction accuracy simultaneously (Galindo et al., 2021). In the work of Hao et al., CFD was used as the modelling method of engine combustion simulation. By using CFD, a new device named fuel split device (FSD) was validated (Hao et al., 2021). As another main method of white box modelling, the thermodynamic model displays another physic-based way to measure the performance of engines. By applying the thermodynamic model, various combustion processes could be described, e.g., multi-zone (Azarmanesh and Targhi, 2021), two-zone (Rakopoulos et al., 2020), and one-zone (Gautam et al., 2022). The high accuracy of these white box modelling methods extremely depends on the time-consuming parameter calibration and relevant expert experience. To keep the high accuracy of the model, more measurement data needs to be introduced (as the grey box model). Severin proposed a modular approach for the diesel engine air path control system based on the grey

* Corresponding author.
E-mail address: j.li.1@bham.ac.uk (J. Li).

https://doi.org/10.1016/j.engappai.2023.107114
Received 15 November 2022; Received in revised form 5 July 2023; Accepted 5 September 2023
Available online 13 September 2023

box model. By the accurate prediction of the grey box model, the significant engine pumping losses (Hanggi et al., 2022). In the research of Zhang et al., a grey box model is presented to analyse the swirl characteristics of a combustion system and help to achieve the efficient identification of faulty combustors (Zhang et al., 2020). Compared to the white box model, the grey box model simplifies the working process of engines. However, more data should be introduced to keep the high prediction accuracy of grey box modelling. To further accelerate the product development and save more experimental costs, the data-driven black box modelling methods are widely researched and used. The black box modelling methods establish representational relationships between inputs and outputs by fitting data. By using various advanced artificial intelligence technologies, black box modelling is rapidly and widely applied in different engineering applications. Meysam et al. presented an efficient hybrid deep learning model to improve the accuracy of daily solar radiation prediction (Alizamir et al., 2023). In the research of Jose M et al., a neural network modelling method is proposed to enhance the prediction ability of the anaerobic membrane bioreactor model, adding the accurate prediction cycles from 25 to 75 (Cámara et al., 2023). Similarly, the neural network model was applied in the failure prediction of mechanical components by Basheer (Shaheen et al., 2023). More research and applications of black box modelling have appeared in the industry, e.g., marine engineering (Coraddu et al., 2021), software engineering (Pachouly et al., 2022), and civil engineering (Arab et al., 2022). Meanwhile, the black box modelling methods have been widely applied in the entire auto industry, e.g., the design of supervisory controllers (J. Li et al., 2020b; Zhou et al., 2022a), driving pattern identification (J. Li et al., 2020a), fuel consumption prediction (Barbado and Corcho, 2022), emission prediction (Shin et al., 2020), and battery degradation prediction (Q. H.F. Zhou et al., 2021). In engine development, Li et al. propose a geometric neuro-fuzzy transfer learning for in-cylinder pressure modelling of a diesel engine fueled with raw microalgae oil. It helps reduce the operation time by 41.5% whilst keeping the prediction accuracy at a high level (Li et al., 2022a). To reduce experimental efforts, the Gaussian distributed resampling technique is developed to quantify the air mass flow through the engine (Li et al., 2022c). Saeid et al. combined a black box model and a physical model to predict the soot from a 4.5-liter compression ignition engine, achieving better prediction performance than the physical model (Shahpouri et al., 2021a). In addition to the applications mentioned above, black box modelling has been applied in the prediction models of various engine performance indicators, such as volumetric efficiency (Li et al., 2022d), emission, and knock density (Aliramezani et al., 2022).

The quality of the dataset notably influences the prediction performance of the data-driven black box modelling. The high quality of the dataset has been proven to support the data-driven model and further improve the prediction performance (Li et al., 2023). As the advanced technologies of data pre-processing, the various feature selection methods are applied to enhance the quality of the initial dataset, which is divided into three kinds in terms of 1) filter, 2) wrapper, and 3) embedded methods (Rong et al., 2019). Compared with the other two methods, filters assess the importance of model inputs rapidly based on the external criteria and save huge calculating time, displaying superior potential in the practical application (Chen et al., 2019; Xie et al., 2023). In the current research, various filters are applied to simplify the black box model structure and improve the modelling efficiency to remove the irrelevant and redundant inputs that cause the model's overfitting (Ma et al., 2023). Wu et al. proposed an optimised ReliefF to enhance target identification accuracy on the application of radar infrared combined sensors. By analysing the results, this proposed filter could achieve higher class separability and lower feature redundancy (Wu et al., 2022). Zhou et al. also proposed an optimised ReliefF algorithm combining the Decision Tree to develop the feature selection (H. F. Q. Zhou et al., 2021). Michel et al. used a filter based on γ-metric to detect the atrial fibrillation automatically. The feature selection by using this filter could help doctors to diagnose heart disease in real time (Michel et al., 2021). Because of different working principles, the assessments of the model inputs could not be kept the same by using different filters. The single filter was proven not to keep the robust performance in different cases. To address this, multi-filter applications and further filter fusion are widely used for engineering items. In the research of Chaudhary et al., an attack detection system to identify anomalous activities in the fog enabled IoT network was proposed. This system applied four different filters to achieve cooperative identification. In these filters, the minimum-redundancy-maximum-relevancy (mRMR) brought the best classification accuracy (Chaudhary et al., 2022). Omuya et al. developed a hybrid filter model for feature selection based on principal component analysis and information gain. This hybrid filter improves the classification accuracy significantly on the breast cancer data set, which is a public dataset (Odhiambo Omuya et al., 2021). Similarly, in the research of Balogun, a rank aggregation-based multi--filter feature selection method is proposed for software defect prediction. This method was proven to improve performance compared to the single filter (Balogun et al., 2021).

Summarising the background mentioned above, the filters own the superior potential to identify the important parameters for the modelling targets and help to reduce the model scale while the high prediction accuracy is kept (L. Li et al., 2020). Meanwhile, the requirement of calculating time in filters is significantly lower than those in other feature selection methods. The rapid solving process promotes the practical application of filters in the industry (Hassani et al., 2021; Zheng et al., 2023). These advantages of filters superbly match the urgent demand for the rapid and efficient development of dedicated engines. Though some researchers made attempts to apply the filters to the data-driven modelling of engines to optimise the balance of model complexity and performance (Kuzhagaliyeva et al., 2021; Mohammad, A., Rezaei, R., Hayduk, C., Delebinski, T., Shahpouri, S., & Shahbakhti, 2022; Shahpouri et al., 2021b), the cooperative fusion of different filters is still rarely introduced into the modelling process. Meanwhile, the combustion process of the internal combustion engine (ICE) is described as the extremely nonlinear and complex physical phenomenon, including various relevant parameters and performance indicators (Aliramezani et al., 2022). The generic feature selection methods for different performance indicators of ICEs are scarce. For each specific performance indicator, the feature selection method is designed purposely, bringing the huge time cost, and further hampering the rapid development of novel ICEs. Based on the pain points presented above, a generic feature selection method for the modelling of different engine performance indicators needs to be proposed and be valuable both in research and industry application.

It is clear from the literature review that the further development of engine prediction systems needs to overcome the following research gaps:

1). There are performance biases of different filters in different applications, bringing potential risks in the practical application.
2). For a dataset with insufficient characterisation, the random precision issues caused by data absence will be serious and hamper the effectiveness of the modelling method.
3). The development of assistive technologies has not been able to match the rapid growth of the industry, limiting the application of effective generic modelling methods for engine performance indicators.

To systematically address the technical challenges mentioned above, this paper proposed a K-fold filter-collaborated feature selection (KFFC-FS) approach to identify the key parameters of different engine performance indicators, i.e., volumetric efficiency, thermal efficiency, and fuel consumption. The existing feature selection approaches always neglect the robustness of identification in different cases and present unsatisfactory generality, which increasing the potential risk of usage in the practical applications. These negative consequences significantly

hamper the development of the feature selection methods in industry. To avoid these consequences, the proposed approach utilises the filter-collaborated feature selection method and the K-fold cross validation method to avoid the performance biases of different filters in different independent engine performance indicators, which reduce the potential risk of usage in industry. Based on the robust key parameter identification provided by the proposed approach, the development of the dedicated hybrid engine is accelerated. In this approach, the feature importance calculated by five filters, i.e., Random Forest (RF), ReliefF, Neighbourhood Component Analysis (NCA), F-test, and minimum redundancy maximum relevance (mRMR), are utilised to reassess the importance rank by considering the rank-based and weight-based ways, which avoids the feature overestimation caused by the single filter. Based on this rank, different feature combinations are used as the modelling dataset to train the MLP networks. For the independent engine performance indicators, the K-fold cross validation method ensures that all data is trained to avoid the random precision issue for the key parameter identification. Meanwhile, the K-fold cross validation method is introduced into the network training process to prevent overfitting. These two methods strengthen the robustness of key parameter identification in the proposed generic approach. The experimental validation is based on an experimental test bench with a BYD 1.5L gasoline engine. A comparative study is carried out in terms of 1) the feature importance ranks based on multiple filter collaboration, 2) the validation for the process robustness of feature importance ranking, and 3) the K-fold key parameter identification and related robustness validation. Three main contributions of this paper are drawn from the comprehensive investigation:

1). Compared to the specific filter feature selection method for the specific targets, a more generic filter collaborated feature selection method is originally designed to allow the feature importance calculated by five filters to reassess the importance rank by considering rank-based and weight-based ways, which avoids the feature overestimation caused by the single filter.
2). To avoid the identification bias caused by the random precision issue, a K-fold cross validation method is introduced in the entire feature selection process, strengthening the reliability of key parameter identification. Compared to the experimental method to increase sample size, this K-fold cross validation method save more experimental cost and calibration.
3). By using the experimental test bench with a BYD 1.5 L gasoline engine, the generality of this approach in the three studied engine performance indicators is verified.

Following by Introduction, the paper outline is organised into four main sections. Section 2 describes the main procedures of the proposed KFFC-FS approach. In Section 3, the data collection is described, including the Introduction of the working processes and the description of experimental data. Section 4 carries out a comparative analysis of the proposed approach in terms of 1) the feature importance ranks, 2) the performance of MLP networks, and 3) the key parameter identification. The conclusion is summarised in Section 5.

2. Methodology

Generally, adding the irrelevant and redundant model inputs expands the training dataset size and increases the training time significantly. Meanwhile, the irrelevant and redundant inputs create more noise and further mislead the model decisions. To address these issues, various feature selection methods are applied to remove these misleading features and assist the models in keeping the most appropriate balance between training time and training performance. In this research, the KFFC-FS approach is proposed to robustly identify the key parameters of dedicated hybrid engine performance indicators. Two main procedures of the proposed approach are followed as 1) the filter-collaborated feature importance ranking; 2) the K-fold robust key parameter identification. The workflow of the whole research is shown in Fig. 1.

2.1. Filter-collaborated feature importance

The filter is used as a feature selection method to provide external criteria to rapidly assess different features. These criteria reflect the importance of different features for the modelling targets. By comparing the importance, the importance-based feature ranks are obtained and applied as guidance for further key parameter identification. Due to different work principles, different filters calculate different importance for each feature and display different feature importance ranks in the same modelling cases. To avoid misleading guidance from the single filter, multiple filters are collaborated and provide a robust feature importance rank for different cases.

2.1.1. Single filters

In this research, five common filters, i.e., RF, ReliefF, NCA, F-test, and mRMR, are introduced to assess the importance of each feature and provide the importance-based feature ranks.

Random Forest (RF) Algorithm: The RF algorithm is an ensemble learning method for regression and classification, which combines a multitude of decision trees to ignore outliers and correct for decision trees' habit of overfitting (BREIMAN, 2001). Because of the better prediction accuracy and generalisation compared with the decision tree, the RF algorithm is introduced in this case. The procedures of Random Forest are as follows: 1) Some samples are randomly introduced into a bag consisting of many decision trees for training. 2) The rest of the samples are used as the out-of-bag (OOB) samples to assess the significance of the independent variables. 3) The OOB samples are randomly selected to replace the samples in the bag. The following change in error quantifies the feature importance. 4) The more pronounced the change in error, the higher the feature importance.

ReliefF Algorithm: ReliefF algorithm is an optimised Relief algorithm that is not sensitive to noise and quickly assigns weights to features based on the relevance of individual features and categories (Kononenko et al., 1997). The procedures of the ReliefF algorithm are as follows: 1) the ReliefF algorithm randomly selects one sample R at a time from the training set. 2) the ReliefF introduces k nearest-neighbour samples (H) of the same kind as R and the same number of samples m in the different kinds. 3) the distances between the nearest neighbour samples and R under the same feature are compared. If this distance is less than the distance between R and other different samples, then the feature distinguishes the different kinds of nearest neighbours. 4) the weight corresponding to the chosen feature should be increased. The formula for updating the weights is:

$$W(A) = W(A) - \sum_{j=1}^{k} d\left(A, R, H_j\right) \Big/ (mk) + \sum_{c \in C(R)} \left[\frac{p(C)}{1 - p(C(R))} \sum_{j=1}^{k} d\left(A, R, M_j(C)\right) \right] \Big/ mk \quad (1)$$

where R means the chosen sample; H means the samples in the same kind of chosen sample R; m means the number of the samples in the different kinds; k means the number of the nearest-neighbour samples; $d(A, R, H_j)$ denotes the difference between sample R and sample H_j in feature A, expressed as the Euclidean distance between them; $M_j(C)$ denotes the j_{th} nearest neighbour sample in class C.

Neighbourhood Component Analysis (NCA): NCA algorithm is based on the KNN algorithm with the Mahalanobis distance and learns the transformation matrix by continuously optimising the accuracy of the K-nearest neighbour (KNN) classification. Compared to the common KNN with other distance matrices, by introducing the Mahalanobis distance to keep the scale of features consistent, the NCA algorithm avoids the

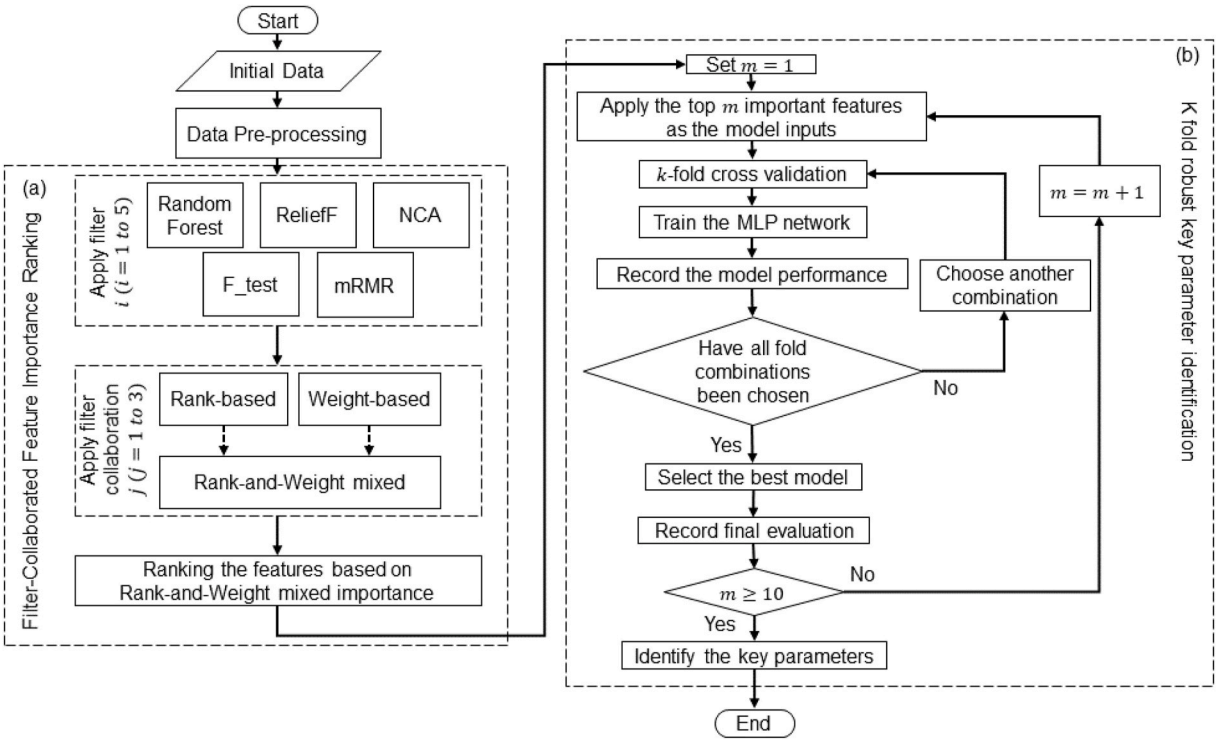

Fig. 1. The workflow of KFFC-FS including (a) the filter-collaborated feature importance ranking; (b) the k fold robust key parameter identification.

weights' calculation error (Jacob Goldberger et al., 2005). The full procedures of the NCA algorithm are similar to the ones of the ReliefF algorithm as follows:

1) The NCA algorithm introduces A as the Mahalanobis distance metric with the specific expression.
2) The NCA algorithm calculates the nearest-neighbour distribution for sample i.

$$P_{i,j} = \frac{exp\left(-\left|\left|Ax_i - Ax_j\right|\right|_2^2\right)}{\sum_{k \neq i} exp\left(-\left|\left|Ax_i - Ax_k\right|\right|_2^2\right)}, P_{i,i} = 0 \tag{2}$$

where k means the number of samples; P means the probability of nearest neighbour distribution; i and j mean the samples; k means the number of the nearest-neighbour samples.

3) The NCA algorithm Optimises the objective and obtains the weights of the specific features by averaging the regression values of the final samples. The formula for the optimisation objective is:

$$f(A) = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \sum_{j \in C_i} P_{i,j} \tag{3}$$

where $f(A)$ denotes the function of the Mahalanobis distance metric; P means the probability of nearest neighbour distribution; n means the total samples; i means the chosen sample; C means the class of i; j means the sample in C.

F-test: F-test is a statistical test method used to compare models (Fisher, 1923). For feature selection, F-test is applied to compare the model with features and constants and the one only with constants. By comparing the least square errors in both models, the key features are selected. The feature that brings smaller least square errors presents higher relevance and importance for the target variables. Due to the F-test feature selection highly depends on the relevance, which doesn't capture strong nonlinear relationships, F-test could not obtain the feature importance accurately and further identify key parameters efficiently.

Minimum Redundancy Maximum Relevance (mRMR): The minimum redundancy maximum relevance (mRMR) is a feature selection method based on mutual information (MI) (Hanchuan Peng, Member, IEEE, Fuhui Long, 2005). The core measures of mRMR are the related V_r calculated by F-statistic method and the redundancy W_r calculated by the Pearson correlation coefficient. By combining the relevance and the redundancy, the mutual information quotient (MIQ) could be expressed as follows:

$$(MIQ)_x = \frac{V_x}{W_x} \tag{4}$$

where x denotes the x_{th} feature; V_x denotes the relevance of the x_{th} feature; W_x denotes the redundancy of the x_{th} feature. By applying a greedy search, the *MIQ* of different features show the feature importance to help identify the key parameters of the target variable.

2.1.2. Multi-filter collaboration

Due to the different nonlinear relationships between model inputs and engine performance indicators studied, the single filter owns the risk of providing misleading guidance by displaying a wrong importance rank in different modelling cases. To avoid the negative influence caused by the possible misleading guidance, five studied filters collaborate to provide a robust importance-based feature rank for different performance indicators. These collaborations are shown as follows:

Rank-based feature importance: the rank-based feature importance rank depends on the feature ranks obtained using different single filters. The procedures are as follows:

1) A feature with higher importance achieves a higher score in each feature rank provided by different single filters. For example, the feature, which is *No*.1 in a feature rank, achieves 10 points. The one which is *No*.2, achieves 9 points. The rest of the features achieve the corresponding points in descending order. Based on this, the

importance of each feature is distinguished as a linear form. The unit gap of feature importance is kept the same, equal to 1. By using this type of collaboration, the extremely wrong assessment of features is further modified.

2) All weight-based scores of each feature among the feature importance ranks provided by all the single filters are combined. The combined formula is as follows:

$$I_r^k = \sum_{i=0}^{F} (S)_i^k \tag{5}$$

where I_r^k means the feature importance calculated in the rank-based filter collaboration; k means the k_{th} feature; F, $(F = 1\ to\ 5)$, means the different filters for feature selection, S means the feature score calculated by the single filter.

Weight-based feature importance: the weight-based feature importance rank depends on the normalised feature importance calculated by different single filters. The range of normalisation is $[0, 1]$. The opposite of the rank-based assessment, the normalisation of feature importance could scale the importance of each feature from different ranks into the same size. Meanwhile, the gap between the initial importance is retained in equal proportions. Based on this, the extremely correct assessments of feature importance in different ranks are kept. After the normalisation, the normalised importance of each feature in different ranks is combined. The combined formula is as follows:

$$I_w^k = \sum_{i=0}^{M} \left[n(I)_i^k \right] \tag{6}$$

where I_w^k means the feature importance calculated in the weight-based filter collaboration; I means the feature importance calculated by the single filter method as the form of the weight; n means the normalisation of the feature importance in the range of $[0, 1]$.

Rank-and-Weight mixed feature importance: The rank-based feature importance rank modifies the extremely wrong assessments of feature importance to some extent. Meanwhile, the weight-based feature importance rank keeps the extremely correct assessments of feature importance. Based on this, the further Rank-and-weight-mixed collaboration combining these two manners could strengthen the robustness of importance assessment in different cases and provide a robust feature importance rank to guide further key parameter identification. The multiple-combined formula is shown as follows:

$$I_h^k = \sum \left[n\left(I_r^k\right) + n\left(I_w^k\right) \right] \tag{7}$$

where I_h^k means the feature importance is calculated by further collaboration based on the above-mentioned collaborations.

2.2. Model structure and training

Based on the guidance of the filter-collaborated feature importance rank, the K-fold robust key parameter identification is introduced in this research. The K-fold cross validation approach provides a more robust training process for the MLP networks. Different combinations of features from the initial dataset are used as the model inputs of the MLP networks. By comparing the model performance, the key parameters of different targets are identified.

2.2.1. K-fold cross validation

For validating the performance of the prediction model, different cross validation approaches are widely applied (Geisser, 1974). In this research, the K-fold cross validation approach is applied to validate the model performance. This validation approach randomly divides the initial data into k groups, then train the prediction model on the $k-1$ groups and then test it on the k_{th} group. Ultimately with k partitions, the k prediction accuracies are recorded. By comparing these prediction accuracies, the model with highest accuracy is chosen as the best model. Based on this approach, the robustness of the model is further strengthened. In this way, a limited data set is effectively extended (Vanwinckelen and Blockeel, 2012). Fig. 2 displays the working process of the K-fold cross validation in this research. By using this validation approach, the selected prediction model is re-trained by the training data to achieve an efficient and robust modelling performance. In this paper, the K-fold cross validation method ensures that all data is trained in the MLP network during the feature selection process, preventing the identification bias caused by the random precision issue. This is different from the existing feature selection methods and provides stronger reliability.

2.2.2. Network specification

To search for the most appropriate balance between the prediction performance and training time, the features are combined as different combinations of model inputs based on the feature importance rank mentioned above. In this research, the MLP network is used as the prediction model to predict the studied engine performance indicators. Due to the superior ability to efficiently handle non-linearities with multidimensional discrete inputs, the MLP network is an efficient tool for modelling the dedicated hybrid engine performance indicators. Meanwhile, the layer connection of the MLP network is simplifier than other complex artificial neural networks, saving more computational time in the practical application. In our previous research (Li et al., 2022b), MLP networks with specific structures have been proven to own superior computing efficiency and prediction accuracy. Based on this, the same fully connected MLP network construction with two hidden layers is applied in this research. The related structure hyperparameters of the applied MLP network are shown in Table 1. Considering the prediction errors caused by the network training bias, all training tasks are repeated ten times separately. After the repetitive training, the mean performance measurement metrics are applied as the results to be analysed comparatively.

To search for the most appropriate balance between the prediction performance and training time, the features are combined as different combinations of model inputs based on the feature importance rank mentioned above. In this research, the MLP network is used as the prediction model to predict the studied engine performance indicators. Due to the superior ability to efficiently handle non-linearities with multidimensional discrete inputs, the MLP network is an efficient tool for modelling the dedicated hybrid engine performance indicators. Meanwhile, the layer connection of the MLP network is simplifier than other complex artificial neural networks, saving more computational time in the practical application. In our previous research (Li et al., 2022b), MLP networks with specific structures have been proven to own superior computing efficiency and prediction accuracy. Based on this, the same fully connected MLP network construction with two hidden layers is applied in this research. The related structure hyperparameters of the applied MLP network are shown in Table 1. Considering the prediction errors caused by the network training bias, all training tasks are repeated ten times separately. After the repetitive training, the mean performance measurement metrics are applied as the results to be analysed comparatively.

3. Experimental data collection

The experimental data in this research were collected from a test bench with a 4-cylinder, 1.5L gasoline engine. The engine was run under steady-state conditions at different operating points that covered engine torque and speed range. Considering the experimental setup and computational time, the entire research is under steady-state conditions. In the experimental dataset, 2732 samples, which include three studied engine performance indicators and other relative parameters, were recorded at different engine running conditions. The range of specifications at different operating points was as follows: 1000–6000 rpm for

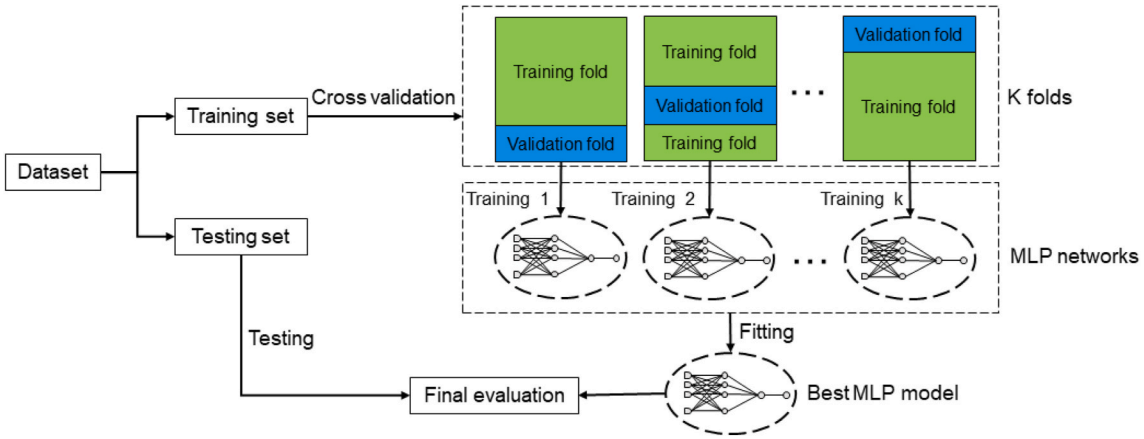

Fig. 2. The working process of the K-fold cross validation approach.

Table 1
Structure hyperparameters of the applied MLP network.

Hyperparameters		Determination
Learning algorithm		backpropagation
Activation function		tanh
Optimisation algorithm		gradient descent
Number of the hidden layers		2
Number of neurons in each hidden layer		100
Number of iterations		1000
Bias		1
Threshold on gradient norm		1e-3
Learning rate		0.01
Tykhonov hyperparameter		0.01
Momentum hyperparameter		0.5
Data split ratio	Training	75%
	Testing	15%
	Validating	10%

Table 2
Studied features of the dedicated hybrid engine.

Index	Parameter	Unit
1	S	rpm
2	i_{VVT}	CAD
3	P_{int}	kpa
4	P_{EGR}	%
5	P_{EXH}	kpa
6	T_{int}	$^{\circ}C$
7	T_{EGR}	$^{\circ}C$
8	A_{spr}	Degree
9	V_r	%
10	A_{inj}	Degree
11	VE	%
12	TE	%
13	FC	g/h

engine speed, 1.5–135 Nm for engine torque, and 0–100% for EGR positions. Engine speed variables at each steady-state point were obtained from the average of 600 points sampled at 10 Hz. The exact procedure of the experiment is referred to in the study by Li et al. (2022c). The principal diagram of the dedicated hybrid engine applied in this research is shown in Fig. 3.

The performance indicators of the engine are complex and nonlinear, which are influenced by relevant parameters, engine speed (S), variable valve timing with intelligence (i_{VVT}), intake manifold pressure (P_{int}), exhaust manifold pressure (P_{EXH}), intake manifold temperature (T_{int}), EGR position (P_{EGR}), EGR temperature (T_{EGR}), Spark angle (A_{spr}), Relative air volume (V_r), Inject angle (A_{inj}) and other parameters. Considering the practical application in the industry, these original physical features are more worthy of being studied than the interacted features. Meanwhile, these parameters were studied in our previous research. Based on these, the paper aims to find the key parameters from these original ten features to develop the modelling of dedicated hybrid engine performance indicators using the proposed approach. Table 2 shows the involved features and their feature numbers.

In this paper, volumetric efficiency (VE), thermal efficiency (TE), and fuel consumption (FC) are set as the target indicators of the data-driven modelling of the dedicated engine. The ten features mentioned above are considered as the contending candidates to be selected by the proposed feature selection approach. The selected features, as the key parameters of these target indicators, are used as the inputs of the prediction model based on the MLP network. By applying the prediction model, the prediction values of volumetric efficiency, thermal efficiency, and fuel consumption are obtained and further compared with the real values of these three indicators to assess the performance of the model. The configuration of the entire modelling process is shown in Fig. 4.

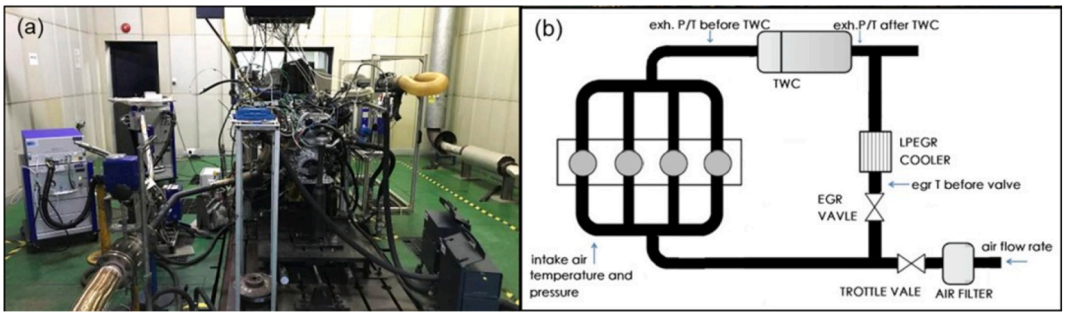

Fig. 3. Dedicated hybrid engine: a) testing bench and b) principal diagram (Li et al., 2022c).

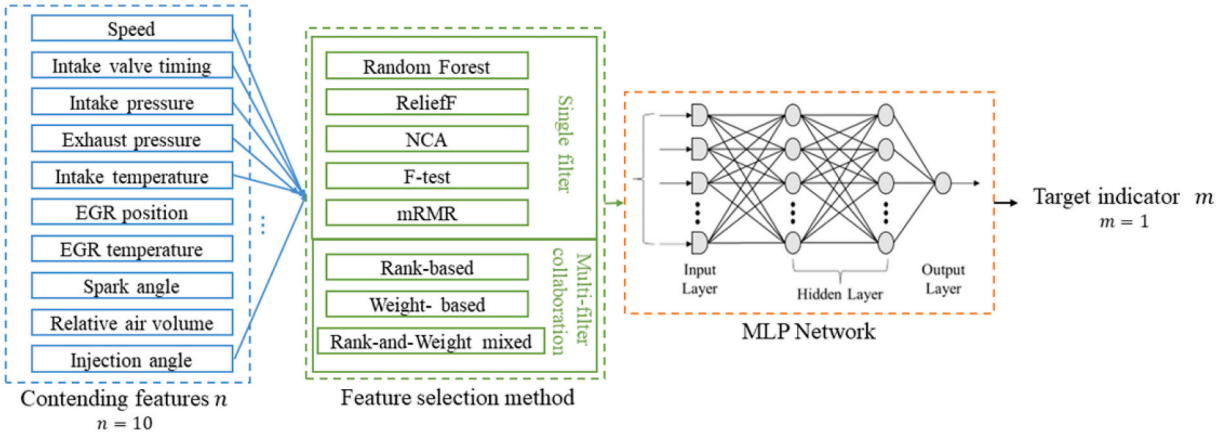

Fig. 4. The input-output configuration of the entire modelling process.

4. Result analysis and discussion

A comprehensive analysis is organised into three aspects: 1) the feature importance rank, 2) the adaptability of ranking approaches, and 3) the K-fold key parameter identification.

4.1. Evaluation of filter-collaborated feature importance

The filter-collaborated feature importance provides a guide for further key parameter identification. The collaborated importance rank depends on the importance assessment of studied features by using the five filters, i.e., RF, ReliefF, NCA, F-test, and mRMR. By using VE modelling as the example, this section investigates the feature importance ranks from different single filters and the procedure of further filter-collaborated feature importance ranks. Table 3 enumerates the importance rank of each feature obtained by single filters. The ranks are based on the order of strongest to weakest feature importance. The filter measures calculate different importance for each feature to cause a significant difference between the ranks of feature importance. However, by observing ranks, some commonalities can be found. The low importance of P_{EGR} and T_{EGR} shows that these two parameters are irrelevant parameters of VE modelling. The reason for their low importance is speculated that they influence exhaust air directly, not intake air, while VE is an index that reflects the intake air. On the contrary, V_r shows the highest importance because they influence the intake air volume directly. Based on these commonalities, the authors speculated that the parameters influencing intake air directly show high importance to be considered the relevant parameters of volumetric efficiency modelling.

As mentioned above, the single filter cannot keep the correct assessment of feature importance for different engine performance indicators. Multiple filter-collaborated feature importance rank is applied to avoid misleading guidance for further K-fold key parameter identification. Fig. 6 shows the combined importance calculated by a) the rank-based, b) the weight-based, and c) the rank-and-weight mixed collaboration.

By observing the combined feature importance shown in Fig. 5, the feature importance ranks based on the rank-based and the weight-based importance are similar. However, there are still minor differences between the two feature importance ranks. For fixing the possible incorrect importance assessments, the rank-and-weight mixed collaboration is used to further combine the feature importance and strength of the robustness. For VE modelling, V_r is proven as the most important parameter in each rank obtained by cooperative filtrations, while P_{EGR} and T_{EGR} shows the lowest importance.

For other performance indicators, the rank-and-weight mixed feature importance is summarised as a comprehensive and robust feature rank for assisting the further K-fold key parameter identification. Table 4 displays the filter-collaborated feature importance ranks for three performance indicators. For all engine performance indicators, the filter-collaborated feature importance ranks are significantly different. It proves that the relationships between performance indicators and the studied features are complex and nonlinear. The same feature displays the highest importance for an indicator whilst it is the end of the feature importance rank for another indicator. To address these issues, the rank-and-weight mixed feature importance is introduced to create robust feature ranks for all indicators.

Table 3
Feature importance ranks for volumetric efficiency modelling based on single filters.

Filter	Importance rank
RF	[2 10 9 1 5 3 4 8 7 6]
ReliefF	[2 1 9 3 8 7 6 5 4 10]
NCA	[8 2 9 1 5 7 6 10 4 3]
F-test	[5 9 10 3 6 1 8 7 4 2]
mRMR	[9 1 6 3 5 2 8 10 7 4]

4.2. Comparison of feature selection approaches

As the regression models, the data-driven models of the engine performance indicators could be assessed by different measurement metrics. In this paper, the comprehensive analysis of the feature selection approaches consists of the regression indicators and the industrial indicators. The mean absolute percentage error (MAPE), the root mean squared error (RMSE), and the mean coefficient of determination (R^2) are applied to assess the regression performance of the data-driven modelling process with different feature selection approaches. Considering the practical application of the proposed approach, the predicting pass rate is applied to measure the qualification of the prediction results. Under the established determination criteria, the forecasts with less than 5% relative error are satisfactory. The rate of satisfactory results to the total sample is considered to be the predicting pass rate. Combining the comparison among these different indicators, a comprehensive analysis is displayed.

By using VE modelling as an example, the feature combinations are chosen as the network inputs in order of the feature importance rank. After the repetitive training, MAPE, RMSE, and R^2 are used to measure the model performance, as shown in Fig. 6.

As shown in Fig. 6, the Random Forest algorithm shows an inefficient performance on feature selection for volumetric efficiency modelling, whilst mRMR selects key parameters precisely. It proves that there are significant differences among these mentioned filters. In contrast, the approach, based on the rank-and-weight mixed feature importance,

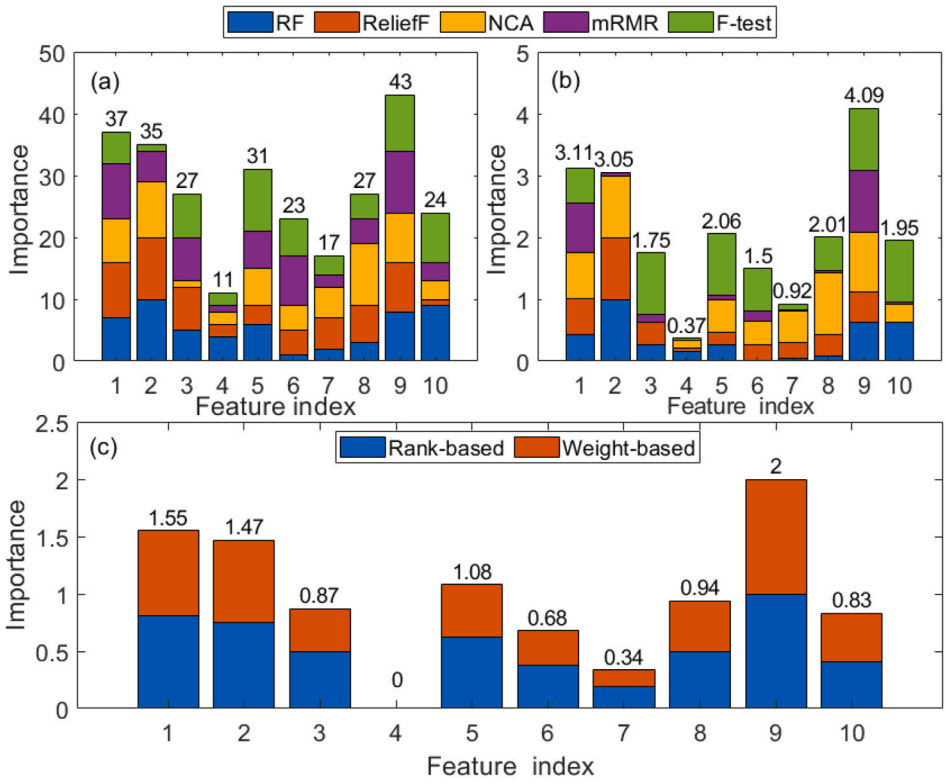

Fig. 5. Combined importance calculated by a) the rank-based, b) the weight-based, and c) the rank-and-weight mixed collaboration.

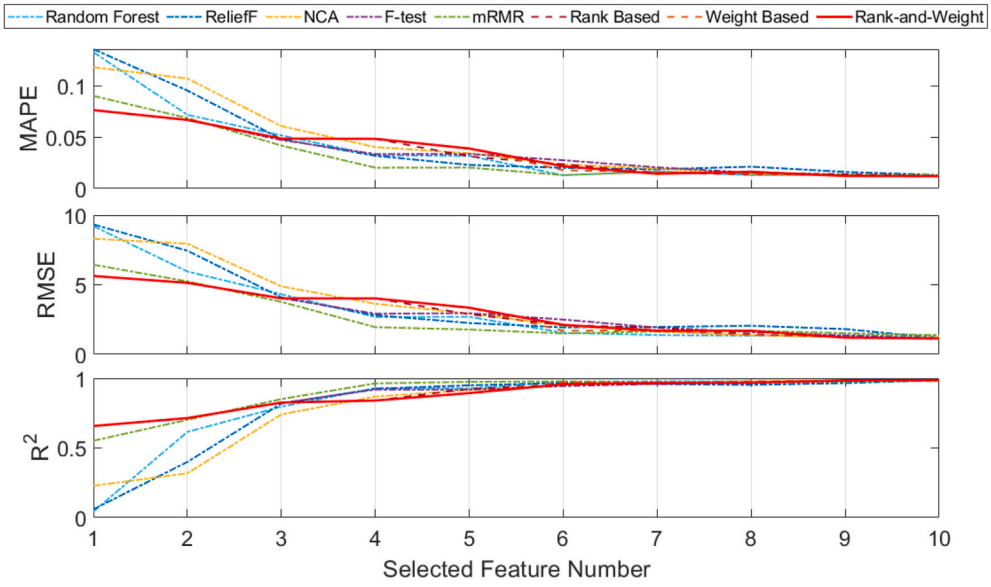

Fig. 6. The performance of the MLP network for volumetric efficiency modelling, including a) MAPE, b) RMSE, and c) R.^{2}.

Table 4
Rank-and-weight mixed feature importance ranks for three performance indicators.

Engine performance indicator	Importance rank
VE	[9 1 2 5 8 3 10 6 7 4]
TE	[8 10 1 7 3 5 6 9 4 2]
FC	[5 1 9 10 3 8 4 2 6 7]

could provide a correct importance rank to some extent. Based on this importance rank, the most important key parameters could be identified since the beginning of feature selection. Thus, while choosing fewer parameters, this approach selects the most relative parameters for VE, which are trained in the prediction models to achieve lower MAPE, lower RMSE, and higher R^2.

In real industrial applications, robustness is an important indicator to measure the performance of a novel approach. Considering this, the entire robustness during feature selection based on the feature importance ranks is also assessed. Based on each feature importance rank, the

average RMSE of the model with different input combinations is applied to reflect the robustness. These average RMSE values for different indicators are shown in Table 5. By using the underlines, the lowest average RMSE obtained by the guidance from the single filter and the filter-collaborated feature importance ranks are marked.

By observing Table 5, the performances of single filters in different cases display significant differences. For VE modelling, mRMR shows the best mean performance. However, for thermal efficiency and fuel consumption modelling, Random Forest shows the best performance. In contrast, based on the rank-and-weight mixed importance, the feature rank helps the MLP network to keep strong robustness in all these three modelling cases, almost achieving the lowest RMSE. The RMSE in each case is normalised and summed to be the overall RMSE among these three different cases. By comparing the overall RMSE of different filters, the rank-and-weight mixed feature importance is proven to robustly provide the correct feature rank for further K-fold key parameter identification.

As shown in Table 5, the proposed feature selection approach displays strong robustness for different engine performance indicators by quantizing the regression performance of the models as average RMSE. To compare the specific performance of different feature selection approaches more comprehensively, the average values of the regression indicators and the industrial indicator mentioned above are presented in Fig. 7. To directly display the superior performance of the proposed feature selection approach on different indicators, The reciprocals of MAPE and RMSE are used to replace MAPE and RMSE. Based on it, the areas of the quadrilaterals represent the overall performance of different feature selection approaches.

Fig. 7 evaluates the overall performance for each feature selection approach in the models of different engine performance indicators. Combining the results in Table 5, the single filter is further proven to keep the weak robustness in the models of different engine performance indicators. For example, mRMR displays superior overall performance on the regression indicators and the industrial indicator for volumetric efficiency modelling, while losing its advantage for thermal efficiency and fuel consumption modelling. Even the cooperation considered by the rank-based and weight-based ways could not maintain the overall performance in the models of different performance indicators. Though, the proposed mixed approach could not display the best performance in all models, it provides an importance rank with the strongest robustness to the modelling process and keeps the overall performance of models for different indicators at a satisfactory level.

4.3. Non-K-fold vs K-fold robust key parameter identification

In this paper, the K-fold cross validation approach mentioned above is used to fit the MLP network and find the best MLP network model. Based on the previous research, the value of k is set to 9 (Zhou et al., 2022b). Compared to the normal modelling approach, the K-fold cross validation approach provides a general improvement of the data-driven models for different engine performance indicators. The comparison between the non-K-fold modelling and K-fold modelling is listed in this section. As shown in Fig. 8, the improvement caused by the K-fold cross validation approach in the MLP model of VE trained by the initial dataset (10 inputs) is displayed.

Table 5
The average RMSE for different performance indicator modelling.

	Average RMSE*			
	VE	TE	FC	Overall
Random Forest	3.178	2.549	0.486	0.221
ReliefF	3.487	2.698	0.621	0.717
NCA	3.523	2.766	0.775	1
F-test	2.920	2.580	0.531	0.213
mRMR	2.693	2.576	0.605	0.194
Rank-based	2.953	2.561	0.600	0.270
Weight- based	2.929	2.552	0.465	0.103
Rank-and-Weight mixed	2.920	2.552	0.461	0.095

Note*: The average RMSE is calculated from the MLP networks with different input combinations.

Fig. 8 displays the volumetric efficiency prediction performance comparison between Non-K-fold and K-fold cross validation approaches. Fig. 8 (a) and (b) show the real-time prediction performance by using different approaches. Based on the results, the K-fold cross validation approach improves the prediction accuracy of the MLP model and significantly reduces the fluctuation of the error. A similar conclusion could be found in Fig. 8 (c), the K-fold cross validation could narrow the margin of errors to further improve the prediction performance. In conjunction with Fig. 8 (d), the prediction samples obtained by the K-fold cross validation approach are closer to the baseline, showing superior regression performance. Based on the comprehensive comparison in Fig. 8, the K-fold cross validation approach is proven to own the superior potential for the optimisation of data-driven modelling.

The improvement of the K-fold cross validation approach is not only significant in the MLP model trained by the initial dataset but also for the entire feature selection process. By presenting RMSE and predicting pass rate of the models for all three engine performance indicators, the performance comparison between the non-K-fold modelling and K-fold modelling during the entire feature selection process is presented in Fig. 9.

From Fig. 9, the K-fold cross validation approach is proven to bring a significant improvement in prediction performance during the entire modelling process for different engine performance indicators, reducing the average RMSE effectively (VE: 2.52, TE: 2.42, FC: 0.39). Meanwhile, the improvement in predicting pass rate displays the valuable potential of the proposed approach in the practical application, meeting the higher industrial requirement. Even though for some feature combinations, the optimisation by using the K-fold cross validation approach is not significant, the overall gap between the K-fold approach and non-K-fold approach, which bring the lowest RMSE and higher predicting rate, is obvious. This is useful for key parameter identification both in the lab and the industry.

To keep the most appropriate balance between training time and performance, the parameter-number-related RMSE (E_p) is proposed. To keep the same scalar, the RMSE is normalised in the range [1,10]. When E_p is lower, the selected feature could help the models keep a more appropriate balance between training time and performance. This assessment variable is calculated as follows:

$$\left(E_p\right)_i = i * \left[n(RMSE)_i\right], 1 \le i \le 10 \tag{8}$$

where i means the number of selected features, n means the normalisation process. The normalisation range is $[1, 10]$.

Based on this assessment variable, the performance of K-fold MLP models trained by different feature combinations could be further quantified to assist the key parameter identification. Table 6 shows the parameter-number-related RMSE for different performance indicator modelling.

Compared to the lowest parameter-number related RMSE obtained by the non-K-fold approach for all indicators, the lowest parameter-number related RMSE obtained by the K-fold approach is further reduced (VE: 0.84, TE: 0.40, FC: 0.30). According to the lowest parameter-number related RMSE, the key parameters of different indicators are identified. Combined with the superior performance of the K-fold filter-collaborated feature selection approach for all indicators, the proposed approach is proven to own strong robustness. The aim of robust key parameter identification has been achieved.

5. Conclusions

This paper proposed the novel feature selection approach of KFFC-FS

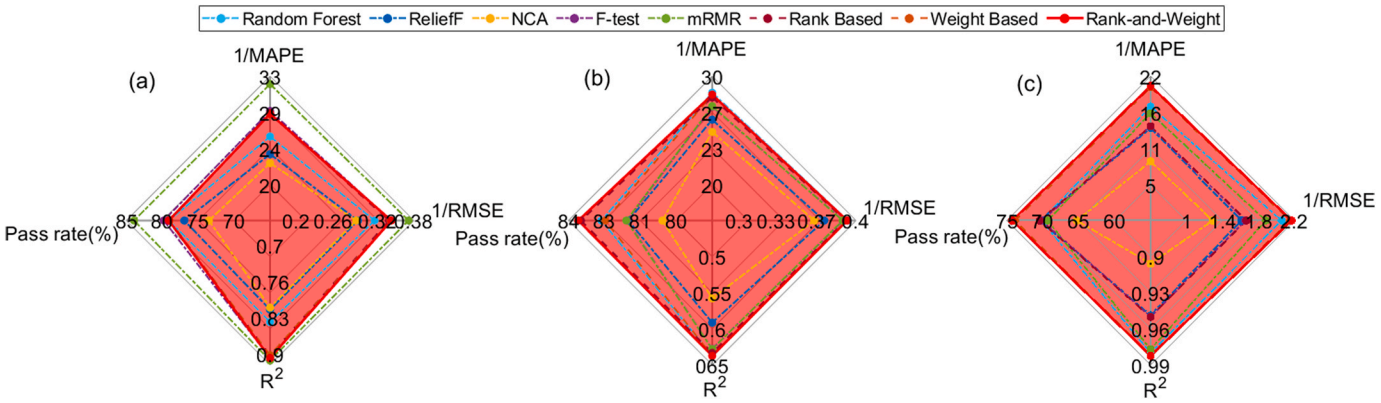

Fig. 7. Analysis of overall performance for each feature selection approach of different engine performance indicators (a) volumetric efficiency, (b) thermal efficiency, and (c) fuel consumption.

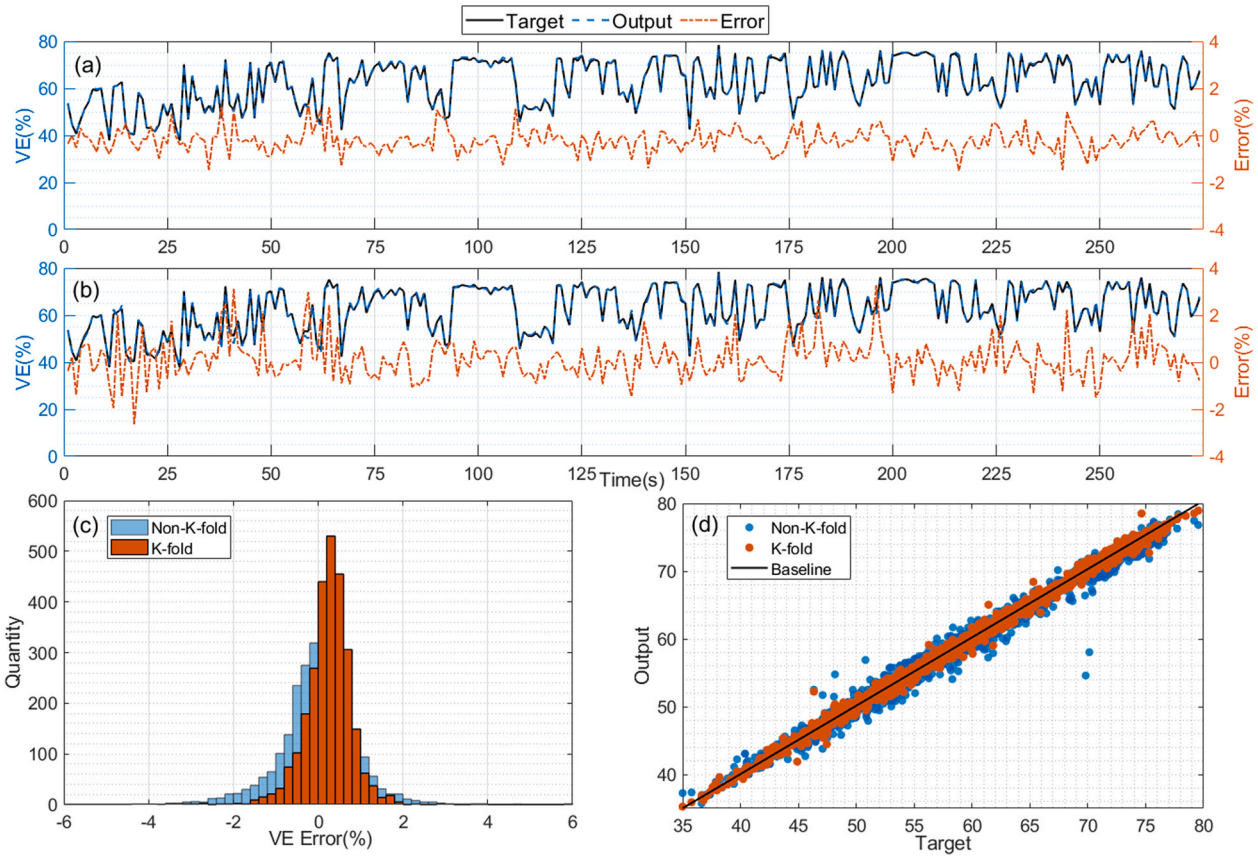

Fig. 8. Volumetric efficiency prediction performance comparison between Non-K-fold and K-fold cross validation approach. (a) Real-time prediction performance by using K-fold cross validation approach. (b) Real-time prediction performance by using Non-K-fold cross validation approach. (c) Error distribution. (d) Regression performance.

for dedicated hybrid engine performance indicators that robustly identify the key parameters of different performance indicators. This approach includes a generic feature selection method and the K-fold cross validation method. The feature importance calculated by five filters is utilised to reassess the importance rank by considering the rank-based and weight-based ways. This effectively prevents the feature overestimation caused by single filters. To minimise the random precision and overfitting issues for studied engine performance indicators, the K-fold cross validation method is introduced in the entire feature selection process. Based on this kind of cross validation, the robustness of the key parameter identification is further strengthened. By validating the experimental data provided by BYD, the proposed approach is comprehensively assessed in three aspects: 1) the feature importance rank, 2) the adaptability of ranking approaches, and 3) the key parameter identification. The contributions from the assessment are as follows:

1) Compared to the single filters, the proposed filter-collaborated feature importance ranking approach could keep the robust influence to provide the correct feature importance ranks in different indicator modelling.
2) The filter-collaborated feature importance ranking method has good adaptability to the studied engine performance indicators and obtained the lowest overall average RMSE (0.095) for these indicators.

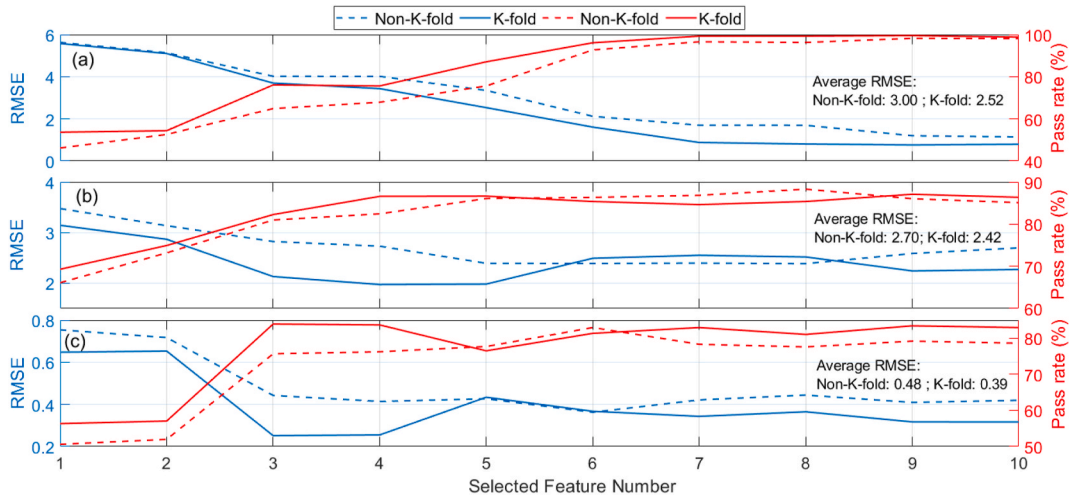

Fig. 9. The RMSE and predicting pass rate of MLP network by using non-K-fold and K-fold cross validation approaches for (a) volumetric efficiency, (b) thermal efficiency, and (c) fuel consumption.

Table 6
The parameter-number-related RMSE for different performance indicator modelling.

Performance indicator	Validation approach	Selected feature number									
		1	2	3	4	5	6	7	8	9	10
VE	**Non-K-fold**	1.00	1.80	2.03	2.71	2.71	1.78	1.48	1.69	1.00	1.00
	K-fold	1.00	1.75	1.87	2.31	2.08	1.32	0.84	0.86	0.90	1.06
TE	**Non-K-fold**	1.00	1.44	1.39	1.54	0.53	0.62	0.76	0.80	2.39	3.58
	K-fold	1.00	1.58	0.66	0.40	0.53	2.99	3.81	4.15	2.76	3.28
FC	**Non-K-fold**	1.00	1.83	0.85	0.87	1.27	0.60	1.64	2.30	1.87	2.31
	K-fold	0.99	2.00	0.30	0.43	2.54	2.16	2.14	2.83	2.21	1.42

3) Compared to the average RMSE (VE: 3.00, TE: 2.70, FC: 0.48) obtained by the MLP networks without the K-fold cross validation approach during the feature selection, the average RMSE (VE: 2.52, TE: 2.42, FC: 0.39) obtained by the MLP network trained by K-fold cross validation approach is reduced by 16%, 10.3% and 18.75% for VE, TE and FC modelling, separately.
4) By using the proposed approach, the key parameters for all performance indicators have been found, with reducing the model inputs by 30%, 60%, and 70% for VE, TE, and FC modelling, respectively.

The research presents a holistic solution to robustly identify the key parameters for the dedicated hybrid engine prediction system. Though the effectiveness of this solution has been proven in the validation by using different engine performance indicators, some aspects of a comprehensive consideration are neglected. These are considered as the research directions in the future. In terms of the practical application, maintaining the strong robustness in the modelling with multiple outputs is worthy of being studied to further save the experimental cost in the development of the dedicated hybrid engine. The impact of the fold quantity should be included in further investigation to discover the optimal fold quantity to further improve the effectiveness of the key parameter identification. Besides, the generalization of the proposed approach should be verified in the more complex input-output configuration, including more contending features and more independent engine performance indicators. These all are worthy to be studied in future work.

CRediT authorship contribution statement

Xu He: Conceptualization, Methodology, Software, Writing – original draft. **Ji Li:** Conceptualization, Resources, Supervision, Writing – review & editing. **Quan Zhou:** Investigation, Formal analysis. **Guoxiang Lu:** Investigation, Funding acquisition. **Hongming Xu:** Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgement

The reported work in this paper is supported by the project entitled AI model for Dedicated Hybrid Engines. The authors acknowledge the funding provided by BYD Auto Ltd, Guangzhou City, China (Grant No.: 1001636). The authors also express gratitude to Birmingham CASE-V Automotive Research and Education Centre.

References

Aliramezani, M., Koch, C.R., Shahbakhti, M., 2022. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: a review and future directions. Prog. Energy Combust. Sci. 88, 100967 https://doi.org/10.1016/j.pecs.2021.100967.

Alizamir, M., Shiri, J., Fard, A.F., Kim, S., Gorgij, A.R.D., Heddam, S., Singh, V.P., 2023. Improving the accuracy of daily solar radiation prediction by climatic data using an efficient hybrid deep learning model: Long short-term memory (LSTM) network coupled with wavelet transform. Eng. Appl. Artif. Intell. 123, 106199 https://doi.org/10.1016/j.engappai.2023.106199.

Arab, M., Akbarian, H., Gheibi, M., Akrami, M., Fathollahi-Fard, A.M., Hajiaghaei-Keshteli, M., Tian, G., 2022. A soft-sensor for sustainable operation of coagulation

and flocculation units. Eng. Appl. Artif. Intell. 115, 105315 https://doi.org/10.1016/j.engappai.2022.105315.
Azarmanesh, S., Targhi, M.Z., 2021. Comparison of laser ignition and spark plug by thermodynamic simulation of multi-zone combustion for lean methane-air mixtures in the internal combustion engine. Energy 216. https://doi.org/10.1016/j.energy.2020.119309.
Balogun, A.O., Basri, S., Mahamad, S., Abdulkadir, S.J., Capretz, L.F., Imam, A.A., Almomani, M.A., Adeyemo, V.E., Kumar, G., 2021. Empirical analysis of rank aggregation-based multi-filter feature selection methods in software defect prediction. Electron 10, 1–16. https://doi.org/10.3390/electronics10020179.
Barbado, A., Corcho, Ó., 2022. Interpretable machine learning models for predicting and explaining vehicle fuel consumption anomalies. Eng. Appl. Artif. Intell. 115, 105222 https://doi.org/10.1016/j.engappai.2022.105222.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35.
Cámara, J.M., Diez, V., Ramos, C., 2023. Neural network modelling and prediction of an anaerobic filter membrane bioreactor. Eng. Appl. Artif. Intell. 118, 105643 https://doi.org/10.1016/j.engappai.2022.105643.
Cervantes-Bobadilla, M., García-Morales, J., Saavedra-Benítez, Y.I., Hernández-Pérez, J.A., Adam-Medina, M., Guerrero-Ramírez, G.V., Escobar-Jímenez, R.F., 2023. Multiple fault detection and isolation using artificial neural networks in sensors of an internal combustion engine. Eng. Appl. Artif. Intell. 117, 105524 https://doi.org/10.1016/j.engappai.2022.105524.
Chaudhary, P., Gupta, B., Singh, A.K., 2022. Implementing attack detection system using filter-based feature selection methods for fog-enabled IoT networks. Telecommun. Syst. 81, 23–39. https://doi.org/10.1007/s11235-022-00927-w.
Chen, L., Zhang, A., Lou, X., 2019. Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning. Expert Syst. Appl. 137, 266–280. https://doi.org/10.1016/j.eswa.2019.02.005.
Coraddu, A., Oneto, L., Ilardi, D., Stoumpos, S., Theotokatos, G., 2021. Marine dual fuel engines monitoring in the wild through weakly supervised data analytics. Eng. Appl. Artif. Intell. 100, 104179 https://doi.org/10.1016/j.engappai.2021.104179.
Demir, U., Coskun, G., Soyhan, H.S., Turkcan, A., Alptekin, E., Canakci, M., 2022. Effects of variable valve timing on the air flow parameters in an electromechanical valve mechanism – a cfd study. Fuel 308, 121956. https://doi.org/10.1016/j.fuel.2021.121956.
Ehsani, M., Singh, K.V., Bansal, H.O., Mehrjardi, R.T., 2021. State of the art and trends in electric and hybrid electric vehicles. Proc. IEEE 109, 967–984. https://doi.org/10.1109/JPROC.2021.3072788.
Fisher, R., 1923. XXI.—on the dominance ratio. Proc. R. Soc. Edinburgh 42, 321–341.
Galindo, J., Serrano, J.R., García-Cuevas, L.M., Medina, N., 2021. Using a CFD analysis of the flow capacity in a twin-entry turbine to develop a simplified physics-based model. Aero. Sci. Technol. 112, 106623 https://doi.org/10.1016/j.ast.2021.106623.
Gautam, P.S., Vishnoi, P.K., Gupta, V.K., 2022. A single zone thermodynamic simulation model for predicting the combustion and performance characteristics of a CI engine and its validation using statistical analysis. Fuel 315. https://doi.org/10.1016/j.fuel.2022.123285.
Geisser, S., 1974. A predictive approach to the random effect model. Biometrika 61, 101–107. https://doi.org/10.1093/biomet/61.1.101.
Goldberger, Jacob, Roweis, Sam, Hinton, Geoff, Salakhutdinov, Ruslan, 2005. Neighbourhood components analysis. Adv. Neural Inf. Process. Syst.
Hanggi, S., Frey, J., van Dooren, S., Diehl, M., Onder, C.H., 2022. A modular approach for diesel engine air path control based on nonlinear MPC. IEEE Trans. Control Syst. Technol. https://doi.org/10.1109/TCST.2022.3228203.
Hao, C., Lu, Z., Feng, Y., Bai, H., Wen, M., Wang, T., 2021. Optimization of fuel/air mixing and combustion process in a heavy-duty diesel engine using fuel split device. Appl. Therm. Eng. 186, 116458 https://doi.org/10.1016/j.applthermaleng.2020.116458.
Hassani, H., Hallaji, E., Razavi-Far, R., Saif, M., 2021. Unsupervised concrete feature selection based on mutual information for diagnosing faults and cyber-attacks in power systems. Eng. Appl. Artif. Intell. 100, 104150 https://doi.org/10.1016/j.engappai.2020.104150.
Koch, L., Picerno, M., Badalian, K., Lee, S.Y., Andert, J., 2023. Automated function development for emission control with deep reinforcement learning. Eng. Appl. Artif. Intell. 117, 105477 https://doi.org/10.1016/j.engappai.2022.105477.
Kononenko, I., Šimec, E., Robnik-Šikonja, M., 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl. Intell. 7, 39–55. https://doi.org/10.1023/A:1008280620621.
Kuzhagaliyeva, N., Thabet, A., Singh, E., Ghanem, B., Sarathy, S.M., 2021. Using deep neural networks to diagnose engine pre-ignition. Proc. Combust. Inst. 38, 5915–5922. https://doi.org/10.1016/j.proci.2020.10.001.
Li, J., Li, Z., Zhou, Q., Zhang, Y., Xu, H., 2018. Improved scheme of membership function optimisation for fuzzy air-fuel ratio control of GDI engines. IET Intell. Transp. Syst. 13, 209–217. https://doi.org/10.1049/iet-its.2018.5013.
Li, L., Zheng, Y., Yang, M., Leng, J., Cheng, Z., Xie, Y., 2020. A survey of feature modeling methods : historical evolution and new development. Robot. Comput. Integrated Manuf. 61, 101851 https://doi.org/10.1016/j.rcim.2019.101851.
Li, J., Zhou, Q., He, Y., Williams, H., Xu, H., 2020a. Driver-identified supervisory control system of hybrid electric vehicles based on spectrum-guided fuzzy feature extraction. IEEE Trans. Fuzzy Syst. 28, 2691–2701. https://doi.org/10.1109/TFUZZ.2020.2972843.
Li, J., Zhou, Q., Williams, H., Xu, H., 2020b. Back-to-back competitive learning mechanism for fuzzy logic based supervisory control system of hybrid electric vehicles. IEEE Trans. Ind. Electron. 67, 8900–8909. https://doi.org/10.1109/TIE.2019.2946571.
Li, J., Zhou, Q., Williams, H., Xu, H., Du, C., 2021. Cyber-physical data fusion in surrogate-assisted strength pareto evolutionary algorithm for PHEV energy management optimization. IEEE Trans. Ind. Inf.
Li, J., Wu, D., Mohammadsami Attar, H., Xu, H., 2022a. Geometric neuro-fuzzy transfer learning for in-cylinder pressure modelling of a diesel engine fuelled with raw microalgae oil. Appl. Energy 306. https://doi.org/10.1016/j.apenergy.2021.118014.
Li, J., Zhou, Q., Williams, H., Lu, G., Xu, H., 2022b. Statistics-guided accelerated swarm feature selection in data-driven soft sensors for hybrid engine performance prediction. IEEE Trans. Ind. Inf. 1–11. https://doi.org/10.1109/TII.2022.3199259.
Li, J., Zhou, Q., Williams, H., Xu, P., Xu, H., Lu, G., 2022c. Fuzzy-tree-constructed data-efficient modelling methodology for volumetric efficiency of dedicated hybrid engines. Appl. Energy 310. https://doi.org/10.1016/j.apenergy.2022.118534.
Li, J., Zhou, Q., Williams, H., Xu, P., Xu, H., Lu, G., 2022d. Fuzzy-tree-constructed data-efficient modelling methodology for volumetric efficiency of dedicated hybrid engines. Appl. Energy 310. https://doi.org/10.1016/j.apenergy.2022.118534.
Li, J., Zhou, Q., He, X., Chen, W., Xu, H., 2023. Data-driven enabling technologies in soft sensors of modern internal combustion engines: perspectives. Energy 272, 1–9. https://doi.org/10.1016/j.energy.2023.127067.
Liu, J., 2022. Gas path fault diagnosis of aircraft engine using HELM and transfer learning. Eng. Appl. Artif. Intell. 114, 105149 https://doi.org/10.1016/j.engappai.2022.105149.
Lu, G., Yang, D., Rong, Y., Gong, Z., Wang, B., 2021. Development of an intelligent thermal management system for BYD DM-i hybrid engine. SAE Tech. Pap 1–10. https://doi.org/10.4271/2021-01-1153.
Ma, H., Peng, T., Zhang, C., Ji, C., Li, Y., Nazir, M.S., 2023. Developing an evolutionary deep learning framework with random forest feature selection and improved flow direction algorithm for NOx concentration prediction. Eng. Appl. Artif. Intell. 123, 106367 https://doi.org/10.1016/j.engappai.2023.106367.
Michel, P., Ngo, N., Pons, J.F., Delliaux, S., Giorgi, R., 2021. A filter approach for feature selection in classification: application to automatic atrial fibrillation detection in electrocardiogram recordings. BMC Med. Inf. Decis. Making 21, 1–17. https://doi.org/10.1186/s12911-021-01427-8.
Mohammad, A., Rezaei, R., Hayduk, C., Delebinski, T., Shahpouri, S., Shahbakhti, M., 2022. Physical-oriented and machine learning-based emission modeling in a diesel compression ignition engine: dimensionality reduction and regression. Int. J. Engine Res.
Odhiambo Omuya, E., Onyango Okeyo, G., Waema Kimwele, M., 2021. Feature selection for classification using principal component analysis and information gain. Expert Syst. Appl. 174, 114765 https://doi.org/10.1016/j.eswa.2021.114765.
Pachouly, J., Ahirrao, S., Kotecha, K., Selvachandran, G., Abraham, A., 2022. A systematic literature review on software defect prediction using artificial intelligence: datasets, Data Validation Methods, Approaches, and Tools. Eng. Appl. Artif. Intell. 111, 104773 https://doi.org/10.1016/j.engappai.2022.104773.
Peng, Hanchuan, Member, IEEE, Fuhui Long, C.D., 2005. Feature selection based on mutual information: criteria of max-dependency. Max-Relevance, and Min-Redundancy 27, 1226–1238. https://doi.org/10.1109/cita.2015.7349827.
Rakopoulos, D.C., Rakopoulos, C.D., Kosmadakis, G.M., Giakoumis, E.G., 2020. Exergy assessment of combustion and EGR and load effects in DI diesel engine using comprehensive two-zone modeling. Energy 202. https://doi.org/10.1016/j.energy.2020.117685.
Rong, M., Gong, D., Gao, X., 2019. Feature selection and its use in big data: challenges, methods, and trends. IEEE Access 7, 19709–19725. https://doi.org/10.1109/ACCESS.2019.2894366.
Shaheen, B., Kocsis, Á., Németh, I., 2023. Data-driven failure prediction and RUL estimation of mechanical components using accumulative artificial neural networks. Eng. Appl. Artif. Intell. 119, 105749 https://doi.org/10.1016/j.engappai.2022.105749.
Shahpouri, S., Norouzi, A., Hayduk, C., Rezaei, R., Shahbakhti, M., Koch, C.R., 2021a. Soot emission modeling of a compression ignition engine using machine learning. IFAC-PapersOnLine 54, 826–833. https://doi.org/10.1016/j.ifacol.2021.11.274.
Shahpouri, S., Norouzi, A., Hayduk, C., Rezaei, R., Shahbakhti, M., Koch, C.R., 2021b. Hybrid Machine Learning Approaches and a Systematic Model Ignition Engines.
Shin, S., Lee, Y., Kim, M., Park, J., Lee, S., Min, K., 2020. Deep neural network model with Bayesian hyperparameter optimization for prediction of NO x at transient conditions in a diesel engine. Eng. Appl. Artif. Intell. 94, 103761 https://doi.org/10.1016/j.engappai.2020.103761.
Vanwinckelen, G., Blockeel, H., 2012. On estimating model accuracy with repeated cross-validation. 21st Belgian-Dutch Conf. Mach. Learn. 39–44.
Wu, Y., Liu, G., Li, Y., Jiang, M., 2022. Filter-based feature ranking technique for target recognition by radar infrared combined sensors. IET Radar, Sonar Navig. 16, 182–192. https://doi.org/10.1049/rsn2.12175.
Xie, J., Sage, M., Zhao, Y.F., 2023. Engineering Applications of Artificial Intelligence Feature selection and feature learning in machine learning applications for gas turbines : a review. Eng. Appl. Artif. Intell. 117, 105591 https://doi.org/10.1016/j.engappai.2022.105591.
Zhang, Y., Martínez-García, M., Kalawsky, R.S., Latimer, A., 2020. Grey-box modelling of the swirl characteristics in gas turbine combustion system. Meas. J. Int. Meas. Confed. 151, 107266 https://doi.org/10.1016/j.measurement.2019.107266.
Zhao, K., Jia, Z., Jia, F., Shao, H., 2023. Multi-scale integrated deep self-attention network for predicting remaining useful life of aero-engine. Eng. Appl. Artif. Intell. 120, 105860 https://doi.org/10.1016/j.engappai.2023.105860.
Zheng, Z., Ali, M., Jamei, M., Xiang, Y., Karbasi, M., Yaseen, Z.M., Farooque, A.A., 2023. Design data decomposition-based reference evapotranspiration forecasting model: a soft feature filter based deep learning driven approach. Eng. Appl. Artif. Intell. 121, 105984 https://doi.org/10.1016/j.engappai.2023.105984.

Zhou, Q., Wang, C., Sun, Z., Li, J., Williams, H., Xu, H., 2021. Human-knowledge-augmented Gaussian process regression for state-of-health prediction of lithium-ion batteries with charging curves. ASME. J. Electrochem. En. Conv. Stor. 18 (3), 030907. https://doi.org/10.1115/1.4050798 (April 29, 2021).

Zhou, H.F., Zhang, J.W., Zhou, Y.Q., Guo, X.J., Ma, Y.M., 2021. A feature selection algorithm of decision tree based on feature weight. Expert Syst. Appl. 164, 113842 https://doi.org/10.1016/j.eswa.2020.113842.

Zhou, Q., Li, Y., Zhao, D., Li, J., Williams, H., Xu, H., Yan, F., 2022a. Transferable representation modelling for real-time energy management of the plug-in hybrid vehicle based on k-fold fuzzy learning and Gaussian process regression. Appl. Energy 305. https://doi.org/10.1016/j.apenergy.2021.117853.

Zhou, Q., Li, Y., Zhao, D., Li, J., Williams, H., Xu, H., Yan, F., 2022b. Transferable representation modelling for real-time energy management of the plug-in hybrid vehicle based on k-fold fuzzy learning and Gaussian process regression. Appl. Energy 305. https://doi.org/10.1016/j.apenergy.2021.117853.