Machine learning and Bayesian optimization for performance prediction of proton-exchange membrane fuel cells

Soufian Echabarri, Phuc Do, Hai-Canh Vu, Bastien Bornand

▶ **To cite this version:**

Soufian Echabarri, Phuc Do, Hai-Canh Vu, Bastien Bornand. Machine learning and Bayesian optimization for performance prediction of proton-exchange membrane fuel cells. Energy and AI, 2024, 17, pp.100380. 10.1016/j.egyai.2024.100380 . hal-04611875

HAL Id: hal-04611875
https://hal.science/hal-04611875

Submitted on 14 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Energy and AI 17 (2024) 100380

Contents lists available at ScienceDirect

Energy and AI

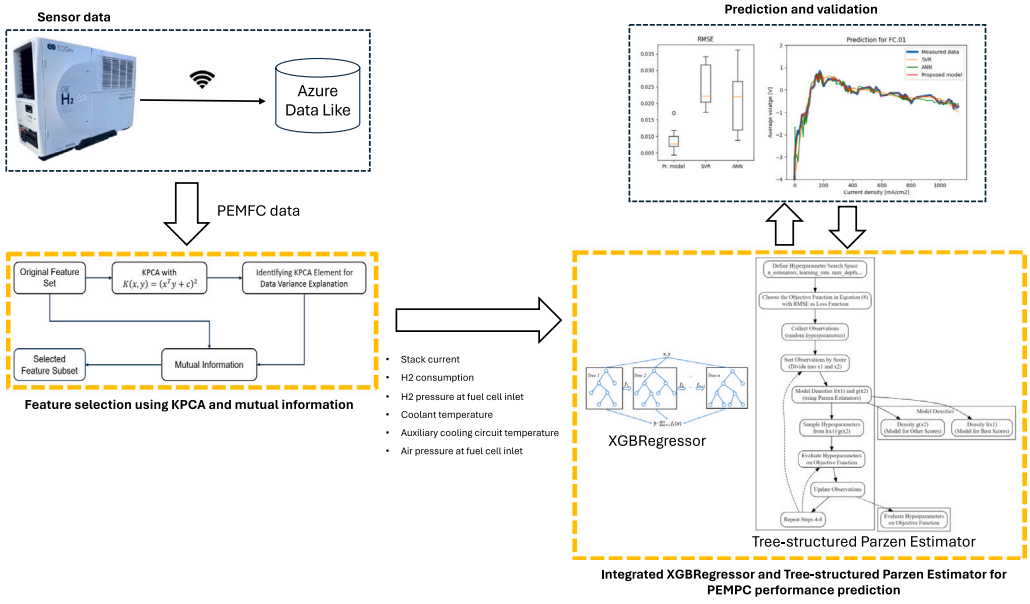
journal homepage: www.elsevier.com/locate/egyai

Machine learning and Bayesian optimization for performance prediction of proton-exchange membrane fuel cells

Soufian Echabarri [a,c], Phuc Do [a,*], Hai-Canh Vu [b], Bastien Bornand [c]

[a] Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France
[b] Université de Technologie de Compiègne (UTC), CS 60319, CEDEX, 60203 Compiègne, France
[c] Energy Designer and IT, EODev, Issy-Les-Moulineaux, 92130, Paris Region, France

GRAPHICAL ABSTRACT



HIGHLIGHTS

- An artificial-intelligence-based approach for PEMFCs performance prediction.
- A new feature selection method based on KPCA and mutual information.
- XGBRegressor and Tree-structured Parzen are jointly used for predicting the polarization curve.
- A comparison study with conventional machine learning prediction models.

ARTICLE INFO

Keywords:
Proton-exchange membrane fuel cell

ABSTRACT

Proton-exchange membrane fuel cells (PEMFCs) are critical components of zero-emission electro-hydrogen generators. Accurate performance prediction is vital to the optimal operation management and preventive

* Corresponding author.
E-mail address: phuc.do@univ-lorraine.fr (P. Do).

https://doi.org/10.1016/j.egyai.2024.100380
Received 14 March 2024; Received in revised form 11 May 2024; Accepted 28 May 2024
Available online 8 June 2024
2666-5468/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Hydrogen
Machine learning
XGBRegressor
Tree-structured Parzen estimator
Polarization curve
Performance prediction

maintenance of these generators. Polarization curve remains one of the most important features representing the performance of PEMFCs in terms of efficiency and durability. However, predicting the polarization curve is not trivial as PEMFCs involve complex electrochemical reactions that feature multiple nonlinear relationships between the operating variables as inputs and the voltage as outputs. Herein, we present an artificial-intelligence-based approach for predicting the PEMFCs' performance. In that way, we propose first an explainable solution for selecting the relevant features based on kernel principal component analysis and mutual information. Then, we develop a machine learning approach based on XGBRegressor and Bayesian optimization to explore the complex features and predict the PEMFCs' performance. The performance and the robustness of the proposed machine learning based prediction approach is tested and validated through a real industrial dataset including 10 PEMFCs. Furthermore, several comparison studies with XGBRegressor and the two popular machine learning-based methods in predicting PEMFC performance, such as artificial neural network (ANN) and support vector machine regressor (SVR) are also conducted. The obtained results show that the proposed approach is more robust and outperforms the two conventional methods and the XGBRegressor for all the considered PEMFCs. Indeed, according to the coefficient of determination criterion, the proposed model gains an improvement of 6.35%, 6.8%, and 4.8% compared with ANN, SVR, and XGBRegressor respectively.

1. Introduction

Renewable fuels are essential in the current energy mix and are increasingly used due to new energy policies aimed at significantly reducing greenhouse gas emissions. Fuel cells, particularly proton exchange membrane fuel cells (PEMFCs), are an efficient and clean way of utilizing energy. They can convert chemical energy into electrical energy and have several advantages, such as being environmentally friendly, having a high energy conversion efficiency, producing low noise, and operating at low temperatures. In this context, Energy Observer Development (EODev) aims to expand the use of renewable fuels, particularly hydrogen, as an energy carrier to realize a low-carbon society. The zero-emission electro-hydrogen generator (GEH2) is the most compact and efficient hydrogen generator available commercially in terms of power output (Fig. 1) [EODev products, Blue Diamond Machinery Brochure]. GEH2 is composed of *(a)* a PEMFC operating on dihydrogen; The PEMFC requires auxiliary systems such as pumps, cooling systems, and power supplies for operation; *(b)* a 44 kWh battery to optimize the service life of the PEMFC and satisfy customer requirements; *(c)* power conversion and control systems to ensure the smooth operation of the group. The main components of GEH2 and their characteristics are shown in Fig. 1.

EODev currently adheres to a systematic maintenance plan comprising more than 50 operations to ensure the long-term smooth operation of GEH2. Given the ever-changing and unpredictable nature of operational situations, as well as the gradual wear and tear of components over time, the current maintenance plan is inadequate to meet the real needs of GEH2. To reduce the maintenance cost while maintaining the proper functioning of GHE2, EODev aims to apply the predictive maintenance approach to deploy maintenance actions only at the right component and at the right time. For this purpose, the performance of GEH2's key components should be estimated to enable predictive maintenance decision-making. In that way, we focus on predicting the performance of PEMFC, a crucial component of GEH2.

The performance of a PEMFC refers to its ability to efficiently convert chemical energy, typically from hydrogen and oxygen, into electrical energy. This efficiency is generally assessed through various metrics and one of the most common ones is the polarization curve [1, 2] which is also known as the I–V curve. The polarization curve represents the relationship between the current density (I) and the voltage (V) output of the fuel cell under different operating conditions, such as fuel cell temperature, pressure at the fuel cell inlet, and air flow. It characterizes the electrochemical behavior of the fuel cell and provides insights into its efficiency and performance under different load conditions. Quantifying PEMFC performance involves analyzing the shape and characteristics of the polarization curve, like the voltage at which maximum power is achieved (peak power point), the slope of the curve, and any deviations from ideal behavior, such as voltage losses due to ohmic, activation, or mass transport losses.

A machine learning approach based on XGBRegressor and Bayesian optimization is developed to estimate PEMFC performance. The proposed approach is tested and validate though a real industrial dataset including 10 PEMFCs. When compared to XGBRegressor and conventional approaches such as artificial neural network (ANN) and support vector machine regressor (SVR), the proposed approach provides better results for all the considered PEMFCs.

In the next section, related works on the existing works on PEMFC performance prediction are discussed to identify the related research gaps and highlight the scientific contributions of this study.

2. Related studies and scientific contributions

2.1. Existing studies pertaining to PEMFC performance prediction

According to the authors of Wang et al. [1], Ding et al. [2], the polarization curve is selected as the focal point for the performance prediction model because of its ability to encompass the crucial properties of PEMFCs, including the current density, voltage, and other significant factors. Currently, three main approaches are employed to analyze the performance of PEMFCs: model-driven, hybrid, and data-driven approaches.

The model-driven approach forecasts the performance of PEMFCs based on physical and mathematical models of the associated electrochemical, transport, and thermal processes. These models can simulate PEMFC performance under a range of operational conditions and do not require a significant amount of data to construct. However, they require a comprehensive understanding of the underlying operational mechanisms and the interactions between components, as well as the incorporation of temporal and spatial elements. Zhao et al. [3] reviewed physics-based models for real-time control of PEMFCs and compared 1D physical models incorporating transport and electrochemical phenomena. Shi et al. [4] developed a mathematical model for a passive fuel cell fed with an e-fuel and examined the effects of various structural and operating conditions. Hasan et al. [5] introduced a numerical method to predict the lifetime and deterioration of membranes in PEMFCs, with emphasis on mechanical fatigue failure as a typical degradation mechanism that can result in PEMFC failure. Krishan et al. [6] investigated the correlation between electrode performance and drying techniques by constructing a dynamic two-dimensional physical continuum model that incorporated the sensitivity of the microstructure parameters of the catalyst layer. Kishimoto et al. [7] developed a numerical methodology for predicting electrochemical characteristics that considers various features, such as the current–voltage behavior, macroscopic properties, and impedance. Singh et al. [8] created a dynamic chemical degradation model to investigate the effect of membrane degradation on PEMFCs. Danilov and Tade [9] developed a new technique for estimating the cathodic and anodic charge transfer coefficients from PEMFC voltage-current curves. Kim et al. [10] formulated an equation to fit the cell potential to the current density data of PEMFCs under different conditions. This equation

PERFORMANCES	
Power output - ESP ISO rating	110 kVA / 88 kW
Power output - PRP ISO rating	100 kVA / 80 kW
Voltage output	400 VAC/480 VAC
Frequency output	50 Hz – 60 Hz
Operating temperature	-10°C to 40°C (without derating)
Protection index	IP 43

MAIN COMPONENTS	
Fuel cell brand	Toyota
Fuel cell type	PEM
Fuel cell efficiency	50 %
Battery type	LiFePO4
Battery capacity	44 kWh
Battery discharge rate	2C
External hydrogen expansion system included	

Fig. 1. Zero-emission electro-hydrogen generator [EODev GeH2 Specifications].

included an exponential term that accounted for the effects of mass transport, thus allowing slope changes and rapid potential drops to be captured. Meanwhile, Guinea et al. [11] developed a voltage-current model that considers the electron leakage current density to achieve accurate matching performance using gradient optimization methods and rotation.

The hybrid approach predicts the performance of PEMFCs based on physical models and historical data. Bressel et al. [12] proposed a novel approach using an extended Kalman filter-based observer to accurately estimate both the health status and degradation dynamics. Yue et al. [13] presented an online method for detecting and forecasting the degradation of PEMFCs. To establish a degradation indicator, they employed a nonlinear regression approach independent of the operating conditions and a multistep window-sliding echo-state network model to predict the future evolution of the identified degradation indicators. Pan et al. [14] introduced a hybrid methodology that combined a model-based adaptive Kalman filter with a data-driven nonlinear autoregressive exogenous model (NARX) neural network to predict the degradation of PEMFCs. The overall degradation trend was captured using an empirical aging model and an adaptive Kalman filter, whereas the intricate degradation specifics are depicted using the NARX neural network. Hu et al. [15] proposed a hybrid method for predicting the probability of performance degradation in PEMFCs to extend the service life and reduce maintenance costs. Wang et al. [16] presented a new method that combined the benefits of machine learning methods and semi-empirical models to predict the degradation of a PEMFC system comprising 300 cells. Zhou et al. [17] combined a physical aging model with time-delay neural networks to predict the deterioration of PEMFCs. A physical aging model was used to remove nonstationary trends from the original data, and the linear component was filtered using an autoregressive and moving-average model. Subsequently, the remaining nonlinear model was used to train the delayed neural networks, which were used to obtain the final prediction. Cheng et al. [18] proposed a method to enhance the precision of prognostic results in cases involving uncertain characterizations. They used the least-square support vector machine for initial prognostics and subsequently employed a regularized particle filter to determine the final probability distribution of the remaining useful life for PEMFCs.

The application of model-driven and hybrid approaches requires a certain level of physical knowledge regarding the system behavior, which may be difficult to gain for some complex applications. In this context, a data-driven approach for predicting PEMFC performance based purely on historical data has been extensively developed because of its remarkable flexibility and robust predictive capabilities [19]. For example, Wilberforce and Olabi [20] employed an artificial neural network (ANN) to predict the current and voltage of a PEMFC, thus minimizing the power required for fuel pumping and reducing the net losses in the cell. Legala et al. [21] conducted a comparative study between an ANN and a SVR (support vector machine regressor) for predicting variables such as cell voltage and membrane resistance. They showed that the ANN performed better than the SVR, particularly for multivariate output regression tasks. However, the SVR performed better in simpler regressions and reduced the computational load while maintaining high accuracy. He et al. [22] proposed an autoencoder-LSTM network model to predict the progress and mechanisms of PEMFC degradation during vehicle operation. In this method, a health indicator representing the PEMFC degradation states was employed, and LSTM analysis was performed. Kheirandish et al. [23] proposed a method for predicting the performance of a PEMFC system of a commercially available electrical bicycle using a support vector machine. Huo et al. [24] used a combination of Random Forest algorithm for feature selection and Convolutional Neural Networks (CNNs) for predicting the polarization curve of PEMFCs. Chen et al. [25] developed an insulation variation model using a data-driven long short-term memory neural network to identify insulation resistance value anomalies caused by deionizer failure in fuel cell vehicles. Falcão et al. [26] used a feed-forward ANN with three layers to predict the influence of relative humidity of the gases and fuel cell temperatures on the polarization curve. Li et al. [27] proposed a framework that combines a state-of-the-art meta-heuristic algorithm with a machine learning technique to predict performance and optimize parameters of PEMFCs. They developed a three-dimensional model to serve as the framework's data source, then construct a prediction model using the Kernel Extreme Learning Machine. Han and Chung [28] combined an ANN and an SVR to predict the stack performance of PEMFCs while considering the effects of different PEMFC operating conditions. Hong [29] used a deep belief network to build a model to predict the performance and maximize the power density of a PEMFC. Meanwhile, the authors of Zheng et al. [30] used long short-term memory (LSTM) to predict the performance of PEMFCs under dynamic conditions, particularly for vehicle applications. Chen et al. [31] used a gradient backpropagation neural network to predict the aging evolution of PEMFCs. The parameters of this model were adjusted using an evolutionary algorithm, which included a mental evolutionary algorithm, particle swarm optimization, and a genetic algorithm. Zuo et al. [32] developed a recurrent neural network model with an attention mechanism to optimize prognostic and health management predictions, thus promoting a more accurate anticipation of output voltage deterioration in PEMFCs. A comparison of recent research on polarization curve prediction based on data driven methods and the proposed method is summarized in Table 1.

According to the literature review and the comparative Table 1, existing models for estimating PEMFC performance have the following limitations: (1) the majority of models rely on data gathered from fuel cell aging test bench. Despite the meticulousness of these experiments,

Table 1
Comparisons of recent research works and the proposed method.

References	Feature selection	Prediction on various dynamic operating conditions	Number of PEMFCs used for model validation	Validation on real data	Prediction accuracy
Wilberforce and Olabi [20]	No	Yes	1	No	High
Legala et al. [21]	No	No	2	No	High
Han and Chung [33]	No	Yes	2	No	High
Hong [29]	No	Yes	1	No	High
Huo et al. [24]	Yes	No	140	No	Moderate
Chen et al. [31]	No	No	3	No	Moderate
Zuo et al. [32]	No	No	2	No	Medium
Kheirandish et al. [23]	No	No	1	No	Medium
Li et al. [27]	No	Yes	1	No	High
Proposed method	Yes	Yes	10	Yes	High

the data obtained may not fully represent the genuine behavior of the PEMFC when operating under real-world conditions; (2) in the case of dynamic operating conditions, the selection of parameters that control the performance of PEMFCs (relevant features) has been made on the basis of experience without any explanation; (3) the performance of the existing machine learning models was not robust enough and still limited in exploring complex operational data of real applications; (4) the optimization of the models' hyperparameters, a key factor in machine learning performance, has rarely been investigated. A new explainable and robust model that exploits the complex real-life operational data is required to provide a better prediction of PEMFC performance.

2.2. Scientific contributions

In this study, we propose a new explainable solution for the selection of the relevant feature and an efficient prediction approach based on the XGBRegressor and Tree-structured Parzen Estimator (TPE). The main contributions of this study are summarized as follows:

1. A new feature selection method based on KPCA and mutual information is developed to select relevant features for controlling the performance of PEMFCs. The method provides useful information for interpreting the results and understanding the factors that affect the polarization curve;
2. An efficient and robust method based on XGBRegressor (a powerful machine learning approach for the regression problem) and TPE (a new Bayesian method for machine learning optimization) is proposed to predict the polarization curve of PEMFCs, taking into account their dynamic operational conditions;
3. The proposed model is compared to XGBRegressor and two popular machine learning-based methods (ANN and SVR) in predicting PEMFC performance, on a complex industrial dataset provided by EODev.

The remainder of this paper is organized as follows: Section 3 presents a description of the PEMFC system and introduces the problem statement and the proposed methodology. Section 4 describes the selection of PEMFC characteristics using the proposed feature selection method. Section 5 presents a detailed description of the proposed prediction model based on the XGBRegressor and TPE. Section 6 presents the evaluation of the performance of the proposed method using the actual polarization curve data of 10 PEMFCs. Furthermore, a comparison study conducted using three established machine learning regressors, the XGBRegressor, the artificial neural network (ANN), and the support vector machine regressor (SVR), to predict the polarization curve is presented. Finally, conclusions inferred from this study are presented in Section 7.

3. System description and problem formulation

In this section, we describe the PEMFC system structure and its principle of working flows, as well as how the data was collected from different GEH2s under different operating conditions. Then, we describe the problem statement and introduce our proposed methodology.

3.1. PEMFC system structure and principle of working flows

The underlying structure of a PEMFC entails two electrodes, namely the anode and cathode, separated by a solid membrane serving as an electrolyte (Fig. 1). Hydrogen fuel passes through a network of channels to reach the anode, where it undergoes dissociation into protons and electrons. Specifically, hydrogen molecules dissociate into protons (H^+) and electrons (e^-) through the reaction:

$$H_2 \rightarrow 2H^+ + 2e^-$$

These protons migrate across the membrane towards the cathode, while the electrons travel through an external circuit connecting the two electrodes, thus generating electrical current. Concurrently, the oxidant, typically air in this study, follows a similar pathway through channels to reach the cathode. At the cathode, oxygen combines with the electrons from the external circuit and the protons migrating through the membrane, resulting in the formation of water according to the reaction:

$$1/2O_2 + 2H^+ + 2e^- \rightarrow H_2O$$

The overall cell reaction, which summarizes the electrochemical processes occurring within the PEMFC, can be expressed as:

$$H_2 + 1/2O_2 \rightarrow H_2O$$

As a result, the PEMFC produces water, DC electricity, and heat. The working principle of PEMFC is illustrated in Fig. 2.

The PEMFC system includes additional components, such as a pump for cooling water, a valve for cooling water temperature control, an air compressor, and a converter an inverter. Each of them plays a crucial role in the PEMFC system, contributing to its efficiency, reliability, and overall functionality. Furthermore, the PEMFC structure is designed to optimize the performance of the PEMFC system inside the GEH2. A detailed description of the PEMFC and its operation within the GEH2 will be addressed in the following section.

3.2. PEMFC system and its monitoring data

The data collected from several GEH2s for various components and sub-components provides useful information to better understand and analyze the PEMFC mechanisms that allow improving and optimizing the GEH2 performance.

Each GEH2 generator is equipped with an internet-connected data logger that acquires more than 1000 signals sampled at 100 ms and emitted by both the sensors and control units. Regarding the PEMFC system, 22 variables are monitored, including the stack current, stack tension, fuel cell temperature, etc. Table 2 summarizes all the monitored variables. Subsequently, the polarization curves are obtained using the acquired data after they are downsampled to 30 s time intervals, resulting in a database of 12,000 points. Ten identical PEMFCs deployed worldwide by EOdev provided experimental data for validation and training of our model under various conditions.

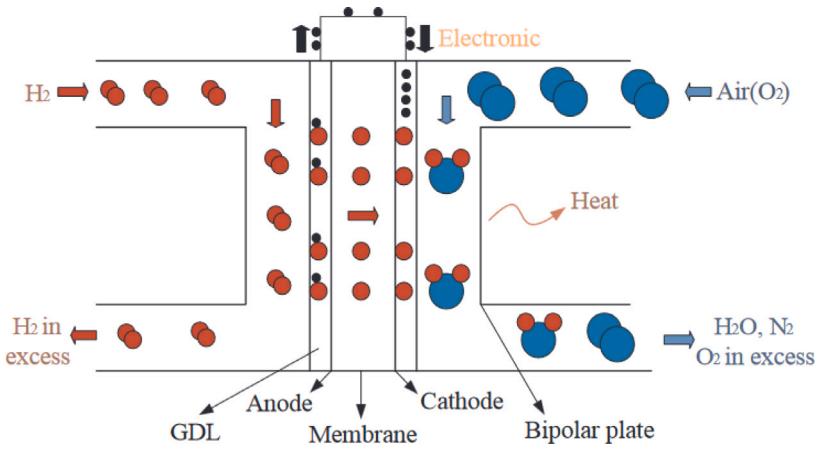


Fig. 2. The working principle of PEMFC [34].

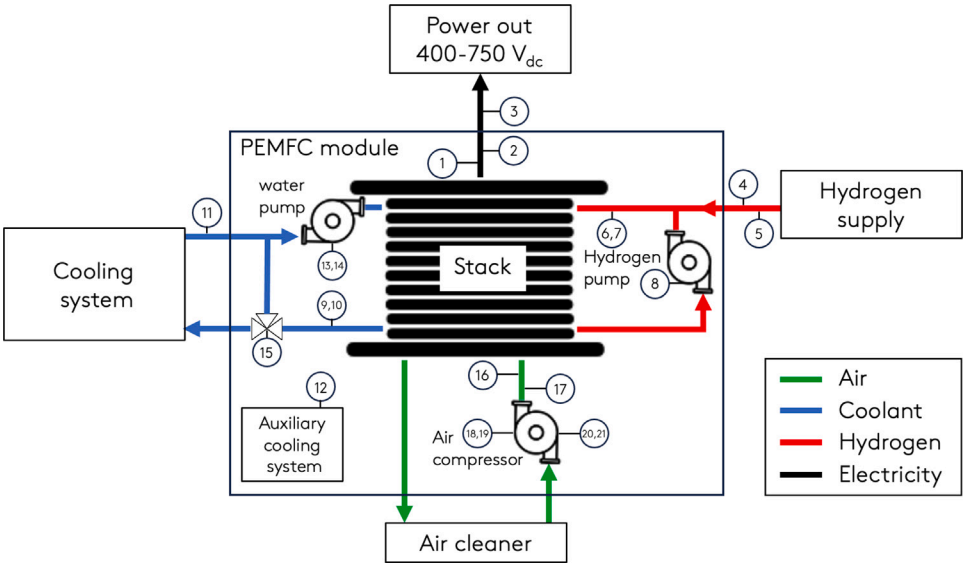


Fig. 3. Sensors location in the PEMFC module, components and auxiliary systems. The numeric labels correspond to the sensor reference in Table 2.

To comprehend the parameters outlined in Table 2 and their placement within the PEMFC system, we illustrate in Fig. 3 the sensors (represented by numeric labels) and the interdependencies among the sub-components.

3.3. Problem statement and proposed methodology

The main objective of this study is to predict the evolution of the PEMFC's performance, i.e., the polarization curve representing the relationship between PEMFC current density and voltage, from recorded real dataset mentioned in the previous section. In that way, the focus of this study is centered on comprehending and addressing the complex and nonlinear relationships that are inherent in the electrochemical reactions of PEMFCs. Therefore, a comprehensive model must be established that captures the intricate dynamic behavior of PEMFCs, incorporating various operational factors such as temperature, pressure, and airflow. Additionally, the model must accurately capture the nonlinear relationships between these parameters. Such relationships are often challenging to model accurately using conventional methods.

To overcome these challenges, we propose to use a machine learning model, namely XGBRegressor, that allows effectively modeling complex dynamic behavior by learning the nonlinear relationships between input variables and output variable. However, it is important to note that the predictive quality of XGBregressor is highly dependent on its estimated hyperparameters, so the selection of the best hyperparameters is a crucial and important step in our methodology. Therefore, it is proposed to integrate Tree-structured Parzen Estimator (TPE) into XGBregressor to optimize its hyperparameters. This approach allows us to search for the best hyperparameters and update them in a effective manner based on the model's performance. Unlike traditional methods such as random search and grid search, which can be limited in their capacity to thoroughly investigate the hyperparameter space, especially when complex dependencies are present, TPE concentrates on promising areas of the hyperparameter space, thus significantly decreasing the number of evaluations needed to find an optimal configuration.

In addition, due to the complexity and dynamic nature of PEMFC electrochemical reactions, careful identification of the operational variables that significantly influence performance is essential. Therefore, relevant variables need to be selected as inputs to the XGBRegressor model to achieve accurate and robust predictions. To address this, a combination of KPCA and mutual information was used for feature selection. This approach aims to overcome the limitations of traditional approaches by capturing nonlinear relationships between variables and identifies complex interactions that may affect model performance. Indeed, KPCA transforms data into a nonlinear feature space, providing a better representation of complex interactions, while mutual information quantifies the dependency between variables, aiding in the selection of influential features. This feature approach enhances

Table 2
Operating variables of the PEMFC.

Sensor reference	Variable name	Description	Unit	Type
1	mes_u_stack	Stack tension	V	measurement
2	mes_i_stack	Stack current	*A*	measurement
3	mes_p_fc_net	Fuel cell net power	W	measurement
4	mes_d_fc_h2In	Instantaneous H2 consumption	mg	measurement
5	mes_pr_fc_h2Mid	H2 supply pressure	kPa	measurement
6	mes_pr_fc_h2Low	H2 pressure at fuel cell inlet	kPa	measurement
7	req_pr_fc_h2Low	H2 pressure at fuel cell inlet	kPa	request
8	req_n_h2Pump	H2 pump speed	rpm	measurement
9	mes_t_fc	Fuel cell temperature	°C	measurement
10	mes_t_fc_out	Coolant temperature at fuel cell outlet	°C	measurement
11	mes_t_rad_out	Coolant temperature at radiator outlet	°C	measurement
12	mes_t_aux_out	Auxiliary cooling circuit temperature at radiator inlet	°C	measurement
13	mes_n_wp	Water pump speed	rpm	measurement
14	req_n_wp	Water pump speed	rpm	request
15	mes_prc_3wVlv_opn	3 way valve opening rate	%	measurement
16	mes_d_aSup	Airflow	NL/ min	measurement
17	mes_pr_aSup	Air pressure at fuel cell inlet	kPa	measurement
18	mes_t_aComp_inv	Air compressor inverter temperature	°C	measurement
19	mes_t_aComp_mot	Air compressor internal temperature	°C	measurement
20	mes_n_aComp	Air compressor speed	rpm	measurement
21	req_n_aComp	Air compressor speed	rpm	request
22	st_fc_mod	Fuel cell mode (start/stop/generating/off)	–	state

the understanding of the parameters controlling the PEMFC system, improving overall performance understanding.

Due to the complex and ever-changing nature of electrochemical reactions in PEMFCs, it is crucial to carefully identify the operational variables that have a significant impact on PEMFC performance. Therefore, selecting the appropriate variables to use as inputs for the Proposed XGBRegressor prediction model is a critical step in achieving accurate and reliable predictions. To address this challenge, the following section combines Kernel Principal Component Analysis (KPCA) and mutual information to select the most relevant features.

In conclusion, the proposed methodology for predicting the PEMFC performance can be divided by two steps: (1) incorporating KPCA and mutual information for feature selection and (2) combining XGBRegressor machine learning model and TPE for the PEMFC performance prediction. These methodological choices address the challenges arising from the complexity of electrochemical reactions and the dynamic behavior of PEMFCs, ultimately improving the quality, robustness, and relevance of the predictive model. The detailed description of the proposed methodology will be presented in Sections 4 and 5.

4. Features selection using KPCA and mutual information

Feature selection, integral to the method of dimensionality reduction, aims to decrease the dimensionality of data by eliminating irrelevant or redundant variables. On the other hand, feature extraction techniques achieve dimensionality reduction by combining variables. These two dimensionality reduction techniques will be discussed in detail in Section 4.1.

PEMFCs exhibit intricate electrochemical reactions involving multiple nonlinear relationships between their operating variables and the average PEMFC stack voltage. Feature selection yields insignificant results, particularly when variables that are supposed to control the polarization curve, such as current, temperature, and pressure, are not selected. For extraction methods, model accuracy is more important than interpretability. In this study, we propose a selection method based on KPCA and mutual information to select relevant features while ensuring the interpretability of the model. This is discussed in more detail in Section 4.2.

4.1. Fundamentals of feature selection and feature extraction

Feature selection

Feature selection involves selecting the most relevant features from a dataset [35]. The main advantage of this method is that it reduces the dimensionality of the dataset while conserving information to improve the model performance by reducing overfitting and improving interpretability. This reduces the complexity of the model, thus rendering it easier to understand and interpret [36]. The overall feature selection process is illustrated in Fig. 4.

Feature selection can be accomplished via several approaches, including filter, wrapper, embedded, and ensemble methods. Among them, filter methods are considered the oldest and are also known as open-loop methods. They involve evaluating the relevance of features with respect to the target variable independent of the model. Moreover, feature characteristics are typically measured based on dependence, information, consistency, and distance [37]. By contrast, wrapper methods, also known as closed-loop methods, are based on the performance of learning algorithms. In these methods, features are evaluated using a machine learning model and the most relevant features are searched based on the performance accuracy [38]. Embedded methods, which is similar to wrapper methods, are different in that they perform feature selection during model training by incorporating features into the feature extraction algorithm. This implies that the features are selected during the model implementation process [39]. Finally, the ensemble method involves creating multiple feature subsets and combining the results to obtain a more robust outcome. This approach relies on several subsampling techniques, wherein a specific feature selection method is applied to different subsamples and their resulting features are merged

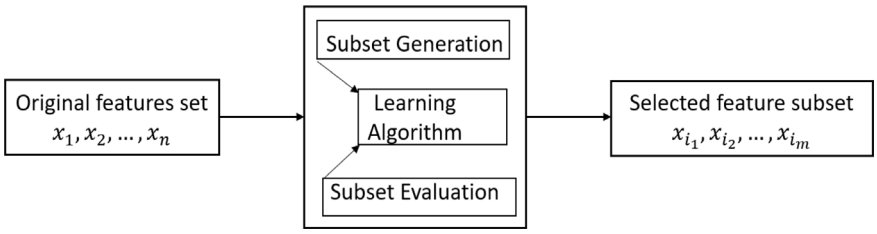


Fig. 4. Process of feature selection.

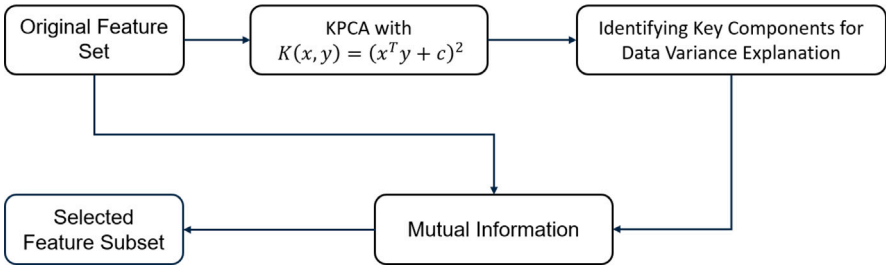


Fig. 5. PEMFC features selection.

to form a more stable subset. In summary, each of these methods has its own advantages and disadvantages. In Dash and Liu [37], a detailed explanation pertaining to the selection of the best method to adapt to our data is provided by highlighting the advantages and disadvantages of each approach.

Feature extraction

Feature extraction involves transforming the original data into a new set of features that is more representative of the underlying patterns of the data. The most well-established methods are principal component analysis (PCA), KPCA, multidimensional scaling, locally linear embedding, or independent component analysis. Feature extraction can be useful when many features are present in the data, some of which are highly correlated, because it can reduce the number of features without excessive information loss. For more details on feature extraction, please refer to Elhadad et al. [40], Aziz et al. [41].

4.2. Proposed feature selection method

As mentioned above, a PEMFC undergoes intricate electrochemical reactions involving multiple nonlinear relationships between the operating variables of the PEMFC as inputs and the average PEMFC stack voltage as the output. To select the relevant features, we propose applying KPCA to extract KPCA components that explain the data and then calculate the mutual information between these KPCA components and all PEMFC variables to extract the relevant variables as described in Fig. 5.

KPCA [42] is an extended form of PCA that relies on kernel techniques to perform nonlinear dimensionality reduction. The basic idea behind KPCA is to transform the source data into a high-dimensional feature space through a nonlinear mapping function and then perform PCA in that feature space. This technique allows KPCA to capture the nonlinear relationships between data points that cannot be detected via linear PCA. The steps for reducing the dimensionality via KPCA are outlined as follows:

- Construct the kernel matrix K. In our study, we choose the polynomial kernel

$$K_{i,j} = \kappa\left(\mathbf{x_i}, \mathbf{x_j}\right) = (x_i^T x_j + 1)^2. \tag{1}$$

- Compute the gram matrix $\tilde{K}$ using to the following equation:

$$\tilde{K} = K - \mathbf{1_N}K - K\mathbf{1_N} + \mathbf{1_N}K\mathbf{1_N}, \tag{2}$$

where N is the number of data points and $\mathbf{1_N}$ is the $N \times N$ matrix with all elements equal to $1/N$.

- Find the vector a_k by solving the following equation:

$$\tilde{K}a_k = \lambda_k N a_k, \tag{3}$$

where $a_k = [a_{k1}, a_{k2}, \ldots, a_{kN}]^T$ are the eigenvalues of $\tilde{K}$ and λ are the corresponding eigenvectors.

- Finally, compute the kernel principal components $y_k(x)$

$$y_k(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{v}_k = \sum_{i=1}^{N} a_{ki}\kappa\left(\mathbf{x_i}, \mathbf{x_j}\right). \tag{4}$$

To ensure the reliability of our KPCA, the input data were normalized before the KPCA was applied. After applying KPCA to the data, the number of KPCA components that explained the main variance in the data was determined.

To select the relevant variables that affect the polarization curve of the PEMFC and determine the parameters that significantly affect the first KPCA element, we applied the mutual information method. In probability and information theories, the mutual information of two random variables is used to quantify the statistical dependence. If the variables are independent, then the mutual information is zero; however, it increases with the statistical dependence. Mutual Information is mathematically defined as follows:

- in the discrete case:

$$I(X;Y) = \sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}, \tag{5}$$

- in the continuous case:

$$I(X;Y) = \int_{\mathbb{R}}\int_{\mathbb{R}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}dxdy, \tag{6}$$

where $P(x,y)$, $P(x)$ and $P(y)$ represents the densities of (X,Y), X and Y, respectively.

We applied mutual information between the KPCA components that represented the data and the operating variables of the PEMFC. The variables selected using the proposed hybrid approach are shown in Table 3.

We compare our proposed selection technique with other techniques such as the Pearson correlation and mutual information (filter methods), recursive feature elimination–random forest and genetic algorithm (wrapper methods), and auto-encoder as well as Lasso and ridge regressors (embedded methods) in Section 6.

Table 3
Selected variables with the proposed method.

Sensor reference	Variable name	Description
2	mes_i_stack	Stack current
4	mes_d_fc_h2In	Instantaneous H2 consumption
6	mes_pr_fc_h2Low	H2 pressure at fuel cell inlet
10	mes_t_fc_out	Coolant temperature at fuel cell outlet
12	mes_t_aux_out	Auxiliary cooling circuit temperature at radiator inlet
17	mes_pr_aSup	Air pressure at fuel cell inlet

5. Integrated XGBRegressor and TPE model based prediction

In this section, we present a method to predict the PEMFC performance based on the relevant variables identified in the previous step using the XGBRegressor and TPE. Additionally, evaluation criteria are presented at the end of this section to evaluate the performance of the proposed model.

5.1. XGBRegressor and tree-structured Parzen estimator

5.1.1. XGBRegressor

Extreme Gradient Boosting (XGBoost) is a library that provides an efficient implementation of the gradient boosting ensemble algorithm based on decision trees. The XGBRegressor is a version of XGBoost designed to perform regression tasks. The objective function of XGBoost at the tth iteration is defined as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right) + \Omega\left(f_t\right), \tag{7}$$

where l is a differentiable convex loss function; x_i and y_i are the observation vector and the actual value of observation i, respectively; f_t is the prediction function of tree t; and $\hat{y}_i^{(t)}$ is the prediction of the observation ith in the t-th iteration. The second term Ω is a regularization that penalizes the regression tree functions and is defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2, \tag{8}$$

where T is the total number of leaves in the tree, w is the leaf weights, and γ and λ are hyperparameters control the regularization strength.

As shown, the function $\mathcal{L}^{(t)}$ cannot be optimized using traditional optimization techniques in Euclidean space. Therefore, this function must be transformed into a function in the Euclidean domain. Hence, a second-order Taylor approximation was applied to obtain a new form of the objective function as follows:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2}h_i f_t^2\left(\mathbf{x}_i\right)\right] + \Omega\left(f_t\right), \tag{9}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$. By removing the constant terms, we obtain the following simplified form in step t:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2}h_i f_t^2\left(\mathbf{x}_i\right)\right] + \Omega\left(f_t\right). \tag{10}$$

For more details regarding the construction of the next learner and the method to measure the quality of a tree structure, please refer to Chen et al. [31].

5.1.2. Tree-structured Parzen estimator

In this study, we aim to estimate the hyperparameters of the XGBRegressor using the TPE [43]. The TPE is a Bayesian-based method for tuning model hyperparameters. Let θ and y be the hyperparameter and loss function of the model, respectively. After selecting a new set of hyperparameters, the improvement (EI) of the model can be expressed as follows:

$$EI_{y^*}(\theta) = \int_{-\infty}^{y^*}\left(y^* - y\right)p(y \mid \theta)dy, \tag{11}$$

where y^* is a control parameter.

To tune the hyperparameters, TPE simulates $p(y \mid \theta)$ by simulating $p(\theta \mid y)$ and $p(y)$ indirectly.

Therefore, we replace $p(y \mid \theta)$ in Eq. (11) and EI is expressed as:

$$EI_{y^*}(\theta) = \int_{-\infty}^{y^*}\left(y^* - y\right)\frac{p(\theta \mid y)p(y)}{p(\theta)}dy, \tag{12}$$

where $p(\theta \mid y)$ is the probability density defined as the piecewise function in terms of y:

$$p(\theta \mid y) = \begin{cases} l(\theta) & \text{if} \quad y < y^* \\ g(\theta) & \text{if} \quad y \geq y^* \end{cases}, \tag{13}$$

where $l(\theta)$ and $g(\theta)$ are two probability densities formed by loss values less than and greater than y^*, respectively. Therefore, if we consider

$$\gamma = p\left(y < y^*\right) \tag{14}$$

we obtain

$$p(\theta) = \int_{\mathbb{R}} p(\theta \mid y)p(y)dy = \gamma l(\theta) + (1-\gamma)g(\theta). \tag{15}$$

Furthermore, EI can be written as follows:

$$EI_{y^*}(x) = \frac{\gamma y^* l(\theta) - l(\theta)\int_{-\infty}^{y^*} p(y)dy}{\gamma^{l(\theta)} + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1}. \tag{16}$$

Based on Eq. (16), to determine the hyperparameters that yield the highest EI, the TPE algorithm assesses the hyperparameters using the ratio of $g(\theta)/l(\theta)$ and selects the hyperparameters θ^* that yield the maximum EI.

5.2. Hyperparameters estimation using TPE

The process of estimating the XGBRegressor hyperparameters using TPE is depicted in Fig. 6 in which solid lines depict the core, non-iterative steps that establish the framework for the TPE algorithm. They define the initial setup (specifying the search space and objective function) and the actions taken after each iteration (evaluating newly sampled configurations and updating observations). In the oher hand, dashed lines indicate a loop that encompasses multiple steps within each iteration. It essentially signifies that steps 4 through 8 are repeated for a fixed number of iterations. These steps involve sorting observations, modeling densities, sampling new hyperparameters, evaluating them, and updating the observations. All the process steps illustrated in Fig. 6 are specified as follows:

- **Step 1:** In this initial step, we start by establishing the potential range of values for the hyperparameters to be set in the XGBRegressor model. These parameters typically include the number of estimators, maximum depth, learning rate and colsample_bytree. This space of hyperparameters forms the basis for exploring the TPE algorithm.
- **Step 2:** To quantify the model efficacy, the objective function represented in Eq. (7) is utilized and the root mean square error is designated as the loss function. This directs the optimization process by minimizing its value.
- **Step 3:** This entails randomly sampling a set of hyperparameter configurations from the defined search space in step 1. These configurations represent the initial set of "observations" used by the TPE algorithm.

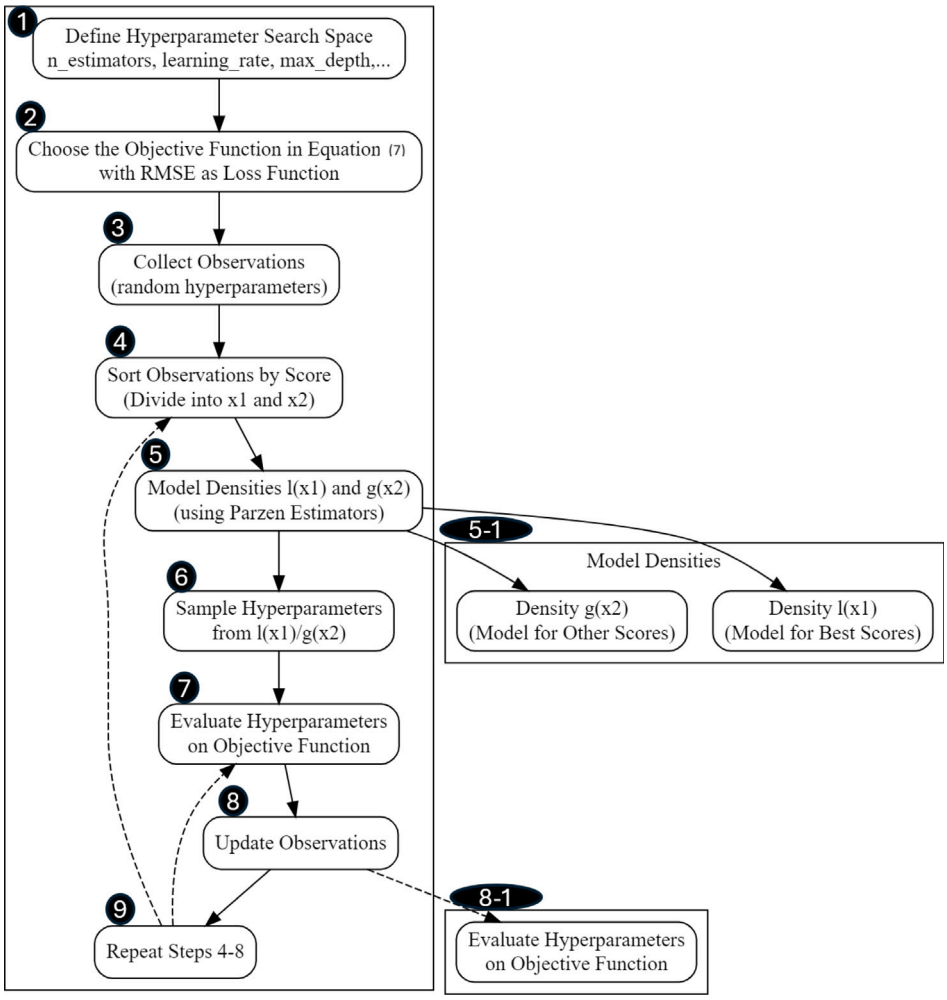


Fig. 6. Process for estimating XGBRegressor parameters using TPE.

- **Step 4:** Here, the performance of each randomly chosen hyperparameter configuration is assessed using the designated objective function. The observations are then sorted based on their scores and segregated them into two groups, where the first group contains the best-performing configurations (x_1) and the second contains the remaining configurations (x_2).
- **Step 5:** The TPE algorithm employs a technique known as Parzen density estimation to approximate the underlying probability distributions of the two sets of observations, (x_1) and (x_2). These densities, denoted by $l(x_1)$ and $g(x_2)$ (as shown in **step 5-1**), respectively, represent the likelihood of encountering a specific hyperparameter configuration within each group.
- **Step 6:** This step leverages the modeled densities to strategically select new hyperparameter configurations. As shown in Eq. (16), EI is proportional to $g(x_2)/l(x_1)$. Therefore, we obtained sample hyperparameters from $l(x_1)$ and evaluated them relative to $g(x_2)/l(x_1)$. The set that minimized this ratio and corresponded to the highest EI was selected. This approach steers the search towards regions in the hyperparameter space that are more likely to yield superior performing models.
- **Step 7:** The newly sampled hyperparameter configurations are assessed using the objective function, similar to how the initial random samples were evaluated in step 4.
- **Step 8:** Incorporate the newly observed performance metrics into the Bayesian model. By updating the probability distributions of hyperparameters based on observed outcomes, the algorithm adapts its search strategy to focus on promising regions of the hyperparameter space (**step 8-1**).
- **Step 9:** This iterative process, encompassing steps 4 through 8, is repeated for a fixed number of iterations. The core principle behind TPE is its iterative refinement of the search space. By prioritizing regions with promising configurations and progressively incorporating new information, TPE aims to efficiently locate optimal hyperparameters for the XGBregressor model.

Finally, to evaluate the performance of the proposed model, three measures were used: the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These three metrics are expressed mathematically as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_i - \hat{V}_i)^2}. \tag{17}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|V_i - \hat{V}_i|. \tag{18}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(V_i - \hat{V}_i)^2}{\sum_{i=1}^{n}(V_i - \overline{V})^2}, \tag{19}$$

where n is the number of observations, V_i the average observed PEMFC stack voltage, $\hat{V}_i$ the predicted value, and $\overline{V}$ the mean value of the average PEMFC stack voltage observed.

Detailed results and discussions obtained from 10 PEMFCs datasets will be addressed in Section 6.

Table 4
Results of different feature selection methods in terms of RMSE and R^2.

		XGBRegressor		ANN		Number of
		RMSE	R^2	RMSE	R^2	selected features
Filter method	Mutual Information	0.0501	0.8717	0.0588	0.7354	4
	Correlation	0.1065	0.5919	0.1633	0.5615	7
Wrapper method	RFE-Random Forest	0.0318	0.9376	0.0487	0.8766	9
	Genetic algorithm	0.0914	0.7590	0.0766	0.6918	12
Embedded method	Auto-encoder	0.0421	0.9054	0.0549	0.7422	8
	Lasso	0.0652	0.8577	0.0789	0.6700	13
	Ridge	0.0682	0.8505	0.0734	0.6692	9
Proposed method		**0.0112**	**0.9888**	**0.0362**	**0.9240**	6

Table 5
Hyper-parameters estimated by the proposed XGBRegressor model for the 10 PEMFCs.

Hyper-parameters	FC.01	FC.02	FC.03	FC.04	FC.05	FC.06	FC.07	FC.08	FC.09	FC.10
n_estimators	1600	1500	1800	400	1600	800	1400	1200	200	400
max_depth	8	12	8	6	2	2	2	4	2	12
learning_rate	0.1335	0.0607	0.1095	0.0905	0.0119	0.0375	0.1252	0.1167	0.1398	0.1477
colsample_bytree	0.55	1	0.7	0.6	1	0.5	0.6	0.8	0.85	0.9

6. Results and discussions

We applied the proposed model to predict the polarization curve using real data acquired from 10 distinct PEMFCs associated with different GEH2s. To assess and contrast the effectiveness of the proposed model, we conducted benchmarking against XGBRegressor alone, as well as two widely used machine learning-based methods (ANN and SVR) for predicting PEMFC performance.

ANNs are biological neural networks that develop structures of the human brain. Similar to the human brain with interconnected neurons, ANNs feature interconnected neurons in various layers of the network. The performance of ANNs has been demonstrated in several applications, including regression problems [44,45]. In this study, the ANN was designed with an input layer comprising six variables, two hidden layers comprising 64 and 32 neurons each, and an output layer comprising a single neuron. The activation function for the first two layers was a ReLU and that for the third layer was linear. The loss function selected was the RMSE, which was minimized using the Adam optimizer. The training data were segmented into batches of size 32 and the model was trained over 50 epochs. A schematic illustration of the ANN, where the variables selected in the previous selection were used as feature vectors (inputs) to predict the PEMFC polarization curve, is shown in Fig. A.1 (Appendix A).

The SVR is an extension of the support vector machine (SVM) applied to regression analysis [46,47]. Its purpose is to identify a regression function that predicts continuous values by maximizing the margin between the predicted and actual values while controlling the complexity of the model. We focus on the data points closest to the margin, known as support vectors, to construct the regression function. The SVR can use different kernel functions such as the radial basis function to capture nonlinear relationships. It solves an optimization problem that balances prediction errors and model regularization, similar to SVMs, for classification. In this study, the SVR used to predict the polarization curve comprised a Gaussian kernel, an epsilon error tolerance of 0.025, a regularization parameter C of 5, and a kernel-independent term of 0.01. Fig. A.2 (Appendix A) shows a schematic illustration of the SVR model used to predict the polarization curve of the PEMFC.

As discussed in Section 4.2, KPCA and mutual information were jointly used to select the relevant features. To demonstrate the effectiveness of this method, we compared it with different feature selection methods. The results of this comparison are presented in Table 4. As shown, the proposed feature selection method yielded better results than the other methods presented in the table, including the XGBRegressor and ANN prediction models.

As shown in Table 4, feature selection methods lead to different numbers of selected features primarily due to algorithmic differences, scoring metrics, search strategies, etc. Indeed, the algorithms of some methods prioritize features based on distinct criteria. Some methods may focus on individual feature importance, while others may consider feature interactions or correlations. Feature selection methods often employ different scoring metrics to evaluate feature relevance. Metrics such as information gain, correlation coefficients, or model performance metrics, may lead to different feature selections. The search strategies employed by feature selection methods, such as forward selection, backward elimination, or recursive feature elimination, can influence the number of selected features. Therefore, the evaluation of prediction quality presented in Table 4 allows determining whether the selected features are optimal, taking into account both the number and the specific features selected.

6.1. Hyper-parameters tuning

The different hyperparameters of the proposed XGBRegressor model, such as the number estimators, maximum depth, learning rate, and colsample_bytree, were first optimized for each PEMFC using the TPE (see Section 5.2). The optimization process is illustrated in Fig. 6. The representative hyperparameters of the 10 PEMFCs are presented in Table 5.

6.2. Validation and discussion pertaining to application of proposed method

To evaluate and validate the performance of the proposed method, we compared its prediction results with real data provided by 10 PEMFCs, as well as with XGBregressor, SVR, and ANN. As outlined in Section 3, the PEMFC system comprises 22 measured parameters, some of which directly influence its performance, as detailed in Table 3. In Table 6, we present the possible range of values for these parameters. It is crucial to emphasize that each PEMFC operates under unique conditions, resulting in varying parameter values within the range presented in Table 6.

The results on four first PEMFCs (FC.01, FC.02, FC.03 and FC.04) are shown in Figs. 7–8.

Considering the variability in PEMFC performance resulting from diverse operating conditions, our evaluation incorporates these conditions, as well as not all the GEH2 units operate for the same duration. This reflects the various scenarios encountered in real-world applications. These results not only demonstrate the robustness of our methodology, but also demonstrate the flexibility and the adaptability

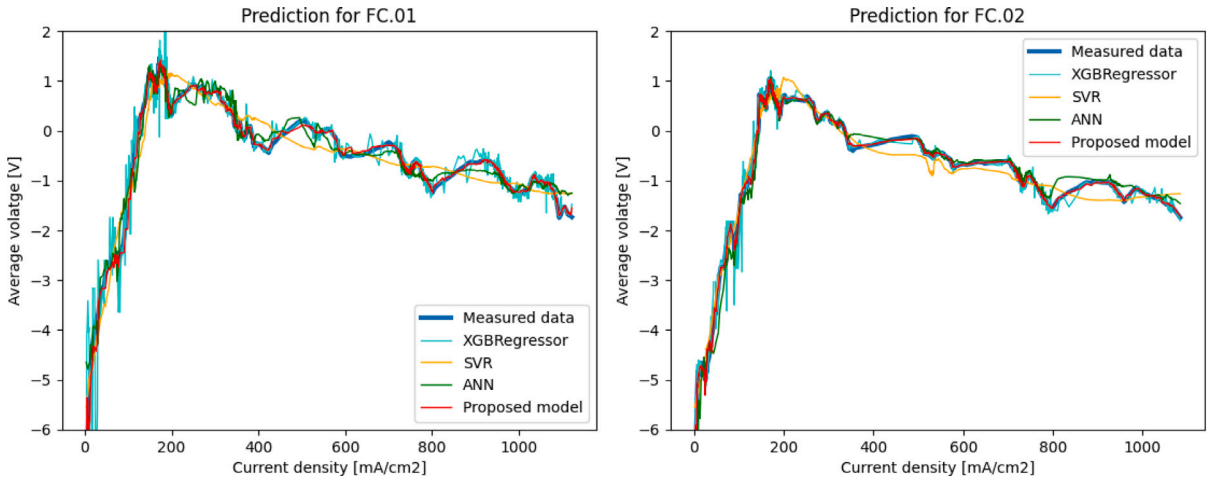


Fig. 7. Polarization curve prediction for FC.01 and FC.02.

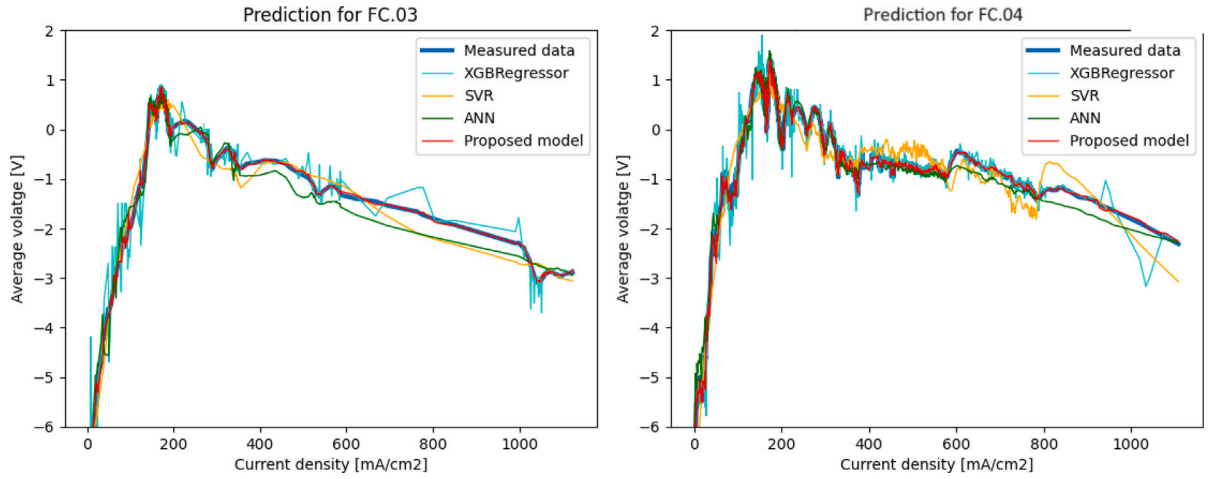


Fig. 8. Polarization curve prediction for FC.03 and FC.04.

Table 6
Range of PEMFC parameters..

Parameter	Unit	Range
Stack current	A	0–290
Instantaneous H2 consumption	mg	0–585
H2 pressure at fuel cell inlet	kPa	50–282
Coolant temperature at fuel cell outlet	°C	15–69
Auxiliary cooling circuit temperature at radiator inlet	°C	16–61
Air pressure at fuel cell inlet	kPa	80–206

of the proposed method to various sets of PEMFC units operating under different operating conditions.

Similar results on the six remaining PEMFCs are shown in Appendix B. The metrics (RMSE, MAE, and R2) used to evaluate the results were computed for 10 PEMFCs, as shown in Table 7.

The results confirmed that the proposed model outperformed the XGBRegressor, ANN, and SVR models. In fact, our predicted values were almost identical to the measured voltages, whereas the curves predicted by XGBRegressor, ANN, and SVR deviated significantly from the actual ones, particularly in cases where the current density exceeded 180 mA/cm^2. Moreover, the proposed model was more robust than the three benchmark models because it consistently yielded favorable performance when applied to different PEMFCs. Based on the PEMFC specifications, values below 180 mA/cm^2. are outside the normal operating range and are thus considered transient values that occur temporarily at the beginning and completion of the process. Moreover, downsampling the data to 30 s time intervals may have reduced the coherence in the dataset for such transient values, thus causing the models to be less effective in predicting complex phenomena occurring at low current densities.

To estimate the performance of our model more precisely, three evaluation metrics (i.e., RMSE, MAE, and R^2) were used, and the results are listed in Table 7. Finally, the box plots in Fig. 9 further support our conclusions.

As shown by the results presented in Table 7, and Fig. 9 our method consistently yielded results that were better than those yielded by XGBRegressor, ANN, and SVR. For example, the mean R^2 value of the proposed approach was 0.9917, whereas those of SVR, ANN, and XGBRegressor were only 0.9312, 0.9281, and 0.9457, respectively. In addition, as shown Table 7, the proposed model showed significant improvement in terms of all the prediction effectiveness metrics compared with the three benchmark models. Moreover, when compared with the ANN model, the proposed model provided lower RMSE and MAE values by 59.47% and 71.15% respectively, and a higher R^2 value by 6.85%. When compared with the SVR model, the proposed approach gave lover RMSE and MAE values by 64.06% and 84.45% respectively, and a higher R^2 value by 6.49%. Meanwhile, the proposed method indicated lower RMSE and MAE values by 57.40% and 71.69%, respectively, and a higher R^2 value by 4.8%, compared with the XGBRegressor model.

PEMFCs have the same capabilities and characteristics but operate under different conditions. Therefore, the predictive quality of the

Table 7
Polarization curve results of the three methods in terms of RMSE, MAE, and R^2.

	Proposed model			SVR			ANN			XGBRegressor		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
FC.01	0.0076	0.0024	0.9951	0.0173	0.0109	0.9745	0.0223	0.0099	0.9580	0.0219	0.0044	0.9535
FC.02	0.0068	0.0022	0.9943	0.0181	0.0128	0.9599	0.013	0.0071	0.9792	0.0119	0.0030	0.9786
FC.03	0.0080	0.0028	0.9954	0.0306	0.0268	0.9323	0.0262	0.0103	0.9505	0.0328	0.0209	0.8841
FC.04	0.0078	0.0027	0.9894	0.0211	0.0152	0.9241	0.0109	0.0059	0.9361	0.0243	0.0147	0.9282
FC.05	0.0119	0.0042	0.9919	0.0341	0.0299	0.9336	0.022	0.0091	0.9651	0.0271	0.0180	0.9458
FC.06	0.0112	0.0052	0.9888	0.0222	0.0154	0.9554	0.0198	0.0095	0.9647	0.0133	0.0154	0.9572
FC.07	0.0060	0.0026	0.9905	0.0216	0.0189	0.8753	0.0395	0.0247	0.6640	0.0261	0.0048	0.9443
FC.08	0.0089	0.0032	0.9954	0.0327	0.0237	0.9336	0.0362	0.0145	0.9240	0.0178	0.0107	0.9456
FC.09	0.0070	0.0021	0.9907	0.0250	0.0211	0.8816	0.0104	0.0051	0.9769	0.0144	0.0021	0.9693
FC.10	0.0171	0.0030	0.9855	0.0342	0.0186	0.9423	0.0272	0.0081	0.9634	0.0270	0.0120	0.9508
Mean	**0.0092**	**0.0030**	**0.9917**	**0.0256**	**0.0193**	**0.9312**	**0.0227**	**0.0104**	**0.9281**	**0.0216**	**0.0106**	**0.9457**

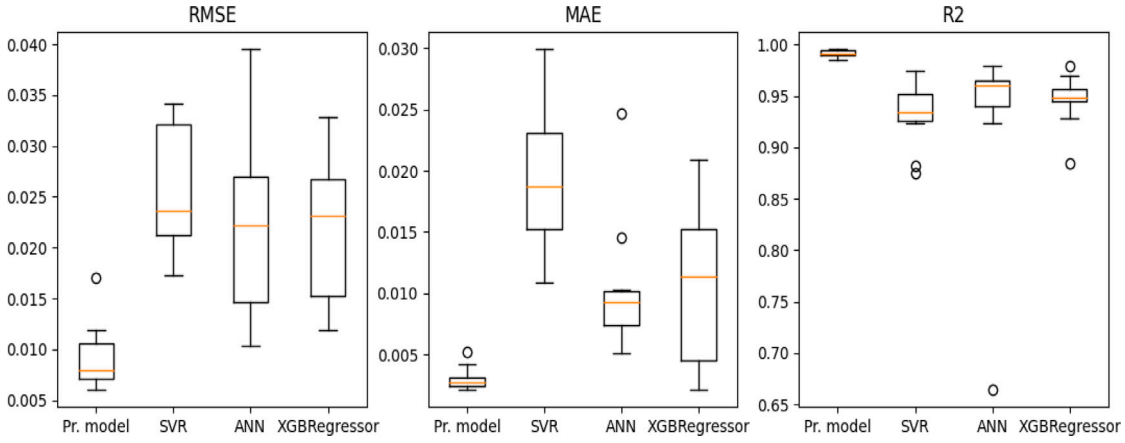


Fig. 9. Box-plot of the four models in terms of RMSE, MAE, and R^2.

proposed approach was assessed based on the operating conditions, as not all were GEH2s operating during the same conditions. Table 7 shows that these conditions can slightly affect the predictive quality of the proposed model; however, the predictive quality remains better than that of the XGBRegressor, ANN, and SVR models, which were significantly affected by the operating conditions.

In conclusion, the proposed model performed better in predicting the polarization curves of all tested PEMFCs. In addition, the different tests, which were realized by considering various operating conditions, demonstrated the effectiveness of the TPE in estimating the XGBRegressor hyperparameters. Finally, from a practical perspective, our proposed model can support both offline and online predictions of PEMFC performance owing to its reasonable processing time. Online prediction of the polarization curve allows one to detect short-term anomalies, such as the stack poisoning or malfunction of auxiliary systems. This provides a direct interpretation of the electrochemical performance of the PEMFC, thus enabling the proactive management of its operation to maximize efficiency and lifetime while minimizing unexpected interruptions. Subsequently, the developed model can be integrated into a control system (such as a GEH2 management system) to automatically control certain decisions, such as adjusting the operating parameters in response to volatile conditions. These advantages render the proposed model a promising option for performance prediction in industrial applications.

7. Conclusion

In this study, a prediction approach based on the XGBRegressor with a tree-structured Parzen estimator was proposed for estimating the proton-exchange membrane fuel cells performance of zero-emission electro-hydrogen generators. As proton-exchange membrane fuel cells feature complex electrochemical reactions with multiple nonlinear relationships between the inputs (operating variables of the fuel cell) and outputs (average fuel cell stack voltage), we combined kernel principal component analysis and mutual information for feature selection. Using a dataset comprising 10 proton-exchange membrane fuel cells, the performance and effectiveness of the proposed approach in predicting the polarization curve under different operating conditions were tested and validated. In addition, compared with XGBRegressor and with two well-established machine learning regressors (artificial neural network and support vector machine regressor) that are widely used to predict the polarization curve based on three performance metrics, the proposed method performed significantly better in terms of all the prediction effectiveness metrics. For example, the coefficient of determination of the proposed model showed average improvements of 6.35%, 6.80%, and 4.8% compared with artificial neural network, support vector machine regressor, and XGBRegressor, respectively. Although the obtained results demonstrated the performance and effectiveness of the proposed model, some aspects warrant further investigation: (1) The proposed model was evaluated on a database comprising 10 proton-exchange membrane fuel cells operating in the nominal state (but under different operating conditions). It would be interesting to assess the robustness of the proposed model under other operating conditions, such as stress or overload conditions; (2) the proposed model should be extended to predict the performance of other types of fuel cells, such as solid oxide or alkaline fuel cells.

CRediT authorship contribution statement

Soufian Echabarri: Writing – original draft, Methodology, Investigation, Conceptualization. **Phuc Do:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Hai-Canh Vu:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis. **Bastien Bornand:** Writing – review & editing, Validation, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgment

This work was financed in part by the ANRT (Association nationale de la recherche et de la technologie) with a CIFRE n° 2023/0482.

Appendix A. Schematic of ANN and SVR architectures

We present here the schematic of ANN and SVR architectures for PEMFC polarization curve prediction. Recall that both models have the following variables input variables: mes_i_stack (Stack current) mes_d_fc_h2In (instantaneous H2 consumption), mes_pr_fc_h2Low (H2 pressure at fuel cell inlet) mes_t_fc_out (coolant temperature at fuel cell outlet), mes_t_aux_out (auxiliary cooling circuit temperature at radiator inlet), and mes_pr_aSup (air pressure at fuel cell inlet). Configuration and the hyperparameters of ANN and SVR are described in Section 6.

Appendix B. Comparison results on other PEMFCs

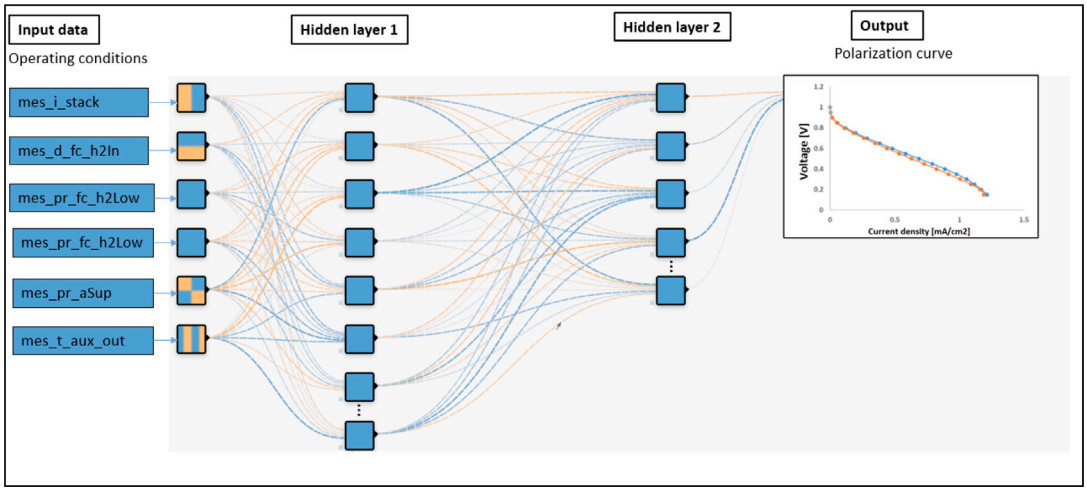


Fig. A.1. Schematic of ANN architecture for PEMFC polarization curve prediction.

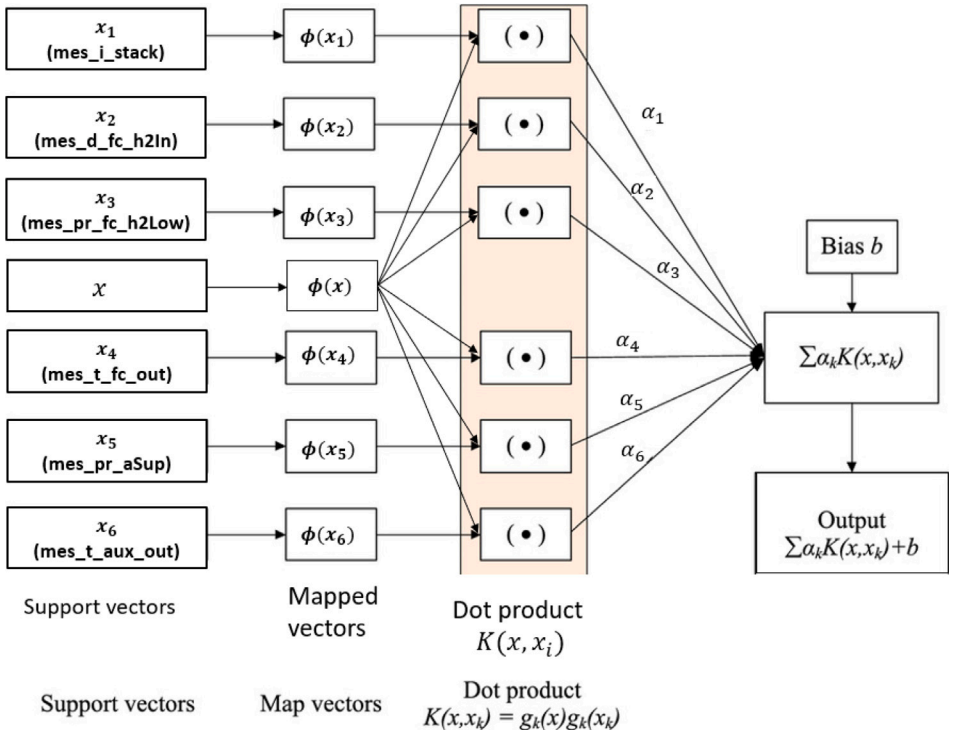


Fig. A.2. Schematic of SVR architecture for PEMFC polarization curve prediction.

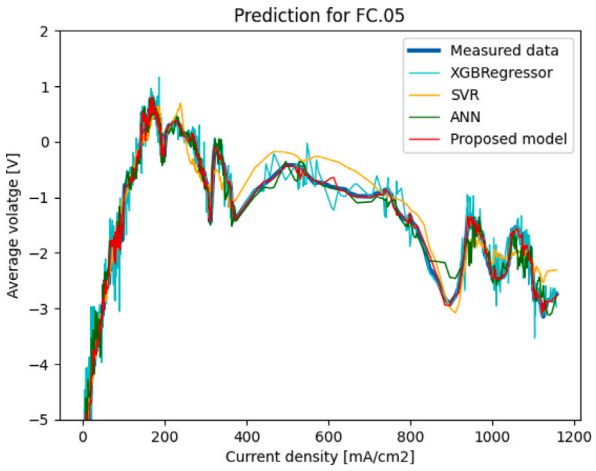


Fig. B.3. Polarization curve prediction for FC.05.

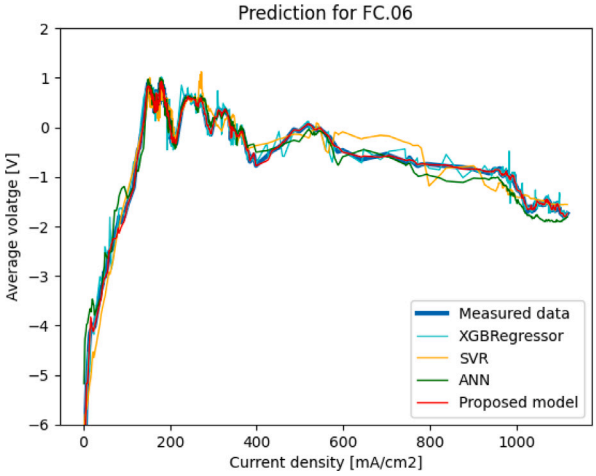


Fig. B.4. Polarization curve prediction for FC.06.

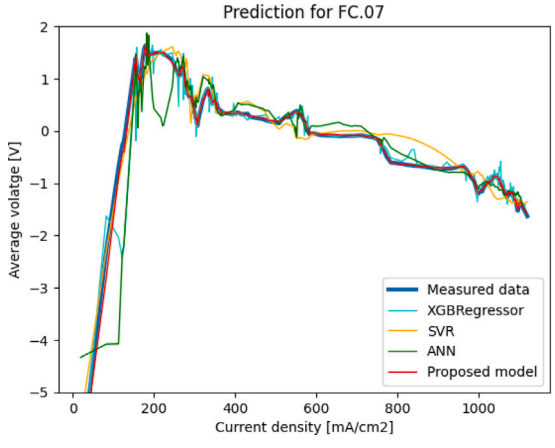


Fig. B.5. Polarization curve prediction for FC.07.

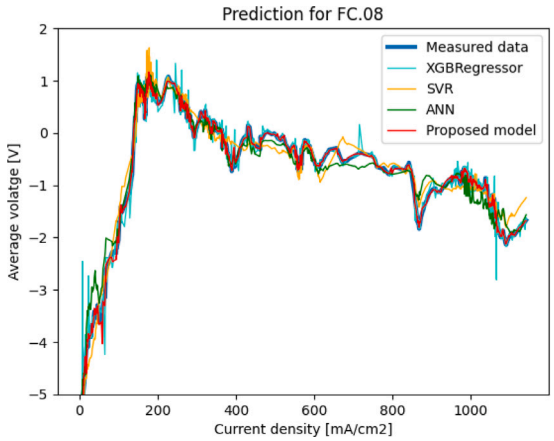


Fig. B.6. Polarization curve prediction for FC.08.

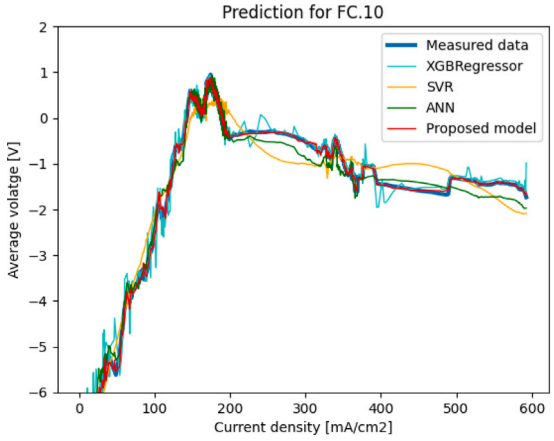


Fig. B.7. Polarization curve prediction for FC.09.

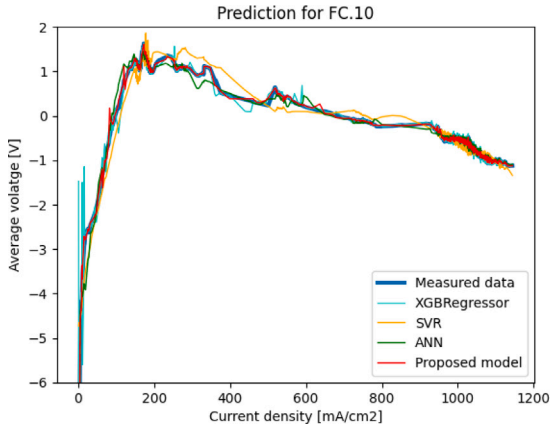


Fig. B.8. Polarization curve prediction for FC.10.

As mentioned in Section 6.2, PEMFCs have the same capacities and characteristics but operate under different conditions. The following figures show prediction results for other PEMFCs operating under different conditions and over different lengths of time. The aim is to demonstrate the adaptability of the proposed method to different scenarios encountered in real-world applications (see Figs. B.3–B.8).

References

[1] Wang B, Xie B, Xuan J, Jiao K. AI-based optimization of PEM fuel cell catalyst layers for maximum power density via data-driven surrogate modeling. Energy Conv Manag 2020;205:112460.

[2] Ding R, Wang R, Ding Y, Yin W, Liu Y, Li J, et al. Designing AI-aided analysis and prediction models for nonprecious metal electrocatalyst-based proton-exchange membrane fuel cells. Angew Chem 2020;132(43):19337–45.

[3] Zhao J, Li X, Shum C, McPhee J. A review of physics-based and data-driven models for real-time control of polymer electrolyte membrane fuel cells. Energy AI 2021;6:100114.

[4] Shi X, Huo X, Esan OC, Pan Z, Yun L, An L, et al. Mathematical modeling of fuel cells fed with an electrically rechargeable liquid fuel. Energy AI 2023;14:100275.

[5] Hasan M, Chen J, Waldecker J, Santare M. Predicting fatigue lifetimes of a reinforced membrane in polymer electrolyte membrane fuel cell using plastic energy. J Power Sources 2022;539:231597.

[6] Krishan T, Asaduzzaman RM, Thomas J, Pawel G, Tobias M, Andreas FK. Experimental and numerical study on catalyst layer of polymer electrolyte membrane fuel cell prepared with diverse drying methods. J Power Sources 2020;461:228169.

[7] Kishimoto M, Kishida S, Seo H, Iwai H, Yoshida H. Prediction of electrochemical characteristics of practical-size solid oxide fuel cells based on database of unit cell performance. Appl Energy 2021;283:116305.

[8] Singh R, Sui P, Wong KH, Kjeang E, Knights S, Djilali N. Modeling the effect of chemical membrane degradation on PEMFC performance. J Electrochem Soc 2018;165:3328–36.

[9] Danilov V, Tade M. An alternative way of estimating anodic and cathodic transfer coefficients from PEMFC polarization curves. Chem Eng J 2010;156:496–9.

[10] Kim J, Lee S, Srinivasan S. Modeling of proton membrane fuel cell performance with an empirical equation. J Electroanal Chem 1995;142:2670–4.

[11] Guinea D, Moreno B, Chinarro E, Guinea D, Jurado J. Rotary-gradient fitting algorithm for polarization curves of proton exchange membrane fuel cells (PEMFCs). Int J Hydrogen Energy 2008;33:2774–82.

[12] Bressel M, Hilairet M, Hissel D, Bouamama B. Extended Kalman filter for prognostic of proton exchange membrane fuel cell. Appl Energy 2016;164:220–7.

[13] Yue M, Li Z, Roche R, Jemei S, Zerhouni N. Degradation identification and prognostics of proton exchange membrane fuel cell under dynamic load. Control Eng Pract 2022;118:104959.

[14] Pan R, Yang D, Wang Y, Chen Z. Performance degradation prediction of proton exchange membrane fuel cell using a hybrid prognostic approach. Int J Hydrogen Energy 2020;45:30994–1008.

[15] Hu Y, Zhang L, Jiang Y, Peng K, Jin Z. A hybrid method for performance degradation probability prediction of proton exchange membrane fuel cell. Membranes 2023;13:426.

[16] Wang Y, Wu K, Zhao H, Li J, Sheng X, Yin Y, et al. Degradation prediction of proton exchange membrane fuel cell stack using semi-empirical and data-driven methods. Energy AI 2023;11:100205.

[17] Zhou D, Al-Durra A, Zhang K, Ravey A, Gao F. Online remaining useful lifetime prediction of proton exchange membrane fuel cells using a novel robust methodology. J Power Sources 2018;399:314–28.

[18] Cheng Y, Zerhouni N, Lu C. A hybrid remaining useful life prognostic method for proton exchange membrane fuel cell. Int J Hydrogen Energy 2018;43:12314–27.

[19] El-Brawany MA, Ibrahim DA, Elminir HK, Elattar HM, Ramadan E. Artificial intelligence-based data-driven prognostics in industry: A survey. Comput Ind Eng 2023;184:109605.

[20] Wilberforce T, Olabi A. Proton exchange membrane fuel cell performance prediction using artificial neural network. Int J Hydrogen Energy 2021;46:6037–50.

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

[22] He K, Liu Z, Sun Y, Mao L, Lu S. Degradation prediction of proton exchange membrane fuel cell using auto-encoder based health indicator and long short-term memory network. Int J Hydrogen Energy 2022;47:35055–67.

[23] 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.

[24] 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 Manage 2021;243:114367.

[25] Chen Y, Zhang J, Zhai S, Hu Z. Data-driven modeling and fault diagnosis for fuel cell vehicles using deep learning. Energy AI 2024;16:100345.

[26] Falcão D, Pires JCM, Pinho C, Pinto A, Martins FG. Artificial neural network model applied to a PEM fuel cell. In: International conference on neural computation, vol. 2, SCITEPRESS; 2009, p. 435–9.

[27] Li H-W, Qiao B-X, Liu J-N, Yang Y, Fan W, Lu G-L. A data-driven framework for performance prediction and parameter optimization of a proton exchange membrane fuel cell. Energy Convers Manage 2022;271:116338.

[28] Han I, Chung C. Performance prediction and analysis of a PEM fuel cell operating on pure oxygen using data-driven models: A comparison of artificial neural network and support vector machine. Int J of Hydrogen Energy 2016;41:10202–11.

[29] Hong W. Performance prediction and power density maximization of a proton exchange membrane fuel cell based on deep belief network. J Power Sources 2020;228:154.

[30] Zheng L, Hou Y, Zhang T. Performance prediction of fuel cells using long short-term memory recurrent neural network. Int J Energy Res 2021;45:9141–61.

[31] Chen K, Laghrouche S, Djerdir A. Aging prognosis model of proton exchange membrane fuel cell in different operating conditions. Int J Hydrogen Energy 2020;45:11761–72.

[32] Zuo J, Lv H, Zhou D, Xue Q, Jin L, Zhou W, et al. Deep learning based prognostic framework towards proton exchange membrane fuel cell for automotive application. Appl Energy 2021;281:115937.

[33] Han I-S, Chung C-B. Performance prediction and analysis of a PEM fuel cell operating on pure oxygen using data-driven models: A comparison of artificial neural network and support vector machine. Int J Hydrog Energy 2016;41(24):10202–11.

[34] Chen K, Laghrouche S, Djerdir A. Degradation model of proton exchange membrane fuel cell based on a novel hybrid method. Appl Energy 2019;252:113439.

[35] Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends 2020;01:56–70.

[36] Padmaja D, Vishnuvardhan B. Comparative study of feature subset selection methods for dimensionality reduction on scientific data. In: IEEE 6th int. conf. on advanced computing. 2016, p. 31–4.

[37] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1:131–56.

[38] Zhao H, Min F, Zhu W. Cost-sensitive feature selection of numeric data with measurement errors. J Appl Math 2013;2013.

[39] Zebari D, Haron H, Zeebaree S. Security issues in DNA based on data hiding: A review. Int J Appl Eng Res 2017;12:6940–8.

[40] Elhadad M, Badran K, Salama G. A novel approach for ontology-based dimensionality reduction for web text document classification. Int J Softw Innov 2017;5:44–58.

[41] Aziz R, Verma C, Srivastava N. Dimension reduction methods for microarray data: A review. AIMS Bioeng 2017;4:179–97.

[42] Wang Q. Kernel principal component analysis and its applications in face recognition and active shape models. 2012, arXiv preprint arXiv:1207.3538.

[43] Shen K, Qin H, Zhou J, Liu G. Runoff probability prediction model based on natural gradient boosting with tree-structured parzen estimator optimization. Water 2022;4:545.

[44] Han I, Shin H. Modeling of a PEM fuel cell stack using partial least squares and artificial neural networks. Korean Chem Eng Res 2015;53:236–42.

[45] Tsadiras AK, Papadopoulos C, O'Kelly ME. An artificial neural network based decision support system for solving the buffer allocation problem in reliable production lines. Comput Ind Eng 2013;66(4):1150–62.

[46] Zhou Y, Zhang Y, Pang R, Xu B. Seismic fragility analysis of high concrete faced rockfill dams based on plastic failure with support vector machine. Soil Dyn Earthq Eng 2021;144:106587.

[47] Yang C-C, Shieh M-D. A support vector regression based prediction model of affective responses for product form design. Comput Ind Eng 2010;59(4):682–9.