International Journal of Hydrogen Energy 127 (2025) 462–483

Contents lists available at ScienceDirect

International Journal of Hydrogen Energy

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

AI-driven surrogate modelling for simulating hydrogen production via proton exchange membrane water electrolysers

Mohammad Abdul Baseer [a,b], Harjeet Singh [a], Prashant Kumar [b,c,a], Erick Giovani Sperandio Nascimento [a,b,c,d,*]

[a] Surrey Institute for People-Centred Artificial Intelligence, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom
[b] Global Centre for Clean Research (GCARE), School of Sustainability, Civil and Environmental Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom
[c] Institute for Sustainability, University of Surrey, Guildford, GU2 7XH, United Kingdom
[d] Stricto Sensu Department, SENAI CIMATEC University, Salvador, Bahia, Brazil

ARTICLE INFO

Handling Editor: Suleyman I. Allakhverdiev

Keywords:
H_2 production
Proton exchange membrane water electrolysis
Machine learning
Deep learning
Performance metrics
Wilcoxon signed-rank test

ABSTRACT

We developed and evaluated an AI-based surrogate model to simulate Hydrogen (H_2) production via Proton Exchange Membrane Water Electrolysis (PEMWE). A variety of Machine Learning (ML) and Deep Learning (DL) models were tested and fine-tuned using real-world PEMWE datasets from multiple sources, ensuring model robustness and accuracy. The models included ML algorithms such as k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Category Boosting (CB), Light Gradient Boosting (LGB), and Gradient Boosting (GB), with DL models Long Short-Term Memory (LSTM), Multi-Layer Perceptron (MLP), and one-dimensional (1D) Convolutional Neural Networks (CNN). Among these, the 1DCNN model demonstrated superior performance, achieving an R-squared (R^2) of 0.998944, a Mean Squared Error (MSE) of 488.82, a very low Normalised Root Mean Square Error (NMSE) of 0.001055, and Pearson correlation of 0.999472, having MLP also performing exceptionally well. Conversely, the LSTM model performed the poorest. Unlike many prior studies, we employed cross-validation techniques to rigorously validate model performance and the Wilcoxon Signed-Rank Test for statistical validation, establishing the robustness and reliability of 1DCNN predictions. Comparatively, the MLP model performed well but failed to pass the Wilcoxon test, indicating its predictions were not statistically different from any other models. These findings underscore the novelty of the proposed 1DCNN-based surrogate model, which stands out in its ability to accurately simulate PEMWE behaviour for H_2 production. This model can simulate PEMWE processes in a fraction of the time required by traditional methods, providing valuable operational insights and advancing technologies across H_2 production, energy sectors, transportation, and sustainable energy systems.

1. Introduction

The transition to a sustainable energy system is increasingly driven by H_2, garnering attention for its unique properties as an energy carrier. It stems from two key characteristics, a remarkable energy density and the potential for carbon-zero emission [40]. PEMWE emerging as one of the most promising technologies for clean H_2 production and is valued for its high efficiency, ability to operate at low temperatures, and compatibility with renewable energy sources [1]. Among the various water electrolysis technologies, PEMWE has gained significant attention due to its advantages, such as compact design, quick response, and tolerance of large variations in power input [2]. [42] emphasises the need for low-cost, efficient electrocatalysts for a sustainable H_2, economy. However, high costs and inefficiencies related to electrocatalysts, hinder scalability and the broader adoption of PEMWE for industrial-scale H_2 production. Studies have shown that ML models can effectively manage PEMWE degradation and optimise operational conditions. For instance, Hayatzadeh et al. [3] examined the effects of

* Corresponding author. Surrey Institute for People-Centred Artificial Intelligence, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom.
E-mail addresses: ma04676@surrey.ac.uk (M.A. Baseer), hs01697@surrey.ac.uk (H. Singh), p.kumar@surrey.ac.uk (P. Kumar), erick.sperandio@surrey.ac.uk (E.G. Sperandio Nascimento).

https://doi.org/10.1016/j.ijhydene.2025.04.098
Received 13 January 2025; Received in revised form 31 March 2025; Accepted 5 April 2025
Available online 15 April 2025

operating conditions, such as temperature and current density, on PEMWE performance, focusing on optimising catalyst loading using Support Vector Regression (SVR) and Artificial Neural Network (ANN) models. Their findings underscored the effectiveness of ANN, particularly for larger datasets, with an R^2 value of 0.99925 for Iridium-black catalyst predictions. Moreover, Mao et al. [4] developed advanced control strategies to enhance the robustness of PEMWE systems, highlighting the role of AI in overcoming operational challenges. Despite significant technological advances, PEMWEs face efficiency limitations, particularly when operating across diverse conditions [2]. Recent modelling approaches, such as the simulink-based mathematical models developed by [43], analyse the impact of input factors like temperature, pressure, and current on system efficiency. Although their work offers insightful information about how PEMWE behaves in different scenarios, it does not use AI or ML techniques, which are necessary for additional optimisation and increased forecast accuracy using AI-driven models.

AI and ML techniques provide interesting solutions to these performance issues. These algorithms can be used to develop predictive models that optimise system parameters, make intelligent control mechanisms possible for dynamic operation, and make use of large data sets to raise electrolyser efficiency overall. AI-driven surrogate modelling has appeared as a viable substitute to get beyond these restrictions, providing quick and accurate predictions. In this study, we present a systematic methodology for building an AI-powered surrogate model designed to simulate H_2 production in PEMWEs. The first step in this process is the development of a comprehensive dataset that includes a variety of PEMWE system-relevant operating conditions and design parameters. Our goal is to create an AI-surrogate model that can accurately predict PEMWE performance for H_2 production by utilising ML/DL techniques. In this study, we integrated Explainable AI techniques, specifically SHAP (SHapley Additive exPlanations), to interpret and validate model predictions. By incorporating SHAP-based explainability, we not only achieved high predictive accuracy with the 1DCNN model but also enhanced the model's interpretability. This novel method offers researchers an effective way to expedite the development and optimisation of H_2 production methods, marking a substantial advancement in PEMWE simulation. A simplified diagram of the PEMWE process is shown in Fig. 1. It emphasises the flow of ions and electrons between the cathode and anode as well as the separation of H_2 and oxygen (O_2) across the membrane. This method of producing H_2 is especially relevant to the global emission reduction goals set as part of efforts to promote sustainability. Emission reduction targets for major regions by 2030 and 2050 were covered by [44]. The emission reduction targets for the US, UK, EU, Japan, and South Korea are shown in Fig. 2, with the US and UK exhibiting a slight lead over other countries.

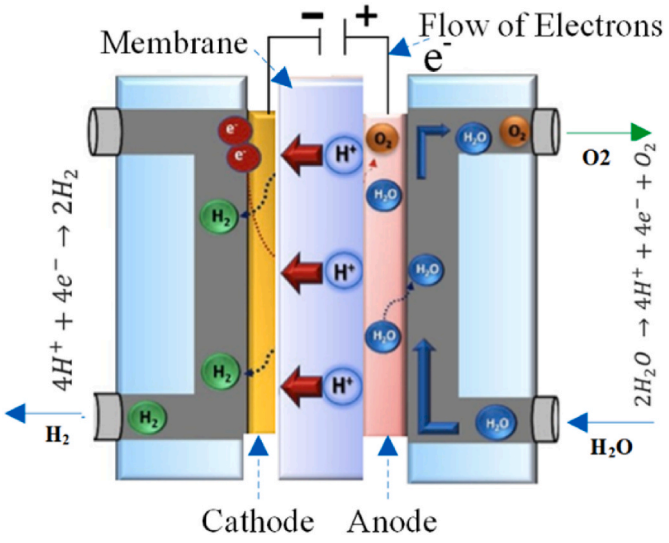


Fig. 1. Simplified diagram of PEMWE [41].

2. Related studies

2.1. Process of PEMWE

PEMWE functions by electrochemically splitting water (H_2O) into H_2 and O_2, utilising a PEM to facilitate ion transport. The system comprises an anode, a cathode, a PEM, and electrocatalysts that enhance reaction efficiency. The process is illustrated in Fig. 1, highlights the key components and mechanism involved. According to Li et al. [2], the fundamental reactions (1–2) governing this process are as follows:

At the anode, water molecules undergo the oxygen evolution reaction as shown:

$$2H_2O \rightarrow 4H^+ + 4e^- + O_2 \quad (1)$$

This reaction splits water into protons (H^+), electrons (e^-), and oxygen gas. Protons (H^+) migrate through the PEM, while electrons travel through an external circuit, generating an electric current. At the cathode, the hydrogen evolution reaction occurs as shown:

$$4H^+ + 4e^- \rightarrow 2H_2 \quad (2)$$

H_2 gas is then collected at the cathode, while O_2 is released at the anode. This method of producing H_2 is especially relevant for the transition to green energy. Compared to alkaline water electrolysis, PEMWE offers advantages such as: Higher H_2 purity, due to the selective transport of protons through the PEM. Rapid response time, making it highly suitable for renewable energy integration (solar, wind, etc.). Compact and scalable design, reducing the need for large electrolyte management systems. High current density operation, enabling higher H_2 production rates. However, PEMWE faces challenges such as catalyst cost, membrane durability, and efficiency limitations, necessitating ongoing research into AI-driven modelling, predictive optimisation, and material advancements.

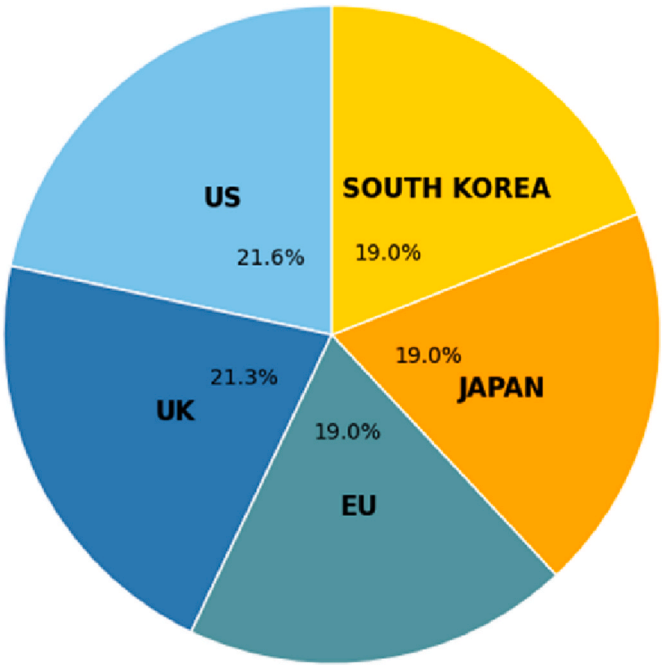


Fig. 2. Emission reduction targets for major regions by 2050 as a share of the total commitment (100 %) [44].

2.2. Machine Learning in PEMWE optimisation

Previous studies have explored ML models to optimise the design and performance of PEMWE systems. Mohamed et al. [5] presented two ML approaches to predict the H_2 production rate and cell current density in PEM electrolyser cells. Their studies trained five, ML models such as ANN, Polynomial Regression, SVM, k-NN, and DT by using 15 input parameters and found that ANN model demonstrated the best performance, with a MAE of 6.44 for H_2 production and 0.04 for current density. In continuation of the studies, Mohamed et al. [6] developed ML models using polynomial and logistic regression to predict the optimal design of PEM electrolyser cells. The models were trained on 148 samples and validated on a test set of 16 samples, predicting 11 design parameters based on input factors such as H_2 production rate, cathode area, anode area, and cell design type. The models achieved an accuracy of 83.6 % and a MAE of 6.825. A custom-made PEM electrolyser cells was fabricated based on the predicted parameters, and its performance showed excellent agreement with simulation results, within a negligible experimental uncertainty a MAE of 0.615.

Recent advancements in data-driven modelling have demonstrated the effectiveness of ML in optimising PEM systems. For instance, Zhang et al. [7] proposed a data-driven approach combining a Genetic Algorithm-Backpropagation neural network with Particle Swarm Optimisation (PSO) to optimise the operating parameters of PEM fuel cells. Their study identified inlet temperature, operating pressure, and relative humidity as critical factors influencing power density, achieving a 3.3 % increase in power density through optimised conditions. The surrogate model exhibited excellent predictive performance, with correlation coefficients of 0.99896 and 0.99815 for the training and test sets, respectively. While their study focused on PEM fuel cells, the data-driven surrogate modelling approach provides valuable insights into optimising electrochemical systems, which can be extended to PEMWE systems for H_2 production. Rui et al. [45] applied ML models to predict optimal design parameters for PEMWE. The study utilised k-NN and DTR to predict 17 key design parameters, including H_2 production rate, electrode area, and flow-field patterns, using 1062 data points. The model achieved a MSE of 0.31, showing high reliability compared to experimental results. The research focused on designing commercial-scale PEM electrolysers with H_2 production rates ranging from 50 to 3000 mL/min, contributing to a reduction in the time and cost needed for developing large-scale electrolysers.

2.3. Challenges in PEMWE optimisation

Despite the promising applications of ML in optimising H_2 production processes, challenges persist due to the limited availability of relevant datasets. Dinçer et al. [8] introduced a hybrid Q-learning and molecular fuzzy-based model to optimise water electrolysis investments for green H_2 production. Their study found that electrolyser lifespan and production capacity were the two most critical performance factors. They further identified PEM and alkaline water electrolysis as the most effective strategies for increasing green H_2 yield. [46] provided insights into the application of ML-based optimisation for high-pressure electrolysis systems. Their study emphasised the role of AI in enhancing efficiency, reducing energy losses, and predicting long-term system performance. The integration of explainable AI methods, such as SHAP analysis, allowed for a deeper understanding of parameter interactions in PEM electrolysis. Arjmandi et al. [9] applied DT, SVM, and Regression models to predict anode-side parameters like water feed rate, catalyst loading, and current density. The DT model emerged as the most effective, achieving 100 % accuracy, while SVR showed moderate improvements in accuracy, ranging from 0.79 to 0.82. These results illustrate that ML models can efficiently predict anode-side parameters, facilitating H_2 production optimisation.

Recent studies have also explored the integration of renewable energy sources with PEMWE systems. [47] demonstrated the effectiveness of DL models in estimating H_2 yield for solar-powered PEMWE systems. Their Agnostic Deep Neural Network model achieved an R^2 of 96.26 %, confirming the high predictive capability of DL for H_2 production forecasting. This study highlights the potential of DL models in capturing complex non-linear relationships in H_2 production systems, making them suitable for real-time optimisation and control. Urhan et al. [10] developed an ML-based approach for H_2 production using PEM electrolysers integrated with solar and wind energy systems. Their study, leveraging 10 years of meteorological data, identified an optimal system configuration comprising 548 kW Photovoltaic, 1040 kW wind turbines, a 600 kW electrolyser, and 600 kg H_2 storage. This system produced 40, 000 kg of green H_2 annually at a total net present cost of $8,351,442, demonstrating the feasibility of renewable energy integration in H_2 production.

Purnami et al. [11] demonstrated the successful use of AI-based adaptive systems to optimise H_2 production in a dynamic magnetic field-assisted electrolysis system. By employing a Double Deep Q Network, their study achieved real-time optimisation of operational parameters, significantly improving the efficiency of the electrolysis process. This AI-driven approach allowed for dynamic adjustments, leading to reduced energy consumption and enhanced H_2 production. Hybrid models have also gained traction in addressing PEMWE challenges. Rezk et al. [12] and Bensmann et al. [13] continuously explored hybrid models, such as physics-informed neural networks, which combine traditional models with Adaptive Neuro-Fuzzy Inference System (ANFIS) to enhance simulation accuracy. Tawalbeh et al. [14] utilised ANN to predict H_2 production rates by training their model on operational parameters such as cell voltage, current, power, temperature, and water flow rate. The model, optimised through the Levenberg–Marquardt Backpropagation (LMBP) algorithm, achieved a high R^2 of 0.9989 and a MAE of 0.012, significantly outperforming RF and SVM. This study highlights ANN in offering precise control over PEMWE operational conditions, thus improving H_2 production efficiency. The transient hydrogen mass flow rate of a PEM electrolyser system was also accurately predicted by Biswas et al. [15] using ANN modelling with different LMBP algorithms and time delay structures. The best-performing model achieved an R^2 of 0.9013 and a MSE of 0.003371. They concluded that such models could increase the PEM electrolyser system's efficiency and lower their power consumption.

Chen et al. [16] introduced an innovative Ladder of Knowledge-Integrated ML framework that enhances model robustness by integrating domain-specific knowledge at three levels: Interpolation, Extrapolation, and Representation. Their framework, which included models like SVR, DT, and ANN, demonstrated an improvement in interpolation accuracy by up to 30 %. The ANN model achieved superior performance in predicting PEMWE performance under diverse conditions, with an NMSE of 3.238 and an R^2 of 0.988, outpacing both SVR and DT. This framework offers a structured approach to improving ML precision in predicting system degradation and overall performance. [49] focused on the membrane electrode assembly, a critical PEMWE component, optimising its performance using ML models such as GB. They achieved a notable R^2 value of 0.943 for predicting current density at 1.9V. Moreover, black-box interpretation methods like SHAP were employed to interpret ML results, offering valuable insights into the relationships between membrane electrode assembly design parameters and performance. This study underscored the reliability of ML in optimising membrane electrode assembly design, reducing experimental time and costs. Rehman et al. [17] introduced an AI-based surrogate model to optimise H_2 liquefaction processes. Their model, developed using ANN and optimised with PSO, achieved a prediction error of 4 % for the minimum internal approach temperature and 0.04 % for specific energy consumption. The surrogate model significantly reduced the computational time required for optimisation by over 99.99 %, demonstrating to the AI-driven techniques in improving H_2 liquefaction efficiency.

2.4. Research gaps

While previous studies have made significant strides in applying ML and DL to PEMWE optimisation, several research gaps remain.

1. There are a limited dataset availability poses a major challenge, as many studies rely on small or synthetic datasets, which restrict the generalisability and robustness of their models.
2. There is a lack of comprehensive model comparisons, as few studies have systematically evaluated the performance of multiple ML and DL architectures for PEMWE simulation.
3. Insufficient robustness and reliability are common issues, with many existing models lacking rigorous validation and testing across diverse operating conditions.

2.5. Our work and contributions

Our work demonstrates the development of an AI-based surrogate model that imitates H_2 production by PEMWE to get beyond these limitations. Using real-world PEMWE datasets from multiple sources, we developed, evaluated, and analysed 10 different ML/DL methods to ensure their generalisability, robustness, and reliability. Therefore, our approach is innovative in five important ways.

1. The development of an AI-surrogate model to simulate the production of H_2 by developing, testing, and comparing 10 ML/DL architectures such as GB, RF, 1DCNN, LSTM, and MLP. By selecting the best AI method able to capture both non-linear relationships that are often ignored by models, this method improves the accuracy of PEMWE simulations.
2. The combination of two datasets of real-world PEMWE for H_2 production from multiple sources, yielding 1210 samples, providing a more comprehensive and robust dataset for training, and assessing the models, thereby improving its generalisability and performance in real-world applications.
3. The full and comprehensive assessment of the developed models by using statistical metrics like R^2, MSE, NMSE, MAE, and Pearson correlation to assess their performance in simulating H_2 production from PEMWE. Also, to ensure the reliability and trustworthiness of the results, the employment of Wilcoxon Signed-Rank Test, was to test whether the models would be statistically different from each other so that their predictions could be trusted as reliable or not. The 1DCNN model emerged as the top performer, while MLP, also showed strong results, demonstrating their suitability for accurate H_2 simulation, as confirmed by the Wilcoxon Signed-Rank Test.
4. Enhancing model robustness and interpretability through Explainable AI by applying SHAP analysis. By integrating SHAP-based explanations, our study ensures model transparency, enabling better system optimisation and decision-making.
5. The advancement in model robustness by addressing limitations in previous studies that relied on minimal datasets. This improvement ensures reliable predictions across diverse operating conditions and enhances the generalisability of the proposed AI models for accurately simulating PEMWE behaviour for H_2 production.

By integrating these key aspects, our proposed model becomes more accurate, robust, trustworthy, and scalable for industrial applications, reducing computational costs and the time required for experimentation when compared to conventional physical and data-driven models.

To achieve this, the research sets the following objectives, which aim to.

1. Develop ML and DL models to accurately simulate PEMWE behaviour under a wide range of conditions.
2. Use these models to predict H_2 production rates based on various input parameters, ensuring precise and reliable outputs.
3. Conduct a comparative analysis of different ML and DL models, including k-NN, SVM, DT, RF, CB, GB, LSTM, 1DCNN, and MLP, to determine the most effective and reliable approach for PEMWE simulation.
4. Leverage real-world datasets with numerous input parameters to enhance model accuracy and predictive capabilities.
5. Validate and fine-tune the developed models to ensure performance and robustness in predicting PEMWE behaviour of H_2 production rates for industrial applications.

The remainder of this paper is organised as follows. Section 3 describes the methodology, detailing the dataset preparation, model development, and evaluation metrics used to assess the performance of the proposed ML/DL architectures. Section 4 presents the results and discussion, highlighting the performance of the models and their comparative analysis, with a focus on their accuracy, robustness, and reliability in simulating PEMWE behaviour for H_2 production. Finally, Section 5 concludes the paper, summarising the findings, and outlining future research directions to further advance the field of PEMWE optimisation.

3. Methodology

This study uses a methodical way to develop and optimise an AI-based surrogate model for PEMWE-based H_2 production prediction. The methodology goal is to estimate H_2 flow rates with accuracy and robustness. The method starts with the initialisation see (section 3.1) of two datasets, one from Mohamed et al. [5] and the other from Rui et al. [45], as shown in Fig. 3. Together, these datasets are designated as (X_1) and (X_2), respectively to a comprehensive dataset (X), which forms the foundation for the development of the model. The next critical stage is data pre-processing, which involves cleaning the datasets to address missing values. For numerical data, missing values are filled using mean imputation with SimpleImputer from sklearn.impute. Additionally, the data is normalised using StandardScaler from scikit-learn [18] to ensure consistency across all variables and models. Categorical variables, such as membrane type, anode/cathode materials, and electrolysis type, are transformed using one-hot encoding with the OneHotEncoder from scikit-learn to prepare the data for ML algorithms. The next step is exploratory data analysis, which involves creating statistical summaries and generating visualisations using pandas [48] and Matplotlib [19] to highlight patterns and similarities in the data. This step aids in identifying significant features that might influence the model's performance. After data preparation, the dataset is split into training (847 samples) and testing (363 samples) sets using a 70/30 % ratio with the train_test_split function from scikit-learn. This split ensures that the models are trained on a substantial portion of the data while retaining a separate subset for testing to evaluate performance on unseen data. Several ML and DL models are then developed, including MLP, 1DCNN, LSTM, implemented using TensorFlow [20] and Keras [21] and ensemble models like RF and GB. Each model is trained and fine-tuned through experiments by adjusting hyperparameters such as the learning rate, the number of estimators, and layer configurations.

To effectively capture local dependencies in the input features, a 1DCNN model is defined in the following phase. The model's first layer is a 1D convolutional layer with 64 filters and a kernel size of 3, designed to extract important features from the input data. The ReLU activation function introduces non-linearity, enabling the model to learn complex patterns and correlations among features. To reduce the dimensionality of the feature maps, a max pooling layer with a pool size of 2 follows the convolutional layer, which decreases computational complexity and helps prevent overfitting by down sampling the data. To learn higher-level representations, the collected features are subsequently flattened and passed through a fully connected layer that has 128 neurons and ReLU activation. Finally, the output layer consists of a single neuron designed for regression, producing a continuous value representing the

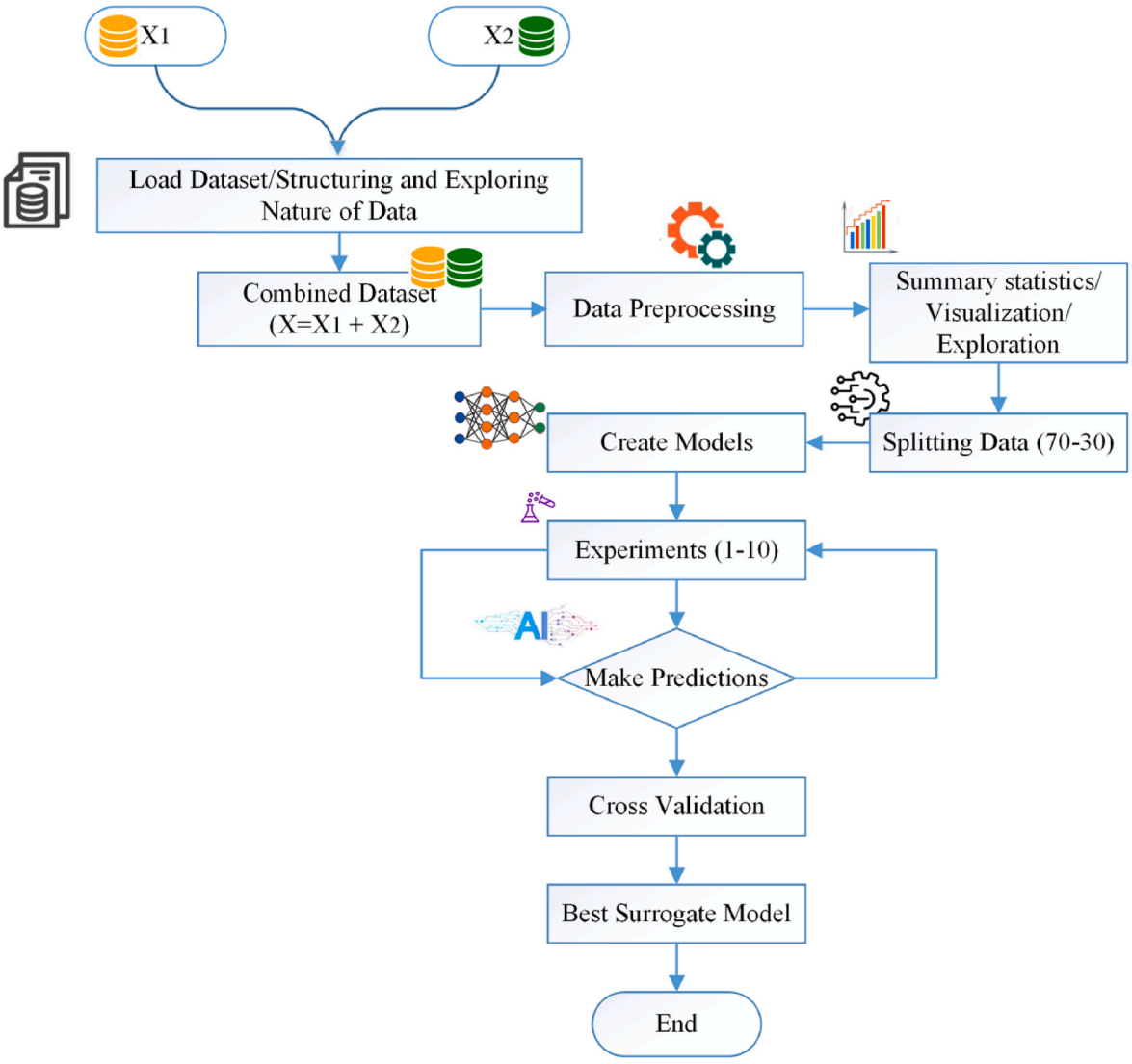


Fig. 3. Flowchart of model development and selection process.

predicted H_2 flow rate. The performance of the trained 1DCNN model is evaluated by comparing the actual and predicted values for H_2 production after the test set has been passed through it. Five-fold testing is used to evaluate performance by dividing the test set into equal halves and producing fold-specific assessments of actual vs. predicted outcomes. To improve visualization, a plot is also created for the complete test set. To evaluate the accuracy of the model, evaluation metrics are calculated, such as R^2, MSE, NMSE, MAE, and Pearson correlation.

To ensure robustness, a five-fold cross-validation approach is applied. During cross-validation, each fold of the dataset is split into an 80/20 % for training and validation, and the best models' performance is trained and assessed across different folds. All metrics are computed across folds, and their mean and standard deviation are calculated, to evaluate whether it is generalising well when exposed to different portions of the data. The total cross-validation time for the 1DCNN model was approximately 0.58 s, highlighting its computational efficiency despite the complexity of DL. SHAP analysis was conducted to identify the most influential features affecting H_2 production. Then, all models are further validated using the Wilcoxon Signed-Rank Test to confirm whether they are statistically different from each other to ensure whether their predictions are reliable and trustworthy or not.

3.1. Data description

The dataset used for this research consists of 1210 samples and 26 features, compiled from Mohamed et al. [5] and Rui et al. [45]. Key features such as electrolysis type, electrode area, anode flow area, cathode flow area, membrane type, catalysts, water flow rates, and H_2 flow rates. The categorical columns, such as electrolysis type, anode type, cathode type, anode/cathode gas diffusion electrode, membrane type, and others, were transformed into numerical values using one-hot encoding. This encoding process allows ML algorithms to interpret categorical data properly. The dataset provided by Rui et al. [45] comprises 1062 data points, with 357 values obtained from direct experimental measurements and 705 extracted from previously published studies. It includes 5 input features, which serve as predictive variables for 17 electrolyser parameters. These parameters cover key aspects such as electrode materials, gas diffusion layers, catalyst compositions, current density, operating voltage, and environmental conditions (e.g., pressure, temperature, and water flow rate). Additionally, the dataset incorporates data from both single-cell and bipolar PEMWE configurations, with variations in electrolyte composition, including deionized water, acidic solutions, and choline chloride aqueous solutions. The diverse range of conditions captured in this dataset provides a robust foundation for analysing PEMWE performance across different

operational settings. Table 1, presents various parameters and their corresponding values or ranges for electrolysis experiments, including details about the cell components, operating conditions, and performance metrics. The parameters cover aspects such as electrolysis type, electrode materials, catalyst types, electrolyte solutions, cell design, voltage, current density, power output, flow rates, temperature, and pressure.

To further understand the relationships between key features in the dataset, a correlation heatmap was generated, as shown in Fig. 4. This visualization provides valuable insights into the interrelationships among input variables, helping to identify potential dependencies that could influence H_2 flow rate predictions. Notably, parameters such as **power (W), power density (W/cm^2)**, and **water flow rate (ml/min)** exhibit strong positive correlations with H_2 production, with correlation coefficients of 0.99, 0.83, and 0.67, respectively. These findings align with fundamental electrochemical principles, as higher power input, increased water flow, and larger anode areas are expected to enhance H_2 generation.

3.2. Model development

The implementation and evaluation of several ML and DL models for predicting H_2 flow rates in PEMWE systems are the main topics of this section. To improve performance, each model was meticulously fine-tuned using hyperparameter optimisation. Among the models used are the following.

Table 1
Parameters and value ranges for PEM electrolysis [6].

Parameter type	Category/Value range
Electrolysis type	PEM
Anode type	Carbon plate, Ni foam, Titanium, 316L_stainless_steel_felt316L_stainless_steel,
Cathode type	Carbon plate, Titanium 316L_stainless_steel_felt316L_stainless_steel
Anode gas diffusion electrode	Porous carbon paper, Ni foam, Titanium, Titanium felt Titanium mesh, a double layer felt, a single layer felt
Cathode gas diffusion electrode	Porous carbon paper, Ni foam, carbon paper, porous titanium, a wet proofed non-woven carbon cloth
Electrode area(cm^2)	6.0–100
Cathode flow area (cm^2)	4.0–100
Anode flow area (cm^2)	4.0–100
Membrane type	Nafion117, Nafion115, Aquivion™_E79-05S_50_μm_ (Solvay_Solexis)
Cathode catalyst	DNA, Pt/C, MoS2, platinum, 0.5 mg Pt/cm2, 40 wt%_Pt/C,
Anode catalyst	MoS2, WSe2, iridium oxide (iro2), iro2 Ruo, 1.5 mg ir/cm2
Catholyte	H_2SO_4_0.1mol, choline_0.5 M, DI water, steam water
Anolyte	H_2SO_4_0.1mol, H_2SO_4_0.1mol, DI water, steam water
Cell design type	Single or bipolar
Cell design number	1–20
Cell voltage(V)	0.5–32
Cell currents des(A/cm^2)	0.000241–2.0
Power(w)	0.0–1300
Power density(W/cm^2)	0.000361–4.41
Water flow rate(ml/min)	1–1000
Hydrogen flow rate (ml/min)	0.0–5000
Temperature cell(k)	298–360
Pressure (atm)	1.0–3.0
Electrode shape	Rectangular, round
Flow type A (number)	1.0–28.0
Flow type C (number)	1.0–28.0

3.2.1. Support Vector Machines

An algorithm for regression tasks that determines the best hyperplane to divide data into distinct classes. SVMs can recognise intricate patterns because they can use kernel functions to handle both linear and non-linear relationships [22]. SVMs effectively capture non-linear relationships and reveal complex correlations between input variables and output rates for H_2 prediction. In this study, we evaluated the SVM model, feature scaling was not applied to better understand the model's performance on raw data.

3.2.2. K-nearest neighbors

A non-parametric technique called the k-NN algorithm classifies instances according to the dominating class of their nearest neighbors [23]. Because of its versatility and ability to handle both classification and regression tasks, it is a good choice for finding local trends in data. Using the premise that instances in feature space that are adjacent to one another are likely to have similar results, k-NN predicts H_2 flow rates in this context by using the features of the nearest, most similar data points.

3.2.3. Decision trees

DTs are tree-structured models that predict values or classes at the leaf nodes by repeatedly dividing the input into subsets at each node [24]. For applications including both regression and classification, DTs provide interpretability. They successfully reveal intricate patterns in the input parameters when used for H_2 prediction. To predict H_2 flow rates, we assessed a DT model in this investigation. Hyperparameters were adjusted to enhance the model fitting performance while reducing the possibility of overfitting. Several metrics were computed to assess the model's performance.

3.2.4. Random Forest

A powerful ensemble learning method, RF builds several DTs and aggregates their predictions. Even with complicated and sophisticated data, this method improves the overall model's accuracy and robustness across a range of tasks while reducing overfitting problems that may result from individual DTs [25]. To enhance the predicted performance, the model was set up with hyperparameters, such as the maximum depth, number of trees, and feature selection technique. More accurate predictions of future H_2 rates are made possible by RF algorithm's exceptional ability to identify and model complex correlations between input data and the goal output variable for the H_2 flow rate.

3.2.5. Gradient Boosting

One important optimisation method for training ML models with lots of parameters, like neural networks, is GB. It improves accuracy and model convergence during training by iteratively fine-tuning the parameters to minimise a loss function [38]. By methodically adjusting the parameters, GB improves the accuracy of PEM H_2 rate prediction.

3.2.6. Light Gradient Boosting

It is a fast and efficient GB framework developed by Microsoft, which uses a leaf-wise tree growth strategy and histogram-based decision trees to handle large datasets with high-dimensional features. Its optimisations, such as gradient-based one-side sampling and exclusive feature bundling, make it highly scalable and effective for tasks like classification, regression, and ranking [26].

3.2.7. Category Boosting

Yandex created the CB algorithm, which effectively manages categorical features without requiring a lot of pre-processing. With little feature engineering, it provides robust performance and ease of use for a range of ML applications, including regression, classification, and ranking [27].

3.2.8. Convolutional Neural Networks

CNNs are a class of DL models designed to capture spatial hierarchies

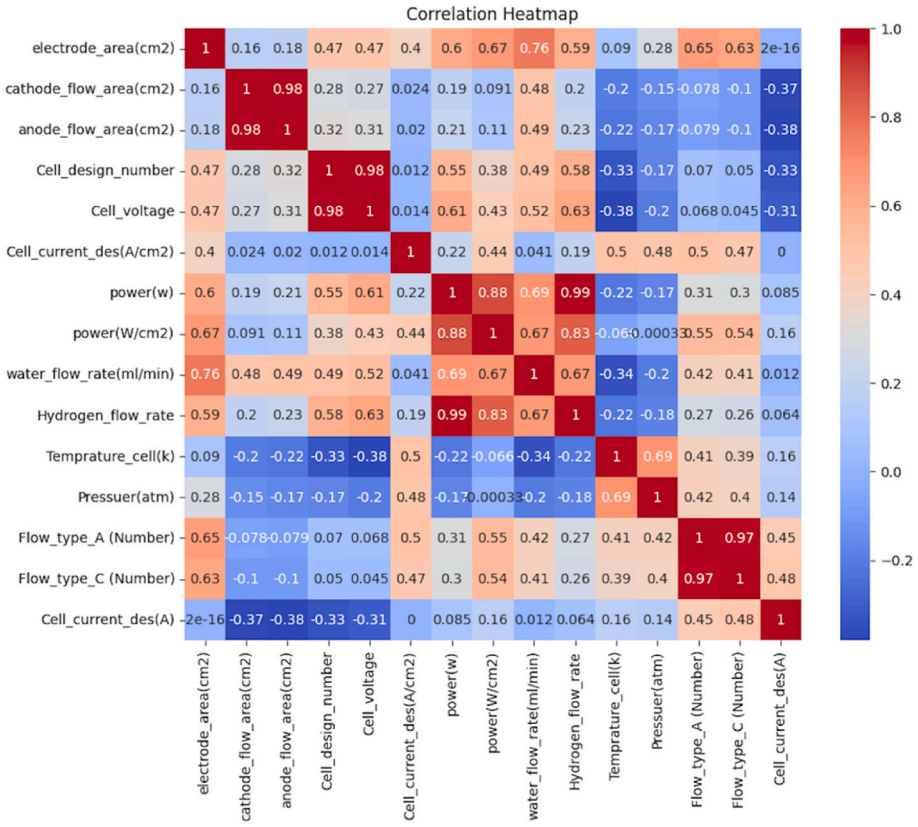


Fig. 4. Correlation heatmap illustrating feature relationships in the PEMWE.

automatically and efficiently in data through the application of convolutional layers. They are particularly effective in tasks involving images or data with spatial relationships, as they can detect local patterns (e.g., edges or textures) and progressively build more complex features, making them suitable for image classification, object detection, and various other ML tasks [28]. In our research, the CNN architecture is defined with convolutional, pooling, and fully connected layers, and the model is trained on the scaled data with early stopping to prevent overfitting. The performance is comprehensively evaluated on the testing data through various regression metrics, and the results are visualised through loss curves and scatter plots for both the full test set and a 5-fold cross-validation approach. Fig. 5 illustrates a multi-layer CNN architecture specifically designed for processing and analysing experimental data from PEMWEs.

3.2.9. *Long Short-Term Memory*

LSTM networks are a specialised type of recurrent neural network (RNN) designed to capture long-range dependencies in sequential data by addressing the vanishing gradient problem. LSTMs use memory cells and gates (input, output, and forget gates) to regulate the flow of information, making them particularly effective in tasks such as time-series forecasting, natural language processing, and speech recognition [30].

3.2.10. *Multi-Layer Perceptron*

An MLP is a class of feedforward ANN that consists of multiple layers of nodes, including an input layer, one or more hidden layers, and an output layer. Each node (or neuron) in the network is fully connected to the nodes in the next layer, and MLPs use non-linear activation functions

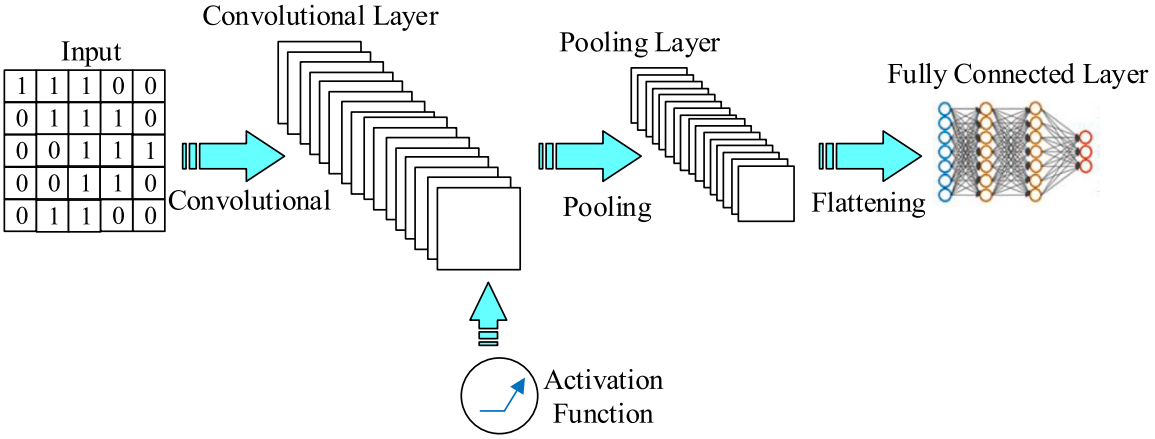


Fig. 5. Leveraging 1DCNN architecture for analysing experimental data from PEMWE [29].

to model complex patterns in data, making them suitable for tasks such as classification, regression, and pattern recognition [31]. In our research, MLP architecture is defined, consisting of two hidden layers with 128 and 64 units respectively, and a single output unit for the regression task; this model is then trained on the scaled training data using the Adam optimiser and mean squared error loss, with early stopping employed to prevent overfitting. The MLP architecture illustrated in Fig. 6 is specifically designed for processing and analysing experimental data from PEMWEs.

We aimed to leverage state-of-the-art AI methods to develop a highly accurate surrogate model for predicting H_2 production rates, pushing the boundaries of efficiency and sustainability in H_2 production. Fig. 7 illustrates the development of an AI-surrogate model for simulating PEMWE.

3.3. Mathematical formulation of 1DCNN on experimental data

In 1DCNN, the convolution operation extracts local features from input variables by applying filters over experimental conditions. The convolutional transformation [51] is given by:

$$Y[i] = \sum_{j=0}^{k-1} X[i+j]K[j] + b \tag{3}$$

where.

- $X[i+j]$: Represents the input feature vector at position $i+j$, which contains experimental parameters such as electrode materials, current density, and cell voltage.
- $K[j]$: Represents the convolutional kernel (filter) of size k, designed to capture dependencies between features.
- b is the bias term and
- $Y[i]$ is the output feature map, which represents transformed feature representations.

After convolution, a max pooling operation is applied to reduce dimensionality while retaining key information. The max pooling operation, widely used in modern DL architectures [52], is given by:

$$Y_p[i] = \max_{k \in s} X[k] \tag{4}$$

where:

$X[k]$: Represents the input feature map within the pooling window.

s: Represents the pooling window size.

$Y_p[i]$: Represents the output feature map after max pooling.

Following feature extraction, the CNN flattens the feature maps and passes them to a fully connected layer, a fundamental component of neural networks from Equations (5) and (6) is given by Ref. [31]:

$$Z = W.A + b \tag{5}$$

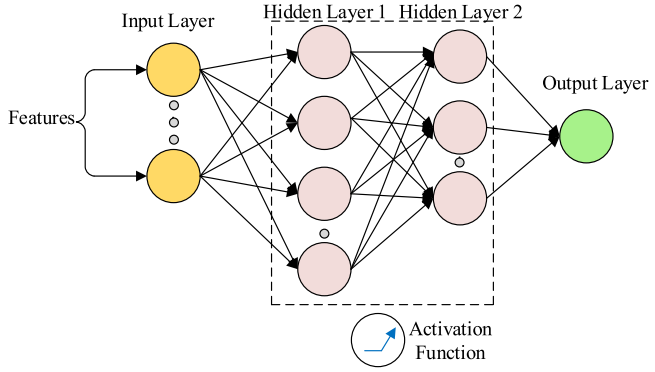


Fig. 6. MLP architecture for analysing experimental data from PEMWE [50].

where.

- Z is the output vector of a fully connected layer in a neural network.
- W is the weight matrix.
- A is the flattened feature vector.
- b is the bias term.

The activation function used is ReLU, defined as:

$$f(x) = \max(0, x) \tag{6}$$

which ensures non-linearity, allowing the model to learn complex relationships between experimental parameters.

3.3.1. Loss function and optimisation

For training, the 1DCNN model minimises the MSE loss function, as defined in Section 3.6.2 (Equation (14)). The MSE measures the average squared difference between the predicted H_2 flow rate ($x_i^!$) and the true H_2 flow rate (x_i) from experimental measurements. Minimising this loss ensures that the model's predictions are as close as possible to the true values.Where:

(x_i): Represents the true H_2 flow rate from experimental measurements.

($x_i^!$): Represents the predicted H_2 flow rate from the 1DCNN model.

n: Represents the number of experimental observations.

The Adam optimiser Equations (7)–(9) from Ref. [32] is employed for weight updates, defined as:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\, g_t \tag{7}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2 \tag{8}$$

$$\theta_t = \theta_{t-1} - \eta \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{9}$$

where.

- g_t: Gradient of the loss function at time step t
- m_t: First-moment estimate (momentum) at time step t
- v_t: Second-moment estimate (adaptive learning rate) at time step t
- β_1 and β_2: Exponential decay rates for the moment estimates
- η: Learning rate.
- ϵ: Small constant to prevent division by zero.
- θ_t: Represents the updated model parameters after applying the Adam update rule.

3.4. Mathematical formulation of MLP on experimental data

The MLP processes input data through a series of fully connected layers, where each layer applies a linear transformation followed by a non-linear activation function. The mathematical formulation of the MLP in Equations 10–12 is given by Ref. [31] is as follows.

1. Input Layer: The input layer receives the feature vector X, which contains experimental parameters such as electrode materials, current density, and cell voltage.
2. Hidden Layers:

The first and second hidden layer computes:

$$Z^{(1)} = W^{(1)}.X + b^{(1)}, A^{(1)} = max\left(0, Z^{(1)}\right) \tag{10}$$

$$Z^{(2)} = W^{(2)}.X + b^{(2)}, A^{(2)} = max\left(0, Z^{(2)}\right) \tag{11}$$

where.

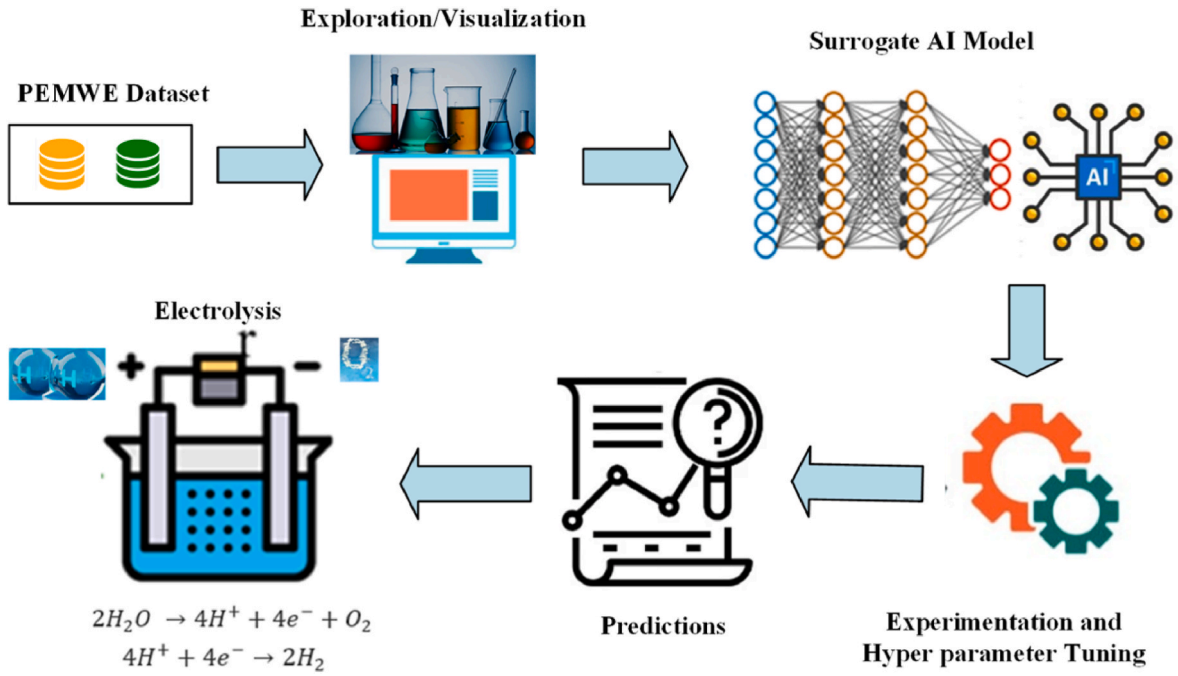


Fig. 7. AI-surrogate model development for PEMWE simulation.

- $Z^{(1)}$: Linear transformation applied to the input before passing it through the activation function.
- $W^{(1)}$ and $W^{(2)}$: Weight matrix of the first and second hidden layer respectively
- $b^{(1)}$ and $b^{(2)}$: Bias term of the first and second hidden layer
- $A^{(1)}$: Activated output using ReLU

3 .Output Layers: Following the feature extraction and transformation in the hidden layers, the MLP passes the final hidden layer's output to the output layer. The output layer computes the predicted value of H_2 flow rate ($x_i^!$) using a linear transformation, which is given by:

$$x_i^! = W^{(3)}.A^{(2)} + b^{(3)} \tag{12}$$

where.

- $x_i^!$: The output layer computes the predicted value of H_2 flow rate.
- $W^{(3)}$: Weight matrix of the output layer
- $b^{(3)}$: Bias term of the output layer
- $A^{(2)}$: Output from the second hidden layer (after applying the activation function).

3.4.1. Loss function and optimisation

The MLP model is trained by minimising the MSE loss function, as defined in Section 3.6.2 (Equation (14)), which measures the average squared difference between the predicted H_2 flow rate ($x_i^!$) and the true H_2 flow rate (x_i) from experimental measurements. For optimisation, the Adam optimiser is employed, as described in Section 3.3.1 (Equations (7)–(9)).

3.5. Hyperparameter tuning

The ML/DL models were configured with carefully selected hyperparameters, including learning rate, number of estimators, hidden units, and optimiser, to optimise their performance. The hyperparameter tuning process was conducted through manual tuning, where key hyperparameters such as the learning rate, batch size, and layer configurations were iteratively adjusted to enhance model performance. This approach allowed for fine-grained control over the model's behaviour and ensured that the selected hyperparameters optimised both training and validation performance while minimising overfitting. The tuning process included an iterative evaluation of hyperparameter variations to assess their impact on model performance, ensuring optimal configurations. The process was guided by cross-validation results, ensuring robustness and generalisability. These models have been categorised as follows: Traditional ML models are detailed in Table 2, ensemble models are outlined in Table 3, and NN models are presented in Table 4. The final hyperparameters for each model were selected based on their performance during cross-validation and are detailed in these tables.

3.6. Evaluation metrics and explainability

The evaluation metrics used to assess the performance of the model emphasise the importance of explainability in understanding the decision-making process of the AI-driven surrogate model. Explainability is critical for ensuring that the model's predictions are consistent with expected physical behaviour and for identifying key parameters that influence H_2 production efficiency. Explainability is a fundamental aspect of AI-driven models, as it provides insights into the underlying decision-making processes. In our study, we employ SHAP analysis to interpret the predictions of the 1DCNN model. SHAP values quantify the contribution of each feature to the model's predictions, enabling us to identify key parameters that influence H_2 production efficiency. This approach enhances the transparency of the model and ensures that its predictions align with expected physical behaviour. By understanding how the model makes predictions, we can verify whether its decisions are consistent with domain knowledge and physical principles.

Table 2
Traditional models hyperparameter configuration

Model	Hyperparameters
SVM	kernel = "rbf, poly, linear", C = 1.0, epsilon = 0.1, gamma = scale
k-NN	n_neighbors = 5, weights = "uniform", leaf_size = 30, p = 2, distance_metrics = ['manhattan', 'Chebyshev', 'minkowski']
DT	max_depth = 15, min_samples_split = 10, min_samples_leaf = 5, max_features = 'sqrt'

Table 3
Ensemble models hyperparameter configuration

Model	Learning Rate	Number of Estimators	Depth	State	Iterations
RF	N/A	200	15	42	N/A
GB	0.1	100	3	42	N/A
LGB	0.05	1000	10	42	N/A
CB	0.05	1000	6	42	1000

Table 4
Neural network models hyperparameter configuration

Model	Learning Rate	Hidden Units/ Layers	Optimiser	Batch Size	Epochs
MLP (keras)	0.001	128,64/2 layers	Adam	4	500
LSTM (keras)	0.001	32/1 layer	Adam	128	600
1DCNN (keras)	0.001	64 filters/1 Conv layer	Adam	128	500

Furthermore, identifying the most influential features allows us to focus on key parameters that can be optimised to improve H_2 production efficiency.

3.6.1. R-Square

R^2 is a statistical metric that quantifies the goodness-of-fit of a regression model. It represents the fraction of the total variation in the dependent variable that can be explained by the independent variables in the model. This measure provides insight into how accurately the model's predictions align with the actual observed data. The R^2 value ranges from 0 to 1, Where a value of 0 indicates that the model explains none of the variability in the data, and value of 1 suggests that the model perfectly explains all the variability [33]. To evaluate the performance of each model, Equation (13) from Galvão et al. [34] and Equations 14–17 from Nascimento et al. [35] were utilised as the key evaluation metrics.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - x_i^{!})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{13}$$

3.6.2. Mean square error

MSE is a key metric in regression analysis that measures prediction accuracy. It calculates the average squared difference between predicted and actual values, emphasising larger errors due to the squaring effect. Lower MSE indicates better model performance, with predictions closer to observed data. While MSE is valuable for model comparison and optimisation, its squared units can complicate interpretation. Despite this limitation, MSE remains a crucial tool in predictive modelling and ML due to its sensitivity to prediction errors and mathematical properties.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i^{!})^2 \tag{14}$$

3.6.3. Normalised square error

NMSE is a refined version of the MSE that scales the error by the variance of the observed data. NMSE effectively measures the model's predictive accuracy relative to the inherent variability in the data. A lower NMSE indicates that the model's predictions align well with the observed values, considering the data's natural spread. This standardised approach makes NMSE particularly useful when comparing models across varying scales or units, offering a more context-aware evaluation of model performance.

$$NMSE = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i^{!})^2}{var(x)} \tag{15}$$

3.6.4. Mean absolute error

MAE is a straightforward metric for assessing prediction accuracy in regression models. It computes the average of the absolute differences between predicted and actual values. Unlike squared error metrics, MAE treats all errors equally, regardless of their magnitude. This characteristic makes MAE less sensitive to outliers and provides a more intuitive interpretation of error in the original units of the dependent variable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_i^{!}| \tag{16}$$

3.6.5. Pearson coefficient correlation

The Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between predicted and observed values in a regression model. It ranges from −1 to 1, where a value near 1 indicates a strong positive correlation, suggesting the model's predictions closely track the observed data's trend. A value near −1 signifies a strong negative correlation and a value close to 0 implies little to no linear relationship.

$$\rho(\curvearrowright) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(x_i^{!} - \overline{x^{!}})}{\sqrt{var(x)var(x^{!})}} \tag{17}$$

From the above equations, where n is the total number of samples, x, $x^{!}$, $\overline{x}$,and $\overline{x^{!}}$ represent the observed, predicted, an average of all observed and predicted values respectively, and $var(x)$ refers to the variance of the input variables, $\rho(\curvearrowright)$ is the Pearson coefficient correlation. These metrics help evaluate not only the accuracy of the models but also their stability and generalisation ability across unseen data.

3.7. Cross-validation

Cross-validation is essential in ML, offering a reliable way to assess a model's capacity to generalise to new data. This technique gives a more accurate view of model performance by simulating how it might perform on previously unseen data. In our methodology, we implement k-fold cross-validation, a particularly rigorous form of this technique. Within the training dataset, we further applied 5-fold cross-validation, ensuring that each fold followed an 80–20 training-validation split. The k-fold cross-validation method involves dividing the dataset into k equally sized segments or "folds". The process unfolds in k iterations, where in each round, k-1 folds are used for training the model, while the remaining fold is reserved for validation as shown in Fig. 8. In our research, we applied 5-fold cross-validation within the training dataset to evaluate the performance of several regression models, including both traditional ML algorithms and DL models. This method helps to account for the variability that can occur due to specific data partitioning methods. We considered multiple performance metrics, including R^2, MSE, NMSE, MAE, and Pearson correlation coefficient. Each model's performance was evaluated across all folds, and we calculated the mean and standard deviation of each metric to ensure robustness. The total cross-validation time recorded for the 1DCNN model, was approximately 0.58 s, demonstrating efficient execution despite the deep learning complexity. This comprehensive approach to model evaluation strengthens the reliability of our results and provides a solid foundation for comparing the effectiveness of various regression techniques.

3.8. Shapiro-Wilk test

The Shapiro-Wilk test was employed to assess whether the residuals or predictions of the models followed a normal distribution. This test is

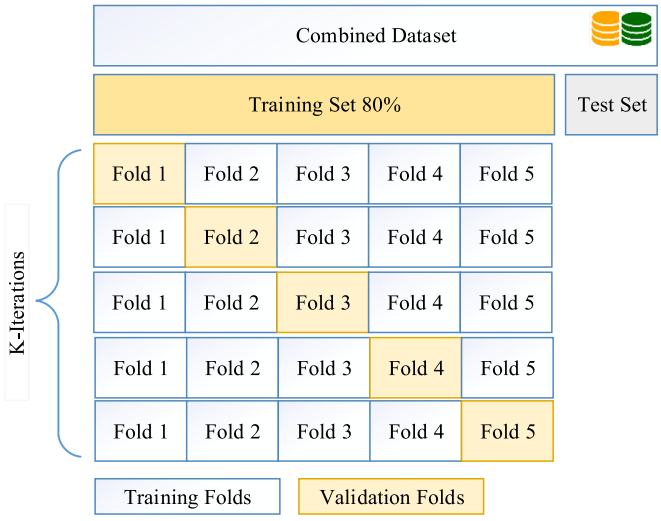


Fig. 8. Dataset split and five-fold cross-validation process [36].

particularly useful for small to moderately sized datasets and is widely recommended for testing normality [37]. The null hypothesis (H0) of the Shapiro-Wilk test states that the data is normally distributed, while the alternative hypothesis (H1) suggests that the data is not normally distributed. A p-value less than the significance level ($\alpha = 0.05$) indicates that the null hypothesis can be rejected, implying non-normality. In this study, the Shapiro-Wilk test was applied to the predictions of all models to determine whether parametric or non-parametric statistical tests should be used for further analysis. The results of this test are presented in Table 8 in the Results section. The findings confirmed non-normality in the data, justifying the use of non-parametric statistical tests, such as the Kruskal-Wallis and Wilcoxon Signed-Rank tests, for comparing model performance.

3.9. Kruskal-Wallis test

The Kruskal-Wallis test is a non-parametric method used to compare the performance of three or more models when the data does not meet the assumptions of normality [53]. In this study, the Kruskal-Wallis test was applied to compare the performance of all models for H_2 predictions. The test was chosen because the Shapiro-Wilk test confirmed non-normality in the data. The results of the Kruskal-Wallis test, including the H-statistic and p-value, are presented in the Results section. This test revealed significant differences between the models, prompting further pairwise comparisons using the Wilcoxon Signed-Rank Test to identify which specific models differed significantly.

3.10. Wilcoxon-signed ranked test

We used the Wilcoxon Signed-Rank Test to assess if the observed performance differences between models were statistically significant or not. According to Galvão et al. [34], this non-parametric test is particularly well-suited for comparing two paired datasets. The statistical significance level for our analysis is set at 0.05. The H0, which asserts that the predictions of the matched samples are similar, was rejected if the test p-value was less than or equal to this cut-off ($p \leq 0.05$). The null hypothesis could not be rejected if the p-value was higher than ($p > 0.05$), indicating that the models' predictions were statistically similar. We were able to carefully evaluate the significance of the differences in the models' performances due to this method.

3.11. Confidence intervals for model performance metrics

We calculated 95 % confidence intervals (CIs) for the performance metrics (R^2, MSE, and MAE) to quantify the reliability and precision of the model predictions. According to Derwent (2023), CIs provide a range of values within which the true metric values are expected to lie with 95 % confidence, accounting for variability in the data. CIs are essential for assessing the stability and generalisability of ML models, especially when comparing multiple models. To compute the CIs, we employed a bootstrapping approach with 10,000 resamples. For each model, we resampled the true and predicted values with replacement and recalculated the performance metrics (R^2, MSE, and MAE) for each resample. This method ensures robust estimates of the CIs, even for non-normally distributed data. The statistical significance of the differences between models was further supported by the CIs. If the CIs of two models for a given metric (e.g., R^2) do not overlap, it suggests a statistically significant difference in their performance. Conversely, overlapping CIs indicate that the models' performances are statistically similar. This approach allowed us to rigorously evaluate the precision and reliability of each model's predictions, providing deeper insights into their comparative performance.

3.12. Error analysis

To ensure a comprehensive understanding of each model's predictive behaviour, an in-depth error analysis was conducted. This approach goes beyond standard evaluation metrics by investigating the distribution and structure of prediction errors through graphical methods. As suggested by [54], analysing error patterns can reveal model limitations that are not captured by scalar metrics such as R^2 or MAE. The analysis utilised error histograms, residual plots, and box plots to assess each model's prediction behaviour. Residuals, defined as the difference between actual and predicted H_2 flow rates and were plotted to inspect for randomness and symmetry around zero. Histograms of the residuals provided insight into error dispersion and skewness, where narrow, symmetric distributions indicated well-calibrated models. Box plots further summarised error variability, highlighting interquartile ranges and identifying extreme values that may suggest overfitting or underfitting.

4. Results and discussion

In this section, we present a comprehensive analysis of the results obtained from the various ML and DL models used to predict H_2 production rates through PEMWE systems. The results are discussed in the context of key performance metrics, including R^2, MSE, NMSE, MAE, and Pearson correlation.

4.1. Performance metrics overview

The results of each model's performance across various evaluation metrics are summarised, showcasing their ability to accurately predict H_2 production. The evaluation includes several ML models, from traditional methods like k-NN and SVM to more advanced ensemble techniques such as GB and RF. Additionally, DL architectures, including 1DCNN, LSTM, and MLP, were assessed for their predictive capabilities. Table 5 provides a detailed analysis of experimental models based on performance metrics, with the best values for each metric highlighted in bold. Both training and testing phases are separately reported for each model, allowing for a comprehensive comparison of their performance.

4.2. Model performance analysis

The 1DCNN model demonstrated outstanding performance, achieving an R^2 value of 0.998944 and an MSE of 488.82 on the test dataset. With a low NMSE of 0.001055 and an exceptionally high

Table 5
Analysis of experimental models based on performance metrics. Best values for each metric are highlighted in bold. Training and testing phases are separately reported for each model.

No	Model	Phase	R^2	MSE	NMSE	MAE	Pearson
1.	SVM	Testing	0.994586	2506.94	0.005414	14.5685	0.997319
		Training	0.991986	2788.56	0.008013	18.2078	0.996119
2.	k-NN	Testing	0.964573	16404.20	0.035426	17.3654	0.984180
		Training	0.983097	5881.96	0.016903	12.8715	0.992485
3.	DT	Testing	0.910855	41278.45	0.089144	44.1893	0.957488
		Training	0.959132	14221.30	0.040868	32.3317	0.979353
4.	LSTM	Testing	−0.050863	486601.00	1.050860	234.3390	0.553064
		Training	−0.133791	394535.00	1.133790	233.3050	0.466299
5.	RF	Testing	0.911555	40954.17	0.088444	28.9703	0.957783
		Training	0.978870	7352.86	0.021130	16.3980	0.990448
6.	GB	Testing	0.989748	4747.07	0.010251	16.5174	0.995113
		Training	0.999542	159.35	0.000457	7.8604	0.999772
7.	1DCNN	Testing	**0.998944**	**488.82**	**0.001055**	**8.7028**	**0.999472**
		Training	0.999352	225.46	0.000647	7.8165	0.999677
8.	LGB	Testing	0.975219	11474.49	0.024780	22.1823	0.988291
		Training	0.996015	1386.74	0.003985	8.7726	0.998012
9.	CB	Testing	0.961827	17675.78	0.038172	18.7181	0.981752
		Training	0.999884	40.44	0.000116	3.7453	0.999942
10.	MLP	Testing	0.997542	1137.74	0.002457	11.2627	0.998773
		Training	0.997845	749.79	0.002154	10.3294	0.998993

Pearson correlation of 0.999472, the model proved to be highly reliable and well-suited for predicting H_2 production rates. During the training phase, the 1DCNN model achieved an R^2 value of 0.999352 and an MSE of 225.46, further confirming its robustness. In contrast, the MLP model achieved an R^2 value of 0.997542 and an MSE of 1137.74 on the test dataset, with an NMSE of 0.002457 and a Pearson correlation of 0.998773. During the training phase, the MLP model achieved an R^2 value of 0.997845 and an MSE of 749.79, demonstrating strong performance. Despite its rapid training time, MLP had slightly lower accuracy than 1DCNN. The LSTM model, however, achieved a negative R^2 value (−0.050863) and a high MSE of 486601.00 on the test dataset, indicating poor predictive performance. During the training phase, the LSTM model also performed poorly, with an R^2 value of −0.133791 and an MSE of 394535.00. This is because LSTM models are specifically designed for sequential or time-series data, whereas our dataset is not time-dependent. In addition to predictive accuracy, the computational efficiency of the models is a critical factor for real-world deployment. Table 6 below compares the training times, inference times, and inference time variability for the 1DCNN, MLP, and LSTM models. All models were trained and tested on the same hardware environment (Tesla T4 GPU with 15.83 GB and 13.61 GB RAM) to ensure fair comparison. The 1DCNN model completed training in ~69.95 s for 500 epochs, while the MLP model trained in ~73.40 s and the LSTM model in ~71.35 s for 600 epochs. However, the inference time is a more critical factor for real-world deployment. As shown in Table 6, the 1DCNN model achieved an average inference time of ~101.89 ms per sample, which corresponds to ~9 predictions per sec. This is significantly faster than conventional physical models, making the 1DCNN model highly suitable for real-time operational use in industrial-scale PEMWE systems. Similarly, the MLP model achieved an average inference time of ~97.67 ms, allowing it to process ~10 predictions per sec. The LSTM model, with a mean inference time of ~125.13 ms, can process ~11 predictions per sec.

Table 6
Comparison of training, average inference times and average inference Std Dev per sample for 1DCNN, MLP, and LSTM models

Model	Training Time (s)	Average Inference Time per sample (ms)	Average Inference Std Dev per sample (ms)	Environment
1DCNN	69.95	101.89	29.58	Tesla T4 GPU (15.83 GB, 13.61 GB RAM)
MLP	73.40	97.67	21.94	
LSTM	71.35	88.99	29.74	

Despite this, we retained the LSTM results in Table 5 to provide a comprehensive comparison of different model architectures. Fig. 9 illustrates the 1DCNN model's predictive capabilities, comparing actual values with predicted values during both the training and testing phases. Fig. 10 clearly shows that the model maintained excellent predictive accuracy throughout the process.

The training and validation loss curves of the 1DCNN model, presented in Fig. 11, exhibit smooth convergence for both loss curves, reflecting efficient learning and consistent performance across both the training and testing datasets, showcasing the proposed 1DCNN capabilities to generalise.

4.3. Model-wise residual behaviour and error insights

To further assess the predictive performance of different models, we conducted an in-depth error distribution analysis, highlighting the scenarios where certain models underperform and identifying potential reasons. The Error histograms, Residual plots, and Box plots (Fig. 12 (a), (b), and (c)) provide insights into the error distribution for each model. The LSTM model exhibited the highest errors among all models, with a significant number of extreme deviations, as seen in the residual plot. This underperformance is attributed to the fact that LSTM is designed for sequential or time-series data, whereas our dataset is not sequential. Consequently, LSTM struggles to learn meaningful patterns, resulting in

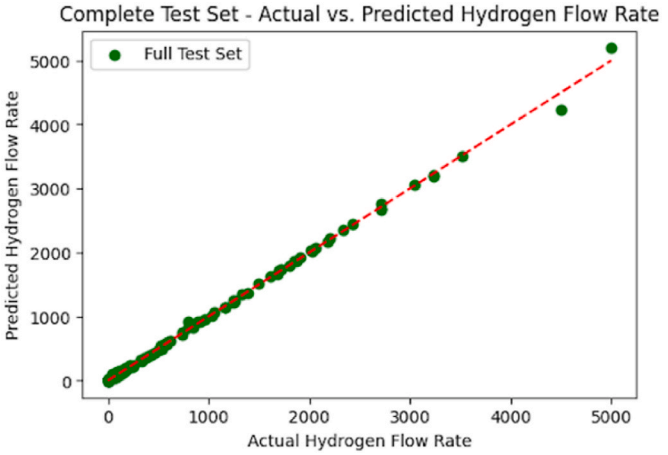


Fig. 9. Actual vs predicted for complete testing for the 1DCNN model.

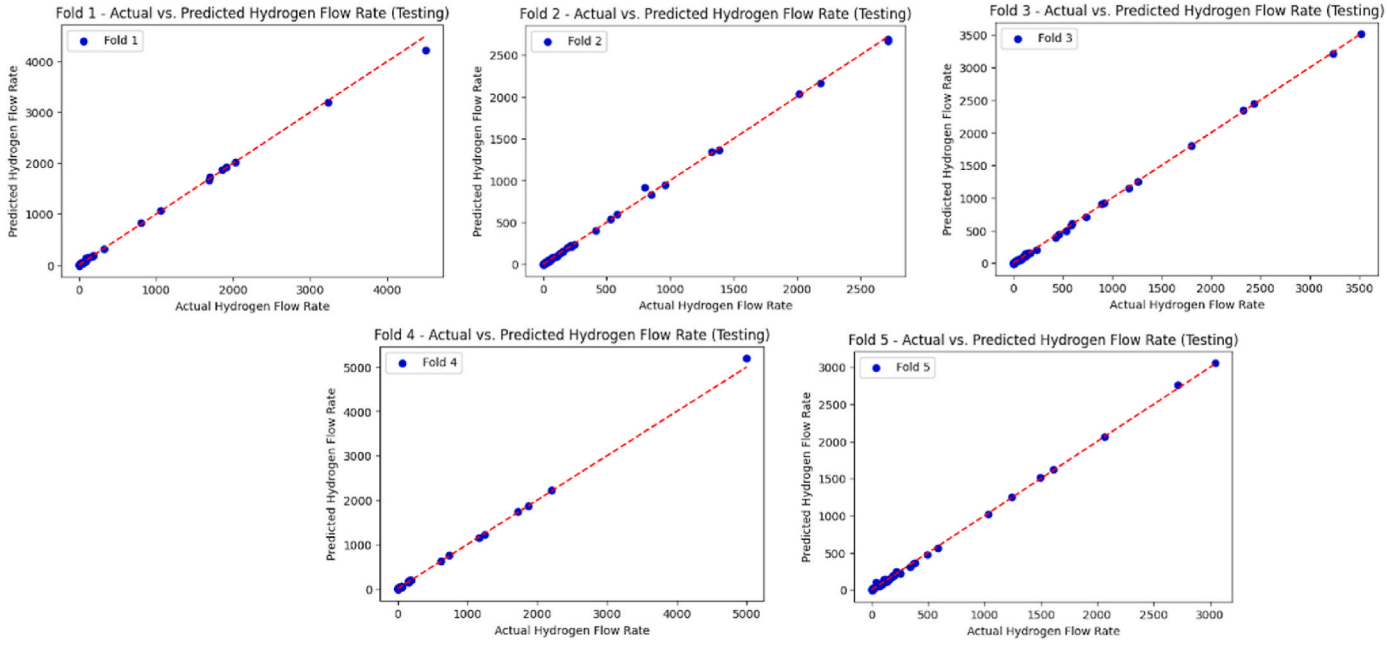


Fig. 10. Actual vs. predicted hydrogen flow rates for each fold in testing during 5-folds for the 1DCNN model.

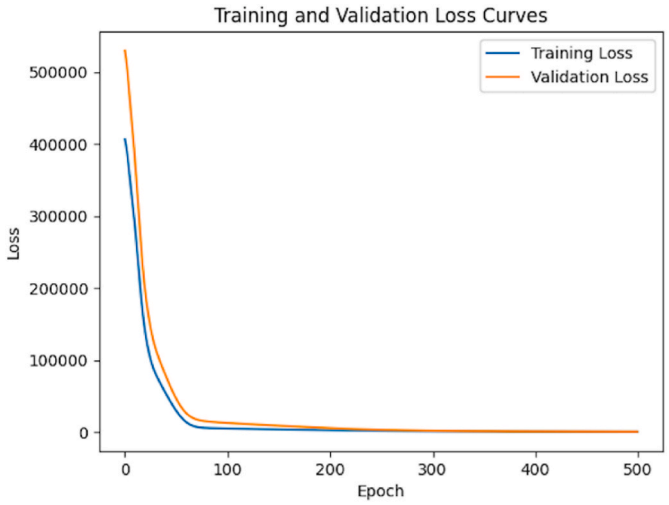


Fig. 11. Loss curve during the training and validation phases for 1DCNN.

a negative R^2 value and high prediction errors. In contrast, models like 1DCNN, which are better suited for non-sequential data, demonstrated superior performance, as discussed in Section 4.4. The k-NN model also showed a wider error spread, indicating inconsistent predictions. k-NN relies on distance-based similarity, making it highly sensitive to feature scaling and the curse of dimensionality. Since our dataset contains multiple features, some with varying magnitudes, k-NN may struggle to identify meaningful neighbors, leading to high variance in its predictions.

Similarly, the DT model demonstrated noticeable performance fluctuations, with a wider error spread and outliers. DT tend to overfit training data when not properly pruned, which leads to poor generalization on unseen data. The residual plot for DT reveals that the model has difficulty capturing the complex relationships between input variables and H_2 production, contributing to its lower predictive accuracy. The SVM model showed moderate errors, performing better than DT and k-NN but worse than 1DCNN, GB, and RF. SVM's performance is affected by feature selection and kernel choice, as it is more effective in datasets with well-defined decision boundaries. Conversely, 1DCNN, GB, and RF emerged as the best-performing models, with tighter error distributions and fewer outliers. 1DCNN's ability to extract spatial correlations in the data, coupled with GB and RF's ensemble learning mechanisms, enables these models to achieve superior generalisation, leading to low MAE and stable residual distributions. Overall, this analysis confirms that models with higher complexity and feature extraction capabilities (1DCNN, GB, RF) demonstrate better predictive performance, whereas models that rely on distance-based or sequential learning (kNN, DT, LSTM) exhibit higher error variance due to their inherent limitations when applied to the given dataset.

4.4. SHAP analysis for 1DCNN model interpretability

To ensure the transparency and explainability of the 1DCNN model's predictions, SHAP analysis was conducted using the gradient explainer framework. The primary goal of this analysis is to identify the most influential features contributing to H_2 production predictions and to assess whether the model's decision-making process aligns with known electrochemical principles governing PEMWE performance. Fig. 13 presents both the SHAP summary and *beeswarm* plots, illustrating the impact of individual input features on the model's predictions. The SHAP summary plot ranks feature by their mean absolute SHAP values, providing a measure of their overall importance. Meanwhile, the *beeswarm* plot offers a granular perspective by visualising how feature values influence individual predictions, with positive (red) and negative (blue) contributions mapped across different samples.

The analysis revealed that power (W) is the most influential factor in the model's predictions, with the highest mean SHAP value, indicating its dominant role in H_2 production. This aligns with fundamental electrochemical principles, as increased power input directly enhances the electrolysis process. Similarly, water flow rate (ml/min) and anode flow area (cm^2) emerged as highly influential parameters, reinforcing the importance of reactant availability and electrode surface area in determining H_2 production. Additionally, cell voltage and temperature (K) exhibited moderate impact, consistent with their role in dictating cell efficiency and reaction kinetics. The SHAP values indicate that higher temperatures and optimised voltages contribute positively to H_2 production, supporting established PEMWE operational insights. Conversely, features such as flow type A (Number) and cell current

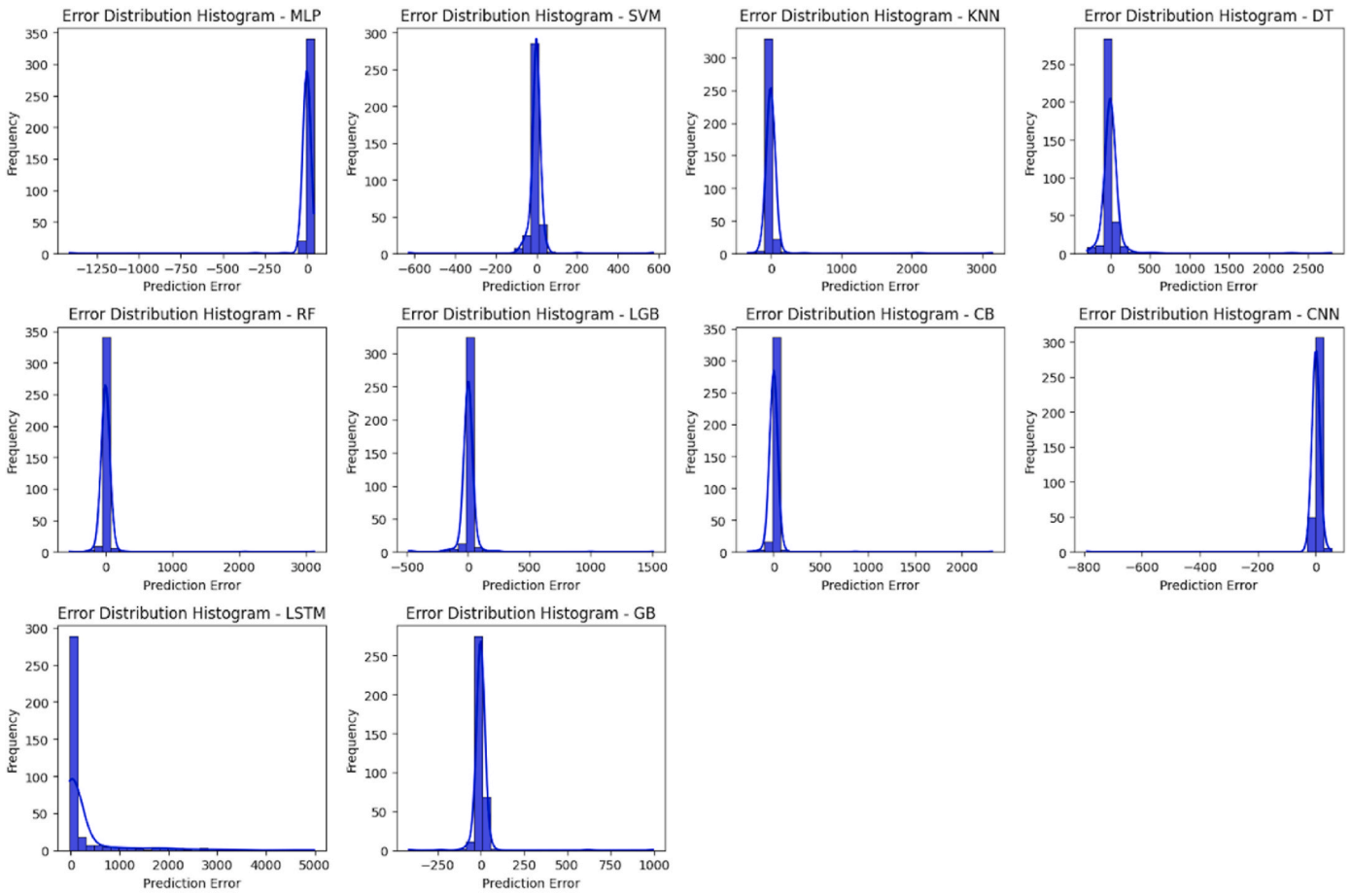


Fig. 12. Error analysis across models: (a) Error histograms, (b) residual plots, and (c) box plots.

design (A) had the lowest mean SHAP values, suggesting their minimal influence on the model's predictions. These findings imply that, while important for specific system configurations, these parameters have limited direct impact on H_2 production under the operating conditions.

The explainability provided by SHAP ensures that the 1DCNN model is not only accurate but also interpretable, offering valuable insights into how AI-driven surrogate models capture complex electrochemical relationships. Understanding these interactions is critical for improving model reliability and aiding experimentalists in optimising PEMWE design and operation. By leveraging SHAP for interpretability, the study ensures that the AI-driven approach not only predicts H_2 production efficiently but also aligns with known physical and electrochemical principles. This enhances trust in the model's reliability and its applicability for real-world PEMWE system optimisation.

4.5. Cross-validation of 1DCNN

The cross-validation process followed the K-fold strategy with 5 splits. Specifically, the training data was split into 5 folds, and for each fold, the data was further divided into training and validation sets with an 80/20 ratio. The 1DCNN model was then trained on 80 % of the training data and validated on the remaining 20 %, where performance metrics such as R^2, MSE, MAE, NMSE, and Pearson correlation coefficient were computed. Each metric includes the count of evaluations, mean, standard deviation (std), minimum (min), and the values at the 25th, 50th and 75th percentiles including differences between mean and testing performance across fold. These results are summarised and presented in a Table 7 to provide a comprehensive evaluation of the 1DCNN model performance.

4.6. Shapiro-Wilk test findings

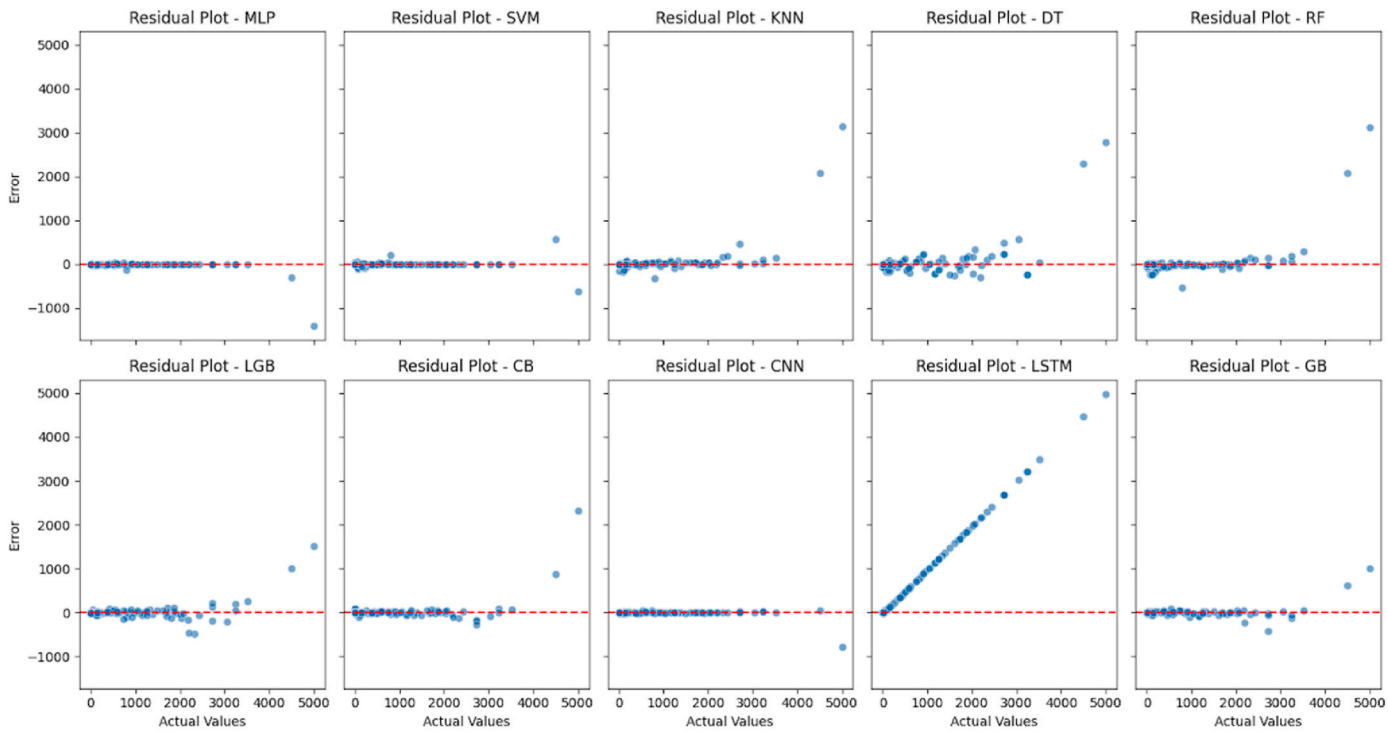
The results of the Shapiro-Wilk test, as discussed in Section 3.8, are presented in Table 8. The test was applied to the predictions of all models to assess whether they followed a normal distribution. All p-values were less than 0.05, confirming that the predictions of all models do not follow a normal distribution. This finding justified the use of non-parametric statistical tests, such as the Kruskal-Wallis and Wilcoxon Signed-Rank tests, for comparing model performance.

4.7. Kruskal-Wallis test results

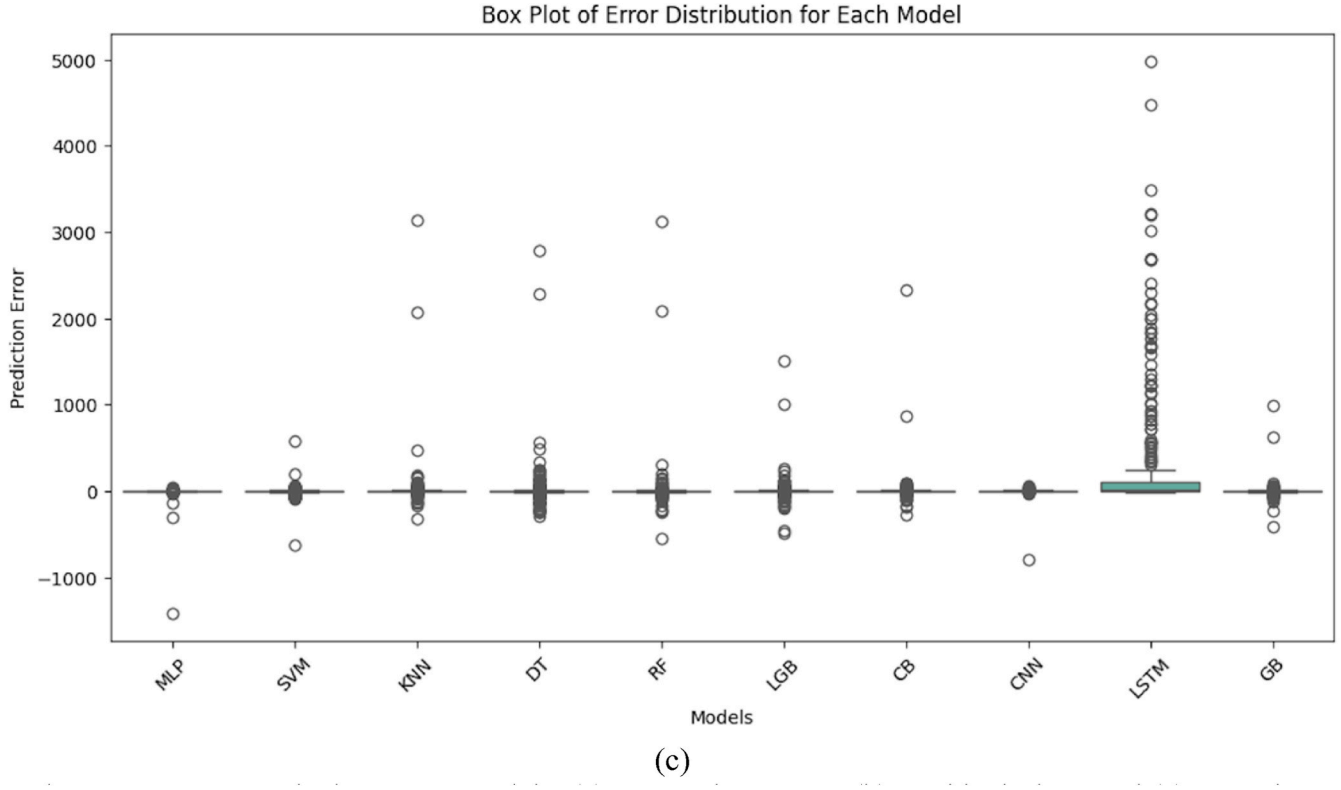
The Kruskal-Wallis test, as discussed in Section 3.9, was applied to compare the performance of all models for H_2 predictions. The test was chosen because the Shapiro-Wilk test confirmed non-normality in the data. The results of the Kruskal-Wallis test are as follows.

- H-statistic: 106.52407586306428
- p-value: 7.4875504437937495e-19

The p-value is less than 0.05, indicating significant differences between the models. This result prompted further pairwise comparisons using the Wilcoxon Signed-Rank Test to identify which specific models differed significantly.



(b)



(c)

Fig. 12. (*continued*).

4.8. Evaluation of prediction statistical discrepancies

The Wilcoxon Signed-Rank Test was employed to assess whether the predictions of the models were statistically different, as discussed in Section 3.10. Table 9, shows a pairwise comparison of the models, with p-values derived from the Wilcoxon Signed-Rank Test. According to the results, models like k-NN, SVM, DT, CB, LGB, GB, RF and MLP show no statistical significant differences in their predictions when compared to 1DCNN and LSTM, as their p-values are greater than 0.05 for most of the comparison, being the only models that have passed the Wilcoxon test. However, the 1DCNN model stands out as not only a more robust but also more accurate model than the LSTM, as the latter had the worst performance among all models. These findings highlight that, although these models like, k-NN, SVM, DT, CB, LGB, GB, RF, and MLP achieved

Fig. 13. SHAP summary and *beeswarm* plot for the 1DCNN model.

Table 7
Performance metrics of 1DCNN cross-validation. "Differences" column presents the difference between the averaged metrics computed across all folds and the metrics computed in the test set.

Metric	Mean	Std	Min	25 %	50 %	75 %	Testing Value	Differences
R^2	0.996234	0.00253	0.99150	0.99584	0.99730	0.99789	0.99894	−0.00270
MSE	1454.076	694.43736	909.19312	1005.92106	1104.23422	1458.65306	488.82570	965.25120
NMSE	0.003765	0.00253	0.00137	0.00210	0.00269	0.00415	0.00105	0.00270
MAE	19.99636	3.01348	16.32914	17.32136	19.62909	22.31241	8.70287	11.29348
Pearson	0.998289	0.00115	0.99619	0.99796	0.99872	0.99920	0.99947	−0.00118

results comparable to 1DCNN, they are not statistically different from each other, meaning that they do not present robust prediction power.

4.9. Analysis of CIs for model performance metrics

The 95 % CIs for the performance metrics (R^2, MSE, and MAE) were calculated to quantify the reliability and precision of the model

Table 8
Shapiro-Wilk test results for normality assessment of model predictions. P-values ≤0.05 indicate that the data does not follow a normal distribution, justifying the use of non-parametric tests for further analysis.

Model	Statistic	P-value
KNN	0.469081	2.02E-31
SVM	0.455871	1.02E-31
DT	0.478613	3.33E-31
RF	0.487613	5.38E-31
CB	0.468621	1.97E-31
LGB	0.463014	1.47E-31
GB	0.457485	1.11E-31
LSTM	0.753276	6.43E-23
MLP	0.445567	6.05E-32
1DCNN	0.449644	7.43E-32

predictions, as discussed in Section 3.11. The results are presented in Table 10. As explained earlier, non-overlapping CIs suggest statistically significant differences in performance, while overlapping CIs indicate statistically similar performances.

4.10. Comparative analysis

To evaluate the effectiveness of our AI surrogate model, we conducted a comparative analysis with the results of previous studies using PEMWE dataset. Table 11 summarises this comparison, highlighting the number of models, dataset sizes, feature counts, statistical validation techniques, and the best performance metrics for each study.

Compared to previous studies, our approach integrates 10 models, making it the most extensive comparative study, trained on a larger dataset (1210 samples, 26 features). In contrast, Mohamed et al.[5] utilised 5 models with 15 features, and Rui et al. [45] employed only 2 models with 5 features. A key distinction in our study is the incorporation of both statistical significance testing and 5-fold cross-validation, ensuring a rigorous evaluation of model generalisability. Although Mohamed et al. [5] reported a lower MAE (6.4383 for testing) compared to our model's MAE of 8.7028, their study did not incorporate cross-validation or statistical analysis. Our approach, by integrating 5-fold cross-validation and statistical significance tests, provides a more comprehensive assessment of model robustness, reducing overfitting risks and ensuring generalisability to unseen data. Rui et al. [45] implemented cross-validation, enhancing model reliability, but their use of only 2 models and 5 features limits predictive power. Additionally, no explicit best metrics were reported for train/test performance, making direct comparisons difficult.

Among other comparative studies, Tawalbeh et al. [14] achieved an exceptionally low-test MAE (0.0012). However, their dataset contained only 450 samples, which may limit generalisability when applied to broader, real-world PEMWE systems. Biswas et al. [15] used a moderate dataset size (1000 samples) and reported a low overall MSE (0.0033). However, like several previous studies, cross-validation and statistical validation were not included, which are essential for ensuring model robustness. [49] applied 9 models on 578 samples with 21 features, providing a detailed exploration of PEMWE design factors. However, they did not report best metrics for train/test, making it difficult to assess direct performance comparisons. Among studies with smaller datasets, Rezk et al. [12] achieved a near-perfect training RMSE (3.4×10^{-6}), but their dataset was very small (17 samples), raising concerns about overfitting and lack of scalability. Similarly, Arjmandi et al. [9] and Bakır et al. [33] reported exceptionally low MAE values (0.0000 and 0.049, respectively), but their datasets were significantly smaller (42/162 samples and 21 samples, respectively), limiting their applicability to large-scale PEMWE systems. These enhancements in our study provide a more comprehensive and reliable assessment of model performance to accurately simulate PEMWE behaviour for H_2 production. Another key distinction of our study is the integration of explainable AI techniques, specifically SHAP analysis, to enhance model interpretability and reliability. Unlike previous studies that primarily focused on predictive accuracy, we employed SHAP analysis to validate the reliability and robustness of our AI surrogate model. This method enables the identification of feature contributions to H_2 production, allowing for a more transparent and interpretable model. As demonstrated in Section 4.4, our SHAP analysis highlights the most influential input variables, such as power (W), water flow rate (ml/min), and anode flow area (cm^2), while revealing features with lower importance, ensuring trustworthiness in our model's predictions. These enhancements in our study, including the use of explainable AI techniques, provide a more comprehensive and reliable assessment of model performance to accurately simulate PEMWE behaviour for H_2 production.

4.11. Discussion

In our study, we systematically evaluated 10 ML/DL models for predicting H_2 production via PEMWE systems. The goal was to identify which models could capture the complex non-linear relationships between input features and outputs, and which would provide the most accurate predictions. The comparative analysis involved several

Table 10
CIs (95 %) for model performance metrics (R^2, MSE, and MAE). Non-overlapping confidence intervals indicate statistically significant differences in model performance, while overlapping intervals suggest statistically similar performances.

MODEL	R^2 95 % CI	MSE 95 % CI	MAE 95 % CI
KNN	(0.91929, 0.99921)	(274.36, 45052.78)	(7.4446, 32.4849)
SVM	(0.99054, 0.99891)	(400.73, 5669.86)	(10.3562, 20.1583)
DT	(0.83457, 0.98720)	(4151.34, 99947.60)	(27.2956, 67.4405)
RF	(0.81898, 0.99626)	(1294.18, 106972.75)	(12.7074, 52.3862)
CB	(0.90656, 0.99843)	(521.19, 49976.89)	(8.6964, 34.4935)
LGB	(0.95437, 0.99495)	(1696.08, 27255.85)	(12.9187, 33.9754)
GB	(0.98077, 0.99857)	(480.89, 11656.27)	(10.6800, 24.2683)
LSTM	(-0.16516, −0.09442)	(308893.40, 763013.43)	(191.0842, 329.2073)
MLP	(0.99564, 0.99912)	(329.81, 2550.47)	(8.3563, 14.8240)
1DCNN	(0.99820, 0.99963)	(134.24, 1022.25)	(6.8643, 11.0919)

Table 9
Wilcoxon signed rank test evaluation for pairwise comparison of model predictions. P-values ≤0.05 are highlighted in bold, highlighting the cases where the models passed the test in a pair-wise comparison.

MODEL	SVM	KNN	DT	RF	GB	1DCNN	LGB	CB	MLP	LSTM
SVM	–	0.133606	0.198408	**0.039364**	0.978067	**0.009592**	0.068011	0.902533	0.079953	**2.91E-32**
KNN	0.133606	–	0.137002	**2.65E-06**	0.108934	**0.027997**	0.853666	0.197364	**0.000463**	**1.04E-42**
DT	0.198408	0.137002	–	**0.008471**	0.063175	**0.033017**	0.605617	0.940631	0.572193	**5.48E-43**
RF	**0.039364**	**2.65E-06**	**0.008471**	–	0.068693	**8.94E-14**	**0.000139**	**0.026465**	0.072663	**2.97E-49**
GB	0.978067	0.108934	0.063175	0.068693	–	**0.000282**	**0.010032**	0.190503	0.317945	**4.62E-50**
1DCNN	**0.009592**	**0.027997**	**0.033017**	**8.94E-14**	**0.000282**	–	**0.026195**	**0.002494**	**5.45E-10**	**5.11E-32**
LGB	0.068011	0.853666	0.605617	**0.000139**	**0.010032**	**0.026195**	–	0.367472	**0.002616**	**4.47E-38**
CB	0.902533	0.197364	0.940631	**0.026465**	0.190503	**0.002494**	0.367472	–	0.197016	**2.53E-40**
MLP	0.079953	**0.000463**	0.572193	0.072663	0.317945	**5.45E-10**	**0.002616**	0.197016	–	**1.40E-43**
LSTM	**2.91E-32**	**1.04E-42**	**5.48E-43**	**2.97E-49**	**4.62E-50**	**5.11E-32**	**4.47E-38**	**2.53E-40**	**1.40E-43**	–

Table 11
Comparative analysis of studies based on number of models, dataset size, features, best metrics on train/test sets, statistical tests, and cross-validation using PEMWE dataset.

Study	No. of Model	Dataset Size	Features	Best Metrics Train/Test	Statistical Test	Cross-Validation
Our work	10	1210	26	MAE 7.8165/8.7028	Yes	Yes
Mohamed et al. [5]	5	1203	15	MAE 5.0006/6.4383	No	No
Rui et al. [45]	2	1062	5	Not reported	No	Yes
Tawalbeh et al. [14]	4	450	7	MAE Not reported/0.0012	No	Yes
Rezk et al. [12]	1	17	3	RMSE 3.4×10^{-6}/0.2308	No	No
Biswas et al. [15]	3	1000	4	MSE Overall 0.0033	No	No
[49]	9	578	21	Not reported	No	Yes
Arjmandi et al. [9]	7	42/162	5	MAE Overall 0.0000	No	No
Bakır et al. [33]	8	21	4	MAE Not Reported/0.049	No	Yes

performance metrics, including R^2, MSE, NMSE, MAE, and Pearson correlation, with a specific focus on understanding the strengths and limitations of each model in different operational contexts. Among the DL models, the neural network, particularly 1DCNN stood out as the best performer, with an R^2 value of 0.998944 and an exceptionally low MSE of 488.82. This model excelled in capturing non-linear relationships and showed strong generalisation across both the training and testing phases. The success of 1DCNN can be attributed to its architecture, which allows it to model complex interactions between input features while maintaining computational efficiency. However, other models also demonstrated impressive performance. The MLP model achieved an R^2 value of 0.997542 and MSE of 1137.74, performing on par with 1DCNN. Similarly, other DL model, such as LSTM, underperformed in this specific application. LSTM, typically well-suited for time-series data, was not able to leverage its strength due to the absence of sequential dependencies in the dataset. The LSTM model's R^2 value was −0.050863 with a very high MSE of 486601.00, indicating its inability to generalise well for the PEMWE data.

To ensure the robustness of our comparative analysis, we assessed the normality of model predictions using the Shapiro-Wilk test, which is widely recommended for small to moderately sized datasets [37]. The H0 assumes normal distribution, while the alternative H1 suggests non-normality. A p-value <0.05 rejects H0, indicating non-normality. In this study, the test confirmed non-normality for all models ($p < 0.05$), justifying the use of non-parametric tests like Kruskal-Wallis and Wilcoxon Signed-Rank for performance comparisons. The Kruskal-Wallis test, suitable for non-normal data (Ates et al., 2023), revealed significant differences between models (H-statistic = 106.524, $p = 7.487 \times 10^{-19}$), prompting pairwise comparisons using the Wilcoxon Signed-Rank Test. To assess the statistical significance of the differences between model performances, the pair-wise Wilcoxon Signed-Rank Test was conducted. Even though the differences in performance metrics between the top-performing model, 1DCNN, and the second-best model, MLP, appear marginal when compared to any other models, they are statistically significant, as confirmed by the Wilcoxon Signed-Rank Test. This indicates that while MLP and RF demonstrated competitive metrics, their predictive capabilities were statistically like 1DCNN, and no model was shown to significantly outperform the others in the PEMWE dataset. The results of this study suggest that 1DCNN is a highly effective model for simulating the behaviour of H_2 production in PEMWE systems, offering superior accuracy and generalisability across a range of operational conditions. In addition to performance metrics, an in-depth error distribution analysis was conducted to understand the prediction behaviour and robustness of each model under varying conditions. The analysis employed residual plots, error histograms, and box plots to visualise the spread and symmetry of prediction errors [54]. These tools revealed that while top-performing models such as 1DCNN and MLP produced tightly clustered and symmetric residuals cantered around zero indicating high predictive accuracy and minimal bias other models like LSTM and k-NN exhibited broader and more skewed error distributions. The LSTM model showed significant deviations, confirming its unsuitability for this non-sequential dataset. On the other hand, the k-NN model's wide error range reflected its sensitivity to high-dimensional feature space and scaling issues. These findings, supported by graphical residual evaluations, complement the numerical metrics, and offer a detailed understanding of each model's generalisation capacity and potential weaknesses. This layer of analysis is essential for identifying not just which models perform best, but also why they perform as they do, thereby enhancing the interpretability and reliability of model selection for PEMWE system simulation.

Its strong performance, coupled with its relatively low computational cost, makes it promising for further optimisation and potential integration into real-time control systems for PEMWE operations. To quantify prediction reliability, we calculated 95 % CIs for R^2, MSE, and MAE. CIs provide a range within which the true metric values are expected to lie with 95 % confidence, accounting for data variability [39]. Using a bootstrapping approach with 10,000 resamples, we computed robust CIs for each model. The CIs revealed statistically significant differences between models. For example, the 95 % CI for R^2 of 1DCNN (0.99820, 0.99963) did not overlap with that of MLP (0.99564, 0.99912), indicating superior performance of 1DCNN. Overlapping CIs, such as between MLP and RF for MAE, suggested statistically similar performances. This rigorous evaluation provided deeper insights into model precision and reliability, highlighting 1DCNN as the top performer.

To gain deeper insights into the 1DCNN model's decision-making process, SHAP analysis was conducted. The SHAP summary and beeswarm plot revealed the relative importance of input features in influencing H_2 production predictions. The analysis highlighted that power (W) was the most influential feature, with the highest mean SHAP value, followed by water flow rate (ml/min) and anode flow area (cm^2). These findings align with the expected physical behaviour of PEMWE systems, where power input and flow dynamics play critical roles in determining H_2 production efficiency. The SHAP analysis also provided instance-level insights into how specific features contribute to individual predictions. For example, higher power inputs were consistently associated with increased H_2 production, while lower water flow rates were linked to reduced efficiency. This granular understanding of feature contributions enhances the interpretability of the 1DCNN model and ensures that its predictions are consistent with domain knowledge. The 1DCNN model demonstrated the highest accuracy ($R^2 = 0.998944$) and required a training time of ~69.95 s, which is slightly faster than the MLP model (~73.40 s) and the LSTM model (~71.35 s). Importantly, the 1DCNN model achieved average inference time of ~101.89 ms per sample, which corresponds to ~9 predictions per sec. This makes the 1DCNN model highly suitable for real-time operational use in industrial-scale

PEMWE systems. This trade-off highlights the importance of balancing accuracy and computational efficiency when selecting a model for real-world applications. The LSTM model, despite its potential in sequential data tasks, performed poorly on this dataset, achieving a negative R^2 value (−0.050863) and require ~71.35s to train. This underscores the importance of selecting model architectures that align with the data characteristics, as LSTM's strength in handling sequential data was not leveraged in this study due to the absence of temporal dependencies in the PEMWE dataset. In terms of inference time, 1DCNN achieved an average inference time of ~101.89 ms per sample, which corresponds to ~9 predictions per sec, with a variability of 29.58 ms indicating a consistent and reliable performance across multiple inference cycles. While the MLP model exhibited a similar average inference time of 97.67 ms with variability 21.94 ms, but it has lower accuracy compared to 1DCNN. However, despite its efficiency, MLP did not achieve the same level of predictive accuracy as 1DCNN, making it less suitable for applications where accuracy is the primary concern. Conversely, the LSTM model recorded an average inference time of 88.99 ms and exhibited 29.74 ms variability, making it computationally inefficient for real-time deployment. Despite its strengths in sequential data processing, LSTM failed to capture meaningful relationships in this structured dataset, further confirming its unsuitability for PEMWE modelling. In industrial applications, where datasets are significantly larger, computational costs could rise substantially. To handle this, optimisation strategies such as batch processing, parallel computing, and distributed DL can enhance efficiency and scalability. Batch processing reduces memory load by dividing data into smaller segments, improving training speed, while parallel computing distributes computations across multiple processors or GPUs, accelerating model execution. These techniques ensure that high-performing DL models remain computationally viable while maintaining predictive accuracy in real-world scenarios.

Our findings align and expand the recent studies by leveraging the ML techniques in optimising PEMWE systems to enhance design parameters, predict performance metrics, and optimise operational conditions. For instance, Zhang et al. [7] used a hybrid Genetic Algorithm-Backpropagation neural network with PSO to optimise PEM fuel cells, achieving a 3.3 % increase in power density. While their study focused on PEM fuel cells, our work extends this approach to PEMWE systems, demonstrating the broader applicability of data-driven methods. Similarly, Dinçer et al. [8] introduced a hybrid Q-learning and molecular fuzzy-based model to optimise water electrolysis for green H_2 production, identifying electrolyser lifespan and production capacity as critical factors. Our study builds on these insights by providing a robust framework for predicting H_2 production under diverse operational conditions. However, the underperformance of more complex DL models like LSTM raises important questions about their applicability in non-sequential, structured data environments. This finding contrasts with the work of [46], who emphasised the role of AI in enhancing efficiency and predicting long-term performance in high-pressure electrolysis systems. Our results suggest that while LSTM models are well-suited for time-series data, they may not be ideal for structured datasets like those used in PEMWE systems. Future work should explore alternative architectures or hybrid approaches to address this limitation.

Recent studies have also explored the integration of renewable energy sources with PEMWE systems, further validating the potential of ML/DL models in this domain. For instance, [47] demonstrated the effectiveness of DL models in estimating H_2 yield for solar-powered PEMWE systems, achieving high predictive accuracy. While their study focused on solar energy, our work extends this approach by evaluating a broader range of operational conditions, demonstrating the superior predictive capability of our 1DCNN model in capturing complex non-linear relationships. This highlights the versatility of data-driven approaches in optimising PEMWE systems under diverse energy inputs. Similarly, Urhan et al. [10] developed an ML-based approach for H_2 production using PEM electrolysers integrated with solar and wind energy systems. Their study identified an optimal system configuration that maximised green H_2 production, demonstrating the feasibility of renewable energy integration in H_2 production. This provides a practical example of how our 1DCNN model could be applied to optimise such systems. By leveraging the predictive accuracy and computational efficiency of 1DCNN, future studies could further enhance the design and operation of renewable energy driven PEMWE systems, particularly in scenarios where energy inputs are variable or intermittent. In addition to renewable energy integration, recent advancements in AI-driven optimisation have shown promising results. Ali Rehman et al. [17] introduced an AI-based surrogate model to optimise H_2 liquefaction processes, achieving significant improvements in prediction accuracy and computational efficiency. Their work underscores the potential of AI-driven techniques to streamline complex processes, which aligns with our findings on the computational efficiency of the 1DCNN model. By adopting similar approaches, future studies could further enhance the optimisation of PEMWE systems, particularly in large-scale industrial applications where computational efficiency is critical.

Previous studies such as, Mohamed et al. [5] evaluated multiple ML approaches including ANN, Polynomial Regression, SVM, k-NN, and DT for predicting H_2 production and cell current density. They focused on predicting optimal PEM electrolyser cells design parameters using polynomial and logistic regression models, validating the results through experimental testing. Similarly, Rui et al. [45] extended this work by applying k-NN and DTR to predict design parameters for commercial-scale applications, contributing to reduced development time and costs. Further advancements include the implementation of advanced optimisation techniques and hybrid approaches. Purnami et al. [11] demonstrated the use of adaptive systems for real-time optimisation in magnetic field-assisted electrolysis, while Rezk et al. [12] and Bensmann et al. [13] explored physics-informed neural networks. For instance, Tawalbeh et al. [14] and Biswas et al. [15] focused on using ANN with Levenberg-Marquardt backpropagation to predict H_2 production rates and flow characteristics. Chen et al. [16] developed a Knowledge-Integrated ML framework that enhances model robustness through domain knowledge integration. [49] focused on membrane electrode assembly optimisation using GB models and interpretation methods like shapely addictive explanations. Arjmandi et al. [9] investigated anode-side parameter prediction using various ML models.

This study highlights the effectiveness of the 1DCNN model for PEMWE modelling, but it is important to recognise potential limitations that could impact its performance. One key limitation is related to the dataset size. Although the model was validated using two independent datasets, the limited availability of large and diverse datasets restricts the model's ability to generalise across a broader range of operating conditions a common challenge in data driven PEMWE studies. Expanding the dataset to include more experimental data from different PEMWE configurations would enhance the model's robustness and adaptability. Furthermore, the current approach relies solely on data-driven feature relationships without explicitly incorporating physical constraints. These limitations have motivated our ongoing efforts to develop a physics-informed that integrates fundamental electrochemical equations while maintaining the computational efficiency of the existing architecture.

Despite these limitations, challenges persist in applying ML to PEMWE systems. Our AI surrogate model, the 1DCNN, achieves superior predictive accuracy and demonstrates robust performance across varying operating conditions. It surpasses traditional models like ANN and SVM, which often rely on smaller and less diverse datasets as mentioned in Table 8. However, the limited availability of comprehensive datasets remains a critical issue, potentially affecting the generalisability of developed models. Integrating physics-based knowledge with ML models has emerged as a promising approach to enhance prediction accuracy and improve model reliability.

5. Conclusion

In this study, we developed an AI-driven surrogate model to simulate and optimise H_2 production in PEMWE systems, utilising advanced ML techniques such as GB, CNN, LSTM, and MLP models. To ensure reliable performance comparisons, thorough analyses were carried out utilising important statistical metrics such as R^2, MSE, NMSE, MAE, and Pearson correlation. The research focused on predicting H_2 production from PEMWE systems by leveraging datasets sourced from experimental studies. Hyperparameter tuning was applied to optimise the performance of each model. The effectiveness of the models was further validated through computational metrics and cross-validation techniques, ensuring their generalisability to unseen data.

Based on the proposed methodology and the results obtained, the following key conclusions can be drawn.

1. The 1DCNN model demonstrated the highest performance across all evaluation metrics, including R^2, MSE, and Pearson correlation. Its superior ability to capture complex non-linear relationships between the input features and H_2 production made it the most suitable for this task.
2. DL models such as LSTM, exhibited varying low performance levels. Even though, MLP performed well, effectively capturing spatial dependencies, and achieving competitive accuracy. However, LSTM, despite its design for temporal relationships, struggled to match the accuracy and stability, particularly in datasets without significant temporal patterns.
3. Ensemble and traditional ML methods such as RF, GB SVM, and k-NN performed well, demonstrating strong generalisation capabilities. However, they fell short of 1DCNN in terms of predictive accuracy and robustness, as indicated by both evaluation metrics, cross-validation performance, and the Wilcoxon test, since 1DCNN consistently outperformed them in every evaluation task.
4. The AI-surrogate model developed in this study demonstrated its capacity to predict H_2 production rates with high precision, robustness, and minimal computational resources.
5. SHAP analysis was applied to enhance model interpretability, identifying key input parameters that significantly influence H_2 production efficiency. The analysis revealed that power (W) was the most critical factor, followed by water flow rate (ml/min) and anode flow area (cm^2). These insights highlight the role of power input and reactant flow rates in determining PEMWE performance, supporting better system optimisation and decision-making.

This study provides a robust AI-driven framework for predicting H_2 production in PEMWE systems. However, several challenges must be addressed for real-world deployment. One critical concern is data privacy, particularly in industrial applications where proprietary operational data may restrict model accessibility. Implementing privacy-preserving techniques, such as federated learning or secure multi-party computation, could enable collaborative model training while safeguarding sensitive data. Since our model was trained on experimental datasets from two different sources, its robustness in real-world settings requires further validation using operational PEMWE data collected from industrial-scale electrolysers. This will ensure the model's predictive capability remains consistent across varying system configurations, operating conditions, and potential uncertainties inherent in large-scale production environments. Hardware and deployment constraints should also be considered. While the 1DCNN model demonstrated high accuracy and computational efficiency, real-time inference in industrial systems may necessitate edge computing solutions for on-site processing or cloud-based architectures for scalability. The system integration remains a challenge, as most existing PEMWE setups lack direct AI compatibility. Developing standardised interfacing protocols for AI-driven process control and optimisation could facilitate smoother deployment in industrial automation frameworks. Beyond PEMWE applications, the proposed AI framework can be extended to other domains. Similar methodologies can be employed in fuel cell performance optimisation, battery health prediction, and biohydrogen production systems. For example, in fuel cell systems, similar AI frameworks have been used to optimise power density and efficiency, as demonstrated by Zhang et al. [7]. By leveraging transfer learning, pre-trained models can be fine-tuned on smaller datasets specific to new domains, reducing the need for extensive retraining and computational resources. The future research should explore the integration of real-time monitoring data to further improve predictive accuracy and enable adaptive control in PEMWE systems. This could involve sensor-driven AI frameworks, where continuous operational data is used to dynamically adjust parameters, improving system efficiency in real-time. Additionally, the development of hybrid AI-physics models will be crucial. Future studies should explore physics-informed ML approaches that integrate data-driven ML techniques with electrochemical models or Computational Fluid Dynamics-based simulations. This integration will enhance model interpretability, ensuring that AI-driven predictions remain consistent with fundamental electrochemical and thermodynamic principles. Furthermore, optimisation strategies incorporating advanced algorithms such as PSO, Genetic Algorithms, Bayesian Optimisation, and Reinforcement Learning will be investigated. These techniques can be leveraged to refine hyperparameters, optimise operational conditions, and improve surrogate model efficiency for real-world deployment in industrial-scale PEM electrolysis systems. These future directions will further enhance the practical applicability of AI-driven models for H_2 production, supporting the wider adoption of PEMWE technology in sustainable energy systems. The compact and efficient design of the 1DCNN model makes it highly suitable for deployment on edge devices, enabling real-time H_2 production monitoring in industrial-scale PEMWE systems.

In conclusion, the AI-driven surrogate model, especially the 1DCNN technique, has proven to be a powerful tool for predicting and optimising H_2 production in PEMWE systems. Among the 10 models assessed in this study which included ensemble approaches and DL architectures including MLP, 1DCNN, and LSTM, the 1DCNN continuously outperformed the others in each evaluation metrics. The results of this study demonstrate the efficacy of 1DCNN through cross-validation using k folds in addition to conventional assessment criteria and it also passed the Wilcoxon Signed-Rank Test with any other models. This study makes a substantial contribution to the field of H_2 production by providing a scalable, effective way to lower computational costs while increasing accuracy.

CRediT authorship contribution statement

Mohammad Abdul Baseer: Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Harjeet Singh:** Visualization, Software, Resources, Methodology, Formal analysis, Data curation. **Prashant Kumar:** Writing – review & editing, Supervision, Resources, Project administration, Methodology. **Erick Giovani Sperandio Nascimento:** Writing – original draft, Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The authors express their sincere appreciation to the Surrey Institute

for People-Centred Artificial Intelligence (PAI) and the Global Centre for Clean Air Research (GCARE) at the University of Surrey, United Kingdom, for their valuable support and resources. Special thanks to Amira Mohamed and Rui Yang for providing the real-time dataset for production of H_2 via PEMWE. We also thank the National Council for Scientific and Technological Development (CNPq, Brazil) for their support, as Erick G. Sperandio Nascimento is a CNPq technological development fellow (Proc. 308963/2022–9). PK acknowledges the support received through the RECLAIM Network Plus (EP/W034034/1) and GP4Streets (APP44894) projects.

References

[1] Wang T, Cao X, Jiao L. PEM water electrolysis for hydrogen production: fundamentals, advances, and prospects. Carb Neutrality 2022;1:21. https://doi.org/10.1007/s43979-022-00022-8.
[2] Li X, Yao Y, Tian Y, Jia J, Ma W, Yan X, Liang J. Recent advances in key components of proton exchange membrane water electrolysers. Mater Chem Front 2024;8:2493–510. https://doi.org/10.1039/d4qm00086b.
[3] Hayatzadeh A, Fattahi M, Rezaveisi A. Machine learning algorithms for operating parameters predictions in proton exchange membrane water electrolyzers: anode side catalyst. Int J Hydrogen Energy 2024;56:302–14. https://doi.org/10.1016/j.ijhydene.2023.12.149.
[4] Mao J, Li Z, Xuan J, Du X, Ni M, Xing L. A review of control strategies for proton exchange membrane (PEM) fuel cells and water electrolysers: from automation to autonomy. Energy and AI 2024;17:100406. https://doi.org/10.1016/j.egyai.2024.100406.
[5] Mohamed A, Ibrahem H, Kim K. Machine learning-based simulation for proton exchange membrane electrolyzer cell. Energy Rep 2022;8:13425–37. https://doi.org/10.1016/j.egyr.2022.09.135.
[6] Mohamed A, Ibrahem H, Yang R, Kim K. Optimization of proton exchange membrane electrolyzer cell design using machine learning. Energies 2022;15:6657. https://doi.org/10.3390/en15186657.
[7] Zhang N, Wang H, Chen W, Zhou H, Meng K, Chen B. Performance prediction and operating parameters optimization for proton exchange membrane fuel cell based on data-driven surrogate model and particle swarm optimization. Int J Hydrogen Energy 2024;69:493–503. https://doi.org/10.1016/j.ijhydene.2024.05.051. ISSN 0360-3199.
[8] Dinçer H, Eti S, Acar M, Yüksel S. Assessment of water electrolysis projects for green hydrogen production with a novel hybrid Q-learning algorithm and molecular fuzzy-based modelling. Int J Hydrogen Energy 2024;95:721–33. https://doi.org/10.1016/j.ijhydene.2024.11.262. ISSN 0360-3199.
[9] Arjmandi M, Fattahi M, Motevassel M, Rezaveisi H. Evaluating algorithms of decision tree, support vector machine, and regression for anode side catalyst data in proton exchange membrane water electrolysis. Sci Rep 2023;13(1):20309. https://doi.org/10.1038/s41598-023-47174-w.
[10] Urhan B, Erdoğmuş A, Dokuz A, Gökçek M. Predicting green hydrogen production using electrolyzers driven by photovoltaic panels and wind turbines based on machine learning techniques: a pathway to on-site hydrogen refuelling stations. Int J Hydrogen Energy 2025;101:1421–38. https://doi.org/10.1016/j.ijhydene.2025.01.017. ISSN 0360-3199.
[11] Purnami P, Nugroho WS, Hamidi N, Winarto W, Schulze AA, Wardana ING. Double deep Q network intelligent adaptive control for highly efficient dynamic magnetic field-assisted water electrolysis. Int J Hydrogen Energy 2024;59:457–64. https://doi.org/10.1016/j.ijhydene.2024.01.321.
[12] Rezk H, Olabi AG, Abdelkareem MA, Alahmer A, Sayed ET. Maximizing green hydrogen production from water electrocatalysis: modeling and optimization. J Mar Sci Eng 2023;11:617. https://doi.org/10.3390/jmse11030617.
[13] Bensmann B, Rex A, Hanke-Rauschenbach R. An engineering perspective on the future role of modelling in proton exchange membrane water electrolysis development. Current Opinion in Chemical Engineering 2022;36:100829. https://doi.org/10.1016/j.coche.2022.100829.
[14] Tawalbeh M, Shomope I, Al-Othman A, Alshraideh H. Prediction of hydrogen production in proton exchange membrane water electrolysis via neural networks. International Journal of Thermofluids 2024;24:100849. https://doi.org/10.1016/j.ijft.2024.100849.
[15] Biswas M, Wilberforce T, Biswas MA. Prediction of transient hydrogen flow of proton exchange membrane electrolyzer using artificial neural network. Hydro 2023;4:542–55. https://doi.org/10.3390/hydrogen4030035.
[16] Chen X, Rex Alexander, Woelke Janis, Eckert Christoph, Bensmann Boris, Hanke-Rauschenbach Richard, Geyer Philipp. Machine learning in proton exchange membrane water electrolysis — a knowledge-integrated framework. Appl Energy 2024;371:123550. https://doi.org/10.1016/j.apenergy.2024.123550. ISSN 0306-2619.
[17] Rehman A, Zhang B, Riaz A, Qadeer K, Min S, Ahmad A, Zakir F, Ismail Mohamed A, Lee Moonyong. Artificial intelligence-based surrogate modeling for computational cost-effective optimization of hydrogen liquefaction process. Int J Hydrogen Energy 2024. https://doi.org/10.1016/j.ijhydene.2024.04.331. ISSN 0360-3199.
[18] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.
[19] Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng May-June 2007;9(3):90–5. https://doi.org/10.1109/MCSE.2007.55.
[20] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. TensorFlow: a system for large-scale machine learning. arXiv preprint arXiv:1605.08695 2016. https://doi.org/10.48550/arXiv.1605.08695.
[21] Chollet F. Keras: deep learning for humans. https://github.com/fchollet/keras; 2015.
[22] Kim M-J, Yun JP, Yang J-B-R, Choi S-J, Kim D. Prediction of the temperature of liquid aluminum and the dissolved hydrogen content in liquid aluminum with a machine learning approach. Metals 2020;10:330. https://doi.org/10.3390/met10030330.
[23] Bishop CM. Pattern recognition and machine learning. Springer; 2006. https://link.springer.com/book/9780387310732.
[24] Sarailidis G, Wagener T, Pianosi F. Integrating scientific knowledge into machine learning using interactive decision trees. Comput Geosci 2023;170:105248. https://doi.org/10.1016/j.cageo.2022.105248.
[25] Blockeel H, Devos L, Frénay B, Nanfack G, Nijssen S. Decision trees: from efficient prediction to responsible AI. Frontiers in Artificial Intelligence 2023;6:1124553. https://doi.org/10.3389/frai.2023.1124553.
[26] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: a highly efficient gradient boosting decision tree. Neural Information Processing Systems 2017. https://doi.org/10.5555/3294996.3295074.
[27] Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. https://doi.org/10.48550/arXiv.1810.11363; 2018.
[28] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324. https://doi.org/10.1109/5.726791.
[29] Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network: a deep learning approach. Procedia Comput Sci 2018;132:679–88. https://doi.org/10.1016/j.procs.2018.05.069.
[30] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
[31] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016. ISBN: 9780262035613.
[32] Kingma DP, Ba JL. Adam: a method for stochastic optimization. Proceedings of the 3rd international conference on learning representations (ICLR). 2015. https://doi.org/10.48550/arXiv.1412.6980.
[33] Bakır R, Orak C, Yüksel A. Optimizing hydrogen evolution prediction: a unified approach using random forests, LightGBM, and Bagging Regressor ensemble model. Int J Hydrogen Energy 2024;67:101–10. https://doi.org/10.1016/j.ijhydene.2024.04.173.
[34] Galvão SLJ, Matos JCO, Kitagawa YKL, Conterato FS, Moreira DM, Kumar P, Nascimento EGS. Particulate matter forecasting using different deep neural network topologies and wavelets for feature augmentation. Atmosphere 2022;13:1451. https://doi.org/10.3390/atmos13091451.
[35] Nascimento EGS, Melo TAC, Moreira DM. A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy. Energy 2023;278:127678. https://doi.org/10.1016/j.energy.2023.127678.
[36] Srivastava A, Wang R, Dinda SK, Chattopadhyay K. Ensemble prediction of mean bubble size in a continuous casting mold using data-driven modeling techniques. Machine Learning with Applications 2021;6:100180. https://doi.org/10.1016/j.mlwa.2021.100180.
[37] Souza RR, Toebe M, Mello AC, Bittencourt KC. Sample size and Shapiro-Wilk test: an analysis for soybean grain yield. Eur J Agron 2023;142:126666. https://doi.org/10.1016/j.eja.2022.126666.
[38] Biau G, Cadre B. Optimization by gradient boosting. In: Daouia A, Ruiz-Gazen A, editors. Advances in contemporary statistics and econometrics. Cham: Springer; 2021. p. 35–54. https://doi.org/10.1007/978-3-030-73249-3_2.
[39] Derwent RG. Global warming potential (GWP) for hydrogen: sensitivities, uncertainties and meta-analysis. Int J Hydrogen Energy 2023;48(22):8328–41. https://doi.org/10.1016/j.ijhydene.2022.11.219.
[40] Van der Spek M, Banet C, Bauer C, Gabrielli P, Goldthorpe W, Mazzotti M, Munkejord ST, Røkke NA, Shah N, Sunny N, Sutter D, Trusler JM, Gazzani M. Perspective on the hydrogen economy as a pathway to reach net-zero CO2 emissions in Europe. Energy Environ Sci 2022;15(3):1034–77. https://doi.org/10.1039/d1ee02118d.
[41] Kumar S, Lim H. An overview of water electrolysis technologies for green hydrogen production. Energy Rep 2022;8:13793–813. https://doi.org/10.1016/j.egyr.2022.10.127. ISSN 2352-4847.
[42] Chen Z, Wei Wei. Bing-Jie Ni, Cost-effective catalysts for renewable hydrogen production via electrochemical water splitting: recent advances. Curr Opin Green Sustainable Chem 2021;27:100398. https://doi.org/10.1016/j.cogsc.2020.100398. ISSN 2452-2236.
[43] Indrajith B, Gunawardane K. Performance analysis of proton exchange membrane (PEM) electrolyser-fuel cell setup using Simulink model. Auckland, New Zealand: IEEE Fifth International Conference on DC Microgrids (ICDCM); 2023. p. 1–6. https://doi.org/10.1109/ICDCM54452.2023.10433593. 2023.
[44] Wan L, Butterworth P. Energy from green hydrogen will be expensive, even in 2050. CRU Group; 2024. Retrieved from, https://sustainability.crugroup.com/article/energy-from-green-hydrogen-will-be-expensive-even-in-2050.
[45] Yang Rui, Mohamed A, Kim K. Optimal design and flow-field pattern selection of proton exchange membrane electrolyzers using artificial intelligence. Energy 2023;264:126135. https://doi.org/10.1016/j.energy.2022.126135.

[46] Höglinger M, Kartusch S, Eder J, Grabner B, Macherhammer M, Alexander T. Advanced testing methods for proton exchange membrane electrolysis stacks. Int J Hydrogen Energy 2024;77:598–611. https://doi.org/10.1016/j.ijhydene.2024.06.118. ISSN 0360-3199.
[47] Mert İ. Agnostic deep neural network approach to the estimation of hydrogen production for solar-powered systems. Int J Hydrogen Energy 2021;46(9):6272–85. https://doi.org/10.1016/j.ijhydene.2020.11.161. ISSN 0360-3199.
[48] McKinney W. Pandas: a foundational Python library for data analysis and statistics. 2011.
[49] Ding Rui, Chen Y, Rui Z, Hua K, Wu Y, Li X, Duan X, Wang X, Li J, Liu J. Guiding the optimization of membrane electrode assembly in a proton exchange membrane water electrolyzer by machine learning modeling and black-box interpretation. ACS Sustainable Chem Eng 2022;10(14):4561–78. https://doi.org/10.1021/acssuschemeng.1c08522.
[50] Meng F, Wang Y. Transformers: statistical interpretation, architectures, and applications. Tech 2023. https://doi.org/10.36227/techrxiv.24638811.v1.
[51] Ige AO, Sibiya M. State-of-the-art in 1D convolutional neural networks: a survey. IEEE Access 2024;12:144082–105. https://doi.org/10.1109/ACCESS.2024.3433513.
[52] Zhao L, Zhang Z. A improved pooling method for convolutional neural networks. Sci Rep 2024;14:1589. https://doi.org/10.1038/s41598-024-51258-6.
[53] Ates EB, Calik E. Public awareness of hydrogen energy: a comprehensive evaluation based on statistical approach. Int J Hydrogen Energy 2023;48(24):8756–67. https://doi.org/10.1016/j.ijhydene.2022.12.070.
[54] Gabbar Hossam A, Hussain Sajid, Hossein Hosseini Amir. Simulation-based fault propagation analysis Application on hydrogen production plant. Process Saf Environ Prot 2014;92(6):723–31. https://doi.org/10.1016/j.psep.2013.12.006. ISSN 0957-5820.