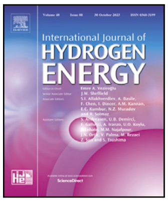
Contents lists available at ScienceDirect

International Journal of Hydrogen Energy

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



Degradation prediction of a fuel cell-based CHP-system under dynamic load using feed forward neural networks

Antonius Tilgner *, Adam Pluta, Heinz Bekebrok, Hendrik Langnickel, Alexander Dyck

Deutsches Zentrum für Luft- und Raumfahrt Institut für Vernetzte Energiesysteme, Carl-von-Ossietzky Straße 15, Oldenburg, 26129, Niedersachsen, Germany

ARTICLE INFO

Keywords:
Fuel cell degradation
Modeling
CHP
Neural networks
FNN
Proton exchange membrane fuel cell

ABSTRACT

A combined heat and power (CHP) system with natural gas as primary energy and a hydrogen fuel cell as power source is modeled by implementing a feedforward neural network (FNN). It is shown that only the load profile and the resulting voltage are already sufficient to provide an accurate prediction of the voltage and its degradation under normal operation, excluding anomalies. The short-term reversible degradation is accurately modeled with high fidelity, whereas the irreversible long-term degradation remains more challenging to predict. The influence on the prediction is analyzed for different input features. Additionally, the size of the training dataset is varied and as a physical parameter the gas composition is incorporated into the model, allowing it to more accurately predict anomalies in the data. The presented approach is further tested with data from a system operated in a residential area.

1. Introduction

Due to the substantial contribution of greenhouse gases such as CO_2 to global warming, an increasing number of nations are implementing concrete plans for carbon neutrality, thus increasing the importance of hydrogen as an alternative energy carrier [1]. Although in many parts of the industry, fossil fuels can be replaced by using the electricity of renewable energy sources, many sectors still require fuel. Hydrogen not only has the potential to satisfy this need in most cases, but can also provide the long-term flexibility needed for an economy that relies on renewable energy sources. As such, hydrogen has a multitude of advantages. Among others, it can be produced in different ways, most importantly by splitting water using electricity, it is easily transportable in pipelines or tanks with significantly lower losses compared to ultra-high voltage (UHV) electrical lines and its conversion to electricity is environmentally friendly, since only water is produced as a by-product [2].

The European Union has also recognized the advantages and potential of transitioning the economy to be built around hydrogen as the main energy carrier and has set itself ambitious goals in terms of the production of green hydrogen [3]. In Germany, the hydrogen core network encompassing 9040 km transport pipelines is set to implement injection and withdrawal capacities of 100 and 87 GW respectively between 2025 and 2032 [4]. A technology that could make efficient use of hydrogen is combined heat and power (CHP). As the name suggests, CHP systems generate heat and electricity simultaneously, resulting in a significant increase in overall efficiency. For combustion-based CHP systems powered by engines or turbines, overall efficiencies range from 65–85%, while the efficiencies when electricity and heat are provided independently range from 45–55% [5].

Another possibility to power CHP systems are fuel cells (FC). These hydrogen powered systems do not produce direct greenhouse gases as a byproduct while also achieving similar and even higher overall efficiencies [6]. Even efficiencies of up to 95% for fuel cell-based systems have been reached [7]. However, as 96% of hydrogen is produced using fossil fuels at this time, the overall carbon footprint remains high [3,6]. However, since hydrogen is expected to be produced primarily with electricity from renewable energy sources, this footprint will diminish in the future. Although complete replacement of gray with green hydrogen may take some time, such bridging technologies play an essential role in facilitating this transition by increasing acceptance of the technology, advancing research and gathering valuable experience in its implementation and application [6]. CHP systems with on-site hydrogen production, e.g. an integrated reformer, can be such a bridging technology that implements fossil fuel as its primary energy carrier, but can quickly and easily be switched towards direct hydrogen usage.

Among others, proton-exchange membrane fuel cells (PEMFC), as investigated in this research, are a common FC-type for CHP systems. They exhibit relatively low operating temperatures between 60 and 90 °C, resulting in rapid start-up times and therefore perform well on dynamic loads [6]. The anode and cathode of PEMFCs are separated by

* Corresponding author.
E-mail address: antonius.tilgner@dlr.de (A. Tilgner).

https://doi.org/10.1016/j.ijhydene.2026.154207
Received 10 July 2025; Received in revised form 20 February 2026; Accepted 23 February 2026
Available online 2 March 2026

List of Acronyms

CHP	Combined Heat and Power
FC	Fuel Cell
ML	Machine Learning
RMSE	Root Mean Squared Error
MLP	Multi-Layer Perceptron
FNN	Feedforward Neural Network
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
PEMFC	Proton Exchange Membrane Fuel Cell
GRU	Gated Recurrent Unit

a thin solid proton exchange membrane, allowing for a very compact and lightweight design. Additional advantages include a high power density compared to other FC types and low operating noise [8]. The disadvantages of PEMFCs include the high sensitivity to CO and the generally high requirements of hydrogen purity, due to their low operating temperatures [9]. Due to their rapid start time and the fact that they can be continuously modulated across a wide power range, PEMFCs are well-suited for small and medium-sized CHP systems [10]. Furthermore, they are less sensitive to start-stop cycles compared to other FC types, making them a good candidate for residential CHP applications [9]. Additional applications of PEMFCs include the automotive sector, specifically heavy duty vehicles, where PEMFC excel due to their energy and power scalability with minimal weight penalty [11].

A drawback of FCs in general is degradation, which causes significant performance loss over time due to factors such as material fatigue or electrochemical instability [12]. Degradation occurs in all parts of the stack, which are directly mechanically interconnected and exposed to physical and chemical conditions such as temperature, pressure, pH value and humidity [13]. These conditions and connections can cause mechanical and chemical stress in the parts, resulting in a loss of voltage that the stack can supply at a given operating point. Degradation can be reversible, as in the case of CO-poisoning of PEMFCs, which can be reversed by reducing the partial pressure of CO, or irreversible, as in the case of mechanical deformations such as holes and cracks in the membrane [9]. Due to the various different mechanisms leading to performance losses, degradation and the resulting end-of-life of a FC can vary significantly depending on the operating conditions. Modeling the degradation proves to be nontrivial as many mechanisms are not yet fully evaluated and intricate knowledge of the internal system parameters is necessary to develop a good physical model. In addition, complex relationships result in substantial computational effort [13].

Another possibility of modeling the mechanisms are data-driven approaches, which rely on statistical or mathematical methods to extract information from the data to deduce relationships and trends within the data. The greatest advantage of data-driven models is that they do not require a thorough understanding of the degradation laws or internal parameters. In order to extract the necessary information, a large amount of representative training data containing the aging processes is required to obtain a well-performing model. By applying mathematical methods to extract information from the data, the inner parameters of the system are learned implicitly. This can offer great benefit, as even structurally identical FCs can potentially exhibit different performances due to tolerances within the manufacturing process [14]. However, due to the individuality of each system, the generalizability of data-driven models is inferior compared to physical models [13].

One possibility is a machine learning (ML) approach using neural networks. These have the potential to greatly reduce the necessary computational power and facilitate real-time data processing once the models have been trained [13]. Although training and application of FNN offer short computation times and easy implementation and handling, they process each datapoint individually, making additional measures necessary when applied to temporal data. Recurrent neural networks (RNN) in comparison are especially suitable for capturing temporal dependencies by maintaining internal states across the temporal inputs. Long-short term memory neural networks (LSTM) and gated recurrent units (GRU) in particular are capable of capturing long-term temporal dependencies, which makes them well suited for timeseries forecasting [15].

Zuo et al. (2019) used LSTM and GRU to obtain promising results in predicting the voltage of a PEMFC operated under a periodic test cycle [16]. Liu et al. (2019) [17] and Liu et al. (2017) [18] used LSTM to predict the remaining useful life of a PEMFC under steady state conditions. Zhang et al. (2025) proposed an approach combining random forest with temporal convolutional networks to predict the degradation of a PEMFC under constant load [19]. For a periodic dynamic load cycle Ma et al. (2018) proposed a Grid-LSTM approach to further optimize existing LSTM methods [20]. Augmenting data-driven models with known physical relationships can help reduce data dependency. Ko et al. (2025) used physics-informed neural networks to model the remaining useful life of a PEMFC, significantly reducing the data dependency [21]. Zerrougui et al. (2025) used physics-informed neural networks to model temperature fluctuations within a PEM based electrolyzer achieving superior results compared to a solely data-driven LSTM approach [22].

LSTM have the disadvantage of slow training and inference times. Due to their simple architecture, FNN offer faster training and inference times and easier implementation. Although many models have been implemented using LSTM, leveraging their ability to capture temporal dependencies, FNN are known to lack this ability and consequently have not yet been employed to process timeseries in the context of degradation of FCs. LSTM models have been shown to be very successful in predicting the voltage of a PEMFC stack for steady-state load and for periodic dynamic load. However, they are unable to handle the varying input of a system under non-periodic dynamic loading. This work proposes a novel application of a simple FNN that aims to capture the temporal dependencies needed to model degradation behavior by means of feature engineering. The aim of this approach is to develop a model that can predict the degradation of the stack for varying load profiles on the basis of past performance data.

A well-performing model acting as a digital twin can improve the operation from a technical and economic point of view in multiple ways. *Predictive maintenance* aims to identify and resolve faults proactively, minimizing their impact on overall system performance [23]. Before making significant changes to the operation of the system, a digital twin can be utilized to analyze and evaluate the impact of the change on the system and address potential problems. For a dynamically loaded FC-system, a favorable operation resulting in a prolonged lifetime might be identified by predicting the performance for different load profiles.

After motivating this work and giving a short introduction into CHP and FC systems, the next section will focus on the methodology. Here, the data basis for the prediction will be explained in detail. Furthermore, the model architecture is elaborated and the feature engineering encompassing the feature selection and the data pre-processing is discussed. Afterwards, information on the hardware used, the training procedure, and the optimization is given. Following the methodology, the results obtained are shown and discussed.

2. Methodology

2.1. Data basis

The data for training and validating the algorithm originates from a natural gas based 5 kW_{el} hydrogen FC CHP system (Fig. 1), which utilized the hydrogen produced in an integrated reformer from desulfurized natural gas to feed a 95-cell FC-stack. During power generation,

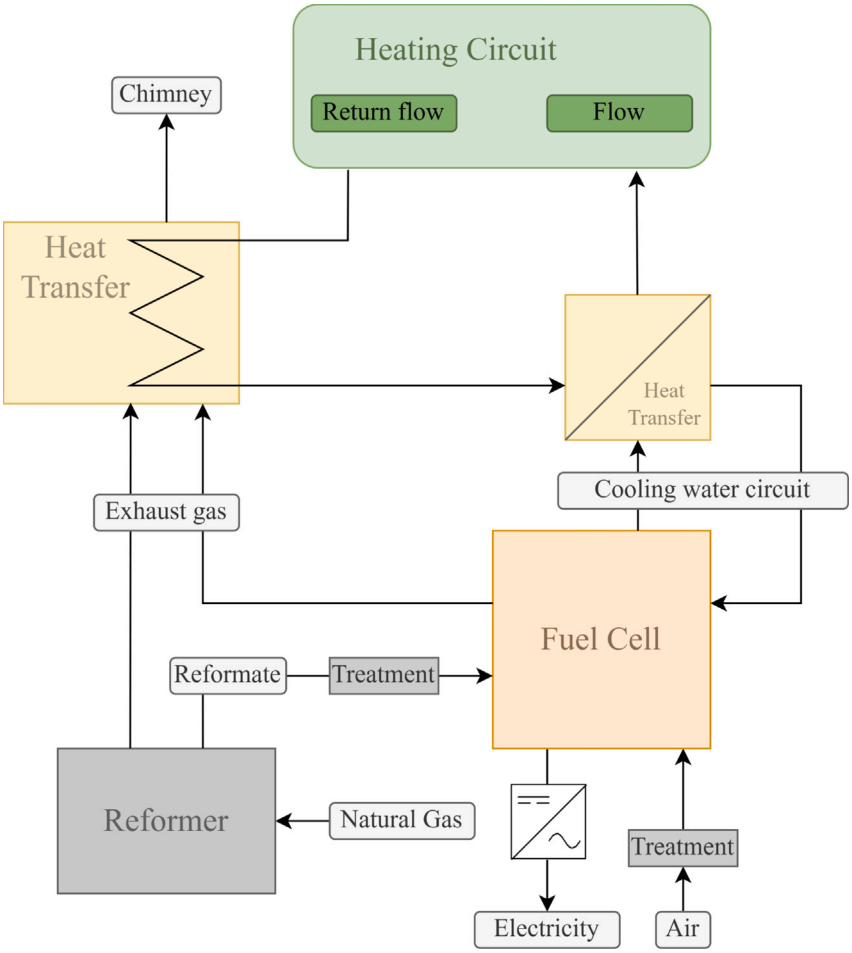


Fig. 1. Schematic flow chart of the regarded CHP system.

a cooling circuit dissipates the waste heat from the FC-stack, which is subsequently transported together with the waste heat of the reformer to be used in a heating circuit. Before entering the FC stack, the ambient air is compressed, which simultaneously heats the air to the required temperature. A humidifier moisturizes the air to prevent the stack from drying.

Within a 20-week testing phase, the current-regulated system was dynamically operated at full load (80 A) and partial load (30 A) in a test cycle according to a test profile, conducted for six consecutive days, followed by one day of idling. Idling describes a state in which the system is not actively generating power but can react to incoming load at any time. During testing data with a sampling frequency of 0.1 Hz was collected, including internal sensors such as temperatures and or pressures, as well as the load profile and the voltage provided by the system (Fig. 2). Additionally, the data recorded in regulated testing conditions, the same type of data was collected for a private customer system outside such an environment.

Within the data, two trends can be identified. First, during longer time intervals, long-term voltage degradation can be observed at each operating point (Fig. 3(a)). Second, a short-term voltage decline can be observed at a given load plateau (Fig. 3(b)). Although long-term degradation is not reversible, short-term degradation is partly reversed during idling between two plateaus, and the voltage regenerates when load is applied again. The exact degradation mechanisms responsible for this behavior have not been thoroughly investigated at the current time.

The composition of the natural gas used during the test runs is measured with a gas chromatograph (*ABBNGC8206*). In contrast to the data sampled with a frequency of 0.1 Hz the gas composition is sampled in a time interval of 5 min and 15 s. Part of this data includes the higher and lower heating values as well as the molar percentages of propane, isobutane, butane, neopentane, isopentane, pentane, hexane, nitrogen, methane, carbon dioxide and ethane.

2.2. Model architecture

LSTM neural networks have been successfully employed to model the voltage degradation of statically operated FCs. Furthermore, for periodic dynamic operation, LSTM yield promising results as well. However, for non-periodic dynamic operation, this approach leads to problems, as the voltage is not only dependent on past voltage developments but very much on the load profile within the time horizon of the prediction. As this profile does not necessarily match the profile of the training data, an approach must be selected that enables the incorporation of the load profile for which the corresponding voltage profile is to be predicted.

This study suggests feed-forward neural networks (FNN) as an alternative. This simple architecture promises short computation times for training and prediction. One of its disadvantages is that the FNN does not take past values into account. Rather, it represents a function approximator that uses the input values of a single time step to model the output value at that time-step (Fig. 4). This results in difficulties in accurately predicting voltage and modeling voltage degradation, as the voltage is subject to degradation effects leading to voltage drops that are in great parts dependent on temporal aspects, such as the operating time. Since FNN cannot directly consider past values, all the information that could have an impact on the voltage degradation must be encoded and fed to the model as an additional input. This

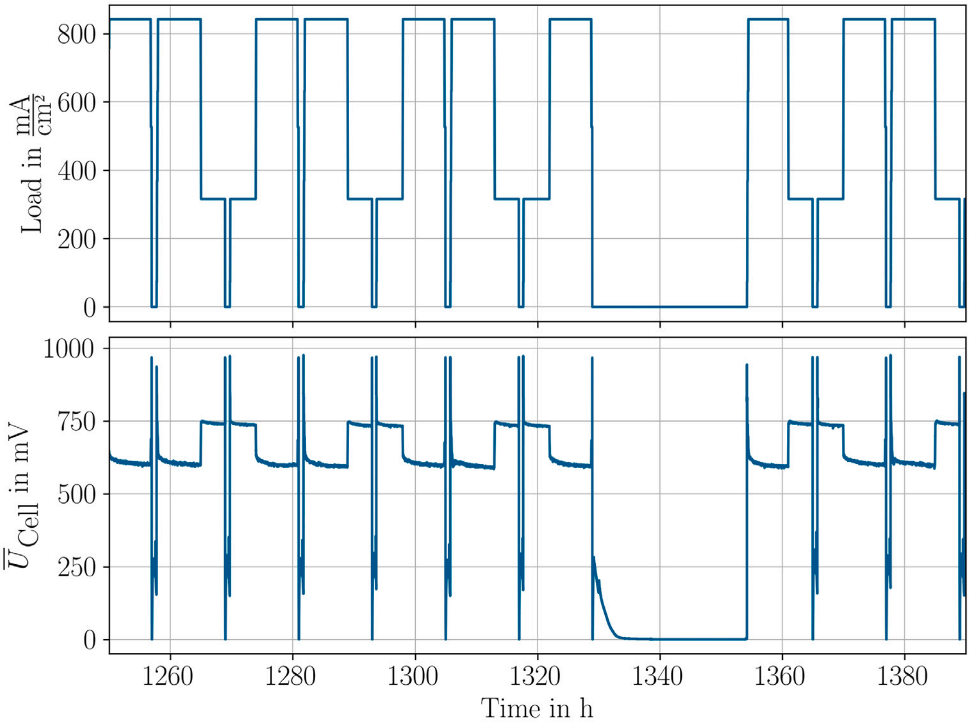
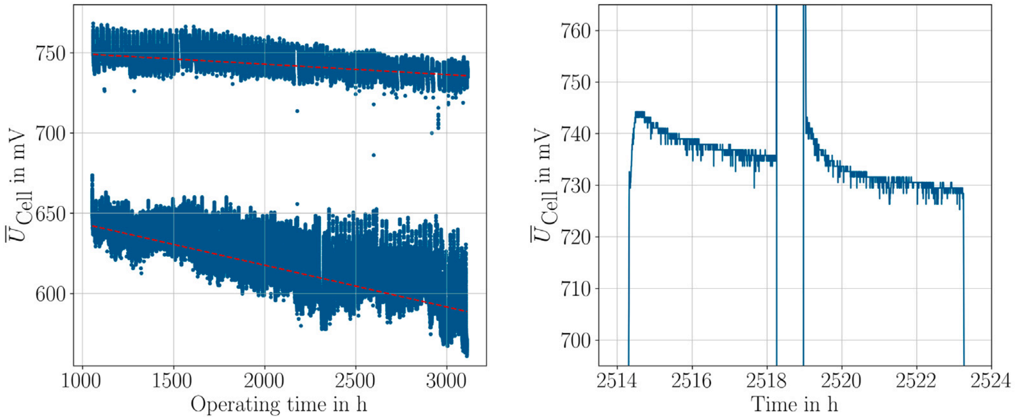


Fig. 2. Magnified example of the testing cycle. In the upper plot the load applied in form of the test profile to the system is depicted. The bottom plot shows the ensuing average voltage provided by each cell.



(a) Visualization of the difference in the long term degradation for the two load levels. The plot shows the voltage with respect to the operating hours and a linear fit for both load levels individually.

(b) Magnified view of the short term degradation. The voltage of the two slopes develops differently despite being generated by the same load.

Fig. 3. Visualization of the short-term and long-term degradation in the data.

ensures that the model incorporates information on previous voltage development.

2.3. Feature engineering

The feature engineering is an integral part of the implementation of any neural network application and serves multiple purposes. Firstly, omitting irrelevant features decreases the dimensionality of the training process, which in turn decreases the computational time and effort. Secondly, this simplifies the learning process, reducing computational effort and accelerating convergence, as the model is no longer distracted by irrelevant features [24].

The process of feature engineering encompasses not only the selection of relevant features which aides the model to learn the desired dependencies but also the pre-processing of the used feature vectors, which further simplifies the training. An example of a common pre-processing step is normalization or the subtraction of the mean.

Feature selection

The voltage is defined as the target feature, while the most crucial input feature is the load profile (Fig. 2). As the model is constrained by its inability to consider information about the past development of the load, the information about historical development must be encoded

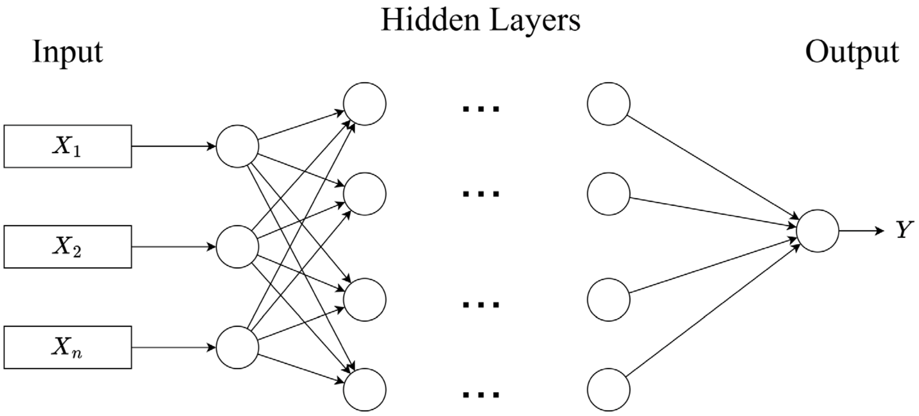


Fig. 4. Schematic flowchart of a FNN as it is used in this work. The model uses the inputs of a given timestep x_i^t to predict the voltage y^t at that timestep.

and passed to the network. This can be achieved with additional dimensionless input features, derived from the load profile. In Fig. 5 the load profile and the derived input features encoding the temporal information are qualitatively shown. Three additional features have been identified that provide the necessary information and are needed to model the voltage of the system:

- **Operating time**: The operating time indicates the duration for which the fuel cell stack has been active. This information is of great importance, as degradation is directly related to the aging process of the FC.
- **Time at load level**: The time, the system has already operated on the current load level. This feature is important to model the short-term decrease in voltage at one level (Fig. 3(b)). The feature vector is implemented as a counter, which increases whenever the load level in the time series remains constant and is set to 0 whenever the load level changes.
- **Time of idling**: The time the system remains on standby before being operated has an impact on the regeneration of the stack. As such, it is crucial for the model to receive information about the duration of idling of the system before reaching the current load level (0, if the load went directly from one level to another), although the voltage between load levels itself is not modeled. The importance of including the system idling can be seen in Fig. 3(b), as after idling, the voltage shortly follows a different pattern compared to before idling.

Data pre-processing

To streamline the training process and achieve superior results in less time. First, data during system idling is removed. This data does not yield any information and is not needed with this approach. Removing it also decreases the training time as fewer data points are processed. The load of the system does not jump to a load level, but rather ramps up, inducing a voltage peak (Fig. 2). These narrow peaks might hinder effective training by dominating the gradient calculation. Furthermore, as the focus of the study lies on the short- and long-term degradation modeling during operation only the data points at the respective load are considered. In addition, the mean voltage at the individual load levels is subtracted before training and added to the prediction after training. In this way, the model is not required to learn the significant voltage differences between the two load levels; rather, it can focus on the slopes of the plateaus. Furthermore, all inputs and outputs are normalized to prevent any feature from being weighted more than any other, further simplifying the training. Lastly, the data is divided into a test and a training dataset with $4/5\text{th}$ of the data being used for training and $1/5\text{th}$ for testing.

2.4. Training and optimization

The implementation of the models is done on a *Nvidia RTX A5000* graphics card, with 128 GB RAM an *Intel(R) Core i9-12900K 3.2 GHz* using the open-source *python* frameworks *TesorFlow* and *Keras*.

Hyperparameter optimization

An essential part of the training of a neural network is hyperparameter optimization. Hyperparameters are parameters that are not learned by the model during training but influence its training and convergence behavior. A good set of hyperparameters is mandatory for a well-performing model. For their optimization the open-source framework *Optuna* is used that employs Bayesian optimization to systematically search the hyperparameter space for a well-performing set. The ranges for each hyperparameter are defined as follows:

- **Number of hidden layers:** The number of layers between the input layer and the output layer as an integer between 1 and 5.
- **Number of hidden units:** For each hidden layer, the number of neurons in that layer is generated dynamically with values between 30 and 600.
- **Initial learning rate:** The initial learning rate ϵ_0 used in Eq. (1) is optimized within the ranges of 10^{-5} and 10^{-1} and sampled logarithmically.
- **Activation function:** The activation function introduces nonlinearity into training by applying it to the value of each neuron. During optimization, it is sampled from a set of four commonly used activation functions: *ReLU, hyperbolic tangent, Leaky ReLU, Sigmoid*.
- **Dropout:** To remedy overfitting, dropout is a common and effective method. With this procedure, after each epoch a fixed ratio of neurons and their connections is ignored and not updated. The dropout ratio is uniformly sampled from values between 0.1 and 0.8.
- **Batch size:** The batch size is the number of datapoints that are being considered during each epoch. It is sampled from values between 50 and 1500 with a step size of 50.
- **Shuffle:** Shuffling the values before drawing batches is a common method to improve training results. Whether to shuffle the dataset or not is also optimized via the hyperparameter tuning. Thus, it is sampled either *True* or *False*.

Learning rate and optimizer

The optimization algorithm calculates the parameter update for the weights and biases for each epoch during training, and thus dictates the training performance and stability. For this application, the Adam optimizer was chosen. It is a very popular optimization algorithm and is a common first choice due to its robustness and versatility, which stem

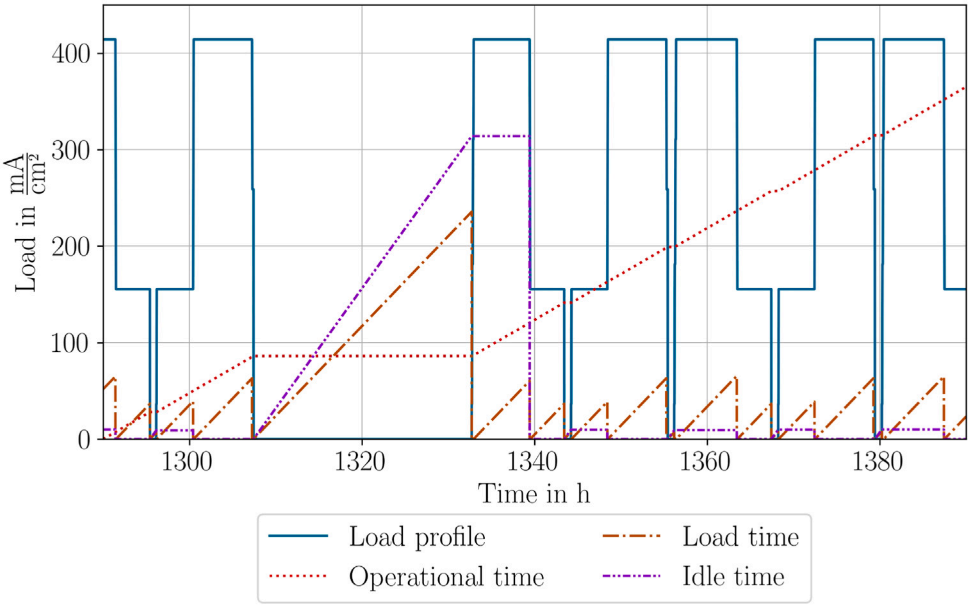


Fig. 5. Qualitative representation of the load profile along the other dimensionless input features. For better visualization, the values for all features except for the load profile are scaled.

from its adaptive learning rate and momentum, which simultaneously prevents the model from settling in an undesirable local minimum in the early stages of training and becoming unstable in later stages of training. Furthermore, it is easy to handle and does not require additional hyperparameters to be tuned [25,26].

Although the Adam optimizer already implements adaptive learning rates, it can be beneficial to use learning rate schedules nonetheless. An exponential learning rate schedule with the initial learning rate ϵ_0 was selected as an optimizable hyperparameter. It is given by

$$\epsilon_t = \epsilon_0 \cdot d \cdot \exp(t/r) \tag{1}$$

where t is the timestep, d is the decay rate and r is the step number, which defines after how many timesteps the learning rate decays by the decay rate once.

Evaluation standard

The root mean squared error (RMSE) is used to evaluate the prediction of the model. Although reversible short-term degradation is a significant effect, irreversible long-term degradation has a minimal impact, which may not be clearly reflected in RMSE. Thus, for the evaluation of this effect, the datapoints are linearly fitted with respect to the operating hours, and the slope of this fit is used for the evaluation.

3. Results

3.1. Individual models

To evaluate the influence on the model performance of each input parameter, a model is trained and optimized after the individual parameters are added subsequently. The models for the different sets of input features are the results of hyperparameter optimization with 150 trials each. Although this is a relatively small number of trials for hyperparameter optimization, it suffices for intercomparison of models with different input features and for identifying trends and tendencies. In the following, the predictions of the five individual models with increasing complexity are shown.

Model 1 — Load profile

Using only the load profile and no additional information as input, the model prediction does not exhibit degradation behavior throughout the operating time or at the individual load levels (Fig. 6). Rather, a constant voltage level is associated with each load level.

Model 2 — Operating hours

After adding the operating hours as input, a long-term voltage decline is modeled. This is observable in Fig. 7(a), where the voltage is plotted with respect to the operating hours for better visibility of the trend. Although the prediction now shows a linear decline approximating the irreversible long-term degradation, the reversible short-term degradation at the individual plateaus is not modeled. Instead of a significant voltage drop, the voltage only decreases according to the linear approximation of the irreversible degradation (Fig. 7(b)).

Model 3 — Plateau time

Incorporating the time the system remains at a load level leads to an approximation of the reversible degradation. Upon closer examination of two plateaus, which are separated by a brief period of idling, it becomes evident that the behavior differs between the two plateaus. Although a voltage decline at the load levels is visible in the prediction (Fig. 8(a)), the difference between separate plateaus cannot be modeled. The idle time is not taken into account and the prediction before and after is nearly identical (Fig. 8(b)).

Model 4 — Idling time

The addition of the idling time as an input leads to a different prediction for two plateaus, which are separated by a short idling period, but are caused by the same load (Fig. 9).

Finetuned model

The models shown have been optimized with 150 trials. To obtain a final model with maximized performance, a hyperparameter tuning with 1000 trials is performed. A higher resolution plot of the prediction together with the measurement can be seen in Fig. 10(a) and the degradation prediction is linearly fitted with respect to the operating hours in Fig. 10(b). As degradation behaves differently at partial and full load, both load levels are fitted independently.

3.2. Hyperparameter comparison and importance

All individual models are optimized independently, and thus a different set of hyperparameters arises for each, listed in Table 1. Additionally, *Optuna* offers the possibility of ranking hyperparameters

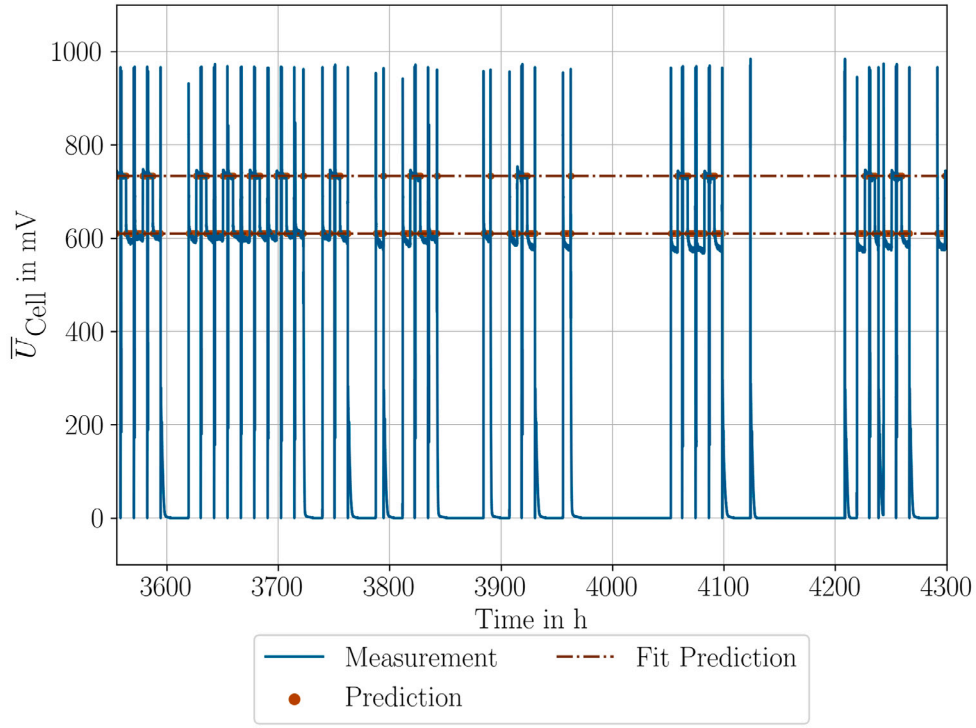
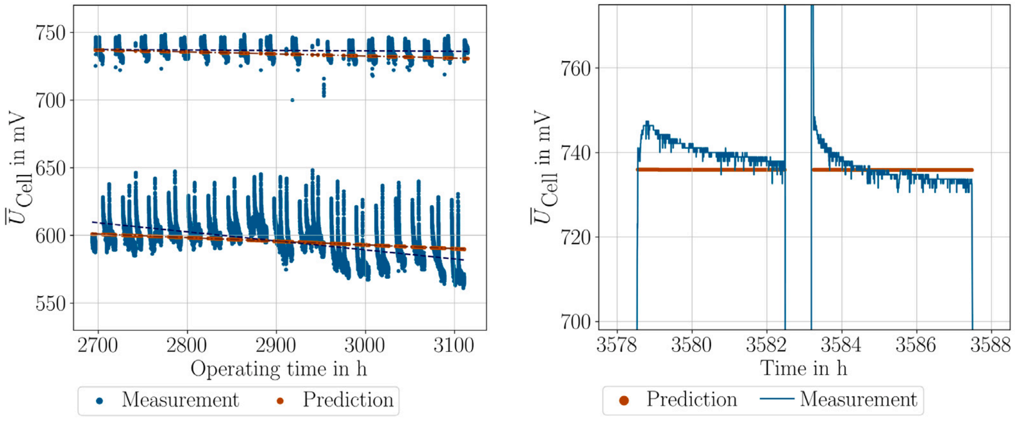


Fig. 6. Magnified view of the voltage prediction for a model trained only with the load profile and none of the additional input features. The degradation cannot be modeled as the information about the age of the stack at any timestep is not provided to the model.



(a) The prediction of model 2 plotted with respect to the operating hours for the test dataset with visible long term degradation modeling. The trendlines show the linear regression of the measurement and the prediction.

(b) Magnified view of the prediction of the model. The reversible degradation is not modeled as no information about the time spent on the load level is provided.

Fig. 7. Visualization of the prediction of the model with the operating hours as additional input.

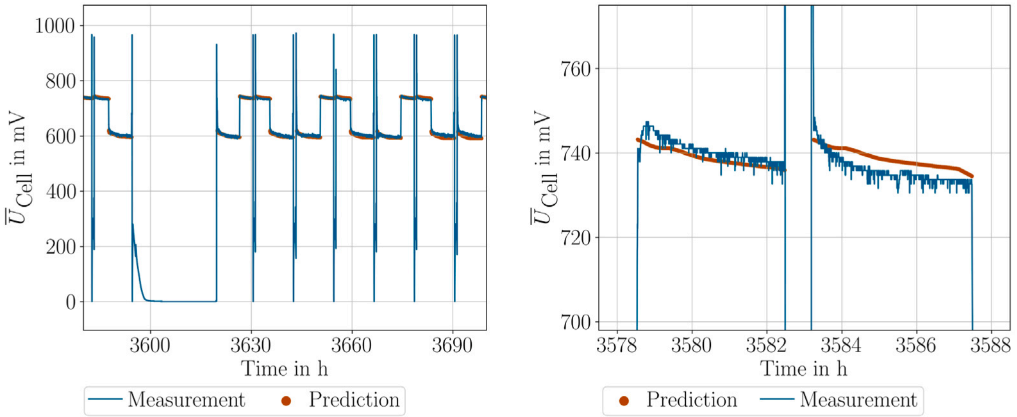
according to their influence on the model performance during optimization. In this context, the influence of a single hyperparameter is the degree to which model performance fluctuates when the hyperparameter is varied. This ranking is depicted in Fig. 11. The regularization in form of the dropout ratio has the greatest influence during optimization followed by the learning rate, the number of hidden layers, and the batch size. All other optimized hyperparameters seem to not strongly influence the performance of the trained model.

3.3. Different stack application

As the internal parameters of each stack can vary, the generalization capabilities of the model are tested by applying it to the data of previous test configurations where a different, but structurally identical stack was installed within the CHP system and comparing the voltage prediction to the measured values. The prediction can be seen in Fig. 12. Although the predicted slope of the long-term degradation is similar to the validation data, the predicted voltage plateaus consistently overestimate the measured values. Furthermore, the predicted short-term degradation slopes of the voltage differ significantly from the validation data.

3.4. Variation of the split ratio

Although a large amount of training data is available for these simulations, this might not always be the case for other simulations.



(a) The prediction of model 3 with respect to the operating hours. The model can now approximate the voltage decline due to the reversible degradation.

(b) Magnified view of the plateaus of the test dataset. The reversible degradation is modeled, but its regeneration is not modeled well.

Fig. 8. Visualization of the prediction of the model with the operating hours and the plateau time as additional inputs.

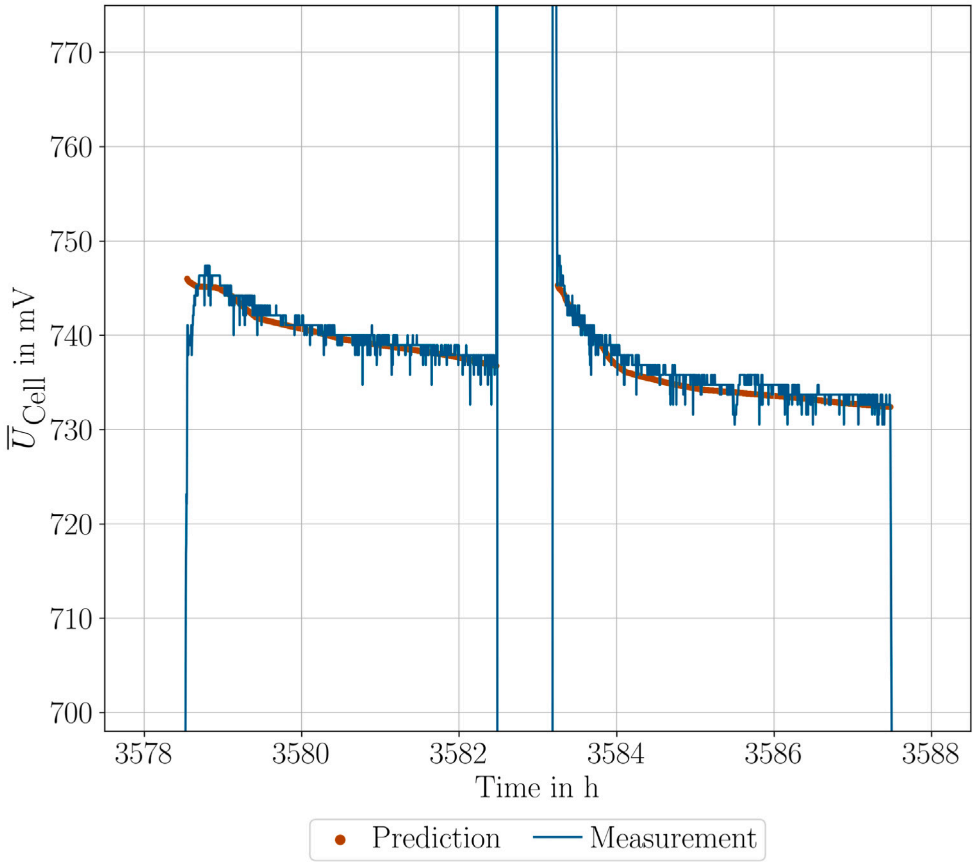
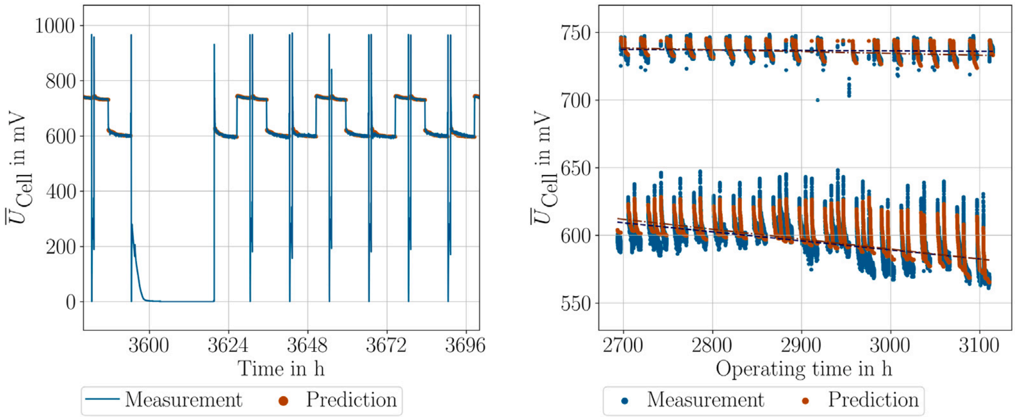


Fig. 9. Magnified view of the plateaus of the test dataset. The discrepancy of the two plateaus due to the regeneration time in between the load levels can be modeled after the incorporation of the idling time.

To test the performance of the model with fewer training data, the split ratio is varied. This is done once for a part of the dataset where few anomalies are present (Fig. 13) and once for a slightly larger part of the dataset where a strong anomaly is present (Fig. 14). Anomalies describe fluctuations within the voltage that are not caused by a change in operation. The split ratios between test and training data are $1/4\mathrm{th}$, $2/4\mathrm{th}$ and $3/4\mathrm{th}$ with the red dashed line indicating the split. For the first dataset (Fig. 13), the predictions are similar for each split ratio and with $1/4\mathrm{th}$ of the dataset used as training data, a voltage decline resembling the long-term degradation is predicted. For the second dataset (Fig. 14), no voltage decline is predicted using only $1/4\mathrm{th}$ of the dataset as training data. This effect vanishes as the size of the training dataset increases. Using $2/4\mathrm{th}$ and $3/4\mathrm{th}$ of the dataset, a clear voltage decline is predicted. The performance of each model is evaluated using RMSE as shown in Table 2.



(a) Magnified view of the prediction of the finetuned model. The prediction is very close to the measured data.

(b) The prediction of the finetuned model plotted with respect to the operating hours for the test dataset with visible reversible degradation modeling. The trendlines show the linear regression of both the measurement and the prediction.

Fig. 10. Visualization of the prediction of the finetuned model with the operating hours, the plateau time and the idling time as additional inputs.

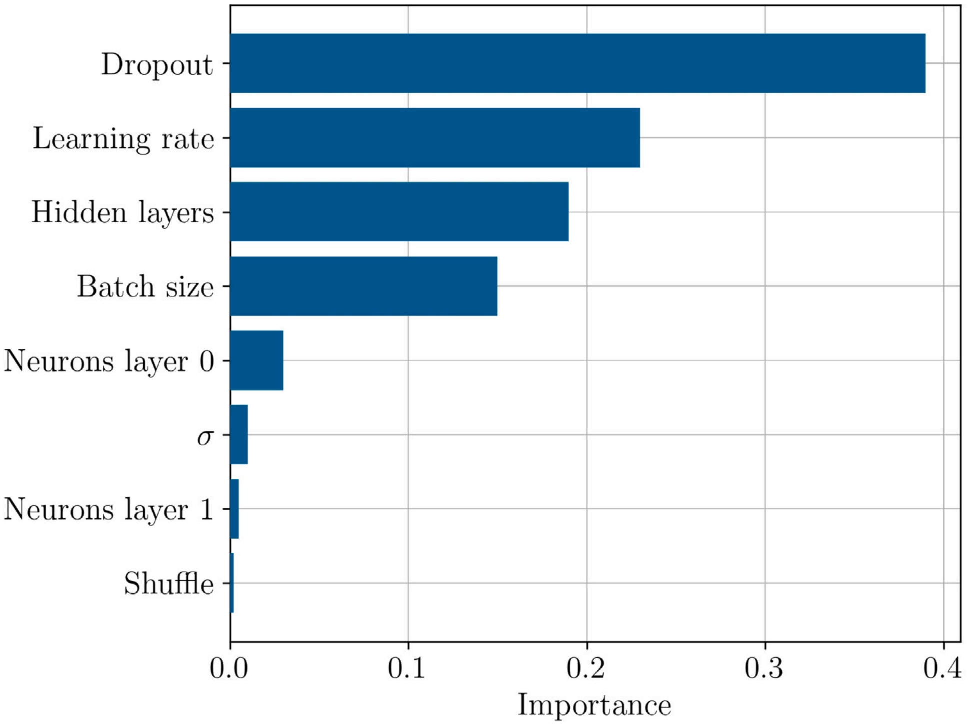


Fig. 11. Ranking of the hyperparameter importance across the whole study.

3.5. Incorporation of the gas composition

All input features on which the model is trained are temporal in nature. As a consequence, voltage fluctuations that are caused by external factors cannot be predicted. Although physical data cannot be used for predictive purposes, as this type of data is not known in the future, it can be used to draw conclusions about the correlation between certain fluctuations and external factors. To test this, the data for the composition of the natural gas during the tests can be used by incorporating it into the training and prediction of the model. In Fig. 15 the prediction of the model can be seen after incorporating the lower heating value and the molar percentage of methane into the model. Performance metrics are displayed together with the other trained models in Table 3.

3.6. Residential application

The data utilized so far was recorded on the test rig in a highly regulated laboratory environment. It follows a predefined test cycle (Fig. 2) which results in a very regular and periodic dataset. This is beneficial for the training procedure and additionally simplifies preprocessing. Data from a commercially used system may be harder to predict, as it might not follow a strict repeating profile. To test the robustness and generalization capabilities of the approach, a model is trained utilizing such data (Fig. 16).

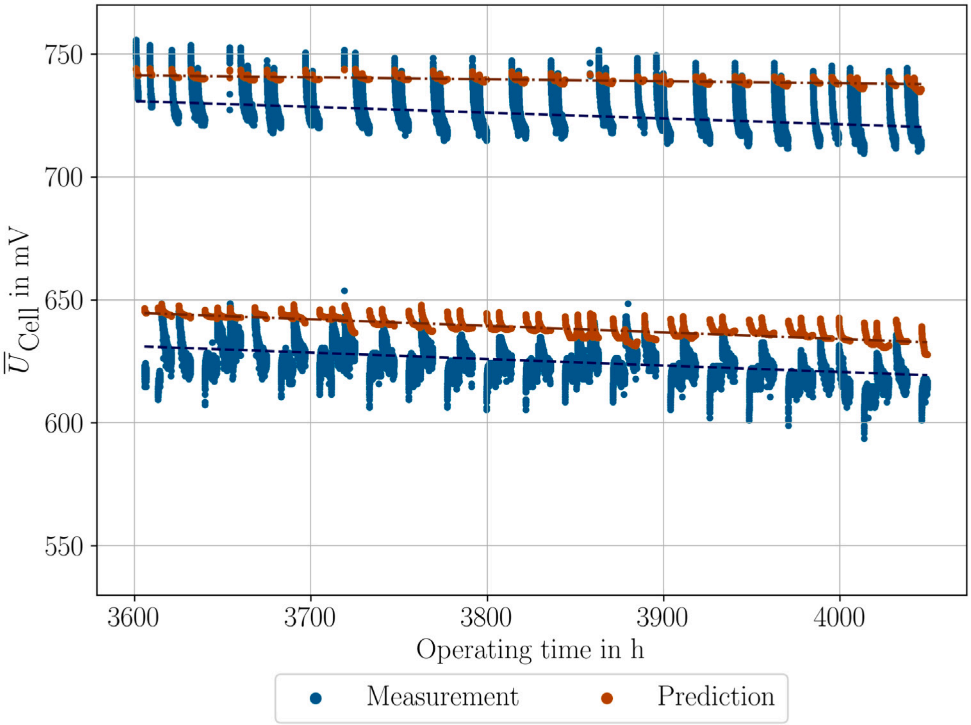


Fig. 12. Prediction of the final model on data from a different stack. The predicted voltage does not correspond to the measured voltage. The trendlines show the linear regression of the measurement and the prediction.

Table 1
Overview of the hyperparameters of the models. The activation function is signified by σ, "LR" corresponds to the learning rate and the architecture to the hidden layers and their respective number of neurons (592-31 e.g. means there are 592 neurons in the first and 31 neurons in the second hidden layer).

Model	1	2	3	4	Tuned
LR$(\cdot 10^{-3})$	0.02	8.77	1.27	0.61	2.12
Architecture	592–31	217	360–261	176–161	466–452
Dropout	0.548	0.683	0.404	0.459	0.524
Shuffle	False	True	False	True	False
Batch size	1050	800	150	850	400
σ	ReLU	leaky ReLU	ReLU	ReLU	ReLU

Table 2
Overview of the performance of the models trained on different amounts of data. To maintain comparability, the errors are calculated for the last quarter of the data and thus on the same data slice independent on the size of the test dataset.

Split ratio	Without anomaly	With anomaly
1/4	6.64 mV	14.28 mV
2/4	6.33 mV	4.98 mV
3/4	4.22 mV	4.02 mV

Table 3
Overview of the evaluation metrics for the different models.

	RMSE [mV]	Full load fit [μV/h]	Partial load fit [μV/h]
Model 1	15.94	0.0	0.0
Model 2	9.67	−26.9	−15.1
Model 3	7.75	−80.2	−29.9
Model 4	7.28	−100.5	−27.3
Finetuned Model	6.07	−73.2	−13.5
Gas comp.	5.32	−46	−20.3
Validation		−66.7	−3.3

4. Discussion

Using solely the load as input (**Model 1**), neither the reversible degradation nor the irreversible degradation can be modeled (Fig. 6). Since degradation is an aging process, the model is unable to predict voltage decline without temporal information of the stack operation. Since only one of two load levels is given at each timestep, only two different constant voltages can be predicted. These are the mean of all data points at the individual load levels.

After introducing the operating hours as an input (**Model 2**), a voltage decrease can be observed in the prediction (Fig. 7(a)). Now, the model is capable of coupling the long-term voltage decline caused by irreversible degradation to the operating hours, and thus approximate it in the prediction. Furthermore, the lack of information about the operating time of each plateau explains the inability of the model to predict reversible degradation. (Fig. 7(b)).

By incorporating plateau time as an input (**Model 3**), reversible degradation is modeled as a function of operating time on a load plateau (Fig. 8(a)). Although a clear voltage decline is visible for each plateau, the model is still unable to differentiate between the individual plateaus. However, a difference in behavior after a short idle time is clearly visible (Fig. 8(b)). As the model cannot distinguish between the plateaus, their prediction is nearly identical.

By adding the idling time as an input (**Model 4**), the model can differentiate between two plateaus and can couple idle time to stack regeneration. Thus, a good prediction of the slopes for the individual plateaus becomes possible (Fig. 9).

With the addition of each new input, the RMSE of the model becomes smaller (Table 3). The irreversible degradation has a very small effect, and consequently the absolute prediction error of the model with respect to the irreversible degradation is very small. In contrast, reversible degradation has a greater impact on the RMSE. Thus, we conclude that RMSE is a good evaluation standard for the quality of the prediction of reversible degradation, while it is less optimal for the evaluation of the prediction of irreversible degradation. For irreversible degradation, the voltage measurement and prediction are linearly fitted

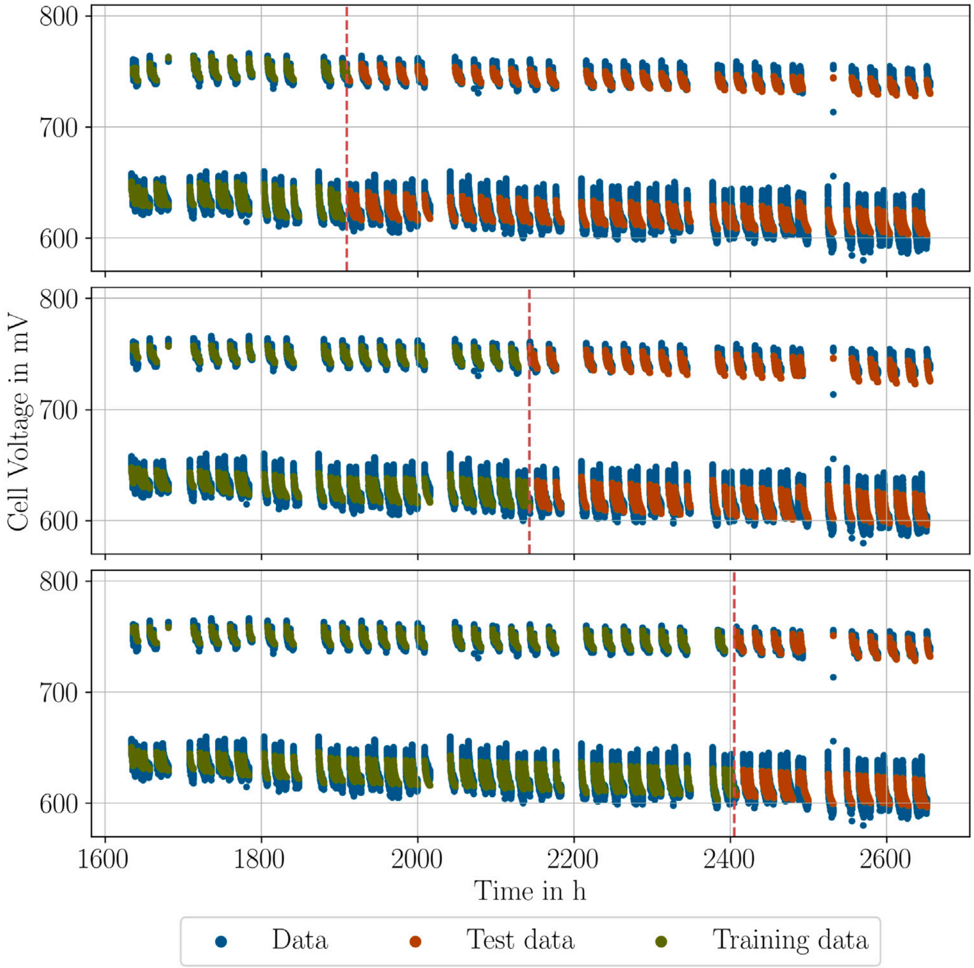


Fig. 13. Variation of the split ratio for a dataset without anomalies. The blue markers represent the whole dataset, while the green markers represent the model prediction on the training dataset and the orange markers represent the model prediction on the test dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

across the test dataset for each load level, and the slopes of these fits are compared (Table 3). Although RMSE becomes better with each input, the same cannot be observed for the fit slopes. Even the fine-tuned model does not approximate the long-term voltage decline well. Because the models are trained by an optimization algorithm based on gradient descent, very small effects are hard to learn. Minor errors lead to a small gradient, resulting in only slight updates to the parameters during each epoch. In later epochs the learning rate drops significantly, further reducing the parameter update.

The prediction performance of the model on a different stack is poor (Fig. 12). The stack behaves differently due to differences in various internal parameters that occur due to tolerances in the manufacturing process [14]. Consequently, the expected poor generalization capabilities of a data-driven model is verified and the conclusion can be drawn that for each stack a new model would have to be trained.

A variation of the split ratio offers insight into the amount of data needed for a good prediction. Considering Fig. 13 and Table 2, the prediction in the first test improves as the size of the training dataset increases. The complete dataset shown in Fig. 13 is very regular, simplifying the training drastically. However, this is not always the case as seen in Fig. 3(a).

A second test was performed with an extended dataset, including an anomaly at the beginning (Fig. 14). When performing the same variation of the split ratio on that dataset, it becomes evident that our model performance does not solely depend on the amount of training data, but on the quality and representativeness of the training dataset as well. The anomaly dominates the first part of the training dataset. Therefore, accurate prediction is only feasible when the training dataset is large enough to capture the degradation behavior and classify the anomaly as non-representative. Using $1/4\text{th}$ of the dataset for training, the model cannot learn the degradation, while with $1/2\text{th}$ the model can predict the degradation behavior well. Increasing the size of the training dataset further beyond $1/2\text{th}$ does not lead to a better prediction. It should be noted that the prediction with $1/2\text{th}$ and $3/4\text{th}$ of the training dataset is better for the larger dataset that includes the anomaly compared to the dataset without the anomaly (Table 2). Due to the black-box nature of neural networks, it is hard to isolate a single reason for this observation. The dataset with the anomaly is 30% larger, which is beneficial during training. This argument is contradicted by the fact that the additional data does not provide additional information about the degradation, as it is mainly comprised of the regeneration anomaly. Furthermore, the performance for $1/2\text{th}$ of the larger dataset is better than that for $3/4\text{th}$ of the smaller dataset, suggesting that the larger overall dataset is not decisive for the better performance. The anomaly within the data might have regularizing effects leading to better generalization on unseen data and thus improving the performance.

With this approach, physical sensor data cannot be used to predict the voltage output of the stack, as it must be fed as an input and is not known in the future. However, training a model with sensor data can be used to gain knowledge about the system's behavior. By training the model with physical sensor data and comparing the prediction with

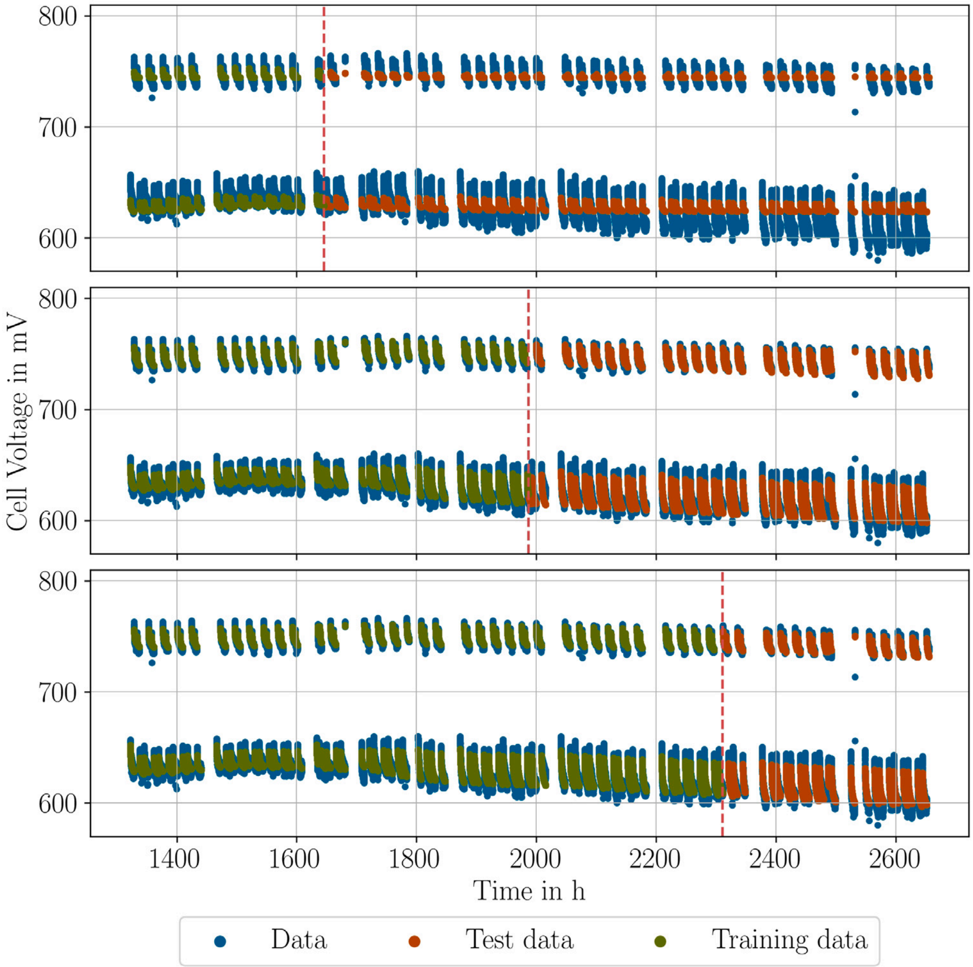


Fig. 14. Variation of the split ratio for a dataset with an anomaly early in the dataset. The blue markers represent the whole dataset, while the green markers represent the model prediction on the training dataset and the orange markers represent the model prediction on the test dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the measurements, conclusions can be drawn about the fluctuations in the output voltage of the system. Incorporating the molar percentage of methane in natural gas and its lower heating value, significantly enhances predictions (Fig. 15), as voltage fluctuations are modeled with greater precision. The naturally fluctuating methane content of the natural gas has a direct effect on the amount of hydrogen in the reformate, which influences the voltage output of the stack. Thus, with the information on the methane content of the natural gas used, the power output of the system can be predicted more accurately.

5. Summary and conclusions

In this study, an approach has been presented to predict the voltage of a FC-based CHP-system under dynamic load employing a finetuned FNN model achieving a high accuracy with a RMSE of 6.07 mV. All essential input features are derived from the load level and encode its development. However, a robust prediction could be achieved without the need for sensor data. For this purpose, individual input features were subsequently added to observe their effect on the model performance. Bayesian optimization, implemented using the *Optuna* framework, identified a set of hyperparameters that yielded a highly accurate model. Due to the poor generalizability of data-driven models, applying the model to data from previous test configurations, including a different stack, yielded poor results.

Furthermore, the split ratio between the training and the test dataset was altered, showing that a well-performing model can already be obtained from a training window of 200 h. However, the amount of data needed for a good prediction depends to a large degree on the quality and representativeness of the training dataset.

In order to predict the deviations from normal operation, the gas composition of the natural gas was introduced as an additional input. Specifically, the molar percentage of methane and the lower heating value were added. The additional information on the gas composition resulted in a good prediction of the voltage fluctuations, suggesting that they are caused by the naturally fluctuating methane content of the natural gas.

Furthermore, a model was trained on the data from a commercially used system outside of a strictly regulated laboratory environment.

Although this study successfully showed how a simple FNN can be implemented to predict the performance of a dynamically operated FC-system, it is limited by the necessity to encode the temporal information in order to deduce dependencies rather than capturing these dependencies directly. While this is sufficient for applications where rather simple load cycles can be expected, for applications exhibiting more complicated load cycles (e.g. applications in mobility) an approach directly capturing the dependencies could be more efficient and better performing. Reservoir computing could be such an approach with minimal computational effort and the possibility for dynamic inputs [27,28]. Successful and reliable degradation predictions of FCs based on the load profile could lead to more economic operation and further implementation of the technology.

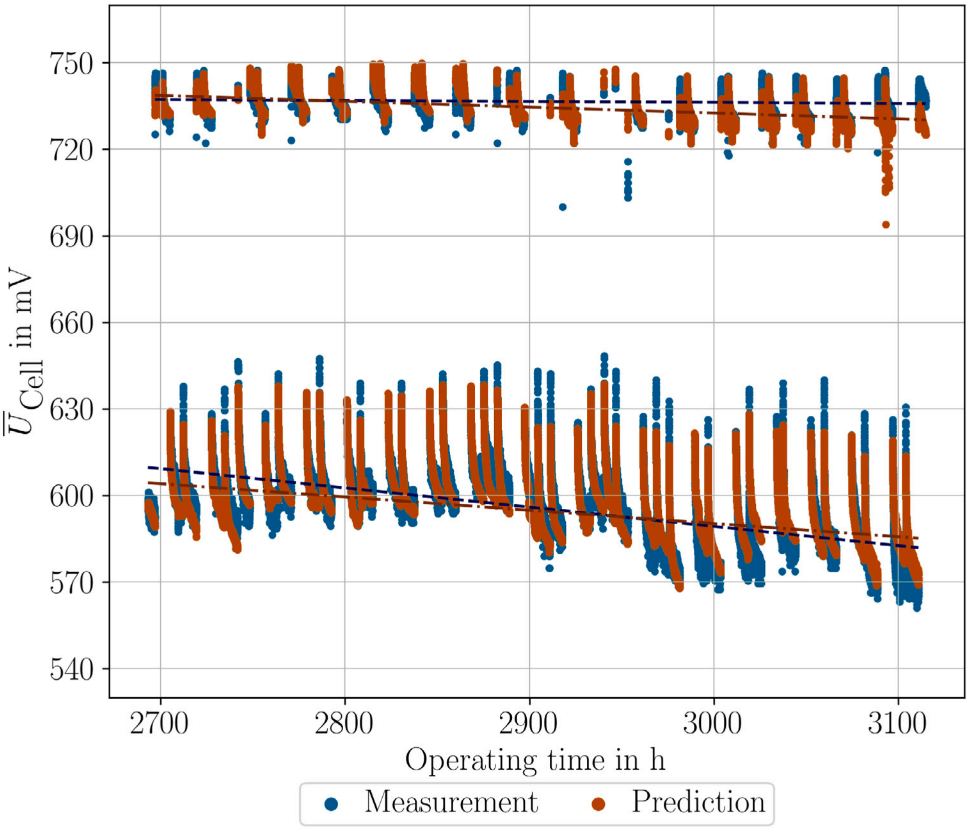


Fig. 15. The model prediction and the measurement data with respect to the operating hours after incorporating gas parameters (the heating value and methane content) into the training. The trendlines show the linear regression of the measurement and the prediction.

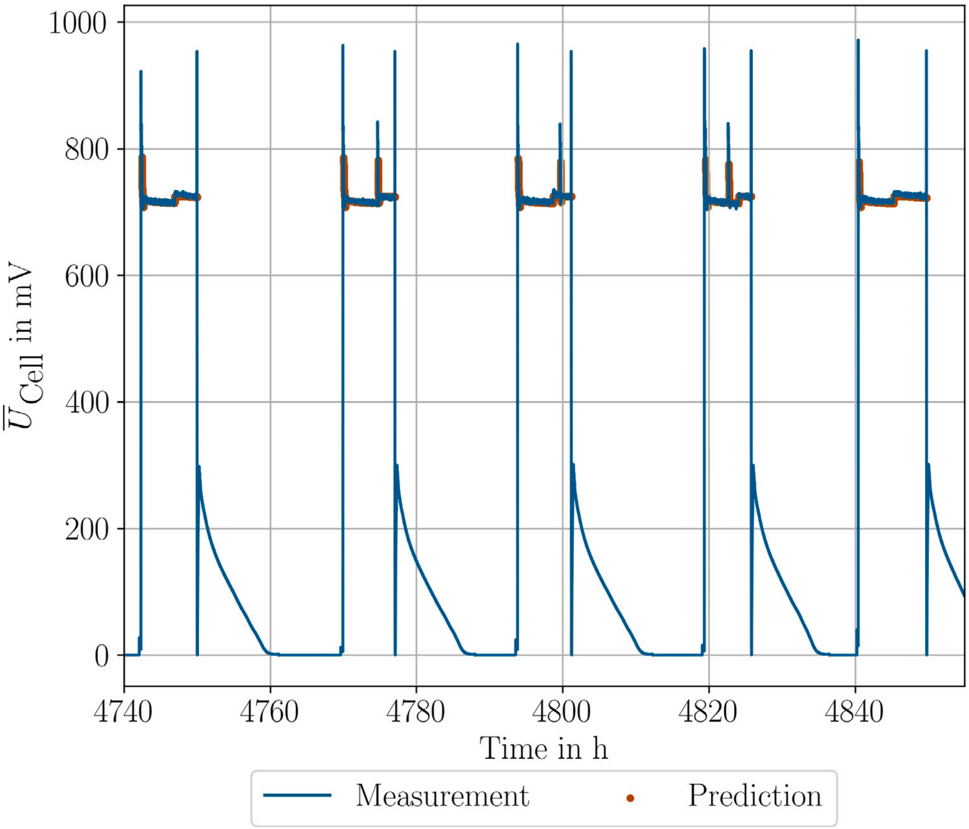


Fig. 16. Application of the model architecture on data of a commercially used system.

CRediT authorship contribution statement

Antonius Tilgner: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Adam Pluta:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Heinz Bekebrok:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Hendrik Langnickel:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Alexander Dyck:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors utilized OpenAI's *ChatGPT* and *DeepL* as programming assistants and in order to improve the style of writing. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

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.

Acknowledgments

We are thankful for the extensive data provided by Schubert GmbH (formerly inhouse engineering GmbH). Especially we are thanking Christoph Hildebrandt, Frank Beckmann, Dr. Janine Ellner and Christoph Schreck for their collaboration. Furthermore, we acknowledge the contribution of Karolin Brucke (German Aerospace Center) and are thankful for the detailed and fruitful discussions in terms of the modeling approach presented in this work. Additionally, the authors thank the German Federal Ministry for Economic Affairs and Energy for financial support within the project SORGLOS-BHKW (reference number 03EN5022C) on the basis of a decision by the German Bundestag.

References

[1] Ishaq H, Dincer I, Crawford C. A review on hydrogen production and utilization: Challenges and opportunities. Int J Hydrog Energy 2022;47(62):26238–64. http://dx.doi.org/10.1016/j.ijhydene.2021.11.149.

[2] Rosen MA, Koohi-Fayegh S. The prospects for hydrogen as an energy carrier: An overview of hydrogen energy and hydrogen energy systems. Energy, Ecol Environ 2016;1(1):10–29. http://dx.doi.org/10.1007/s40974-016-0005-z.

[3] European Commission. A hydrogen strategy for a climate-neutral Europe. Tech. rep. COM(2020) 301 final, European Commission; 2020.

[4] Bundesnetzagentur. Wasserstoff-Kernnetz. Tech. rep., Bundesnetzagentur; 2024. [Accessed: 19 March 2025].

[5] US Energy Department. Overview of CHP technologies. Tech. rep., U.S. Energy Department; 2018.

[6] Yu S, Fan Y, Shi Z, Li J, Zhao X, Zhang T, Chang Z. Hydrogen-based combined heat and power systems: A review of technologies and challenges. Int J Hydrog Energy 2023. http://dx.doi.org/10.1016/j.ijhydene.2023.05.187, S0360319923025284.

[7] Herrmann A, Mädlow A, Krause H. Key performance indicators evaluation of a domestic hydrogen fuel cell CHP. Int J Hydrog Energy 2019;44(35):19061–6. http://dx.doi.org/10.1016/j.ijhydene.2018.06.014.

[8] Daud WRW, Rosli RE, Majlan EH, Hamid SAA, Mohamed R, Husaini T. PEM fuel cell system control: A review. Renew Energy 2017;113:620–38. http://dx.doi.org/10.1016/j.renene.2017.06.027.

[9] Dicks A, Rand DAJ. Fuel cell systems explained. 3rd ed. Hoboken, NJ, USA: Wiley; 2018.

[10] Tullius V. Identifikation CO-induzierter degradationseffekte und entwicklung von regenerationsprozeduren zur steigerung der effizienz von short-stacks auf PEMFC-basis [Ph.D. thesis], Oldenburg: Hochschule Emden Leer; 2019.

[11] Cullen DA, Neyerlin KC, Ahluwalia RK, Mukundan R, More KL, Borup RL, Weber AZ, Myers DJ, Kusoglu A. New roads and challenges for fuel cells in heavy-duty transportation. Nat Energy 2021;6(5):462–74. http://dx.doi.org/10.1038/s41560-021-00775-z.

[12] Liu Z, Xu S, Zhao H, Wang Y. Durability estimation and short-term voltage degradation forecasting of vehicle PEMFC system: Development and evaluation of machine learning models. Appl Energy 2022;326:119975. http://dx.doi.org/10.1016/j.apenergy.2022.119975.

[13] Vichard L, Steiner NY, Zerhouni N, Hissel D. Hybrid fuel cell system degradation modeling methods: A comprehensive review. J Power Sources 2021;506:230071. http://dx.doi.org/10.1016/j.jpowsour.2021.230071.

[14] Park S-K, Choe S-Y. Modeling and experimental analyses of a two-cell polymer electrolyte membrane fuel cell stack emphasizing individual cell characteristics. J Fuel Cell Sci Technol 2008;6(011019). http://dx.doi.org/10.1115/1.2972165.

[15] Das SV, Sanjiv. Deep Learning.

[16] Zuo J, Lv H, Zhou D, Xue Q, Jin L, Zhou W, Yang D, Zhang C. Deep learning based prognostic framework towards proton exchange membrane fuel cell for automotive application. Appl Energy 2021;281:115937. http://dx.doi.org/10.1016/j.apenergy.2020.115937.

[17] Liu J, Li Q, Chen W, Yan Y, Qiu Y, Cao T. Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. Int J Hydrog Energy 2019;44(11):5470–80. http://dx.doi.org/10.1016/j.ijhydene.2018.10.042.

[18] Liu H, Chen J, Hou M, Shao Z, Su H. Data-based short-term prognostics for proton exchange membrane fuel cells. Int J Hydrog Energy 2017;42(32):20791–808. http://dx.doi.org/10.1016/j.ijhydene.2017.06.180.

[19] Zhang T, Hou Z, Li X, Chen Q, Wang Q, Lüddeke C, Wu L, Wu X, Sun W. A novel multivariable prognostic approach for PEMFC degradation and remaining useful life prediction using random forest and temporal convolutional network. Appl Energy 2025;385:125569. http://dx.doi.org/10.1016/j.apenergy.2025.125569.

[20] Ma R, Yang T, Breaz E, Li Z, Briois P, Gao F. Data-driven proton exchange membrane fuel cell degradation predication through deep learning method. Appl Energy 2018;231:102–15. http://dx.doi.org/10.1016/j.apenergy.2018.09.111.

[21] Ko T, Kim D, Park J, Lee SH. Physics-informed neural network for long-term prognostics of proton exchange membrane fuel cells. Appl Energy 2025;382:125318. http://dx.doi.org/10.1016/j.apenergy.2025.125318.

[22] Zerrougui I, Li Z, Hissel D. Physics-informed neural network for modeling and predicting temperature fluctuations in proton exchange membrane electrolysis. Energy AI 2025;20:100474. http://dx.doi.org/10.1016/j.egyai.2025.100474.

[23] Biedermann H. Predictive maintenance: Möglichkeiten und grenzen. In: Biedermann H, editor. Predictive maintenance. 1st ed. TÜV Media; 2018, p. 23–40.

[24] Zhou Z-H. Machine learning. Singapore: Springer; 2021.

[25] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2017, http://dx.doi.org/10.48550/arXiv.1412.6980, arXiv:1412.6980.

[26] Zhang Y, Chen C, Shi N, Sun R, Luo Z-Q. Adam can converge without any modification on update rules. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. In: Advances in neural information processing systems, vol. 35, Curran Associates, Inc.; 2022, p. 28386–99.

[27] Brucke K, Schmitz S, Köglmayr D, Baur S, Räth C, Ansari E, Klement P. Benchmarking reservoir computing for residential energy demand forecasting. Energy Build 2024;314:114236. http://dx.doi.org/10.1016/j.enbuild.2024.114236.

[28] Ma H, Prosperino D, Räth C. A novel approach to minimal reservoir computing. Sci Rep 2023;13(1):12970. http://dx.doi.org/10.1038/s41598-023-39886-w.