Energy and AI 12 (2023) 100229

Contents lists available at ScienceDirect

Energy and AI

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

A novel feature susceptibility approach for a PEMFC control system based on an improved XGBoost-Boruta algorithm

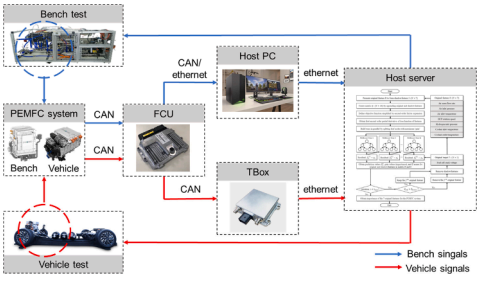
Xinjie Yuan [a,c], Fujun Chen [b], Zenggang Xia [a], Linlin Zhuang [a], Kui Jiao [b,*], Zhijun Peng [b], Bowen Wang [b], Richard Bucknall [c], Konrad Yearwood [c], Zhongjun Hou [a,*]

[a] Shanghai Hydrogen Propulsion Technology Co., Ltd., 1788 Xiechun Road, Shanghai 201804, China
[b] State Key Laboratory of Engines, Tianjin University, 135 Yaguan Road, Tianjin 300350, China
[c] Department of Mechanical Engineering, University College London (UCL), Torrington Place, London WC1E 7JE, United Kingdom

HIGHLIGHT

- Development of the new XGBoost-Boruta algorithm to evaluate feature importance of key BOP features in a PEMFC control system at different current densities.
- Design of a self-adaptive feature selection strategy for PEMFC system performance prediction.
- Comparative analysis between the proposed XGBoost-Boruta algorithm with conventional feature dropout strategy.
- Verification and validation evaluation based on real-time PEMFC bench and vehicle data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Boruta
Extreme gradient boosting (XGBoost)
Feature selection
Proton exchange membrane fuel cell (PEMFC)

ABSTRACT

Data-driven modelling methods are being developed in the quest to achieve more accurate performance prediction of protons exchange membrane fuel cell (PEMFC) systems in response to their complicated physico-chemical phenomena. However, there is little research in this field detailing the pre-processing and selection of balance of plants (BOP) features for the input layer of system performance prediction at different current densities. Furthermore, most of the previous research applies neural networks based on simulation data rather than real-time bench or vehicle operation datasets which leads to low robustness and unreliable practical results. This paper details the application of a novel algorithm denoted XGBoost-Boruta, which utilises the combination of an ensemble learning approach and a wrapping approach, to improve the robustness of feature selection and to increase the accuracy and robustness of PEMFC system performance prediction. By introduction of the Z score and shadow features to eliminate the randomness of conventional ensemble learning methods, seven key controllable BOP variables of the hydrogen anode, air cathode and cooling subsystems are selected as the original input variables to determine their dependency on the stack voltage. Two case studies are presented for verification and validation of the proposed algorithm based on the real-time dataset of bench experimental data and data obtained from heavy truck operation at current densities ranging from 100 to 1500 mA/cm^2. The feature

* Corresponding authors.
E-mail addresses: kjiao@tju.edu.cn (K. Jiao), hou_zhongjun@shpt.com (Z. Hou).

https://doi.org/10.1016/j.egyai.2023.100229

Available online 8 January 2023

selection strategy, based on the proposed XGBoost-Boruta algorithm, largely decreases the RMSE by 23.8% and 14.1% and the R^2 increases by 0.06 and 0.04 of both the bench experimental and the heavy truck validation datasets respectively.

1. Introduction

In line with global sustainability issues, allied to the lowering of greenhouse gas emissions [1], it is desired that the CO_2 content of the atmosphere stabilises at no more than 450 parts per million in 2050 [2]. In 2019, the emissions from the transport sector accounted for about 25% of global CO_2 emissions [3]. The exploitation of the proton exchange membrane fuel cell (PEMFC) as a promising power source that produces zero harmful emissions, is on the increase in the areas of passenger and commercial vehicles [4]. However, the complexity of the physicochemical phenomena and the need for correlation of balance of plants (BOP) in a PEMFC system leads to difficulty in controlling the system by conventional mechanistic modelling methods.

Data-driven modelling methods are gradually being applied to the prediction of the PEMFC system performance to find the relationship between PEMFC system variables as input features and system performance parameters. Novel feature evaluation methods exist for PEMFC fault diagnosis and remaining useful life (RUL) prognosis. For example, a feature evaluation method was designed for the classification of three PEMFC states, including normal, flooding and dehydration to extend PEMFC lifetime [5]. Back propagation neural network (BPNN) and adapted neural fuzzy inference system (ANFIS) were applied for PEMFC prognosis under both static and non-static conditions [6]. However, there is very little research available that provides comprehensive reasons for the choice of features which further complicates the interpretability of neural networks and fails to guarantee robustness. Overall, previous data-driven research into data-driven PEMFC system control strategy bears five main research gaps as illustrated in Fig. 1 [7–13] and [14], the first two of which are data-related problems and the latter three deal with the performance of the prediction algorithm.

From the aspect of the selection of datasets, as illustrated by the first two points in Fig. 1, most research tends to combine uncontrollable PEMFC stack design dependant variables and controllable system BOP features based on empirical or equivalent simulation models. For example, Ding et al. applied random forest (RF) techniques to rank PEMFC stack material features and BOP features, where PEMFC stack design dependant variables, including BET surface area, mesopore ratio and micropore ratio, are selected as top features [7]. Eight structural parameters were selected as input features for electric potential [8]. Although these stack material features are critical to the performance of both the PEMFC stack and system, it is not usual to monitor and detect these features during operation for bench test and vehicle running, therefore these research efforts have limited practical value for the real-time control for vehicle applications. Another dataset-related problem is the accuracy and practicality of the empirical and mathematical equivalent models. Most PEMFC system performance prediction neural networks are established based on simulation models which are derived with a significant number of assumptions that cannot be neglected in vehicle operation and therefore limited in practical values and robustness in the prediction models themselves. For example, in [9], the equivalent model of the dataset is generated by an embedded PEMFC model in MATLAB which assumes that the pressure drops across the flow channels are negligible and the cell resistance remains constant at all the conditions of operation. However, these assumptions do not accommodate actual bench and vehicle operation which requires intensive calibration. Therefore, real-time bench and vehicle datasets are of greater importance for the training of the prediction models as compared with simulation datasets.

The latter three research gaps illustrated in Fig. 1 are related to the methods of input feature selection for PEMFC performance prediction. In recent research, feature selection for a high-dimensional PEMFC system is gradually being recognised as an important data preprocessing step before deriving a data-driven neural network model for the prediction of a PEMFC system. First, most research apply the same sets of input system features in neural network prediction models, covering all the current densities from low to high. However, the performance of a typical PEMFC system is largely influenced by the activation, ohmic and concentration losses at different current densities, which leads to different water transport mechanisms through the membranes with specific priorities of BOP features for real-time vehicle operation [15]. For example, Legala et al. selected current, temperature, cathode pressure, oxygen and hydrogen partial pressure and membrane hydration as input features for the prediction of voltage output [10]. Based on the relatively small semi-empirical model, including approximately 1100 data points, a random dropout technique based on a specific probability value is applied to improve the r-squared of prediction to over 0.99. However, such a neuron disconnection method, based on a fixed probability value over all the ranges of current density risks overfitting and randomness, especially lacks the analysis of prediction performance at different ranges of current density [10]. In addition, there are only two variables of the input layer for the prediction of stack voltage and current, which might lead to overfitting [11]. Furthermore,

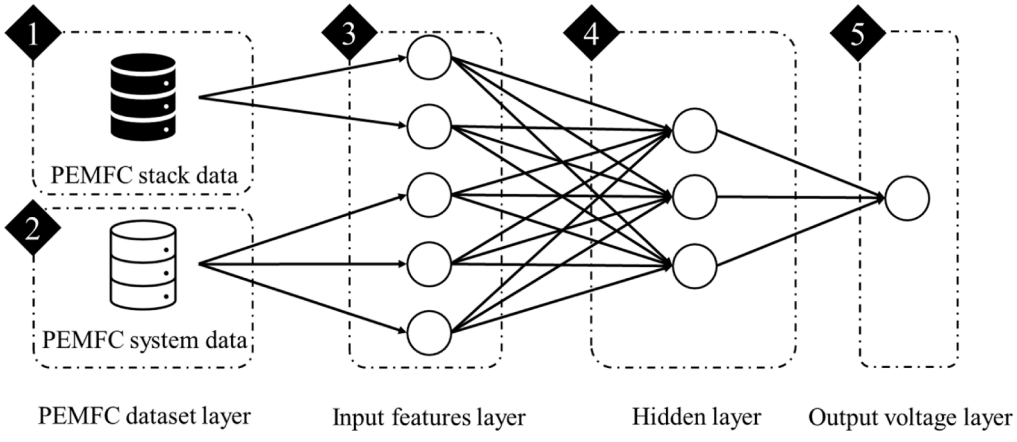


Fig. 1. Five main research gaps in the design of data-driven PEMFC system control.

the selection of current density, as one of the feature inputs, would reduce the influence of other features as voltage largely correlates to current for a semi-empirical PEMFC model [12]. Second, most research tries to adjust hyperparameters of neural networks to improve the performance of prediction, but few clarify the reason for feature choices. For example, Bicer et al. compared different numbers of hidden layers of a conventional neural network model for better prediction performance of the dynamic behaviour of a 6 kW PEMFC [9]. Kheirandish et al. evaluated the influence of difference combinations of the insensitivity function ε and parameters C and γ on the performance of voltage prediction [13]. However, emphasis on the choices of hyperparameters based on one specific simulation data set, rather than the auto-selection of feature inputs, might lead to low robustness, especially under real environmental conditions with changing temperature and pressure during vehicle operation. Third, conventional pre-processing methods, such as random forests, are vulnerable to hyperparameters and the results bear randomness, but vehicle control requires predicable and self-adaptive methods. For example, as a typical bagging method, RF that trains base learners in parallel to minimise variance requires significant amounts of computational resources and the prediction is highly vulnerable to a small change in hyperparameters. Therefore, it is not applicable for a real-time PEMFC system dataset with a pre-defined set of hyperparameters and thereby lacks robustness and accuracy. In comparison, as an advanced boosting ensemble method, extreme gradient boosting (XGBoost) requires limited hyperparameters, including depth and number of trees to train a weak learner and adjust itself iteratively, but still exhibits randomness for each forest [14].

Based on the literature review on the research gaps in the design of data-driven PEMFC system control, the main contribution of this paper is to address the database-related and input feature pre-processing related research gaps for data-driven PEMFC system performance prediction. An innovative data pre-processing algorithm, named XGBoost-Boruta algorithm, is designed to determine the influence of system BOP control features on the performance of an anode self-humidification PEMFC system under different ranges of current density. Boruta is first applied as an ensemble learning method with XGBoost to investigate the influence of controllable BOP features on stack voltage in an anode self-humidification PEMFC system. A one-dimensional convolutional neural network (1DCNN) with fixed hyperparameters to connect the selected key BOP features and voltage output is then applied to verify and validate the accuracy and robustness of the proposed algorithm in an objective, steady and automatic manner based on the real-time data from a hardware-in-loop bench test and the operation of a heavy truck.

This paper is organised as follows. The PEMFC system structure is presented in Section 2. In Section 3, the novel application of the XGBoost-Boruta algorithm for PEMFC system feature selection process is detailed. Section 4 provides key data for the case studies, including specifications of the PEMFC stack and system, bench tests and dataset pre-processing. In Section 5, the proposed algorithm is successfully verified and validated based on the hardware-in-loop (HIL) bench test platform and real-time vehicle data. The proposed XGBoost-Boruta algorithm is compared with the conventional feature dropout strategy, ranking the algorithm as it pertains to accuracy and robustness based on 1DCNN with fixed hyperparameters, followed by the conclusions in Section 6.

2. System illustration

The proposed anode self-humidification PEMFC system is composed of four main parts: a high-power PEMFC stack, independently developed by Shanghai Hydrogen Propulsion Technology Co. Ltd., an air cathode subsystem, a cooling subsystem and a hydrogen anode subsystem, as illustrated in Fig. 2. BOPs are designed to regulate the temperature, pressure and mass flow rate of air, hydrogen and coolant to deliver high performance of the PEMFC stack.

Based on the application of ultra-thin proton exchange membrane technology (8 μm) and the design of efficient hydrogen recirculation, the development of the anode self-humidification system eliminates the requirement for membrane humidifiers and simplifies the system architecture, increases the power density and reduces the total cost [16,17,18]. The mass flow rate, pressure, and temperature of the inlet air, represented by Nodes 1 to 3 respectively in Fig. 2, are controlled by the air compressor, an intercooler, a combination valve and a back pressure relief valve. Hydrogen passes through an injector to regulate the hydrogen inlet pressure before entering the PEMFC stack at Node 5. A hydrogen circulation pump (HCP), with a water-gas separator, is used to

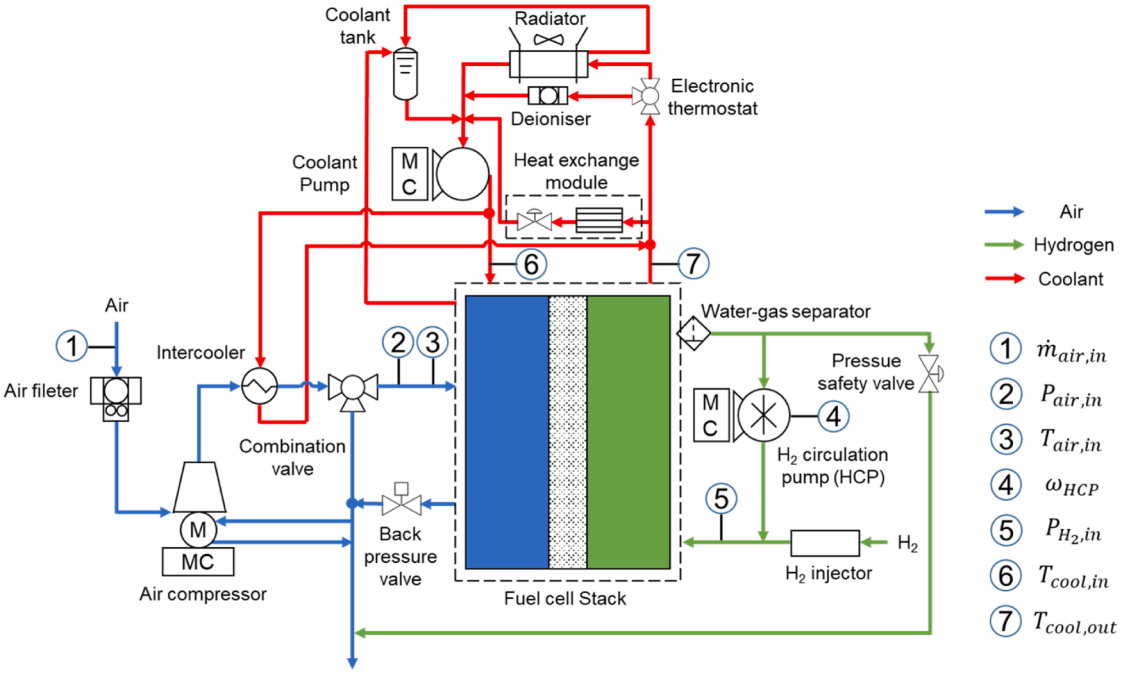


Fig. 2. Schematic diagram of the proposed anode self-humidification PEMFC system.

recirculate the wet exhaust, including unreacted hydrogen enriched with the inlet hydrogen, to maintain the relative humidity of the membrane. The key controllable feature of this recirculation process is the HCP rotation speed, monitored and controlled at Node 4. The cooling sub-system consists of a cooling pump, an electronic thermostat, de-ionizers and a heat exchanger module. As thermal energy of the PEMFC stack is generated from the electrochemical reactions, an electronic thermostat, a heat exchange module and a radiator are used in conjunction to maintain the stack operating temperature [19]. The inlet and outlet temperatures of coolant are critical features in the cooling subsystem monitored at Nodes 6 and 7. Although there are more BOP features in the proposed PEMFC system, the seven features listed above have direct impact on the operation of the PEMFC stack. For example, the electronic thermostat, the coolant pump and the radiator of the coolant subsystem are controlled to attain the required temperature of the coolant and the air and then affect the voltage output indirectly.

Restricted by the characteristics of the tightly coupled subsystems, conventional physical modelling methods are unable to reveal the relationship between the seven critical controllable features and the technical performance of the PEMFC system during the operation of the PEMFC system at different current densities. Therefore, a novel feature susceptibility approach is introduced in Section 3 for the proposed anode self-humidification PEMFC system and other PEMFC systems with similar characteristics of tightly coupled subsystems.

3. XGBoost-Boruta algorithm

In this section, the application of the XGBoost algorithm with Boruta is explained in detail. Generally, the aim of the proposed wrapping method is to evaluate the impact and importance of seven features on the PEMFC stack voltage and select the more relevant features for the PEMFC voltage prediction as presented in Fig. 2.

3.1. Conventional feature selection methods

Conventional ensemble learning methods train multiple weak feature selection learners to further enhance the performance of a single learner [20]. Random forest (RF) and extreme gradient boosting (XGBoost) are two typical bagging and boosting ensemble learning methods used to reduce variance and bias [21] and [22]. Compared with the XGBoost, the RF creates more trees in parallel and as a consequence requires more computational resource, which influences the real-time capability for a large PEMFC system dataset [23]. Besides, the performance of the RF is largely influenced by hyperparameters. However, due to the characteristics of PEMFC degradation, the hyperparameters defined for the training dataset are not applicable during the operation of the system. Furthermore, in a PEMFC system where the correlation between features is high, some key features tend to be removed as they may be deemed to be of low importance leading to errors of judgement. In comparison, the XGBoost algorithm requires a smaller number of

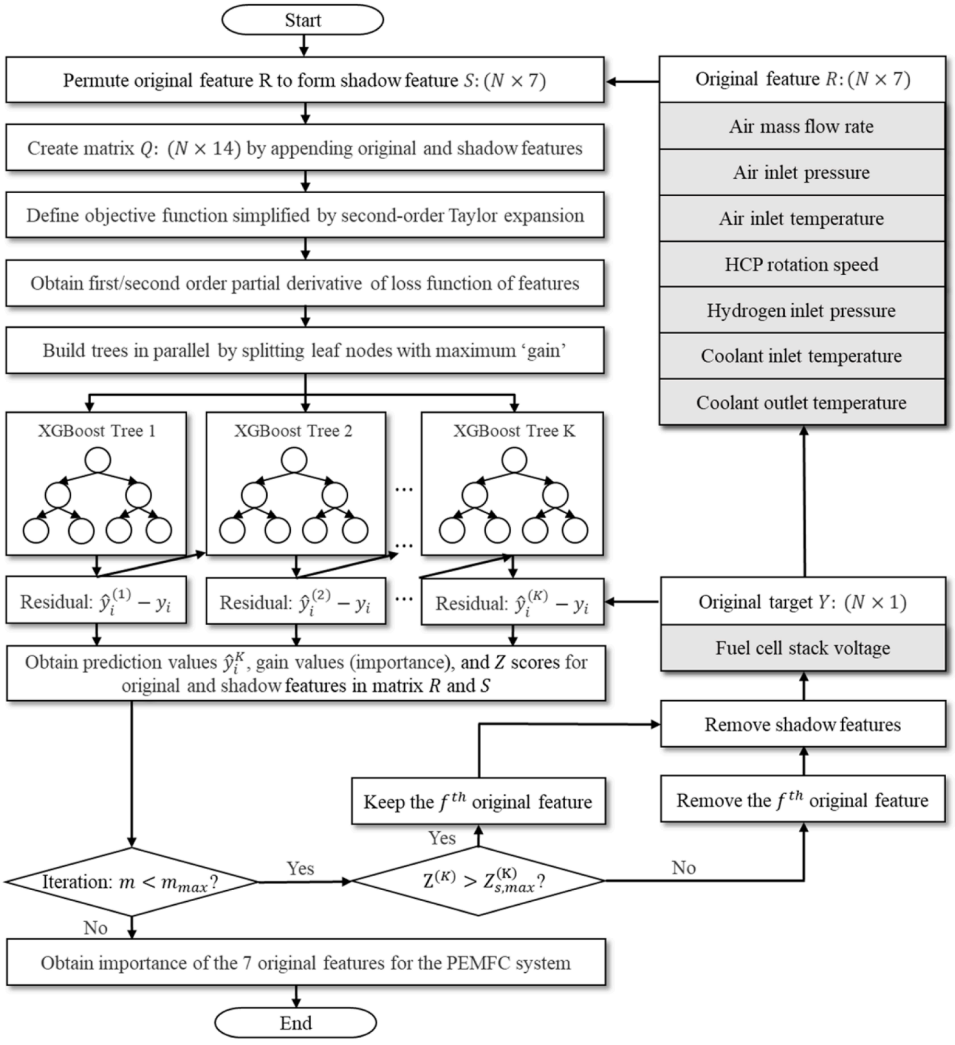


Fig. 3. Flow chart of the application of the proposed XGBoost-Boruta for PEMFC system feature selection.

hyperparameters and has the capability to adjust itself as iteration progresses. Since the self-adapting iteration with different residuals leads to randomness for each forest, wrapping methods are innovatively combined with the ensemble learning method in ranking BOP features of a PEMFC system in this paper. Differing from recursive feature elimination (RFE), which is largely dependant on the performance of XGBoost, Boruta is a fully correlated feature selection method which is able to go through all features that carry key information for prediction instead of focusing on a particular compact subset of features [24] and [25]. Therefore, XGBoost is combined with Boruta to determine the importance of the seven key controllable BOP features of the PEMFC system as illustrated in Fig. 3.

3.2. Mechanism and application of the XGBoost-Boruta algorithm

The XGBoost-Boruta algorithm on the PEMFC system is designed to determine feature importance with high interpretability. The proposed application of the algorithm can be divided into three parts, as presented in Fig. 3: building shadow features, training XGBoost models, and obtaining feature importance iteratively.

1) **Building shadow features.** The original feature matrix R, is composed of N samples with the seven BOP features, is permuted randomly to form a shadow feature matrix S. By appending these two matrices, a new matrix Q $(N \times 14)$ is created as the input feature matrix for the XGBoost model.
2) **Training XGBoost models.** The training process of the XGBoost model follows the principle of the gradient boosting method by minimising the loss and the objective functions [24].

$$Obj^{(K)} = \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(K)}\right) + \Omega(f_K) \tag{1}$$

$$L\left(y_i, \hat{y}_i^{(K)}\right) = \left(y_i - \hat{y}_i^{(K)}\right)^2 \tag{2}$$

$$\hat{y}_i^{(K)} = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

where K is the number of trees set as 50, $Obj^{(K)}$, $L(y_i, \hat{y}_i^{(K)})$, $\hat{y}_i^{(K)}$ and $\Omega(f_K)$are the objective function, the loss function, the prediction value for the K^{th} tree respectively, $f_k(x_i)$ are the penalty term and the predicted value for the K^{th} tree, N is the number of samples, y_i is the real value from the original target Y, T is the number of leaf nodes, ω_j is the weight of the j^{th} leaf node.

With the aim being to minimise the objective function, the XGBoost model uses the second-order Taylor expansion to expand Eq. (1) to Eq. (4) to evaluate the performance of the K^{th} tree.

$$Obj^{(K)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{4}$$

where G_j and H_j are the first and second derivative of the loss function $L(y_i, \hat{y}_i^{(K-1)})$ for the j^{th} leaf node, $\frac{G_j^2}{H_j+\lambda}$ is the gain value of the j^{th} leaf node to determine the split of the tree, and γ is the hyperparameter defaulted to 0 in this paper.

Feature importance is determined by the gains of the leaf nodes before and after the split as shown in Eq. (5).

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{5}$$

where G_L, H_L, G_R and H_R represent gains of left and right nodes after each split.

1) **Obtaining feature importance iteratively.** Boruta applies shadow features and a Z score to enhance the steadiness of a single XGBoost model, the latter of which is defined in Eq. (6), slightly different from the statistical definition [26].

$$Z^{(K)} = \frac{\overline{Gain}}{\mu_{Gain}} \tag{6}$$

where $\overline{Gain}$ and μ_{Gain} are the average value and standard variance of the gains at the K^{th} tree. Original features with Z scores that are smaller than the maximum Z score of the shadow features $Z_{s,max}$ are removed. These three steps are repeated iteratively until the iteration reaches maximum m_{max} which has been set as 50, The pseudo code is given below. Algorithm. 1

The features selected by the proposed XGBoost-Boruta algorithm are set as the input features for the prediction of the PEMFC voltage performance. As the novel contribution of this paper is the feature preprocessing approach, a neural network is required to evaluate the improvement in prediction accuracy. To ensure the repeatability of the results for verification and validation of the proposed XGBoost-Boruta algorithm, a simple one-dimensional convolutional neural network (1DCNN) with fixed hyperparameters is designed as presented in Fig. 4. The 1DCNN includes layers nominated Conv1D, MaxPooling, Flatten, Dense and Dropout. The 1DCNN is set with the number of four dimension kernels to 30, the number of pooling layers to 2, two fully connected layers with 20 neurons and 1 neuron, the dropout rate to 0.5, the number of epochs to 100, the batch size to 32, and the activation function as RELU. The description of the conventional 1DCNN can be found in [27] but are not detailed here.

4. Case study

This section provides the setup information and input data for the verification and validation process for the proposed XGBoost-Boruta algorithm to be applied to the proposed PEMFC system detailed in Sections 2 and 3. The signal transmission for the bench and vehicle tests, as denoted by the blue and the red paths in Fig. 5, demonstrates the verification and the validation processes for the accuracy and robustness of the proposed XGBoost-Boruta algorithm for an anode self-humidification PEMFC system.

The hardware-in-the-loop (HIL) bench test platform (shown by the blue path in Fig. 5) illustrates the signal transmission between the bench and the host server. The fuel cell unit (FCU) receives the system control features through CAN from the PEMFC system and sends them to the proposed algorithm as input feature datasets. The host server with the XGBoost-Boruta algorithm automatically ranks and selects key features for accurate system performance prediction at the different current densities. The results of feature ranking and system performance prediction are used to enhance the command of the influence of the complicated PEMFC system BOP components on the PEMFC system and to further provide the guidance for system calibration at different current densities. Similarly, the red path in Fig. 5 represents the real-time signal transmission for vehicle applications. The telematics-box (T-box) is applied to collect the vehicle dynamic data and upload them to the host server through CAN and ethernet. The hyperparameters of the XGBoost and the 1DCNN remain the same for both the verification and validation processes. The number of trees of the XGBoost is set to 50. The input dataset for the 1DCNN is split into two parts with a ratio of 0.7:0.3, of which the first 70% is used for training while the rest is used for validation.

Sections 4.1 and 4.2 detail the required test information, including the constraints of the input features, the real-time input feature datasets,

Algorithm 1
XGBoost-Boruta feature ranking algorithm.

Algorithm 1: XGBoost-Boruta feature ranking algorithm

Input:
R, original feature matrix
Y, original target matrix
m, number of maximum iterations

Output:
Feature importance of features

Procedure:
1) For i = 1 to m,
Create the shadow feature matrix S by permuting the original matrix R
Fit the XGBoost model with the new matrix N = [R,S]
If Z score of original feature > maximum Z score of S:
Keep the original feature
Else:
Remove the original feature
End If
End for
2) Evaluate the average gain values of features selected

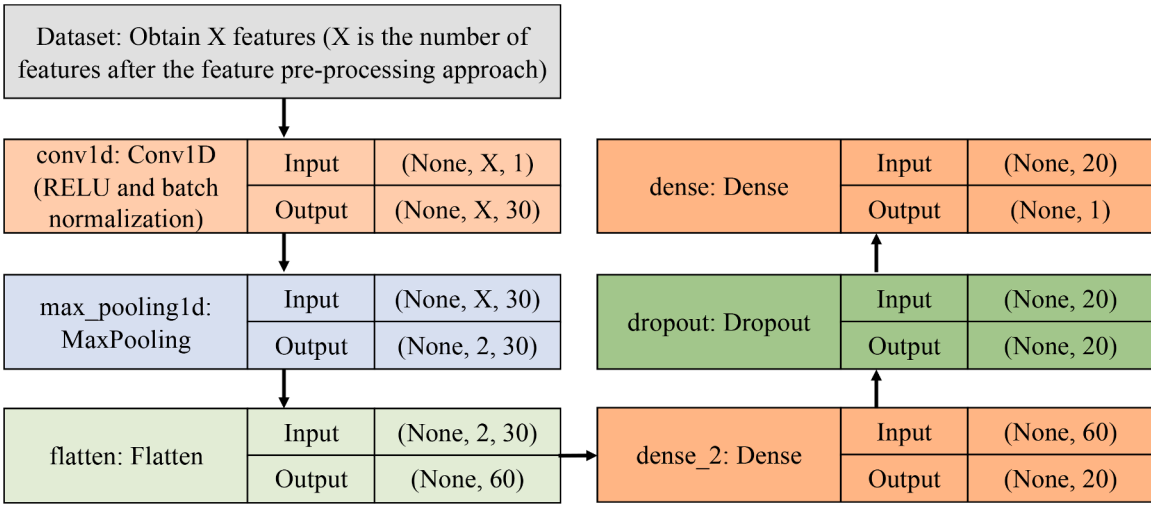


Fig. 4. Visualized structure of the proposed XGBoost-Boruta algorithm and the 1DCNN.

the parameters required for the algorithm and the 1DCNN, and the setup information of the test platform for the HIL and vehicle operation separately. The bench datasets and dynamic vehicle running datasets for verification and validation in Section 4.1 and 4.2 are obtained during normal operation mode without the cold start process. The core strategy for the cold start is rapid ice breaking and heat management to reach the required stack temperature by correction between the desired setpoints and measured process values. Under the fixed control strategy of the cold start process, the proposed feature selection process is not applicable.

4.1. Bench test setup for verification

The PEMFC system is equipped with a high-power FC stack with a stack peak power output and power density of 198 kW and 5.1 kW/L. Each cell of the stack is composed of ultra-thin proton exchange membranes (8 μm), high-activity Platinum-Cobalt (Pt-Co) catalyst (mass activity $>$ 150 $mA/mg(Pt)$ at 0.9V), and ultra-thin metal bipolar plates (0.93 mm). Following the schematic diagram in Fig. 5, the hardware of the HIL test platform is set up as illustrated in Fig. 6(a). The real-time bench test dataset in Fig. 6(b) includes approximately 23,754 data points covering current densities from 104.0 to 1496.0 mA/cm^2 and stack voltages from 275.5 to 382.8 V. As the ambient temperature and pressure are set as constant values in the bench test, these two features provide no information for the PEMFC system prediction and thus are not included in the input features for the algorithm. Therefore, the seven BOP features of the hydrogen, air and coolant subsystem, as described in Section 2, are regarded as input features, the constraints of which are given in Table 1.

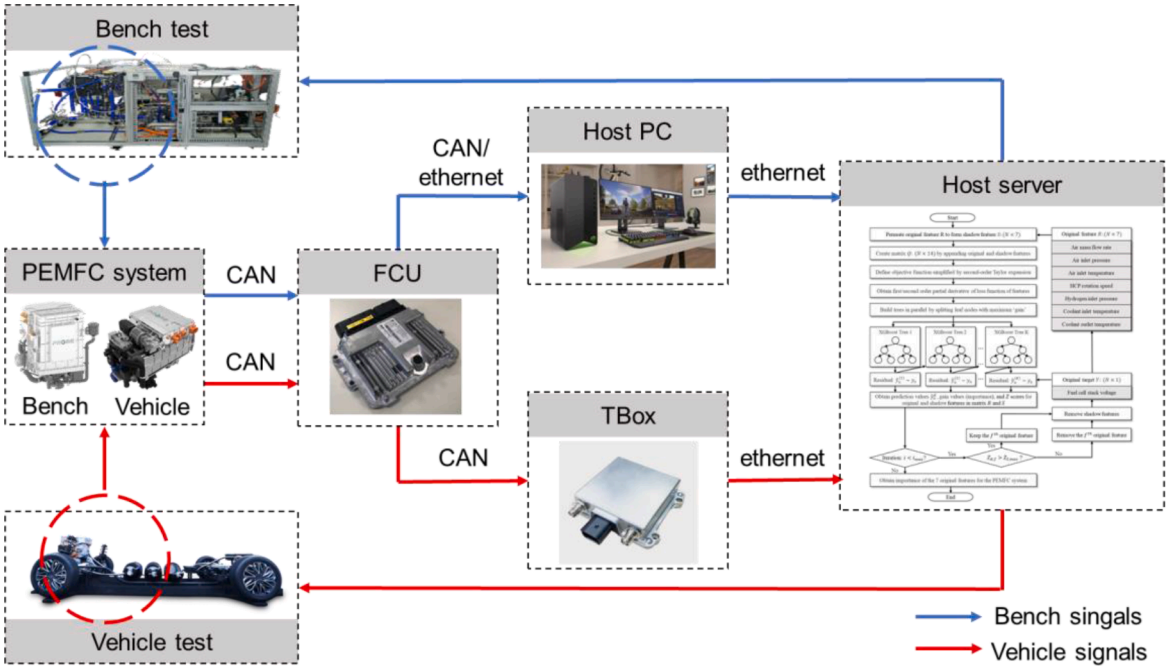


Fig. 5. Schematic diagram of signal transmission of the proposed anode self-humidification PEMFC system with the proposed XGBoost-Boruta algorithm.

4.2. Vehicle test setup for validation

In addition to the verification test setup for the accuracy of the XGBoost-Boruta algorithm, this subsection provides the running information of a heavy truck that will be used for the validation of the robustness of the proposed approach. The heavy truck running in Inner Mongolia province at about 40.8°N, 111.8°E in China (shown in Fig. 7 (a)) is equipped with a PEMFC system with rated stack power output and power density at 140 kW and 3.7 kW/L. The vehicle running dataset includes approximately 5720 data points, all of which are sent to the host server and sampled each sampling period of one second as illustrated in Fig. 7(b). Considering the limited temporal and spatial influence, the variations of these two features are small and are maintained constant for most data points and thus have low influence on the PEMFC performance prediction based on the proposed dataset. Therefore, the seven BOP features, the change of temperature and pressure are taken into consideration for the vehicle validation test. The vehicle validation test is conducted with an increase of current density from 1000.0 to 1485.0 mA/cm^2 and stack voltage from 205.3 to 264.6 V and stack voltage from 205. to 264.6 V. The constraints of inputs are given in Table 2.

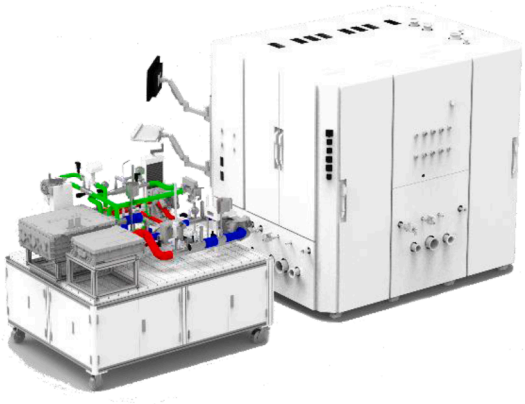
5. Results and discussion

This section analyses the details of the verification and validation results based on the real-time bench and vehicle running data detailed in Section 4. Section 5.1 provides a comparative analysis, focusing on the difference of conventional random forest (RF), XGBoost and the proposed XGBoost-Boruta algorithm applied on the bench data. Based on the ranking and selection of key BOP features, another comparative analysis is conducted to demonstrate the accuracy of the system performance prediction with the features selected by the proposed algorithm. In Section 5.2, following the same feature pre-processing strategy, a similar comparative analysis is performed for the robustness of the system performance prediction based on the vehicle running data.

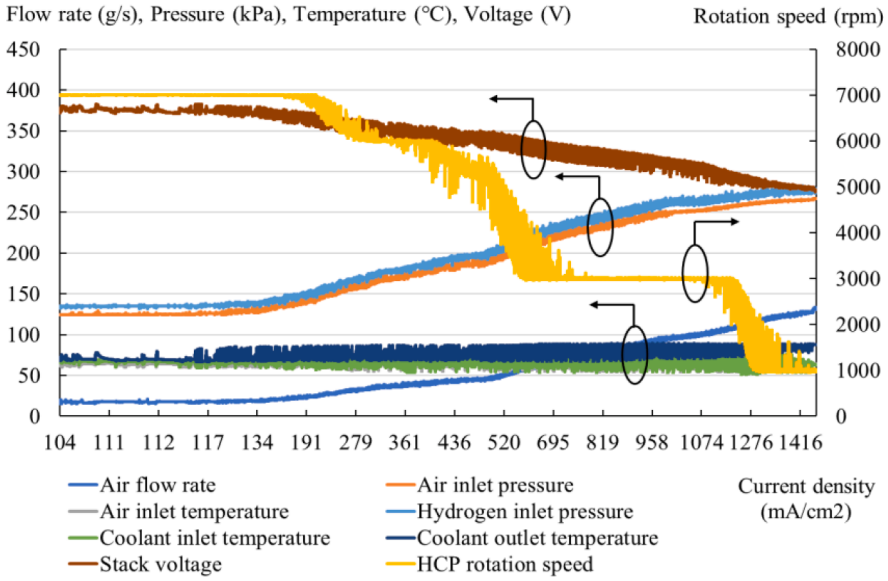
5.1. Verification results and analysis on bench test dataset

The importance heat map obtained for the seven BOP features is based on the RF and the XGBoost algorithm in Fig. 8(a) and (b) respectively. The importance values at current density values of between 100 and 1500 mA/cm^2 shows similar trends. For example, at low current density below 200 mA/cm^2, the hydrogen inlet pressure $P_{H_2,in}$ ranks highest mainly because of its significant influence on water distribution in the membrane and thus has a positive effect on the stack voltage output. The heat map indicates that between 200 and 1100 mA/cm^2, the coolant outlet temperature $T_{cool,out}$, representing the stack operating temperature, is far more significant when compared with the other features. When the current density increases further, the air inlet mass flow rate $\dot{m}_{air}$, the air inlet pressure $P_{air,in}$, and the air inlet temperature $T_{air,in}$ become the more significant features whilst the coolant outlet temperature is less significant, which is caused by there being relatively stable operating temperature at high current density. The importance values of the rotation speed of the HCP ω_{HCP} are low at all the operating conditions, indicating little influence on the output voltage. This is mainly caused by the proposed anode self-humidification system structure. Without using a humidifier at the cathode, the HCP at the anode transports the moisture generated during the stack operation to ensure the desirable humidity of the membrane electrolyte assembly (MEA) [28]. In addition, the HCP affects the hydrogen stoichiometry, but the hydrogen excess coefficient has a smaller impact on the output voltage when compared with the cathode [29,30]. The impact is much smaller when the water uptake is maintained within a reasonable range for the anode self-humidification PEMFC system bench test with appropriate energy and heat management.

However, compared with the ranking result based on the XGBoost approach in Fig. 8(b), the RF approach provides the same trends but different importance rankings in Fig. 8(a). For example, the importance values of coolant outlet temperature in Fig. 8(a) are overall higher than the values in Fig. 8(b) across all current densities, which exposes the common disadvantage of conventional feature selection methods offering imbalanced preferences to related features and may lead to overfitting of the data. In comparison, weightings are allocated from top



(a)



(b)

Fig. 6. Photo of the bench setup and the real-time dataset for verification: (a) Photo of the bench setup (b) Bench test data for the seven selected features and the stack voltage.

Table 1
Details of the seven key features for PEMFC system control features.

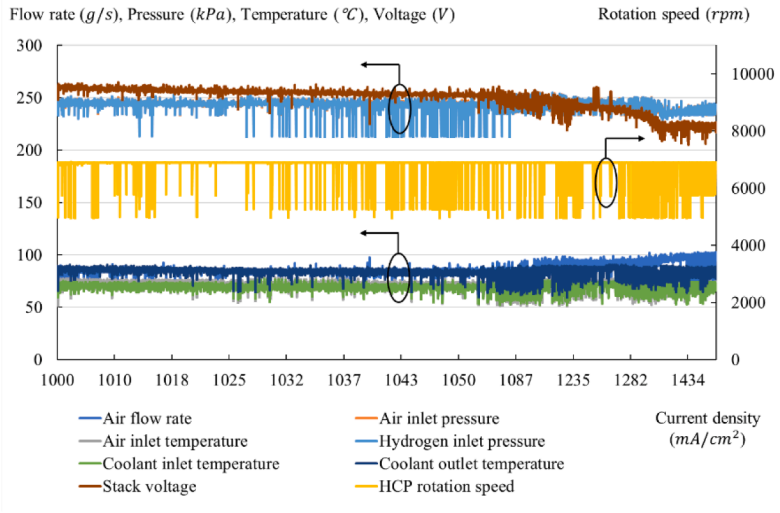
Node	Feature	Symbol	Unit	Range
1	Air flow rate	$\dot{m}_{air}$	g/s	[15.6, 133.8]
2	Air inlet pressure	$P_{air,\ in}$	kPa	[124.4, 267.1]
3	Air inlet temperature	$T_{air,\ in}$	^oC	[55.3, 69.7]
4	HCP rotation speed	ω_{HCP}	rpm	[954.0, 7009.0]
5	Hydrogen inlet pressure	$P_{H_2,\ in}$	kPa	[131.8, 282.4]
6	Coolant inlet temperature	$T_{cool,\ in}$	^oC	[51.9, 75.8]
7	Coolant outlet temperature	$T_{cool,\ out}$	^oC	[66.8, 89.3]

features to related features in an objective manner by the XGBoost approach. In addition to differing from the RF algorithm, which requires repeated attempts to obtain appropriate hyperparameters, the XGBoost algorithm only needs a small number of hyperparameters, which is preferable during dynamic PEMFC vehicle operation. Therefore, considering the difference amongst each PEMFC system and vehicle running conditions, the XGBoost algorithm is more appropriate where the requirement is for high robustness with less impact caused by human judgement.

Furthermore, the large variation of Z scores of the seven features and the maximum Z score of shadow features at each iteration at the current density between 300 and 400 mA/cm^2 in Fig. 9 illustrates that the conventional XGBoost algorithm delivers high randomness, low robustness and low accuracy, which is not acceptable for real-time vehicle control. As shown in Fig. 9, all the upper limits of the Z scores of the seven features are higher than the maximum Z score of the shadow features except HCP rotation speed ω_{HCP}, which demonstrates that in this case, all the other features are marked as important for voltage prediction by the 1DCNN approach.

The feature selection strategy is illustrated in Fig. 10 by comparing the Z scores of the seven real BOP features and the maximum value of the shadow features. The definition of Z score and shadow feature are introduced to minimise the influence of randomness of the conventional XGBoost and the difficulty in the selection of input features. In Fig. 10, all the Z scores in red are higher than the maximum value of the shadow features in blue except the HCP rotation speed. Therefore, the HCP

(a)



(b)

Fig. 7. Heavy truck running in Inner Mongolia Province, China: (a) Photo of the heavy truck; (b) Real-time data for the seven selected features and stack voltage.

Table 2
Details of the seven key features for PEMFC system control features.

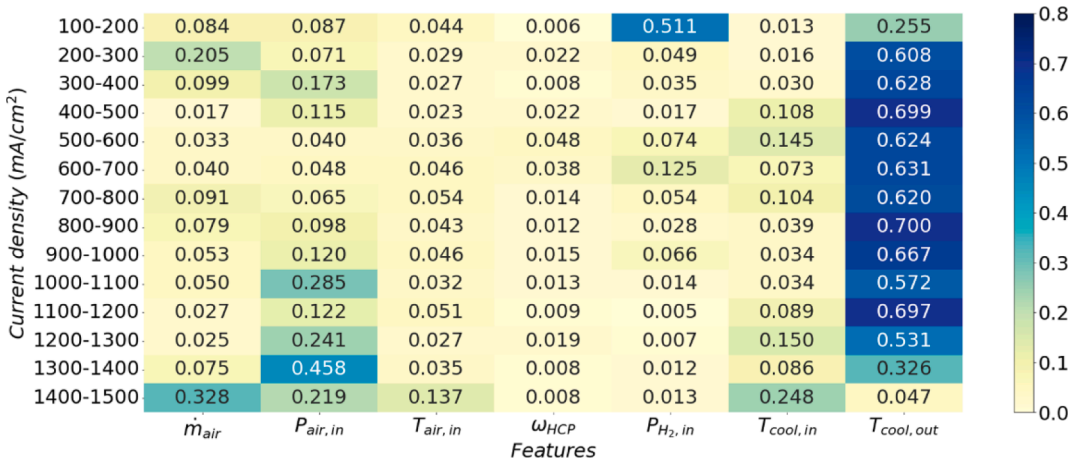
Node	Feature	Symbol	Unit	Range
1	Air flow rate	$\dot{m}_{air}$	g/s	[71.9, 102.5]
2	Air inlet pressure	$P_{air,\ in}$	kPa	[230.3, 255.7]
3	Air inlet temperature	$T_{air,\ in}$	^oC	[51.1, 78.2]
4	HCP rotation speed	ω_{HCP}	rpm	[4980.0, 6920.0]
5	Hydrogen inlet pressure	$P_{H_2,\ in}$	kPa	[213.0, 256.4]
6	Coolant inlet temperature	$T_{cool,\ in}$	^oC	[52.1, 77.1]
7	Coolant outlet temperature	$T_{cool,\ out}$	^oC	[59.9, 90.5]

rotation speed is automatically removed in accordance with the proposed XGBoost-Boruta pre-processing strategy.

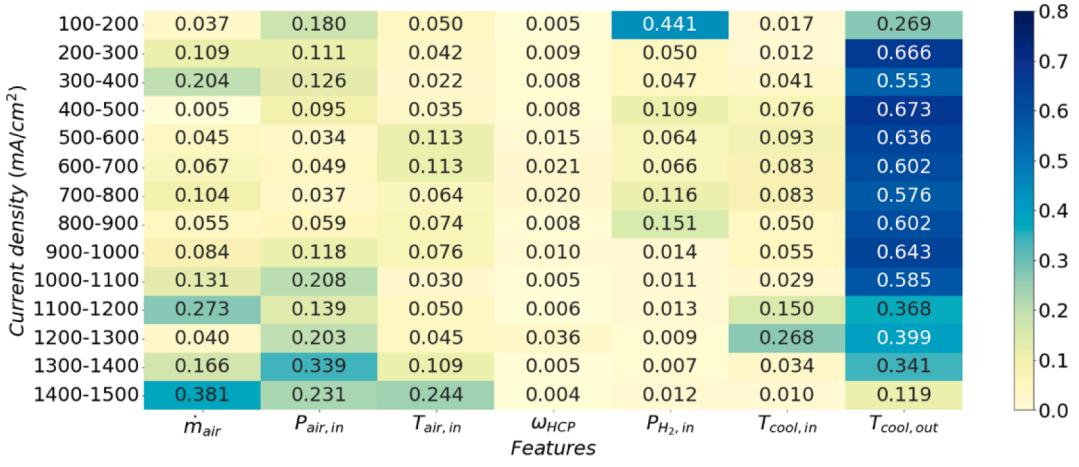
To verify the accuracy of the feature removal strategy, three cases are compared. These cases are as follows: keeping all the features, removing the most important feature manually and following the proposed XGBoost-Boruta algorithm at the current density between 300 and 400 mA/cm^2. A repeatability experiment of 1DCNN based on the proposed XGBoost-Boruta algorithm is conducted to ensure the fair comparison. The ten repeatability results illustrated in Fig. 11 demonstrate that the proposed 1DCNN provides reasonably consistent results and can be used for the comparative analysis of different strategies. The accuracy of the voltage prediction performance by the three strategies varies markedly as visualised in Fig. 12(a) to (c) successively. The difference between keeping all the features and removing the feature by the XGBoost-Boruta algorithm is illustrated in the black dotted rectangular region in Fig. 12 (a) and (b). It is obvious that the degree of coincidence between predicted and experimental datasets in the black dotted oval region in Fig. 12(c) is higher than that in Fig. 12(a). Specific evaluation values are detailed in Table 3. The RMSE of voltage predictions by the proposed XGBoost-Boruta algorithm (1.31), is decreased by 23.8% compared with the RMSE values of 1.72 obtained by keeping all the features. Although the R^2 values of voltage predictions are the same at 0.94 for the training dataset between keeping all the features and removing the feature by the proposed algorithm, the R^2 difference of the validation dataset is approximately 23.8% between 0.92 and 0.86. From Fig. 10, the weighting of the coolant outlet temperature $T_{cool,out}$ is much higher than the values of the other features, so the second strategy is to remove the coolant outlet temperature from the input feature layer of the 1DCNN. It is obvious that the prediction voltage curve in blue is quite different from the experimental voltage data, especially in the black dotted region in Fig. 12(b). The values of RMSE and R^2 get worse at 4.61 and even 0.02 for the validation dataset. Following the proposed XGBoost-Boruta algorithm, the HCP rotation speed is automatically removed by the comparison of the real features and shadow features. The PEMFC voltage prediction performance improves, especially for the validation data where the R^2 increases from 0.86 to 0.92 without human judgement.

5.2. Validation results and analysis on vehicle operation dataset

Based on the verification of the proposed XGBoost-Boruta feature pre-processing strategy on the HIL bench test dataset, the algorithm is



(a)



(b)

Fig. 8. Comparison of the importance values of the seven features based on XGBoost and RF algorithm: (a) RF algorithm; (b) XGBoost algorithm.

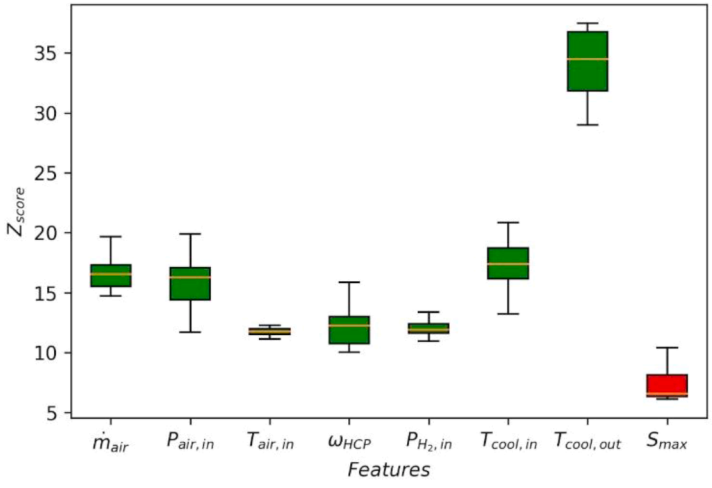


Fig. 9. Box plot of the Z scores of seven features and the maximum Z score of shadow features at each iteration at the current density between 300 and 400 mA/cm^2.

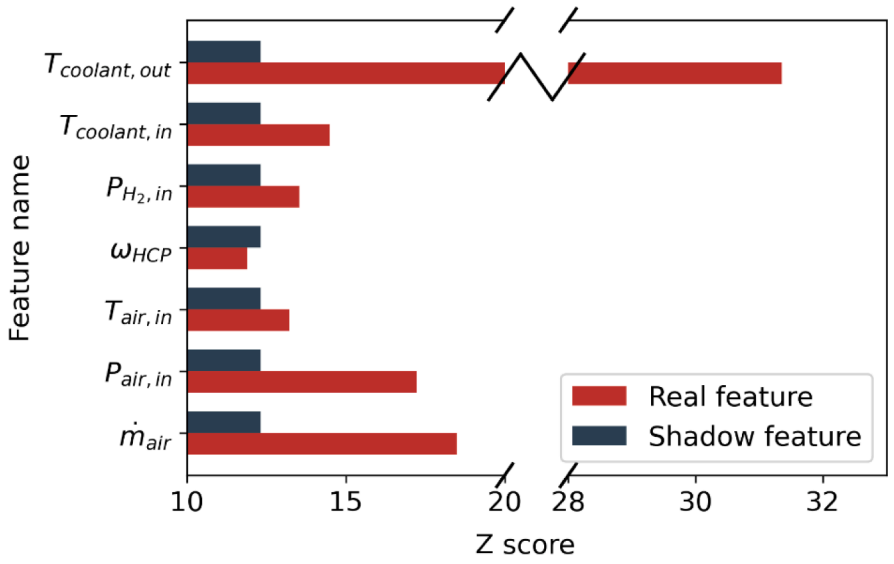


Fig. 10. Bar chart of the Z scores of the seven BOP real features and the maximum shadow feature based on the bench data at the current density between 300 and 400 mA/cm^2.

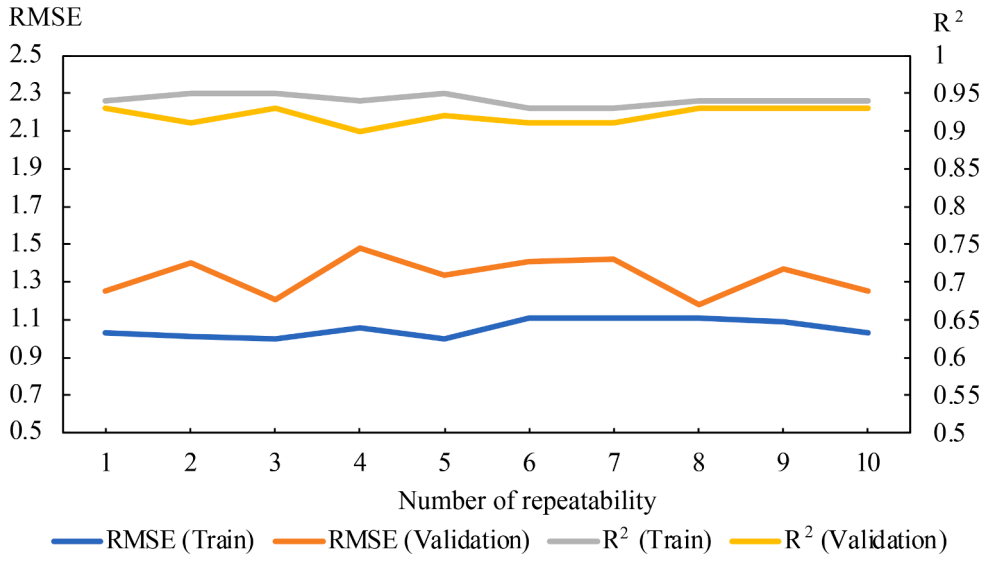


Fig. 11. Ten repeatability results of 1DCNN based on the proposed XGBoost-Boruta algorithm.

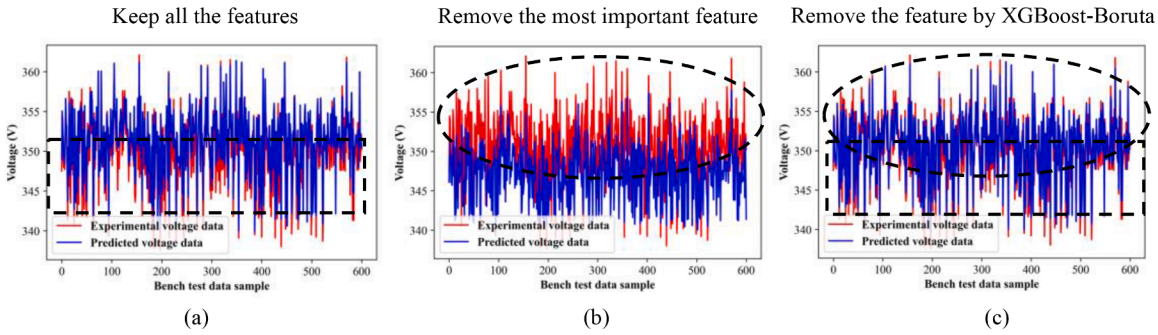


Fig. 12. Part of the validation data prediction visualisation by the XGBoost-Boruta and the 1DCNN algorithm based on the bench verification dataset at the current density between 300 and 400 mA/cm^2: (a) Keep all the features; (b) Remove the most important feature; (c) Remove the features by XGBoost-Boruta.

validated in this subsection utilising the real-time vehicle running dataset to demonstrate the robustness and practical value. Following the real-time signal transmission for PEMFC heavy trucks as illustrated in Fig. 5 in Section 4, the T-box transfers dynamic signals, including the seven BOP features from the FCU to the host server. The seven real BOP features are input into the XGBoost-Boruta algorithm, the Z scores of

Table 3
Comparative results amongst different feature removal strategy based on the HIL bench data at the current density between 300 and 400 $mA\ /cm^2$.

Category	RMSE (Train)	RMSE (Validation)	R^2 (Train)	R^2 (Validation)
Keep all the features	1.10	1.72	0.94	0.86
Remove the most important feature	2.36	4.61	0.71	0.02
Remove the features by XGBoost-Boruta	1.03	1.31	0.94	0.92

which are compared with the maximum Z score of the shadow features as shown in Fig. 13. From Fig. 13, the Z scores of the HCP rotation speed and the air inlet temperature are lower than those of the other five features and the maximum value of the shadow features. These two features are removed by the algorithm as the input features of the 1DCNN. In comparison, the air mass flow rate and air inlet pressure obtain more weighting at high current density between 1000 and 1485 mA/cm^2 than at low current density between 300 and 400 $mA\ /cm^2$, as detailed in the verification case study due to relatively stable operating temperature at high current density.

The predicted voltage data (in blue) and experimental voltage data (in red) are partly visualised in Fig. 14. Compared with the strategy of keeping all the features, illustrated in Fig. 14(a), the degree of coincidence between the predicted and experimental data following the proposed XGBoost-Boruta method in Fig. 14(c) is higher, especially when the voltage is over 250 V in the black dotted rectangular regions. As visualised in the black dotted oval regions in Fig. 14(b) and (c), the prediction accuracy obtained at voltages between 220 and 245 V by removing the most important feature further demonstrates that the proposed strategy performs better than the other two strategies at different voltages. From the specific evaluation values in Table 4, the RMSE value of the validation dataset, after removing the HCP rotation speed and air mass flow rate, is only 3.95, approximately 14.1% and 37.7% lower by than the values obtained under the other two strategies. The R^2 value of the validation dataset obtained by the proposed algorithm is 0.04 and 0.20 higher as well. Overall, the proposed XGBoost-Boruta algorithm performs well for both the relatively steady bench test dataset and the dynamic vehicle running dataset at different ranges of voltage. It demonstrates high accuracy and robustness of the proposed feature selection algorithm.

6. Conclusions

In this paper, a novel application of the combination of an ensemble learning method and a wrapping method named XGBoost-Boruta is proposed to evaluate feature importance of the seven key BOP features in a PEMFC control system, including the air mass flow rate, the air inlet pressure, the air inlet temperature, the HCP rotation speed, the hydrogen inlet pressure, and the coolant inlet and the outlet temperature. First, the proposed algorithm provides stabilised feature ranking at different current densities with low randomness compared with conventional wrapping approaches. The introduction of the definition of the Z score eliminates the randomness of the conventional XGBoost model

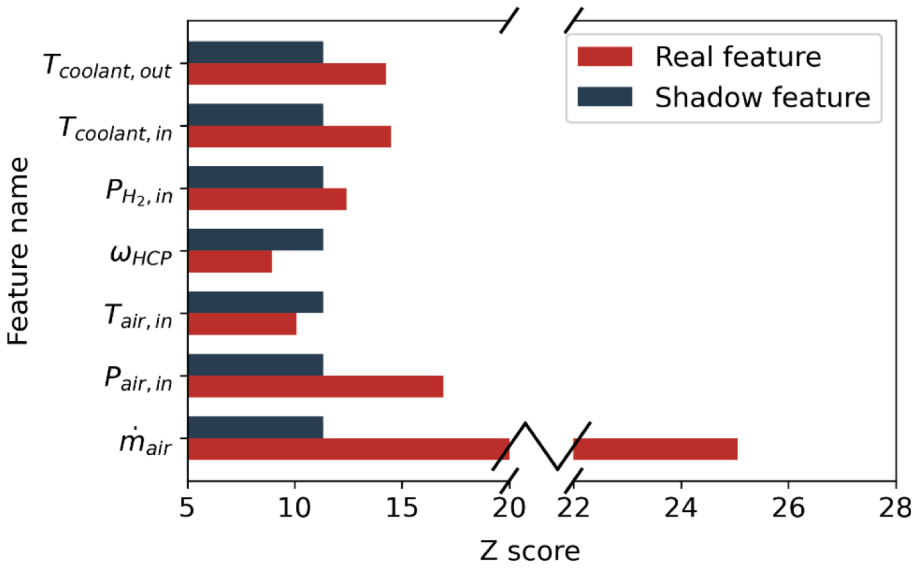


Fig. 13. Z scores of seven BOP features and the maximum shadow feature based on the vehicle dataset at the current density between 1000 and 1485 $mA\ /cm^2$.

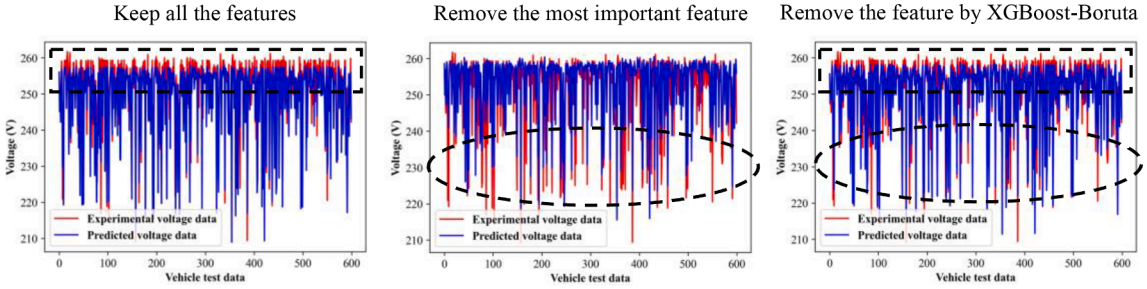


Fig. 14. Part of the validation data prediction visualisation by the XGBoost-Boruta and the 1DCNN algorithm based on the vehicle validation dataset at the current density between 1000 and 1485 mA/cm^2: (a) Keep all the features; (b) Remove the most important feature; (c) Remove the feature by XGBoost-Boruta.

Table 4
Comparative results amongst different feature removal strategy based on the vehicle running data at the current density between 1000 and 1485 mA/cm^2.

Category	RMSE (Train)	RMSE (Validation)	R^2 (Train)	R^2 (Validation)
Keep all the features	4.30	4.60	0.82	0.81
Remove the most important feature	5.56	6.34	0.71	0.65
Remove the features by XGBoost-Boruta	3.86	3.95	0.86	0.85

reducing the impact of human judgement error. At low current density below 200 mA/cm^2, the hydrogen inlet pressure has a significant effect on stack voltage output. When the current density ranges between 200 and 1100 mA/cm^2, the coolant outlet temperature is much more significant. With further increase of current density, the inlet mass flow rate, pressure and temperature of the air are then the significant features. Furthermore, two case studies were conducted for verification and validation based on bench test data and vehicle running data to demonstrate the accuracy and robustness of the proposed XGBoost-Boruta algorithm. By the introduction of the Z score and shadow features to eliminate the randomness of conventional ensemble learning methods, the RMSE values of the voltage validation dataset are 1.31 and 3.95 based on the bench test data and vehicle running data and 23.8% and 14.1% lower than the conventional feature selection approach. The R^2 values also increase from 0.06 to 0.04 to 0.92 and 0.85. The results demonstrate that the proposed XGBoost-Boruta algorithm delivers high accuracy, high robustness and low randomness for both steady bench datasets and dynamic vehicle running datasets. In future work, the hyperparameters of the 1DCNN could be further determined to enhance the prediction performance of the PEMFC voltage output.

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 authors do not have permission to share data.

References

[1] İnci M, Büyük M, Demir MH, İlbey G. A review and research on fuel cell electric vehicles: topologies, power electronic converters, energy management methods, technical challenges, marketing and future aspects. Renew Sustain Energy Rev 2020;137:2021.

[2] G. Reverdiau, A. Le, T. Alleau, T. Aribart, and C.C.E.A. Saclay, "ScienceDirect Will there be enough platinum for a large deployment of fuel cell electric vehicles ?," no. xxxx, pp. 1–13, 2021.

[3] de Almeida SCA, Kruczan R. Effects of drivetrain hybridization on fuel economy, performance and costs of a fuel cell hybrid electric vehicle. Int J Hydrogen Energy 2021. no. xxxx.

[4] Lü X, Qu Y, Wang Y, Qin C, Liu G. A comprehensive review on hybrid power system for PEMFC-HEV: issues and strategies. Energy Convers Manag 2018;171:1273–91.

[5] Mao L, et al. Evaluation method for feature selection in proton exchange membrane fuel cell fault diagnosis. IEEE Trans Ind Electron 2022;69(5):5277–86.

[6] He K, Zhang C, He Q, Wu Q, Jackson L, Mao L. Effectiveness of PEMFC historical state and operating mode in PEMFC prognosis. Int J Hydrogen Energy 2020;45(56):32355–66.

[7] Ding R, Wang R, Ding Y, Yin W, Liu J. Designing AI-aided analysis and prediction models for nonprecious metal electrocatalyst-based proton-exchange membrane fuel cells. Angew Chem Int Ed 2020.

[8] Khajeh-Hosseini-Dalasm N, Ahadian S, Fushinobu K, Okazaki K, Kawazoe Y. Prediction and analysis of the cathode catalyst layer performance of proton exchange membrane fuel cells using artificial neural network and statistical methods. J Power Sources 2011;196(8):3750–6.

[9] Bicer Y, Dincer I, Aydin M. Maximizing performance of fuel cell using artificial neural network approach for smart grid applications. Energy 2016;116:1205–17.

[10] Legala A, Zhao J, Li X. Machine learning modeling for proton exchange membrane fuel cell performance. Energy AI 2022;10:100183.

[11] Saengrung A, Abtahi A, Zilouchian A. Neural network model for a commercial PEM fuel cell system. J Power Sources 2007;172(2):749–59.

[12] Chávez-Ramírez AU, et al. High power fuel cell simulator based on artificial neural network. Int J Hydrogen Energy 2010;35(21):12125–33.

[13] Kheirandish A, Shafiabady N, Dahari M, Kazemi MS, Isa D. Modeling of commercial proton exchange membrane fuel cell using support vector machine. Int J Hydrogen Energy 2016;41(26):11351–8.

[14] Huo W, Li W, Zhang Z, Sun C, Zhou F, Gong G. Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection. Energy Convers Manag 2021;243:114367.

[15] Wang Y, et al. Degradation prediction of proton exchange membrane fuel cell stack using semi-empirical and data-driven methods. Energy AI 2022;11(September 2022):100205.

[16] Cahalan T, Rehfeldt S, Bauer M, Becker M, Klein H. Experimental set-up for analysis of membranes used in external membrane humidification of PEM fuel cells. Int J Hydrogen Energy 2016;41(31):13666–77.

[17] Huang KJ, Hwang SJ, Lai WH. The influence of humidification and temperature differences between inlet gases on water transport through the membrane of a proton exchange membrane fuel cell. J Power Sources 2015;284:77–85.

[18] Wang G, et al. Progress on design and development of polymer electrolyte membrane fuel cell systems for vehicle applications: a review. Fuel Process Technol 2018;179(September 2017):203–28.

[19] Daud WRW, Rosli RE, Majlan EH, Hamid SAA, Mohamed R, Husaini T. PEM fuel cell system control: a review. Renew Energy 2017;113:620–38.

[20] Baradaran R, Amirkhani H. Ensemble learning-based approach for improving generalization capability of machine reading comprehension systems. Neurocomputing 2021;466:229–42.

[21] Nkulikiyinka P, Yan Y, Güleç F, Manovic V, Clough PT. Prediction of sorption enhanced steam methane reforming products from machine learning based soft-sensor models. Energy AI 2020;2:100037.

[22] Li J, Ziehm W, Kimball J, Landers R, Park J. Physical-based training data collection approach for data-driven lithium-ion battery state-of-charge prediction. Energy AI 2021;5:100094.

[23] Kiangala SK, Wang Z. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. Mach Learn Appl 2021;4(December 2020):100024.

[24] Albashish D, Hammouri AI, Braik M, Atwan J, Sahran S. Binary biogeography-based optimization based SVM-RFE for feature selection. Appl Soft Comput 2021;101:107026.

[25] Masrur Ahmed AA, et al. Deep learning hybrid model with Boruta-Random forest optimizer algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. J Hydrol 2021;599(August 2020):126350.

[26] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36(11):1–13.

[27] Liu X, Zhou S, Yan Z, Zhong Z, Shikazono N, Hara S. Correlation between microstructures and macroscopic properties of nickel/yttria-stabilized zirconia (Ni-YSZ) anodes: meso-scale modeling and deep learning with convolutional neural networks. Energy AI 2022;7:100122.

[28] Wang B, et al. Numerical analysis of operating conditions effects on PEMFC with anode recirculation. Energy 2019;173:844–56.

[29] Yuan H, Dai H, Ming P, Wang X, Wei X. Quantitative analysis of internal polarization dynamics for polymer electrolyte membrane fuel cell by distribution of relaxation times of impedance. Appl Energy 2021;303:117640.

[30] Yuan H, Dai H, Ming P, Zhao L, Tang W, Wei X. Understanding dynamic behavior of proton exchange membrane fuel cell in the view of internal dynamics based on impedance. Chem Eng J 2022;431(P2). https://doi.org/10.1016/j.cej.2021.134035. 134035.