

## Article

# Toward a Digital Twin of a Solid Oxide Fuel Cell Microcogenerator: Data-Driven Modelling

Tancredi Testasecca <sup>1</sup>, Manfredi Picciotto Maniscalco <sup>2,\*</sup>, Giovanni Brunaccini <sup>2</sup>, Girolama Airò Farulla <sup>3</sup>,  
Giuseppina Ciulla <sup>1</sup>, Marco Beccali <sup>1</sup> and Marco Ferraro <sup>2</sup>

<sup>1</sup> Department of Engineering, Università degli Studi di Palermo, 90128 Palermo, Italy; tancredi.testasecca@unipa.it (T.T.); giuseppina.ciulla@unipa.it (G.C.); marco.beccali@unipa.it (M.B.)

<sup>2</sup> CNR-ITAE: Istituto di Tecnologie Avanzate per l'Energia "Nicola Giordano", 90128 Palermo, Italy; giovanni.brunaccini@itae.cnr.it (G.B.); ferraro@itae.cnr.it (M.F.)

<sup>3</sup> CNR-INM: Consiglio Nazionale delle Ricerche—Istituto di Ingegneria del Mare, 90146 Palermo, Italy; girolama.airofarulla@cnr.it

\* Correspondence: maniscalco@itae.cnr.it

**Abstract:** Solid oxide fuel cells (SOFC) could facilitate the green energy transition as they can produce high-temperature heat and electricity while emitting only water when supplied with hydrogen. Additionally, when operated with natural gas, these systems demonstrate higher thermoelectric efficiency compared to traditional microturbines or alternative engines. Within this context, although digitalisation has facilitated the acquisition of extensive data for precise modelling and optimal management of fuel cells, there remains a significant gap in developing digital twins that effectively achieve these objectives in real-world applications. Existing research predominantly focuses on the use of machine learning algorithms to predict the degradation of fuel cell components and to optimally design and theoretically operate these systems. In light of this, the presented study focuses on developing digital twin-oriented models that predict the efficiency of a commercial gas-fed solid oxide fuel cell under various operational conditions. This study uses data gathered from an experimental setup, which was employed to train various machine learning models, including artificial neural networks, random forests, and gradient boosting regressors. Preliminary findings demonstrate that the random forest model excels, achieving an  $R^2$  score exceeding 0.98 and a mean squared error of 0.14 in estimating electric efficiency. These outcomes could validate the potential of machine learning algorithms to support fuel cell integration into energy management systems capable of improving efficiency, pushing the transition towards sustainable energy solutions.

**Keywords:** digital twin; energy; solid oxide fuel cell; machine learning; hydrogen



**Citation:** Testasecca, T.; Maniscalco, M.P.; Brunaccini, G.; Airò Farulla, G.; Ciulla, G.; Beccali, M.; Ferraro, M. Toward a Digital Twin of a Solid Oxide Fuel Cell Microcogenerator: Data-Driven Modelling. *Energies* **2024**, *17*, 4140. <https://doi.org/10.3390/en17164140>

Academic Editor: Vladislav A. Sadykov

Received: 17 July 2024

Revised: 7 August 2024

Accepted: 11 August 2024

Published: 20 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The continued rise in carbon emissions, marked by a +1.1% increase in 2023, underscores the ongoing need for intensified efforts to reverse this global trend and achieve the goals set forth in the Paris Agreement [1]. While the expanding deployment of renewable energy capacity worldwide has partially mitigated the rise in emissions, there remains a need for additional systems capable of ensuring consistent and reliable energy production irrespective of weather conditions (such as solar radiation, wind speed, or tidal flows). Consequently, attention is increasingly turning towards systems that can utilise clean fuels or achieve higher efficiencies to facilitate overall emissions reduction. Simultaneously, the green transition involves different key technologies to decarbonise and make more resilient energy production systems. On one hand, hydrogen technologies, such as fuel cells (FCs) or electrolyzers, have a pivotal role in the clean transition in different national strategies to produce, store and use clean hydrogen. On the other hand, the digitalisation wave is also permeating the energy sector, where natural gas (NG) still accounts for almost a quarter of global electricity generation. FCs emerge as a promising solution in this context

due to their high efficiency and versatility in operating with clean and renewable fuels, such as hydrogen or syngas produced from waste or biomass, as well as various hydrocarbons [2]. Among the different FC technologies, solid oxide fuel cells (SOFC) operate at the highest temperatures, which range from 600 °C to 1000 °C. This condition, on the one hand, requires a longer start-up time but, on the other hand, can ensure higher efficiency (mostly if used in a co-generative power system), higher tolerance to corrosion, and can operate with different fuels without the need for apportioning changes to the system [3]. During the last decade, particular attention has also been paid to the development of digital models that can be used to predict cell performance, component degradation, and fault diagnosis, or can act as tools to support component design (thickness, porosity etc.) [4–6]. More recently, the concept of the digital model has been extended to the “digital twin” (DT) which represents a virtual replica created using real-time data collected from sensors embedded in the physical counterpart. In such a way, DTs can bridge the gap between the physical and digital worlds [7].

Electrochemical parameter identification and optimisation are among the most studied aspects in SOFC modelling, through the application of meta-heuristic algorithms. Yang et al. [8] developed an improved genetic algorithm for a tubular dynamic SOFC model to enhance output performance based on cell electrical parameters. Gong et al. [9], proposed an adaptive differential evolution algorithm called IJADE (Improved Jingqiao Adaptive Differential Evolution) to fast and accurately identify SOFC parameters. Jiang et al. [10] used a cooperative barebone particle swarm optimisation method to solve different subproblems that regulate the SOFC model based on six internal variables. Wei and Stanford [11] proposed a novel method to evaluate with high stability and accuracy SOFC conditions for steady state and transient operations. Furthermore, in [12–14], the SOFC parameter definition was tackled through different optimisation methods such as Levenberg–Marquardt backpropagation, extreme machine learning, and the grey wolf optimisation method. Lastly, a DT model with artificial rabbits optimisation was developed by Guo et al. [7] to model a hybrid photovoltaic–SOFC system and define the internal unknown parameters.

Three further studies investigate the inner conditions of SOFC components through digital modelling. Hwang et al. [15,16] used deep learning (DL)-assisted semantic segmentation to gain insight into cathode and anode microstructure (surface and volume fractions, data on two and three-phase boundaries, etc.) starting from focused-ion beam scanning electron microscope images. A DL process was also adopted in [17] in combination with finite element methods (FEM) to foresee the elastic properties of cathode materials. Regarding system operation, a widely investigated assessment is related to fault detection and diagnosis [18–23], by using different machine learning (ML) approaches like support vector regression (SVR), artificial neural network (ANN), or deep neural network (DNN). These processes can effectively determine fault conditions like component degradation or gas leakage.

The prediction of system efficiency by modelling the polarisation curve was conducted by Subotic et al. [24] through the application of ANN. The authors trained the model with both numerical and experimental data, while for validation, only experimental values were used. Performance optimisation was based on operating temperature, fuel mix, and current density. Output parameters, in terms of voltage and thermoelectric efficiency, were also investigated in [6,25–28], including the prediction of voltage level in relation to system degradation [27]. The studies reviewed in this section provide a comprehensive overview of modelling techniques in the literature, focusing on fault detection or efficiency improvement. However, from a high-level perspective, these techniques offered a level of detail that does not significantly impact the practical management of an SOFC, particularly in commercial applications. This insight motivated the current research to estimate fuel efficiency based solely on parameters that can be controlled and adjusted during operation. In this work, the DT of the SOFC is employed to maintain an accurate ML model capable of monitoring ageing, identifying potential faults, and replicating electric efficiency as operational conditions change. The efficiency estimation, which considers electric power

and includes ramp-up and ramp-down phases for the SOFC, is based on a novel approach integrating both RF and DL algorithms. Moreover, the DT will consistently provide the most accurate models, utilising the best available data. This SOFC DT will be part of a broader framework that integrates with other energy systems, optimising both the cell and the entire energy hub, which is based on hydrogen and sustainable energy resources.

In the remaining sections of this paper, Section 2 reports the DT architecture, defining the algorithms used to model the SOFC, including both DL and ML approaches, together with all the validation metrics adopted. Section 3 first describes the experimental setup of the case study and the results from data analysis, including a description of obstacles in data collection. Finally, the models are compared, and validation scores are presented. Conclusions are then reported in Section 4 to summarise the outcomes of the paper and highlight the next steps.

## 2. SOFC Digital Twin Methodology

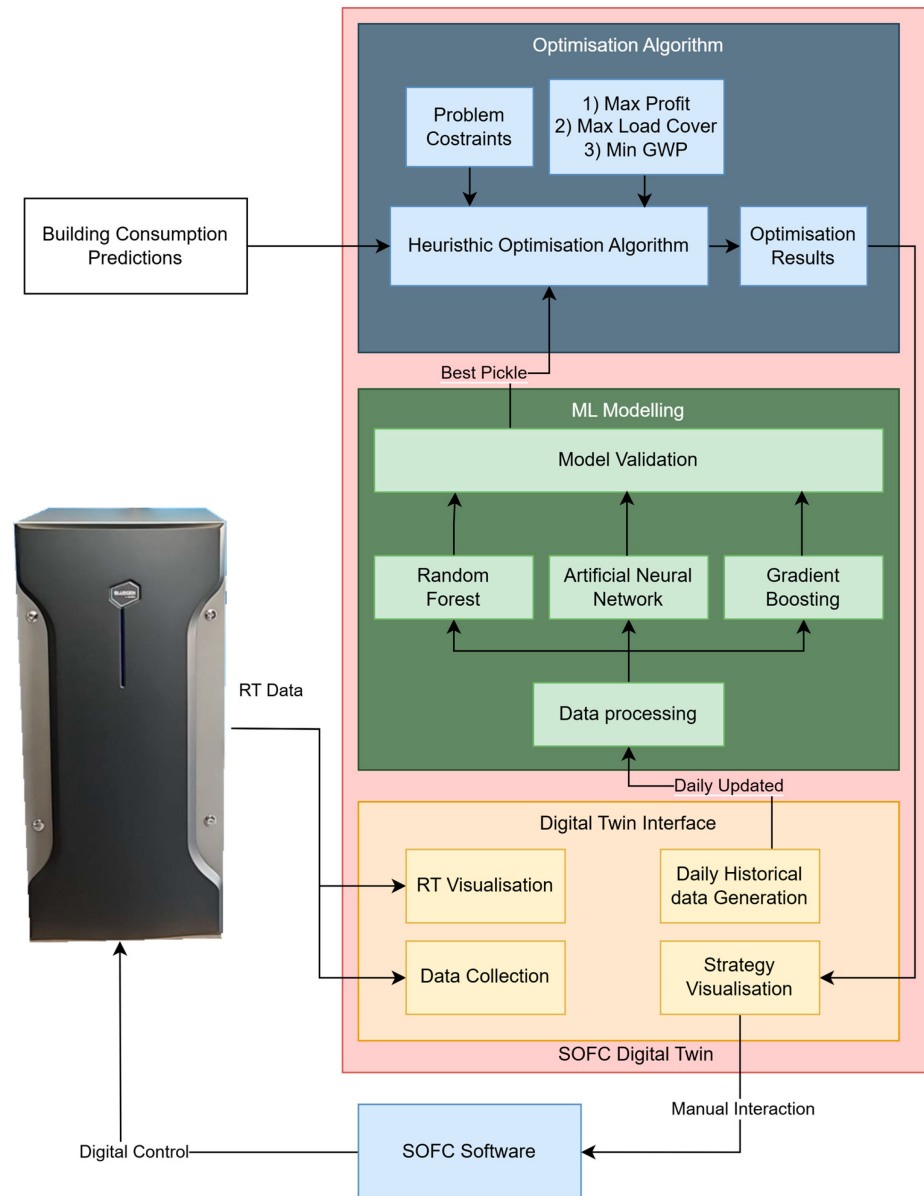
### 2.1. SOFC Digital Twin Scope and Architecture

DT technology is facilitating the transition towards greener and more digitalised energy systems. These synchronised replicas of existing facilities are used for different objectives including increasing energy efficiency [29], reducing carbon footprints [30], and optimising real-time management of buildings [31]. An example of DT architecture is presented in [32] and is composed of different horizontal and vertical layers. In general, according to the Amazon Web Service architecture for DT, the system includes visualisation and application programming interface layers, models for simulation or analytics, a platform for data storing and pre-processing and an integration layer for Internet of Things (IoT) devices, sensors, or actuators. This new technology enables the possibility of optimising the management of existing devices and components and, for this reason, the development of the SOFC DT was started. At the headquarters of CNR-ITAE in Messina, an energy hub based on hydrogen technology will be established, including an electrolyser, smart buildings, electric vehicles (EVs), PV systems, and energy storage. In this framework, the DTs of these systems will be useful to achieve optimal management to maximise carbon footprint reduction and cost savings.

In this work, the presented models are used as the model layer of the DT of a real SOFC powered by natural gas (NG). In particular, the DT being developed is composed of different blocks to produce electric energy efficiently. Focusing on the scheme available in Figure 1, which represents the architecture of the twin, it can be noted how the ML model constitutes one of the cores of the twin. Specifically, the DT comprises three primary blocks; the first is used as an interface for displaying key parameters, and for running and visualising the optimisation results. The interface will then provide users with a tool for real-time monitoring and proactive management of the cell during the optimisation process. The backend includes a second block for data pre-processing and a system for initiating Python scripts to update the ML models. Finally, the third block of the SOFC DT includes the optimisation algorithm, which is used to calculate different operating strategies to modulate the FC on the actual day. To be precise, the algorithm's three objective functions include maximising profit to assess economic benefits, maximising the load cover factor to evaluate the building's autonomy from the grid, and minimising the global warming potential (GWP) to reduce the carbon footprint of the cell. Based on these results, one for each strategy, the operator can choose which one will be visualised on the DT interface and select the correct power modulation to operate the FC. Finally, interaction with the actual solid oxide fuel cell is ensured by proprietary software that provides real-time insights related to the cell performance. This software is also used to adjust power modulation in real time and for the upcoming hours and days.

The present work focuses on the development of ML models that include regressors such as random forest (RF), ANN by multilayer perceptron (MLP) and gradient boosting (GBoost), all updated daily. This frequent updating facilitates three beneficial outcomes: firstly, the models evolve similarly to the actual SOFC, mirroring its ageing process—an

advantage of having a DT of a real system. Secondly, the models enable investigations into how the FC ages, which is crucial for identifying potential malfunctions or quality degradation. Lastly, daily updates ensure that optimisation algorithms are simulated using the most current model, which may vary as more data is accumulated. Indeed, performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and others are calculated to identify the algorithms that best represent the actual behaviour of the FC. In the future, this process could be extended by using all data available in the SOFC proprietary database to increase the prediction accuracy and develop fault detection algorithms.



**Figure 1.** SOFC DT Architecture.

## 2.2. SOFC Deep and Machine Learning Models

As previously stated, in previous works it was established that the functioning points of a SOFC can be difficult to predict. In particular, the use of white- or grey-box approaches involves the knowledge of the physical system of the SOFC, which is difficult to accurately estimate on an hourly basis. In fact, to correctly model the physical behaviour, it is necessary to consider thermodynamic parameters such as the temperature of the air or exhaust gas

and pressure of the NG inlet, as well as electrical parameters such as current and voltage inside the cells. Although these parameters can be monitored, predicting or simulating all of them to quickly calculate the system's electrical efficiency is still a challenging task.

A black-box approach, utilising validated and accurate models, allows for the estimation of one or more outputs based on correlated inputs. The primary goal of this study is to predict the electrical efficiency of the FC based on data collected from 3935 h of experimental testing. Initially, a correlation analysis was performed to identify and understand the relationships between variables. Since indoor air temperature data were unavailable at the time of writing this manuscript, the only future-known parameter (as it would be an output from the optimisation algorithm) is the desired electrical load of the SOFC. Additionally, to enhance the analysis, power increments were calculated and analysed to determine if the measured electric efficiency value was correlated with a period of power increase or decrease. To develop an accurate model capable of estimating energy efficiency based on desired electrical power, it was necessary to engage in a supervised learning framework, where the model is trained to predict specific outputs based on one or more input features. Within this framework, several RF algorithms were employed and compared, including ANNs, GBoost, eXtreme Gradient Boosting (XGBoost), RF, long short-term memory (LSTM), and polynomial regression, with the operational logic of each explained in subsequent sections.

Nevertheless, as mutual correlations may depend on the magnitude of quantities, best practice suggests scaling all the values before training the models. In this study, data was standardised using the Scikit-Learn Python library, which standardises features by removing the mean and scaling to unit variance. Although this preprocessing step enables the evaluation and enhancement of the chosen RF approach, the presence of outliers should be minimised as much as possible, as they can detrimentally affect the quality of the modelling process. Furthermore, for all models employed in this study, the dataset, which comprised over 2500 values, was partitioned into training and testing subsets. Specifically, 90% of the data was allocated for training purposes, while the remaining 10% was consecutively used for testing.

### 2.2.1. ML Algorithms

In the framework of ML algorithms, polynomial regression represents an extension of linear regression that fits a polynomial equation to the dataset. This approach allows the model to capture non-linear relationships between the predictor (independent) and response (dependent) variables, which is a significant advantage over linear models. The flexibility of the model is strictly correlated to the degree of the polynomial. Empirical studies have demonstrated that polynomial regression can effectively model curved trends in data, making it a valuable tool in various scientific and engineering fields where linear models are inadequate. The benefits of its use are related to modelling complex, non-linear relationships with very low computational effort.

RF, GBoost, and XGBoost algorithms belong all together to the family of decision trees. This supervised learning method is used to predict the output by imposing simple decision rules inferred from the input data. The rooted tree is a graph where all the branches and leaves start from a defined origin. Precisely, the input space is divided into smaller portions of the space based on the different hypotheses carried out by the model. GBoost is an ML technique first introduced in 1999 and used for both regressions and classifications [33]. According to [33], GBoost was defined as “an efficient algorithm for converting relatively poor hypotheses into very good hypotheses”. The model, first composed of a weak hypothesis, the first group of decision trees, is updated with additional decision trees capable of reducing model loss. The final model will be composed of an optimised sequence of decision trees capturing even unknown and complex patterns within the data. It must be noted that this algorithm tends easily to overfitting so the estimators and the random states inside the algorithm should be selected accurately. The model considered in this study includes a number of estimators (the decision trees) equal to 200, each with a



depth of 5 and minimum sample values for split and leaf set to 3 and 1, respectively. Finally, including a learning rate set to 0.1, the combination of these hyperparameters led to the most robust model.

XGBoost is one of the most used and up-to-date ML algorithms. It is the evolution of GB, where multiple ML models are combined to increase accuracy and reduce computational effort. It uses parallel and dispersed computing of decision tree and GBoost algorithms. Decision trees are built sequentially to correct errors in the previous step and, to avoid overfitting, it uses a clever penalisation of trees and a proportional shrinking of leaf nodes. In the model presented in this work, the XGBoost Python library was utilised, with 150 estimators selected, a learning rate of 0.2, and a maximum tree depth of 5. Specifically, the number of estimators corresponds to the number of trees iteratively added to the model. The learning rate acts as a shrinkage factor applied to new trees to prevent overfitting. Moreover, although increasing tree depth could enhance model precision by adding complexity to each tree, it was determined that a maximum depth of 5 was optimal for this model. Finally, this choice aligns with the recommendation in [33] to use values between 4 and 8.

In the RF algorithm, multiple decision trees are used to improve the accuracy of the model and avoid overfitting. This method, belonging to the family of ensemble learning methods, overcomes the limitations of decision tree algorithms in having low bias and high variance. In pure RF, the trees are built randomly while, to improve the quality of predictions, in the most recent versions the trees were generated based on input data. This results in a model suitable for regression tasks, even in scenarios where input data are noisy or complex. Moreover, in cases of multiple inputs, RF can be used to assess the relevance of each variable to better understand the problem [34]. In this work, 280 decision trees were utilised, each with a depth of 20, and minimum sample values for split and leaf set to 3 and 1, respectively. It is notable that the number of trees and their depths, particularly the latter, are higher than those used in GBoost and XGBoost. The RF algorithm was chosen because, in boosting processes, new trees are incrementally added, and it is preferable to have simpler individual trees to avoid certain overfitting. Conversely, in the RF algorithm, a single forest composed of diverse individual trees is used, allowing each tree to possess a more complex structure.

### 2.2.2. DL Algorithms

Among the most used classes of regression models based on DL are ANNs, which emulate the functional characteristics of human brain neurons and their synaptic connections. The simplest form of an ANN is the single-layer perceptron, which comprises only an input layer and an output layer. Focusing on research purposes, multilayer perceptrons are widely used, where, as suggested by the name, one or more hidden layers of neurons are modelled between the input and output layers to enhance the model's ability to capture the underlying patterns of the data points it is trained to estimate. In this work, an MLP neural network is structured with six hidden layers, respectively, composed of 10, 52, 50, 100, and 10 neurons. An iterative process of finding different layers was carried out to find the correct values of hyperparameters, which in ANN represent the number of hidden layers and neurons for layers. This process was necessary to finetune the network to maximise the accuracy of the model while reducing the time spent in the iterative process. Additionally, since the internal optimisation process of DL algorithms is affected by the randomness of parameters, different processes were carried out for different random states. Despite this variability, the training consistently converged before reaching 250 epochs, always stopping when the training loss failed to improve for 10 consecutive epochs.

Long short-term memory (LSTM) networks, an advanced variant of recurrent neural networks (RNN), are used to individuate long-term dependencies, a capability encountered across various fields such as thermal comfort or energy demand, and predictive maintenance. The distinctive architecture of LSTM includes memory cells that replace the standard neuron units in the hidden layer of RNNs. Given its capability to handle

long- and short-term dependencies, LSTM is particularly suited for tasks requiring the prediction of time series data. In this study, although the authors were not aware of any existing recurrent dependencies of variables, the LSTM model was selected to investigate the possibility of detecting them. The model architecture, developed using Keras, included an LSTM layer followed by a dense layer with one unit. The LSTM layer includes 100 units (or neurons), iteratively selected to achieve the most accurate model. Secondly, the dense layer is set equal to 1, as required for the regression process of time series value.

### 2.3. Validation Methods

Once the models were defined, to verify the accuracy of each model and its robustness, the parameters for evaluating it must be selected. In this study, it was chosen to consider the following parameters:

- Mean absolute error (MAE)
- Mean absolute percentage error (MAPE)
- Mean squared error (MSE)
- Root mean square error (RMSE)
- $R^2$  Score
- Weighted score (WS)

The MAE is used to measure the average of the absolute errors, providing a linear score without exaggerating outliers. In particular, it measures the average absolute differences between actual and predicted values.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

The MSE measures the average of the squares of the errors, giving more weight to larger errors.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

The MAPE then calculates the average absolute percentage difference between actual and predicted values.

$$\text{MAPE} = \frac{\sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} \quad (3)$$

The RMSE is the square root of the MSE, providing errors in the same units as the target variable, which is useful for interpretability.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

$R^2$  indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

In the current literature, the use of a WS may be suggested when dealing with different metrics and models; however, guidelines for choosing weights for these metrics are not clarified. Therefore, in this work, WS is calculated as the mean value of  $R^2$  and MAPE, as these are the only metrics that can be expressed in a comparable score (Equation (6)). This unified score allows the DT to autonomously evaluate and select the best model.

$$\text{WS} = \frac{[(1 - R^2) + \text{MAPE}]}{2} \quad (6)$$

In these equations:

- $n$  is the total number of observations or data points.
- $y_i$  represents the actual observed values in the dataset.
- $\hat{y}_i$  represents the predicted values corresponding to  $y_i$ .
- $\bar{y}$  is the mean of the observed values ( $y_i$ ).

### 3. Discussion and Results

#### 3.1. Experimental Setup and Data Analysis

The SOFC generator comprises a hot box containing an SOFC stack, a reformer, a burner, a heat recuperator, and a condenser to recover the water for the reforming process. A control and diagnosis board, thanks to numerous sensors that constantly provide readouts of all operational parameters necessary for operations and safety, manages the functions of ancillaries (Balance of Plant) [35]. The experimental setup of the SOFC installed at the CNR-ITAE facilities in Messina, Italy can be seen in Figure 2. It is observable that the SOFC is directly connected to the gas network via a yellow valve, enabling a direct connection to the national NG grid and the potential for integrating a blended mixture of NG and hydrogen. In addition to the electrical connection, this SOFC also can produce heat with a nominal thermal output of 0.75 kW. Accordingly, connections for both inlet and outlet water flow rates are installed, ensuring the proper operation of the cell.



**Figure 2.** SOFC Experimental Setup.

For the sake of completeness, the main operational parameters, according to the producer datasheet, are reported in Table 1.

The SOFC in this study allowed for the collection of cumulative and instantaneous data on an hourly basis. The cumulative outputs included hours of operation, electric power produced, and CO<sub>2</sub> avoided, while the instantaneous data included the SOFC status, electric power, inlet gas flow rate, inlet gas power, electric efficiency, and CO<sub>2</sub> emissions. The status specifically indicated whether the cell was in “Heat Up”, “Power Export”, “Cool Down”, or “Off”. An initial data overview indicated that efficiency depended on whether the FC was increasing or decreasing its power, or operating in a steady state. Therefore, the difference between power at two consecutive timesteps was calculated to distinguish

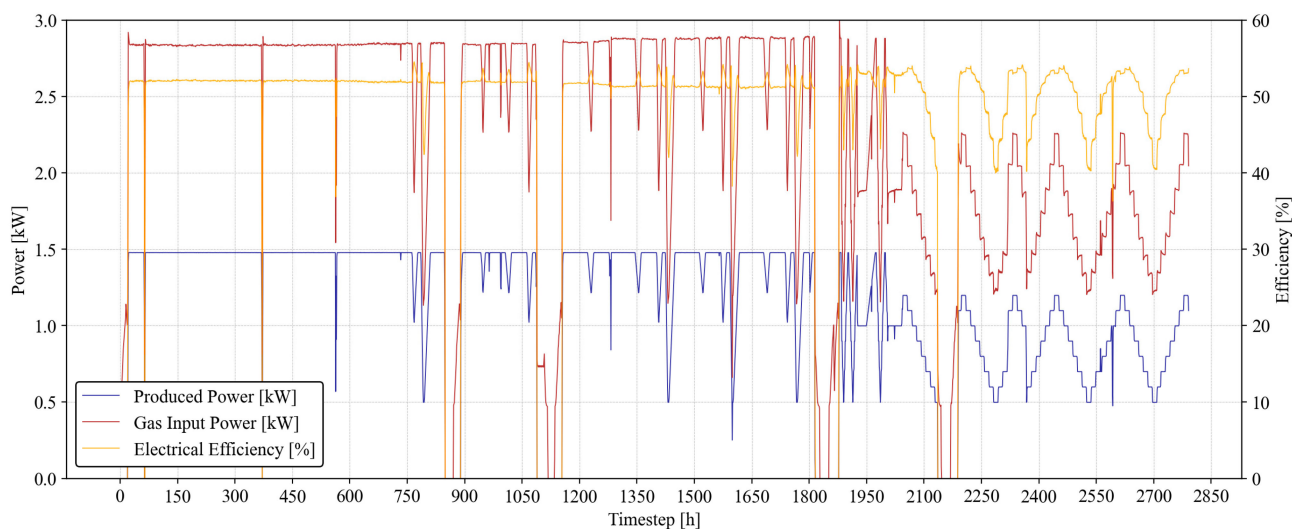


moments of ramp-up, ramp-down, or constant power. The status was updated with this information, generating three distinct time series for power ramping up, ramping down, and remaining constant.

**Table 1.** SOFC Operational Parameters.

Parameter	Value
Nominal electric power in AC [kW]	1.3
Minimum power output in AC [kW]	0.5
Maximum power output in AC [kW]	1.5
Electric efficiency in nominal condition on NG LHV	57%
Cogeneration efficiency in nominal condition on NG LHV	88%
Nominal thermal power output [kW]	0.75
Nominal NG consumption [m <sup>3</sup> /h]	0.24
NG supply pressure [mbar]	17–25
Start-up time [h]	24

The experimental phases extended over six months, accumulating 3935 h of data, which included periods of shutdowns. In Figure 3, the power data of electrical efficiency, NG, and electric power during the periods when the SOFC was operational, which were used for training the model, are shown. Initially, the SOFC was operated at nominal power (1.5 kW) and as observed in the graph, up to hour 750, the efficiency of the FC did not vary significantly. After experimenting with a systematic ramp-up and ramp-down step-like procedure, Figure 3 shows that the FC efficiency decreases as the power production decreases.



**Figure 3.** Behaviour of Power and Efficiency in the Operational Experimental Period.

To develop models capable of predicting energy efficiency, correlations between efficiency and other variables were investigated, as illustrated in Figure 4. The second-order polynomial trendline, which fits the variation in electrical efficiency with respect to electric power, demonstrates that power increments significantly influence the SOFC's efficiency. As seen on the left, during ramp-up phases the cell always has a lower efficiency compared to the case in steady functioning, while considering ramp-down, at lower cell power, the efficiency in steady state is lower than when it operates in ramp-down. Conversely, when the SOFC is operating in power higher than 1000 W, it appears that ramp-down phases achieve lower efficiency. The correlation matrix on the right of Figure 4a reveals how the electric efficiency, as expected, could be easily correlated with produced power, and obviously from gas-related quantities such as gas input and CO<sub>2</sub> emissions, which strictly

depend on the quantity of gas used. This revealed that increasing the power should lead to higher efficiency, as also highlighted in the right graph (Figure 4b). However, it seems that the power increment is not strictly correlated to electrical efficiency, conversely to what is encountered reading the graph on the left. These reasons led to the choice of ML and DL models, which can account for non-linear relations such as XGboost, RF, or ANN.

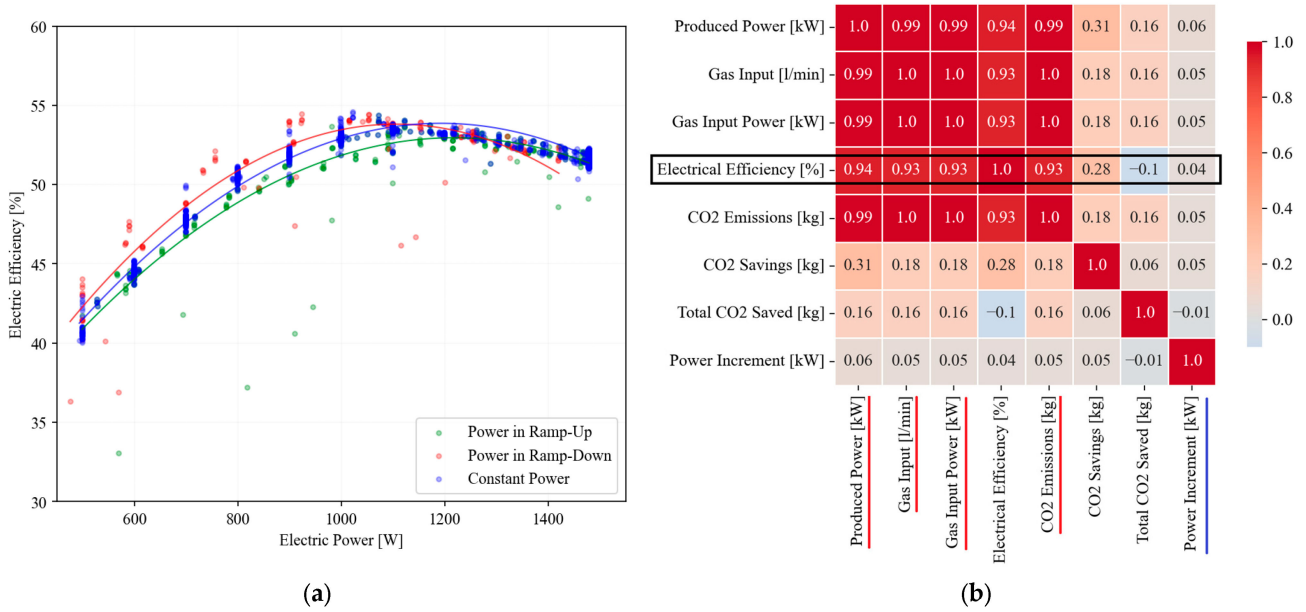


Figure 4. Polynomial Fits on Efficiency–Power Graph (a) and Correlations on Experimental Data (b).

### 3.2. Model Validation

In this work, various algorithms were examined to identify the most accurate model for estimating cell efficiency. Figure 5 depicts the model prediction curves alongside the actual data for the entire dataset. It is evident that all the models can predict efficiency behaviour smoothly, even during efficiency drops due to partialisation. These drops shown in the graph correspond to timesteps when the SOFC was turned off, resulting in very low-efficiency readings. Despite these peaks, the models generally followed the curve accurately, with some exceptions for models such as the ANN, LSTM, and polynomial regression, which displayed more “constant” behaviour compared to other models that better tracked the real data curve. Additionally, the graph shows that when the SOFC operated initially at constant power, the efficiency remained constant, and all the models replicated this behaviour.

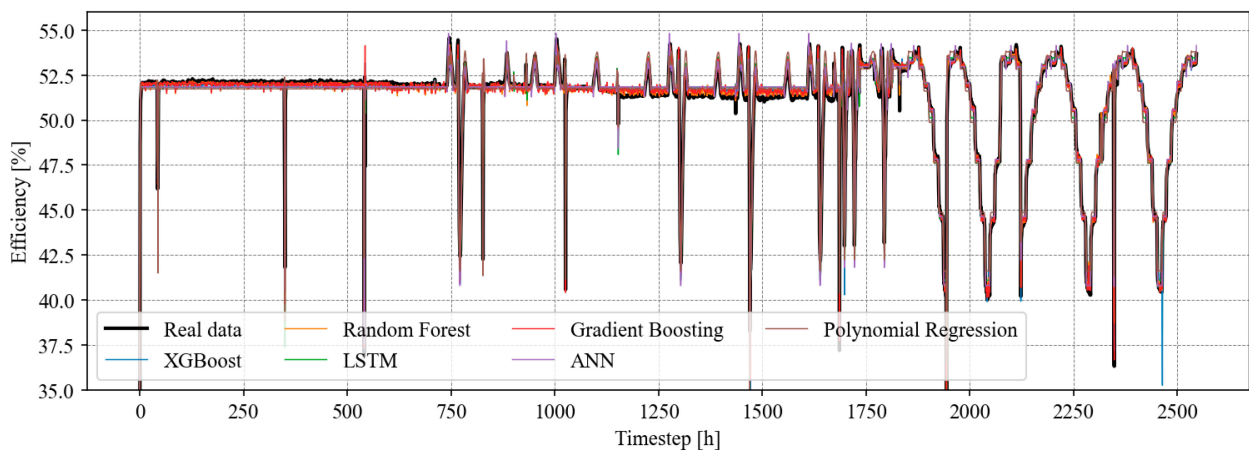
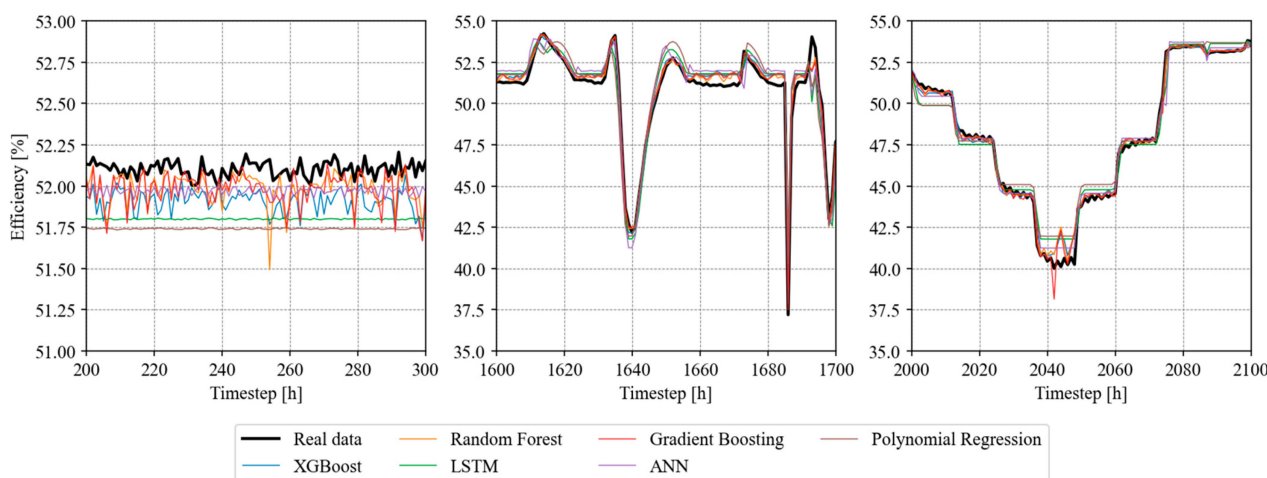


Figure 5. Models and Real Data Behaviour During the Entire Experimental Period.

However, for the sake of clearness, Figure 6 illustrates the models' performance in three different sections, from left to right: a nearly constant zone, a ramp-down section, and a highly variable situation. In the first section, it is evident that LSTM and polynomial regression failed to accurately follow the real efficiency behaviour, despite its limited variation. In contrast, the RF and GBoost models successfully replicated the efficiency trend across the 100 h considered in this time window. In the second section, from hour 1600 to 1700, all the models closely aligned with the real data; however, when the cell operated at an almost constant efficiency of about 51%, an error of 0.7% was observed for almost all the models. Nonetheless, RF, GBoost, and XGBoost remained the most accurate. Finally, in the ramp-up and ramp-down behaviours shown in the right section of Figure 6, all the models, except for polynomial regression, accurately followed the efficiency curve, with decision-tree-based models significantly outperforming the other algorithms.



**Figure 6.** Models and Real Data Behaviour in Different Time Horizons.

Table 2 summarises the scores for all the models under both test and overall conditions. Test conditions represent the validation scores during the training process, while overall scores pertain to the complete dataset. Confirming observations from the graphical comparison between the models and the actual data, XGBoost, GBoost, and RF demonstrated superior accuracy compared to the other models. Specifically, these three models achieved a minimum  $R^2$  of 0.98 for the entire dataset, and even when considering only test data, the scores reached 0.98 for RF and 0.97 for GBoost. Similar MSEs were observed for these models, with the lowest values recorded at 0.10 for the RF for the overall dataset and 0.14 during the validation phase. Regarding the MAPE, the ANN reported a value of 0.04, identical to that of GBoost, XGBoost, and RF, where the worst value was recorded for the polynomial regressor. However, in terms of the RMSE, which can be compared to the MAE due to their similar measurement units, its value was 1.7 times higher than that in the RF algorithm and 1.5 times higher than in XGBoost, which recorded an RMSE of 0.36. As anticipated, polynomial regression and the LSTM algorithm were the least accurate, achieving  $R^2$  values of 0.94 and 0.96, respectively, in the total validation. In this study, even considering the WS, tree-based models such as RF, GBoost, and XGBoost outperformed DL methods, likely due to the tree-based models' superior handling of non-linear relationships and interactions within a small dataset. DL models, while generally powerful as highlighted in the current state of the art (SOA), often require larger datasets for accurate training and may experience overfitting, as evidenced by the performance of the LSTM and ANN in this study.

The primary goal of this paper was to develop a DT capable of self-improvement by utilising the most accurate model based on the current data collected. The accuracy and selection of the best model based on the WS were assessed every 500 h of data collection, simulating an ageing DT. In the final release of the DT, models are expected to be trained almost daily to ensure the best model is available for optimisation purposes. However, for

model validation, a time frame of 500 h was deemed more appropriate. For this reason, Table 3 summarises the scores of all the models, considering WS and  $R^2$  for each time window, revealing the evolution of scores as the dataset increased. Initially, XGBoost was selected as the superior model, achieving a WS of 0.004 with an  $R^2$  of 0.998, surpassing the other models; during this time, the SOFC operated at a constant and maximum power of 1.4 kW. Further analysis indicated that between 1,000 and 2000 h of data collection, the best model was GBoost, with a WS ranging from 0.022 to 0.027 and a peak  $R^2$  of 0.97 encountered after 200 h. Finally, as anticipated from the earlier results, at 2500 h, RF once again proved to be the best model, achieving an  $R^2$  of 0.98 and a WS of 0.031. Regarding artificial neural networks, Table 3 shows that at 1500 h, the  $R^2$  value was 0.6 lower compared to the value at 1000 h. This can be explained by the transition from steady-state operation to a highly variable pattern, where the neural network encountered input and output features it could not accurately predict. However, by 2500 h, the network achieved an  $R^2$  value of 0.95. These results demonstrated how the DT framework presented in this paper can be used to develop the most accurate and “ageing” DT capable of precisely replicating SOFC behaviour across different periods. Furthermore, when used for optimisation purposes, this enables the most accurate results based on the current conditions of the FC. Finally, Table 4 summarises the main findings related to each algorithm presented in this work. It highlights the performance scores of different algorithms and the challenges in fine-tuning the models. For instance, it highlights that ML models achieved better scores and faster training times compared to DL ones.

**Table 2.** Overview of Model Score Results.

Model	MSE		MAE		RMSE		MAPE		$R^2$		WS	
	Total	Test	Total	Test	Total	Test	Total	Test	Total	Test	Total	Test
XGBoost	0.13	0.27	0.16	0.25	0.36	0.52	0.04	0.04	0.98	0.96	0.03	0.04
RF	0.10	0.14	0.17	0.24	0.31	0.38	0.04	0.04	0.99	0.98	0.03	0.03
LSTM	0.29	0.24	0.36	0.36	0.54	0.49	0.07	0.07	0.96	0.96	0.06	0.05
GBoost	0.11	0.17	0.20	0.25	0.33	0.41	0.04	0.04	0.98	0.97	0.03	0.03
ANN	0.28	0.26	0.34	0.34	0.53	0.51	0.04	0.04	0.96	0.95	0.04	0.04
Polynomial Regression	0.41	0.36	0.45	0.44	0.64	0.60	0.09	0.09	0.94	0.94	0.07	0.07

**Table 3.** Model Scores in Different Periods.

Model	500 h		1000 h		1500 h		2000 h		2500 h	
	$R^2$	WS	$R^2$	WS	$R^2$	WS	$R^2$	WS	$R^2$	WS
XGBoost	0.998	0.004	0.876	0.067	0.941	0.038	0.957	0.035	0.975	0.035
RF	0.837	0.085	0.929	0.041	0.943	0.037	0.969	0.028	0.983	0.031
LSTM	0.957	0.030	0.872	0.081	0.883	0.090	0.914	0.080	0.953	0.060
GBoost	0.938	0.034	0.966	0.022	0.958	0.029	0.971	0.027	0.974	0.036
ANN	0.845	0.081	0.961	0.025	0.892	0.062	0.911	0.057	0.950	0.047
Polynomial Regression	0.993	0.009	0.878	0.079	0.859	0.105	0.888	0.099	0.941	0.074

**Table 4.** Comparison of ML Models: Advantages, Disadvantages, and Limitations.

Model	Advantages	Disadvantages	Limitations
XGBoost	High accuracy (MAPE = 0.04) Fast training times (0.19 s)	Requires a lot of memory	Possible overfitting if not properly regularised
RF	Highest $R^2$ (0.99) and WS (0.03) Fast training times (0.54 s)	Poor performance on small dataset ( $R^2_{500\text{ h}} = 0.84$ )	Not always suitable for sequential or temporal data
LSTM	Excellent for time series	Very long training times (6.47 s)	Large amounts of data required

Table 4. Cont.

Model	Advantages	Disadvantages	Limitations
GBoost	Robustness as the number of data variates ( $R^2 > 0.93$ )	Many iterations needed to fine-tune hyperparameters	Can overfit with too many iterations
ANN	Ability to capture complex patterns	High training times (3.72 s)	Large amounts of data for training needed
Polynomial Regression	Very fast training times (0.008 s)	Not as accurate as other models	Sensitive to outliers

#### 4. Conclusions

The need to decarbonise the energy sector has never been more pressing, considering that carbon emissions are still increasing globally and the effects of climate change, especially in Europe, are spreading. To meet the goals of the Paris Agreement, the shift toward sustainable energies and technologies must be expedited, and hydrogen technologies such as SOFCs offer remarkable potential. SOFCs allow for high thermoelectric efficiency when supplied by hydrogen and offer the advantages of producing both heat and electricity with zero emissions; additionally, a CO<sub>2</sub> capture system can produce methane by using green hydrogen and playing a very significant role in this transition. Despite SOFC's benefits from the clean-burning properties of hydrogen, as water is the only byproduct, even NG gives SOFCs high efficiency and extremely low emissions compared to conventional ways of power generation. The integration of hydrogen technologies thus supports a dual strategy toward greater energy security and smaller carbon footprints in the energy mix. In this framework, DTs are changing the way the management and optimisation of energy systems—including SOFCs—are being carried out. A DT is a dynamic, virtual synchronised twin of a physical system using real-time data to simulate, predict, and optimise its performance. For SOFCs, a DT can monitor operational conditions in terms of performance prediction or fault diagnosis and enhance efficiency and reliability.

This paper is aimed at the development of a DT model for a gas-feed SOFC, using ML techniques to get an accurate efficiency estimation of the system under different conditions. The approach that was followed during this paper was to first gather four months of data from a real-world setup of an SOFC. Using this data, several deep and ML models, like ANNs and GBoost, were trained to find a reliable model capable of predicting electrical efficiency for the SOFC using operational parameters. The main findings of this study can be summarised as follows:

- The experimental setting of the study considered the measurement data acquisition during 3935 h of SOFC operation, with parameters like electric power output, gas flow rates, and system efficiency, whose correlation revealed strong dependencies used to conceptualise the models' features.
- Validation results, including the MAE, MAPE, and RMSE, revealed that XGBoost, GBoost, and RF algorithms were accurate and fast enough for real-time efficiency prediction within a DT framework.
- RF Regressor was found to be the best regression, with  $R^2$  almost equal to 0.99 and an RMSE equal to 0.31 using the entire dataset. Such high accuracy of the model potentially supports the integration of SOFCs into energy management systems, improving operational efficiency.
- Additionally, the potential of having a DT was demonstrated by observing how the most accurate model evolved over time based on the data collected, transitioning from XGBoost to GBoost, and finally to RF.

In summary, data-driven modelling techniques and RF have numerous advantages in the development of DTs for SOFCs. Accurate and reliable DTs will play a big role in the energy sector, towards a more digitalised and sustainable future. These virtual models optimise operational efficiency but further provide deeper insight into the behaviour



of the system, enabling proactive maintenance and reducing downtime. Future works will be aimed at enlargement of the dataset, covering more operational scenarios and environmental conditions, including the development of fault detection algorithms. In this framework, testing with blends of hydrogen and NG at various concentrations will also be considered in future works to investigate the effects on cell efficiency. Additionally, optimisation will be integrated into the DT to optimise the operation of the SOFC in an expanded context including photovoltaic systems, energy storage solutions, and buildings. Finally, these integrated systems can realise these dual goals of energy system resilience and sustainability in future decades.

**Author Contributions:** Conceptualization, T.T., G.C., M.B. and M.F.; Methodology, T.T., M.P.M., G.B., G.C., M.B. and M.F.; Validation, T.T.; Investigation, G.C., M.B. and M.F.; Data curation, T.T. and M.F.; Writing—original draft, T.T. and M.P.M.; Writing—review & editing, T.T., M.P.M., G.B., G.A.F., G.C., M.B. and M.F.; Visualization, T.T.; Supervision, G.A.F., G.C., M.B. and M.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Ministero dell'Università e della Ricerca, grant number B53C22010110001.

**Data Availability Statement:** The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- International Energy Agency. CO<sub>2</sub> Emissions in 2023. 2024. Available online: <https://www.iea.org/reports/co2-emissions-in-2023> (accessed on 10 August 2024).
- Rabuni, M.F.; Li, T.; Othman, M.H.D.; Adnan, F.H.; Li, K. Progress in Solid Oxide Fuel Cells with Hydrocarbon Fuels. *Energies* **2023**, *16*, 6404. [[CrossRef](#)]
- Rahmani, M.; Maharluie, H.N. Study of Syngas-Powered Fuel Cell, Simulation, Modeling, and Optimization. In *Advances in Synthesis Gas: Methods, Technologies and Applications*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 493–531.
- Ming, W.; Sun, P.; Zhang, Z.; Qiu, W.; Du, J.; Li, X.; Zhang, Y.; Zhang, G.; Liu, K.; Wang, Y.; et al. A Systematic Review of Machine Learning Methods Applied to Fuel Cells in Performance Evaluation, Durability Prediction, and Application Monitoring. *Int. J. Hydrogen Energy* **2023**, *48*, 5197–5228. [[CrossRef](#)]
- Bozorgmehri, S.; Hamed, M. Modeling and Optimization of Anode-Supported Solid Oxide Fuel Cells on Cell Parameters via Artificial Neural Network and Genetic Algorithm. *Fuel Cells* **2012**, *12*, 11–23. [[CrossRef](#)]
- Wu, Y.; Wu, X.; Xu, Y.; Cheng, Y.; Li, X. A Novel Adaptive Neural Network-Based Thermoelectric Parameter Prediction Method for Enhancing Solid Oxide Fuel Cell System Efficiency. *Sustainability* **2023**, *15*, 14402. [[CrossRef](#)]
- Guo, Z.; Ye, Z.; Ni, P.; Cao, C.; Wei, X.; Zhao, J.; He, X. Intelligent Digital Twin Modelling for Hybrid PV-SOFC Power Generation System. *Energies* **2023**, *16*, 2806. [[CrossRef](#)]
- Yang, J.; Li, X.; Jiang, J.H.; Jian, L.; Zhao, L.; Jiang, J.G.; Wu, X.G.; Xu, L.H. Parameter Optimization for Tubular Solid Oxide Fuel Cell Based on the Dynamic Model and an Improved Genetic Algorithm. *Int. J. Hydrogen Energy* **2011**, *36*, 6160–6174. [[CrossRef](#)]
- Gong, W.; Cai, Z.; Yang, J.; Li, X.; Jian, L. Parameter Identification of an SOFC Model with an Efficient, Adaptive Differential Evolution Algorithm. *Int. J. Hydrogen Energy* **2014**, *39*, 5083–5096. [[CrossRef](#)]
- Jiang, B.; Wang, N.; Wang, L. Parameter Identification for Solid Oxide Fuel Cells Using Cooperative Barebone Particle Swarm Optimization with Hybrid Learning. *Int. J. Hydrogen Energy* **2014**, *39*, 532–542. [[CrossRef](#)]
- Wei, Y.; Stanford, R.J. Parameter Identification of Solid Oxide Fuel Cell by Chaotic Binary Shark Smell Optimization Method. *Energy* **2019**, *188*, 115770. [[CrossRef](#)]
- Yang, B.; Chen, Y.; Guo, Z.; Wang, J.; Zeng, C.; Li, D.; Shu, H.; Shan, J.; Fu, T.; Zhang, X. Levenberg-Marquardt Backpropagation Algorithm for Parameter Identification of Solid Oxide Fuel Cells. *Int. J. Energy Res.* **2021**, *45*, 17903–17923. [[CrossRef](#)]
- Yang, B.; Guo, Z.; Yang, Y.; Chen, Y.; Zhang, R.; Su, K.; Shu, H.; Yu, T.; Zhang, X. Extreme Learning Machine Based Meta-Heuristic Algorithms for Parameter Extraction of Solid Oxide Fuel Cells. *Appl. Energy* **2021**, *303*, 117630. [[CrossRef](#)]
- Wang, J.; Xu, Y.-P.; She, C.; Xu, P.; Bagal, H.A. Optimal Parameter Identification of SOFC Model Using Modified Gray Wolf Optimization Algorithm. *Energy* **2022**, *240*, 122800. [[CrossRef](#)]
- Hwang, H.; Choi, S.M.; Oh, J.; Bae, S.-M.; Lee, J.-H.; Ahn, J.-P.; Lee, J.-O.; An, K.-S.; Yoon, Y.; Hwang, J.-H. Integrated Application of Semantic Segmentation-Assisted Deep Learning to Quantitative Multi-Phased Microstructural Analysis in Composite Materials: Case Study of Cathode Composite Materials of Solid Oxide Fuel Cells. *J. Power Sources* **2020**, *471*, 228458. [[CrossRef](#)]
- Hwang, H.; Ahn, J.; Lee, H.; Oh, J.; Kim, J.; Ahn, J.-P.; Kim, H.-K.; Lee, J.-H.; Yoon, Y.; Hwang, J.-H. Deep Learning-Assisted Microstructural Analysis of Ni/YSZ Anode Composites for Solid Oxide Fuel Cells. *Mater. Charact.* **2021**, *172*, 110906. [[CrossRef](#)]

17. Liu, X.; Yan, Z.; Zhong, Z. Predicting Elastic Modulus of Porous  $\text{La}_{0.6}\text{Sr}_{0.4}\text{Co}_{0.2}\text{Fe}_{0.8}\text{O}_{3-\delta}$  Cathodes from Microstructures via FEM and Deep Learning. *Int. J. Hydrogen Energy* **2021**, *46*, 22079–22091. [[CrossRef](#)]
18. Chen, M.T.; Fu, X.-W.; Deng, Z.-H.; Li, X.; Wu, X.-L.; Xu, Y.-W.; Xue, T. Data-Driven Fault Detection for SOFC System Based on Random Forest and SVM. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2829–2834.
19. Xue, T.; Wu, X.; Xu, Y.; Jing, S.; Li, Z.; Jiang, J.; Deng, Z.; Fu, X.; Xi, L. Fault Diagnosis of SOFC Stack Based on Neural Network Algorithm. *Energy Procedia* **2019**, *158*, 1798–1803. [[CrossRef](#)]
20. Costamagna, P.; De Giorgi, A.; Moser, G.; Pellaco, L.; Trucco, A. Data-Driven Fault Diagnosis in SOFC-Based Power Plants under off-Design Operating Conditions. *Int. J. Hydrogen Energy* **2019**, *44*, 29002–29006. [[CrossRef](#)]
21. Zhang, Z.; Li, S.; Yang, Y. A General Approach for Fault Identification in SOFC-Based Power Generation Systems. In Proceedings of the 2018 Annual American Control Conference (ACC), Milwaukee, WI, USA, 27–29 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3816–3821.
22. Fu, X.; Liu, Y.; Li, X. Source Diagnosis of Solid Oxide Fuel Cell System Oscillation Based on Data Driven. *Energies* **2020**, *13*, 4069. [[CrossRef](#)]
23. Zheng, Y.; Wu, X.; Zhao, D.; Xu, Y.; Wang, B.; Zu, Y.; Li, D.; Jiang, J.; Jiang, C.; Fu, X.; et al. Data-Driven Fault Diagnosis Method for the Safe and Stable Operation of Solid Oxide Fuel Cells System. *J. Power Sources* **2021**, *490*, 229561. [[CrossRef](#)]
24. Subotić, V.; Eibl, M.; Hochenauer, C. Artificial Intelligence for Time-Efficient Prediction and Optimization of Solid Oxide Fuel Cell Performances. *Energy Convers. Manag.* **2021**, *230*, 113764. [[CrossRef](#)]
25. Kang, J.-L.; Wang, C.-C.; Wong, D.S.-H.; Jang, S.-S.; Wang, C.-H. Digital Twin Model and Dynamic Operation for a Plant-Scale Solid Oxide Fuel Cell System. *J. Taiwan Inst. Chem. Eng.* **2021**, *118*, 60–67. [[CrossRef](#)]
26. Huo, H.; Ji, Y.; Kuang, X.; Liu, Y.; Wu, Y. Dynamic Modeling of SOFC Based on Support Vector Regression Machine and Improved Particle Swarm Optimization. In Proceedings of the 11th World Congress on Intelligent Control and Automation, Shenyang, China, 29 June–4 July 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1853–1858. [[CrossRef](#)]
27. Marra, D.; Sorrentino, M.; Pianese, C.; Iwanschitz, B. A Neural Network Estimator of Solid Oxide Fuel Cell Performance for On-Field Diagnostics and Prognostics Applications. *J. Power Sources* **2013**, *241*, 320–329. [[CrossRef](#)]
28. İskenderoğlu, F.C.; Baltacıoğlu, M.K.; Demir, M.H.; Baldinelli, A.; Barelli, L.; Bidini, G. Comparison of Support Vector Regression and Random Forest Algorithms for Estimating the SOFC Output Voltage by Considering Hydrogen Flow Rates. *Int. J. Hydrogen Energy* **2020**, *45*, 35023–35038. [[CrossRef](#)]
29. Hosamo, H.; Hosamo, M.H.; Nielsen, H.K.; Svennevig, P.R.; Svidt, K. Digital Twin of HVAC System (HVACDT) for Multiobjective Optimization of Energy Consumption and Thermal Comfort Based on BIM Framework with ANN-MOGA. *Adv. Build. Energy Res.* **2023**, *17*, 125–171. [[CrossRef](#)]
30. Testasecca, T.; Lazzaro, M.; Sirchia, A. Towards Digital Twins of Buildings and Smart Energy Networks: Current and Future Trends. In Proceedings of the 2023 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv), Milano, Italy, 29–31 May 2023.
31. Hosamo, H.H.; Nielsen, H.K.; Kraniotis, D.; Svennevig, P.R.; Svidt, K. Improving Building Occupant Comfort through a Digital Twin Approach: A Bayesian Network Model and Predictive Maintenance Method. *Energy Build.* **2023**, *288*, 112992. [[CrossRef](#)]
32. Mckee, D. Platform Stack Architectural Framework: An Introductory Guide—A Digital Twin Consortium White Paper. 2023. Available online: <https://www.digitaltwinconsortium.org/wp-content/uploads/sites/3/2023/07/Platform-Stack-Architectural-Framework-2023-07-11.pdf> (accessed on 10 August 2024).
33. Brownlee, J. XGBoost with Python: Gradient Boosted Trees with XGBoost and Scikit-Learn. 2016. Available online: <https://books.google.it/books?id=HgmqDwAAQBAJ> (accessed on 10 August 2024).
34. Louppe, G. Understanding Random Forests: From Theory to Practice. *arXiv* **2014**, arXiv:1407.7502.
35. Palomba, V.; Ferraro, M.; Frazzica, A.; Vasta, S.; Sergi, F.; Antonucci, V. Dynamic Simulation of a Multi-Generation System, for Electric and Cooling Energy Provision, Employing a SOFC Cogenerator and an Adsorption Chiller. *Energy Procedia* **2017**, *143*, 416–423. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.