Computers and Chemical Engineering 186 (2024) 108681

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

A black-box adversarial attack on demand side management

Eike Cramer [a,b,*], Ji Gao [c,b]

[a] Process Systems Engineering (AVT.SVT), RWTH Aachen University, 52074 Aachen, Germany
[b] Institute for Energy and Climate Research – Energy Systems Engineering (IEK-10), 52428 Jülich, Forschungszentrum Jülich, Germany
[c] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, 30318, USA

ARTICLE INFO

Dataset link: https://transparency.entsoe.eu/

Keywords:
Chemical production
Energy systems
Demand side management
Adversarial attacks
Machine learning

ABSTRACT

Demand side management (DSM) contributes to the industry's transition to renewables by shifting electricity consumption in time while maintaining feasible operations. Machine learning is promising for DSM with reasonable computation times and electricity price forecasting (EPF), which is paramount to obtaining the necessary data. Increased usage of machine learning makes production processes susceptible to so-called adversarial attacks. This work proposes a black-box attack on DSM and EPF based on an adversarial surrogate model that intercepts and modifies the data flow of load forecasts and forces the DSM to result in financial losses. Notably, adversaries can design the data modifications without knowledge of the EPF model or the DSM optimization model. The results show how barely noticeable modifications of the input data lead to significant deterioration of the decisions by the optimizer. The results implicate a significant threat, as attackers can design and implement powerful attacks without infiltrating secure company networks.

1. Introduction

Digitalization and Industry 4.0 initiatives have introduced machine learning and data science into chemical production (Lee et al., 2018; Schweidtmann et al., 2021). The increased usage of machine learning and digitalized decision-making in chemical engineering and production opens up new possibilities for external attackers to intervene and manipulate process operations or alternate process designs. In particular, increased usage of machine learning can lead to errors in human-AI interaction (Wen et al., 2023) but also opens the door to *adversarial attacks* (Koay et al., 2023). Adversarial attacks aim to deteriorate the output of machine learning models to force wrong, inferior, or even dangerous decisions (Xu et al., 2020). Adversarial attacks can designed either using white-box or black-box approaches, i.e., with or without knowledge about the details of the machine learning models under attack. Such adversarial attacks pose a threat to production companies with high levels of automation and machine learning-based decision-making (Koay et al., 2023). Most decision-making in chemical engineering and production is based on data. For instance, demand side management (DSM) shifts production in time to gain financial advantages by trading on the auction-based electricity markets (Zhang and Grossmann, 2016). To make profitable decisions, schedulers and their scheduling optimization problems need data to base their decisions on. If this data, e.g., electricity price forecasts, is compromised in a targeted adversarial manner, the potential ramifications of using this data can be catastrophic.

The field of adversarial attacks originated in computer science, but the increased usage of machine learning has led to interest in engineering as well. In particular, there are a number of works discussing adversarial attacks on cyber–physical systems, fault detection systems, control, and electrical systems. These works include Gomez et al. (2022), who discuss the robustness of anomaly detection models in industrial systems. Zhuo et al. (2023) discuss attack and defense strategies in fault detection and classification systems for the Tennessee Eastman process. They conclude that black-box attacks are as potent as white-box attacks when applied to attack fault detection devices. For control applications, Fazlyab et al. (2022) propose a modified training algorithm for improved model and controller robustness towards adversarial attacks. They propose a semidefinite program for the safety verification of neural networks. Since the work by Chen et al. (2019), there are a number of examples of adversarial attacks in electrical engineering. These examples include inducing grid failures in smart grids (Bor et al., 2019; Cui et al., 2020) and attacks on reinforcement learning for scheduling problems (Zeng et al., 2022; Hao and Tao, 2022). Other use cases like load monitoring (Wang and Srikantha, 2021), solar power forecasts (Tang et al., 2021), and wind power forecasts (Heinrich et al., 2023) have been investigated as well. Heinrich et al. (2023) discuss untargeted and semi-targeted attacks for wind power forecasts using different neural network architectures. They find LSTM models to be comparably robust compared to convolutional

* Corresponding author at: Process Systems Engineering (AVT.SVT), RWTH Aachen University, 52074 Aachen, Germany.
E-mail address: eike.cramer@alumni.tu-berlin.de (E. Cramer).

https://doi.org/10.1016/j.compchemeng.2024.108681
Received 9 November 2023; Received in revised form 5 March 2024; Accepted 6 April 2024
Available online 12 April 2024

neural networks that proved to be very vulnerable. Heinrich et al. (2023) further observe that defensive strategies such as adversarial training (Bai et al., 2021) improve the robustness of neural network-based forecasts but only to a certain extent. In general, adversarial training, i.e., including adversarial data in the training dataset (Bai et al., 2021) appears to be the most common approach to defend against adversarial attacks. However, there are some exceptions, e.g., Maiti et al. (2023) that implement invariant checkers, i.e., Boolean operators that only switch in case of unusual actions, for the defense of cyber–physical systems.

The field of process systems engineering (PSE) has seen increased usage of machine learning (Schweidtmann et al., 2021), but adversarial machine learning and the threat to chemical processes are largely missing from the PSE literature. Some exceptions that allude to adversarial machine learning are Tan and Wu (2024), who enforce a Lipschitz constraint to their neural networks to promote robustness, and Addis et al. (2023), who use adversarial dataset augmentation to diversify their training data to better describe the input–output relation of a membrane.

The field of EPF is well established and receives contributions from economics (Weron, 2014; Weron and Ziel, 2019) and engineering (Lago et al., 2021; Cramer et al., 2023). Notably, recent advances in EPF rely on modern machine learning methods like artificial neural networks to find improved prediction quality (Jedrzejewski et al., 2022). Among machine learning methods, artificial neural networks and time series neural networks such as Long Short-Term Memory (LSTM) models serve as the primary methods (Kapoor and Wichitaksorn, 2023; Trebbien et al., 2023b). The majority of works on forecasting day-ahead electricity prices rely on a sequential forecasting methodology by using autoregressive models (Lago et al., 2021; Bozlak and Yaşar, 2024). Notably, this approach contrasts with the actual process of determining prices in day-ahead bidding markets, where all 24 hourly price intervals are established simultaneously (European Power Exchange, 2021). Multivariate forecasting aligns with the underlying structure of the day-ahead market. Ziel and Weron (2018) compare univariate and multivariate forecasting, noting enhanced performance for multivariate forecasting. Ehsani et al. (2024) use neural network-based multivariate forecasting in a sliding window approach to predict electricity prices in Ontario. In our previous work Cramer et al. (2023), we used multivariate probabilistic forecasting to predict the distributions of intraday electricity prices. Other works in EPF investigate the impact of external factors on the price realizations (Wolff and Feuerriegel, 2017; Trebbien et al., 2023a; Shen et al., 2024). Among the different works on impact factors, Trebbien et al. (2023a) present the most rigorous study relying on explainable artificial intelligence to obtain generalizable conclusions.

Most works researching adversarial attacks in industrial systems consider attacks on the lowest level of operation, e.g., in control. However, adversarial attacks can target higher-level planning and scheduling as well. In this work, we investigate how adversarial attacks influence the solution scheduling problems aiming to find optimal DSM schedules. In particular, we investigate the combined decision-making process of electricity price forecasting (EPF) and DSM. Here, we consider a case where the true decision-making process is hidden from the attackers in a black-box attack scenario. To attack the full process, our proposed attack intercepts the data pipeline with the residual load forecasts and manipulates the data before entering the company network. The manipulated data then deteriorates the decisions aiming to induce financial losses for the company. DSM describes the process of shifting production in time to take advantage of these variable electricity prices (Zhang and Grossmann, 2016). Decisions on production schedules are often made using optimization problems that aim at cost minimization while maintaining process feasibility (Zhang and Grossmann, 2016; Schäfer et al., 2019). Here, knowledge of electricity prices is critical to make feasible and profitable decisions and EPF is an omnipresent task to support DSM. The combined decision-making process first predicts the day-ahead electricity prices based on day-ahead residual load forecasts and then solves a subsequent scheduling optimization problem to determine the trading decisions for DSM. In our black-box case, the attacker has no information about the EPF model or the scheduling optimization model. Instead, we sample historical combinations of the EPF input data and trading decisions and train a simple neural network that works as an emulator of the combined decision-making process. We call the resulting model an adversarial surrogate model (ASM) that we then use to design adversarial attacks.

The effects of the attacks are analyzed for two optimization case studies. The first case study is a linear grid-scale electricity storage problem. The second case study considers a mixed-integer linear problem of a chlorine production plant (Brée et al., 2019). In our evaluation, we benchmark the proposed black-box ASM attack against a white-box attack, where the attacker has full knowledge of the EPF model. Both attacks are implemented using two attack heuristics that follow untargeted and targeted strategies, respectively. Using the untargeted attack, we aim to push the prediction away from the true realization. For the targeted attack, we distinguish between the white-box and black-box attacks. in both cases, we aim to dampen, i.e., flatten, the outputs of the models under attack to prevent the scheduling optimizer from making informed decisions. Here, the white-box attack flattens the EPF prediction and the black-box attack directly targets the trading decisions, as the black-box attacker has no access to the EPF forecasts.

The results show that our adversarial attacks lead to significant financial losses. In case of the grid-scale electricity storage, attacks with small data modifications can turn profitable operations into a loss. Notably, the losses induced by the black-box attack are as high as the losses induced by the white-box attacks, i.e., the black-box attack is as potent as the white-box benchmark. The effectiveness of the black-box attack means that potential attackers can launch powerful attacks without having access to the company network.

The remainder of this paper is organized as follows: Sections 2 and 3 introduce the EPF scheme and the considered DSM processes for this work. Next, Section 4 introduces the concept of adversarial attacks and proposes the white-box and black-box attack heuristics. In Section 5, we show an additional evaluation, where we apply the white-box attacks to investigate the effects on the electricity price forecasts. Section 6 analyzes the effects of white-box and black-box attacks on two DSM case studies. Finally, Section 7 concludes this work.

2. Multi-period electricity price forecasting

The day-ahead electricity market is an auction-based market with 24-hourly trading intervals that are settled simultaneously for all 24 intervals (European Power Exchange, 2021). Thus, the machine learning perspective can view day-ahead electricity prices as 24-dimensional data points and the vector of 24 price values can be predicted using multi-period forecasting approaches (Ziel and Weron, 2018; Lago et al., 2021) that output a 24-dimensional vector of day-ahead electricity prices $\mathbf{P}^{\mathrm{DA}} = [P_t^{\mathrm{DA}}\ \forall t \in [1, 2, \dots, 24]]^T$ based on a set of input features $\mathbf{x}$.

Trebbien et al. (2023a) investigate the feature-importance of external input features for day-ahead price realizations and find that the day-ahead forecasts of the residual loads, i.e., renewable electricity production and load demands, have the most significant impact. Based on their findings, we use day-ahead forecasts of the residual loads as input features, i.e., the day-ahead forecasts for photovoltaic $\mathbf{W}^{\mathrm{DA}}_{\mathrm{PV}}$, wind onshore $\mathbf{W}^{\mathrm{DA}}_{\mathrm{Wind,\ on}}$, wind offshore $\mathbf{W}^{\mathrm{DA}}_{\mathrm{Wind,\ off}}$, and load demand $\mathbf{W}^{\mathrm{DA}}_{\mathrm{Load}}$. Furthermore, the day-ahead electricity prices of the previous day $\mathbf{P}^{\mathrm{DA}}_{\mathrm{prior}}$ are included in the input features. Using a multivariate regression model $\mathbf{T}$, the forecasting problem then reads:

$$\mathbf{P}^{\mathrm{DA}} = \mathbf{T}\left(\begin{bmatrix} \mathbf{W}^{\mathrm{DA}}_{\mathbf{PV}} \\ \mathbf{W}^{\mathrm{DA}}_{\mathbf{Wind,on}} \\ \mathbf{W}^{\mathrm{DA}}_{\mathbf{Wind,off}} \\ \mathbf{W}^{\mathrm{DA}}_{\mathbf{Load}} \\ \mathbf{P}^{\mathrm{DA}}_{\mathrm{prior}} \end{bmatrix}\right) \tag{1}$$

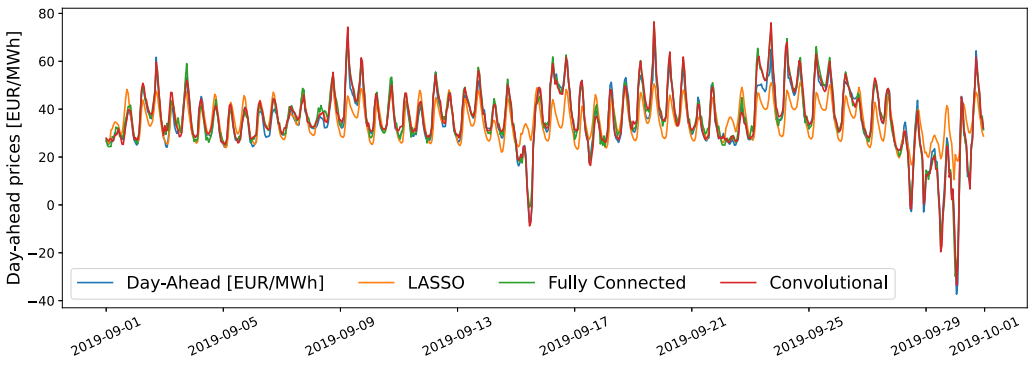

Fig. 1. Predicted day-ahead electricity prices using multi-period LASSO, fully connected, and convolutional regression models. Data for the test month of July 2019.

Table 1
Train and test MSE for the three multi-period forecasting models.

	Train	Test
LASSO	0.42	0.41
Fully Connected	0.02	0.07
Convolutional	0.01	0.05

Multi-period forecasting schemes have proven to be effective for day-ahead forecasting (Ziel and Weron, 2018) and forecasting of electricity prices (Cramer et al., 2023). Furthermore, the multi-period scheme aligns with the simultaneous realization of the day-ahead electricity prices (Lago et al., 2021; European Power Exchange, 2021). This work uses three different model architectures to perform the multi-period EPF. The three models are a LASSO regression model (Ziel et al., 2015), i.e., a linear model with $l1$-penalties on the scale factors, a fully connected neural network, and a convolutional neural network. All models are implemented using the python-based machine learning library *TensorFlow* (Abadi and Agarwal, 2015) trained on data from January 2019 to December 2020. The month of September 2019 is set aside as a test set.

Table 1 lists the training and test losses for the three different models. The two neural networks achieve the most accurate results with train and test losses about an order of magnitude lower than the LASSO model. The test losses for the neural networks indicate low levels of overfitting. However, the test losses still outperform the LASSO model.

Fig. 1 shows the day-ahead electricity prices in comparison to the predicted prices from the multi-period LASSO, fully connected, and convolutional regression models for the test month of July 2019. All models recover the general trends of the day-ahead electricity prices. The two neural networks show accurate results, including days with high or low price peaks. The LASSO model fails to give accurate predictions for days with peak behavior which is resulting from the LASSO penalty restricting the model's output to be dominated by the bias vector.

For more detailed information on the different models and their training, including 5-fold cross-validation, see Appendix A.

3. Scheduling optimization problems for DSM

Many industrial processes offer a certain amount of flexibility that allows their operation to shift to times with lower electricity prices. DSM uses mathematical modeling and optimization to find operation schedules for industrial processes such that the costs for resources like electricity are minimized (Zhang and Grossmann, 2016). This Section briefly introduces the two case studies considered in this work. The first case study called 'Storage Trader' considers the scheduling of grid interactions of a grid-scale electricity storage. The second case study investigates a Mode Switching Chlorine production based on the work by Brée et al. (2019). Both DSM problems are implemented using the python-based modeling library *pyomo* (Hart et al., 2017) and solved using the *gurobi* optimization software (Gurobi Optimization, LLC, 2023).

3.1. Storage trader

First, we consider the optimal scheduling of the charging and discharging actions of grid-scale electricity storage referred to as the 'Storage Trader' problem. This Storage Trader problem is a linear optimization problem that maximizes profits by solving for optimal operation based on electricity price forecasts for the day-ahead period. Due to the linearity of the problem, any changes to the input data, e.g., under adversarial attacks, directly translate to changes in the decisions by the optimizer. Thus, the Storage Trader problem allows for an unbiased interpretation of the results.

The storage can store electricity for short periods of time such that the operators can charge during low-price hours and discharge during high-price hours. The storage is a grid-scale battery storage with a storage capacity of 1.200 MWh (Colthorpe, 2021). The maximum charging and discharging rates are 300 MW. The linear optimization problem reads:

$$
\begin{aligned}
\max_{W_t^{out}, W_t^{in}} \; & = \sum_{t=1}^{24} P_t^{\mathrm{DA}} \left(W_t^{out} - W_t^{in} \right) \Delta t \\
\text{s.t.} \quad & SOC_t = SOC_{t-1} - \frac{1}{\eta} W_t^{out} + \eta W_t^{in} \\
& 0 \le SOC_t \le SOC^{max} \\
& SOC_{t=24} = SOC_{t=0} \\
& 0 \le W_t^{out} \le W^{max} \\
& 0 \le W_t^{in} \le W^{max}
\end{aligned}
\qquad \text{(ST-DSM)}
$$

In Problem (ST-DSM), SOC_t is the state of charge at the tth hour, W_t^{in} and W_t^{out} are the charging and discharging rates, respectively, and $\Delta t = 1\,\mathrm{h}$ is the duration of one day-ahead trading interval. W^{max} is the maximum rate for charging and discharging. The formulation in Problem (ST-DSM) aims to maximize the profits achieved by buying and selling electricity at different points in time. The formulation includes a cyclic constraint to prevent full discharging of the storage at the end of the day. The initial and final state of charge are at 30% of the maximum, and the efficiency η is 90% (Colthorpe, 2021).

3.2. Mode switching chlorine

The Mode Switching Chlorine case study considers the scheduling of a chlorine production plant. The objective of the problem aims to minimize the cost associated with chlorine production. Brée et al. (2019) proposed the model and subsequent scheduling formulations. The plant is designed to switch between two operating modes. The two operating modes are operated via the cathodes used for chlorine

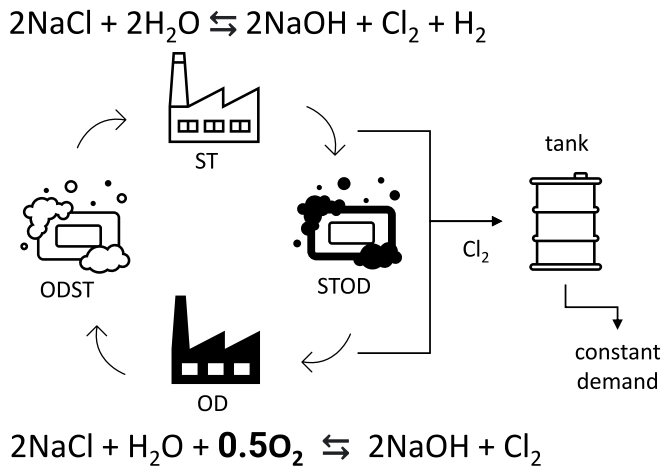

Fig. 2. Mode switching chlorine production with constant chlorine demand (Brée et al., 2019). Operative modes switch between standard cathodes (ST) and oxygen-depolarized cathodes (OD).

oxidation. The standard cathode (ST) uses only sodium chloride and water as educts and produces chlorine and hydrogen that can be sold as a side product. Meanwhile, the oxygen-depolarized cathode (OD) does not produce the hydrogen byproduct and requires additional oxygen for the reaction but uses less electricity per kg of chlorine.

Fig. 2 shows a sketch of the mode-switching operation. The switching between modes is modeled using binary variables. Thus, the Mode Switching Chlorine case study is a mixed-integer linear problem (MILP). Changes to the integer decisions of an optimization problem often indicate major changes in the operation. This makes MILP scheduling optimization particularly sensitive to disturbances. Furthermore, MILPs are prevalent in scheduling optimization and, thus, the Mode Switching Chlorine problem is considered as case study under adversarial attack. For details on the process and its implementation, the reader is referred to the original publication by Brée et al. (2019).

4. Adversarial attacks

Adversarial machine learning describes the field of malicious modifications of machine learning models and their predictions. Prevalent adversarial attacks can be distinguished in *poison* and *evasion* attacks (Demontis et al., 2019). Poison attacks aim to change the training of machine-learning models (Jagielski et al., 2018; Şuvak et al., 2022) and evasion attacks modify input data to fixed models to change the outputs (Biggio et al., 2013). Evasion attacks typically use noise patterns that are added to the input data of a machine-learning model. For instance, the fast gradient sign method (FGSM) (Goodfellow et al., 2015) and its extension to the basic iterative method (BIM) (Kurakin et al., 2018) design adversarial noise using model sensitivities. Such evasion attacks assume readily trained machine-learning models with nonpermutable parameters. In this article, we restrict our analysis to evasion attacks. For simplicity, we use the term adversarial attacks to refer to evasion attacks.

In Section 4.1, we propose two heuristic attack targets and formulate optimization problems to compute the adversarial data modifications. Section 4.2 reviews the FGSM (Goodfellow et al., 2015) and shows how FGSM can compute approximate solutions for the attack design problems for the two heuristic attacks. Finally, Section 4.3 discusses the options for white-box and black-box attacks in the context of EPF in combination with DSM optimization problems.

4.1. Heuristic attack design

Adversarial attacks can either aim to force the model output away from the true realization or to point the prediction toward a specific output. These two approaches reflect heuristic attack designs and are called untargeted and targeted attacks, respectively. Both heuristics assume a trained machine learning model that can be evaluated at will and that also provides gradient information. The modifications to the input data are achieved by adding an adversarial noise pattern to the input data $\mathbf{x} + \Delta\mathbf{x}$. Given a trained machine learning model $\mathbf{T}(\mathbf{x})$ with inputs $\mathbf{x}$ and outputs $\mathbf{y} = \mathbf{T}(\mathbf{x})$, the design of this adversarial noise can be formulated as an optimization problem.

Our untargeted attack aims to push the prediction away from the actual realization $\mathbf{y}^*$. In mathematical terms, we aim to maximize a loss function $\mathcal{L}(\mathbf{x}, \mathbf{y})$ that describes the change to the model (EPF or ASM) output. Meanwhile, we constrain the changes to the input data to small values. A typical choice of objective is the mean-squared-error (MSE) loss (Kurakin et al., 2018):

$$\max_{\Delta\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D} \left(y_i^* - \mathbf{T}(\mathbf{x} + \Delta\mathbf{x})_i\right)^2 \quad \text{s.t.} ||\Delta\mathbf{x}||_p \leq \delta \tag{U-A}$$

In Problem (U-A), D is the dimensionality of the outputs. The p-norm $||\cdot||_p$ of the adversarial noise $\Delta\mathbf{x}$ is limited to a small number δ such that the attack remains difficult to detect. For a given day, the input features $\mathbf{x}$ are constant.

Targeted attacks require a specific target. For the application of DSM, we propose using the historical mean as the target based on the following intuition: DSM applications utilize process flexibility to shift production towards times with lower electricity cost (Zhang and Grossmann, 2016). If the electricity price forecast is flat, the optimizer is unable to find accurate production profiles, which alleviates the advantages gained from DSM. In the black-box case, this intuition extends to the decision by the optimizer as a flat decision profile fails to take advantage of the variable electricity prices. In the following, we refer to this attack heuristic as the dampen attack:

$$\min_{\Delta\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D} \left(\mu_y - \mathbf{T}(\mathbf{x} + \Delta\mathbf{x})_i\right)^2 \quad \text{s.t.} ||\Delta\mathbf{x}||_p \leq \delta \tag{T-A}$$

Problem (T-A) describes the optimization problem for the dampen attack that aims to minimize the MSE between the model output and the historical mean μ_y.

Both attack heuristics apply to the white-box and the black-box cases discussed in this work. In the white-box case, the attack aims to change the electricity price forecasts by the machine learning model $\mathbf{T}$. In the black-box case, the attack aims to change the decisions made by the DSM optimizer via an attack on the ASM $\mathbf{ASM}(\mathbf{x})$.

4.2. Fast gradient sign method

FGSM describes an efficient approach to compute adversarial noise $\Delta\mathbf{x}$ (Goodfellow et al., 2015). Using the loss functions of Problem (U-A) and Problem (T-A), FGSM computes the gradient, i.e., the direction of maximum change in the output, to compute a gradient *ascent* step or *descent* step, respectively. The gradient is normalized via the *sign* function to control the intensity of the attack. With $\mathcal{L}(\mathbf{x}, \mathbf{y})$ as the MSE objective of Problem (U-A) or Problem (T-A), the formula to compute the adversarial noise $\Delta\mathbf{x}$ then reads:

$$\Delta\mathbf{x} = \varepsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \hat{\mathbf{y}})\right) \tag{2}$$

Here, $\mathbf{x}$ are the model inputs, and $\hat{\mathbf{y}}$ is a target value of the outputs $\mathbf{y}$. Eq. (2) describes a step towards the solutions of the optimization problems proposed in Section 4.1. Thus, the attack rate ε is equal to the noise limitation δ if the infinity norm is used:

$$||\Delta\mathbf{x}||_\infty = \varepsilon \cdot \underbrace{||\text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \hat{\mathbf{y}})\right)||_\infty}_{=1} = \delta$$

In regression tasks, the true realization $\mathbf{y}^*$ is unknown to both the operators and the adversaries. Thus, the untargeted attack cannot be computed as shown in Problem (U-A). Instead, the model output $\mathbf{y} = \mathbf{T}(\mathbf{x})$ can be used as a proxy for the true realization. However, using the model output yields a null gradient for the MSE loss function. Thus, the model output is inflated with white noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to compute an untargeted attack pattern. The formula to compute the untargeted adversarial noise then reads:

$$\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{T}(\mathbf{x}) + \epsilon)\right) \tag{3}$$

The dampen attack heuristic described in Problem (T-A) is a minimization problem and, thus, the FSGM describes a gradient descent step:

$$\tilde{\mathbf{x}} = \mathbf{x} - \varepsilon \cdot \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mu_{\mathbf{y}})\right) \tag{4}$$

In the following, the adversarial attacks using feature perturbations based on Eqs. (3) and (4) are called the untargeted and the dampen attack, respectively.

4.3. White-box and black-box attacks on DSM

The attack heuristics proposed in Section 4.1 and the FGSM-based attack designs from Section 4.2 all require readily trained machine learning models to compute gradients. The white-box case in this work attacks only the EPF model. Thus, the white-box scenario assumes the EPF model $\mathbf{T}(\mathbf{x})$ to be known. Hence, the white-box attacks aim to change the electricity price forecasts such that the scheduling optimizer is led towards suboptimal decisions. The white-box attacks discussed in this work only consider the EPF model without considering the reaction of the scheduling optimizer used for DSM. In real applications, the attacker has no knowledge about the machine learning model and companies publish neither internal electricity price forecasts nor DSM optimization models to maintain a competitive advantage. Thus, the black-box case starts in a position without any machine-learning model to compute gradients and design the attacks. Instead, attackers have to rely on black-box attack designs that typically rely on surrogate models to compute data perturbations. These surrogate models describe the same process and, thus, emulate the behavior of the actual machine learning model. The attackers obtain such surrogate models by sampling input and output data and training their own separate machine-learning models describing the same process, i.e., the same input–output relation.

For the combined decision-making process of EPF and DSM the standard black-box case does not apply as the internal electricity price forecasts are unknown, i.e., the output data of the EPF machine learning model is unattainable to the attackers. Instead, we propose to extend the black-box case to include both EPF and DSM. Here, the attacker only has access to the input data to the EPF models, i.e., the residual load forecasts, see Section 2, and the decisions made by the DSM scheduling optimizer, i.e., the grid interaction. The black-box surrogate model then is a shortcut model that maps from the residual load forecasts directly to the trading decisions. Both of these data streams are exchanged with external entities, which opens the possibility for attackers to gain access to the data and the data pipelines without needing to infiltrate the company boundaries. In fact, the historical residual load forecasts are published online, e.g., on the ENTSO-E Transparency Platform (2022). The proposed ASM to emulate the full decision-making process of EPF and DSM reads:

$$\mathbf{W}^{\text{grid}} = \mathbf{ASM}(\mathbf{x}) \tag{5}$$

In Eq. (5), $\mathbf{ASM}(\mathbf{x})$ is the ASM and $\mathbf{W}^{\text{grid}}$ is the vector of grid interactions, e.g., W_t^{out}, W_t^{in} in the Storage Trader problem, see Problem (ST-DSM). In practice, the ASM is implemented as a simple fully connected neural network. Note that the requirements for accuracy of the ASM are comparatively low, as the ASM is not used for decision-making. Instead, the ASM's only purpose is to provide the gradient information to compute the adversarial noise using the FGSM as described in Eq. (2).

Fig. 3 shows a sketch of the black-box attack heuristic proposed in Eq. (5). The attacker intercepts the information streams for the residual loads and the grid interaction, i.e., the buy and sell decisions. Then, the attacker trains the ASM on the collected data. Finally, the trained ASM is used to design data modifications to the input features that enter the company boundaries using either the targeted or the untargeted attack heuristics proposed in Section 4.1. Note that the proposed black-box scheme rests on the assumption that the attackers have access to the databases of external entities or data pipelines to obtain the historical residual load forecasts and the historical trading decisions. The exact action of how these data pipelines are intercepted is beyond the scope of this study.

5. Adversarial attacks on multivariate electricity price forecasting

This Section investigates the effects of adversarial attacks on day-ahead EPF. We include this Section to complete the analysis of our attack designs. Readers only interested in the effects of the ASM-based black-box attack may skip ahead to the next Section.

5.1. Error metrics of attacked forecasts

We apply the heuristic attack methods proposed in Section 4.1 to the three forecasting models discussed in Section 2 for the test month of September 2019. Both the untargeted (Eq. (3)) and the dampen (Eq. (4)) attacks are applied in the white-box format with increasing attack rates ε between 0 and 0.3. Fig. 4 shows the mean-squared-error (MSE) and the mean-absolute-percentage-error (MAPE) for the LASSO regression, the fully-connected neural network, and the convolutional neural network, respectively. Both MSE and MAPE are computed using the unperturbed forecasts $\mathbf{y} = \mathbf{T}(\mathbf{x})$ as the true labels and perturbed forecasts $\tilde{\mathbf{y}} = \mathbf{T}(\mathbf{x}+\Delta)$ as the predictions, i.e., the values in Fig. 4 describe the impact of the attacks on the forecasts independent of any existing forecast errors. The results for the MSE in Fig. 4 show an exponential increase of the error metrics with the attack rate for both the fully connected and the convolutional forecasting models. The untargeted attack leads to larger errors for the fully connected neural network and the dampen attack leads to larger errors in the convolutional neural network. Meanwhile, the LASSO regression model remains at low MSE values even for high attack rates and for both attack heuristics. The MAPE shows higher values for the convolutional neural network compared to the fully connected neural network. The MSE places high penalties on outliers with the square function while the MAPE considers the absolute errors. In other words, the fully connected neural network predicts more outliers while the absolute error induced by the attacks is higher for the convolutional neural network. Overall, the MSE and MAPE values are similar for the untargeted and the dampen attack. Thus, the error metrics indicate no superior attack heuristic.

The LASSO regression has fewer parameters compared to the two neural networks, which decreases the susceptibility to modified data. Furthermore, the scale parameters of the LASSO model are regularized via *l1* penalties which reduces their overall impact on the forecast and places a higher focus on the bias vector. In fact, the results in Fig. 4 indicate that the bias vector dominates the LASSO forecasts. Thus, the LASSO regression is the least vulnerable to adversarial attacks. The fully connected and convolutional neural networks show high susceptibility to adversarial attacks. Table 1 shows higher loss values for the LASSO model compared to the two neural networks. Apparently, there is a trade-off between model accuracy and robustness toward adversarial attacks. In summary, Fig. 4 shows that adversarial attacks can lead to significant changes in the day-ahead electricity price forecasts.

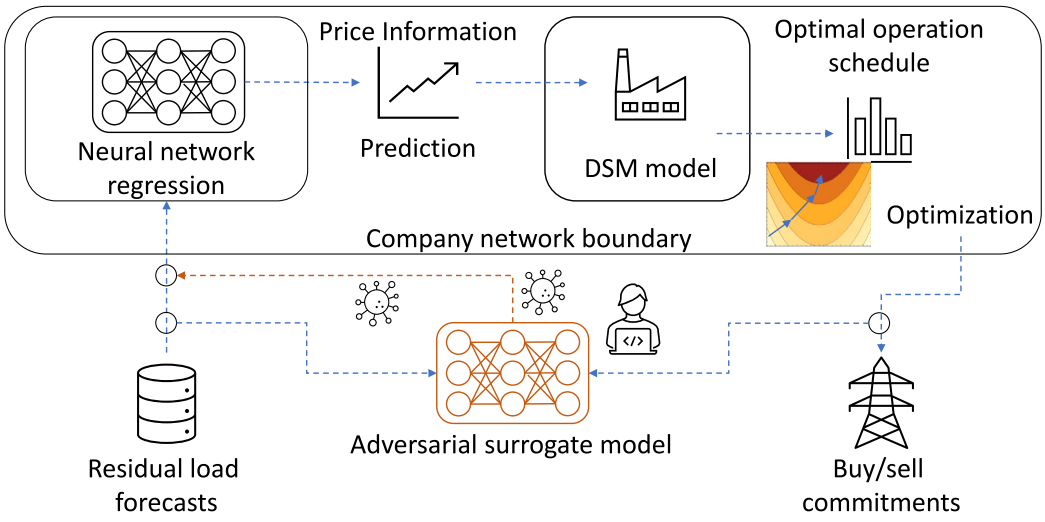

Fig. 3. Adversarial surrogate model (Eq. (5)) for black-box attacks on the full decision-making process.

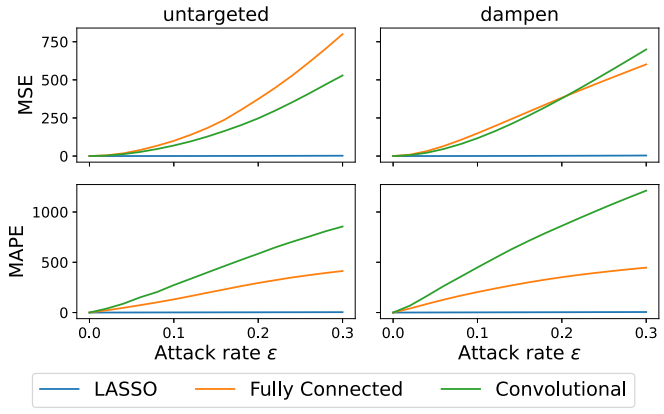

Fig. 4. Mean-squared-error (MSE) and mean-absolute-percentage-error (MAPE) for LASSO Regression, Fully Connected, and Convolutional regression models and for different values of the attack rate ϵ, respectively. The columns show the untargeted and the dampen attack designs derived in Section 4.1. MSE and MAPE are computed using the unperturbed prediction as the label. Metrics computed for the test month of September 2019. Visualization using *seaborn* (Waskom, 2021).

5.2. Perturbation of input features

Besides causing high values in error metrics, the adversarial attack should be difficult to detect, i.e., the changes to the input data should be minimal and difficult to notice by the human eye. Fig. 5 shows the input and output data for the convolutional neural network on September 3rd, 2019 and for different attack rates using the dampen attack heuristic.

For low attack rates of $\epsilon < 0.05$, the changes to the input data are barely noticeable visibly. For higher attack rates, the input data becomes noisy. In particular, the nighttime hours of the solar day-ahead forecasts show uncharacteristic fluctuations and the wind forecasts show discrete jumps with high frequency. To make the attacks even more difficult to detect, the solar nighttime hours can be constrained to zero. Trebbien et al. (2023a) observe that the day-ahead wind and load forecasts have the largest impact on the day-ahead electricity price. Fig. 5 reflects this observation as the values for wind onshore and offshore as well as load show the largest perturbations.

The last row in Fig. 5 shows the forecasts for the day-ahead electricity prices. There is a noticeable change in the outputs even for low attack rates that show hardly visible perturbations to the input data. For high attack rates over 0.05, the day-ahead price prediction loses its typical structure.

In conclusion, the adversarial attack heuristics proposed in Section 4.1 induce significant changes to the outputs of multivariate forecasting models by using small and barely noticeable changes to the input data. Of the three considered models, the linear LASSO regression is the least susceptible to adversarial attacks. The vulnerability of the two neural networks is significantly higher compared to the LASSO regression. The results of MSE and MAPE evaluation do not indicate either heuristic to be superior.

6. Adversarial attacks on demand side management

This Section investigates the effects of the adversarial attacks on DSM, i.e., downstream decision-making problems. Section 6.1 investigates white-box attacks with full knowledge of the EPF models, and Section 6.2 investigates the black-box attack proposed in Section 4.3. Finally, Section 6.3 discusses the results and draws general conclusions.

6.1. White-box attacks

For each day in the test month of September 2019, the full decision-making processes of the Storage Trader and the Mode Switching Chlorine case studies are solved for the untargeted and dampen attack heuristic proposed in Section 4.1. First, the modified data is fed into the forecasting model. Then the outputs are used as parameters to solve the downstream optimization problems. In the next step, the scheduling decisions are applied to the process and the actual profits and actual costs are computed for the Storage Trader and the Mode Switching Chlorine case studies, respectively. In other words, the actual profits and costs are computed for the true realizations of the day-ahead electricity prices.

Fig. 6 shows distribution intervals of the actual profits and the actual costs obtained for the Storage Trader and the Mode Switching Chlorine case studies as well as the untargeted and the dampen attack heuristics, respectively. The distribution intervals are estimated over the results of each day in the test month. In all cases, the attack rates vary between 0 and 0.3 with 0.02 intervals. In all four cases shown in Fig. 6, the adversarial attacks lead to a worsening of the respective objectives for the forecasts by the two neural networks. Namely, the Storage Trader returns lower profits, and the Mode Switching Chlorine plant has a higher cost. The profits in the Storage Trader case study regress significantly and even reach negative values for attack rates over 0.2 for the dampen attack heuristic. Meanwhile, the profits

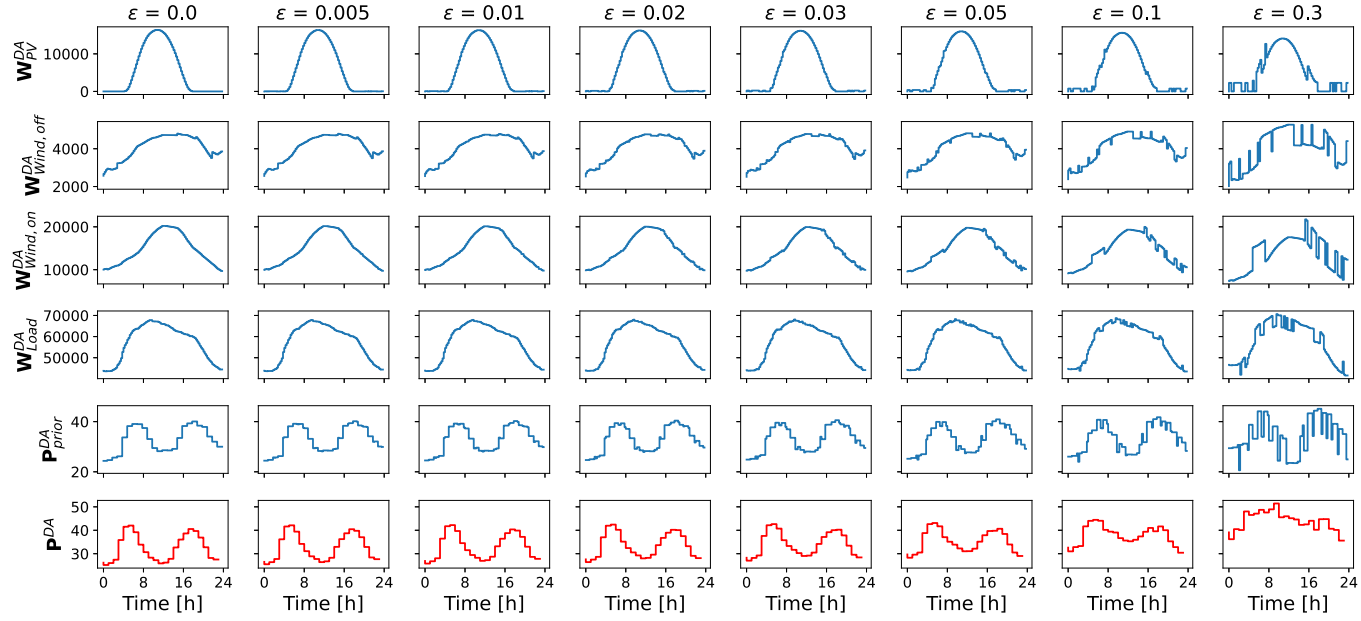

Fig. 5. Changes to inputs and outputs of the convolutional neural network induced by the dampen attack (Eq. (3)) for different attack rates. Inputs are depicted in blue, and outputs are depicted in red. Data shown for September 3rd, 2019. Power values in [MW] and price values in [EUR/MWh].

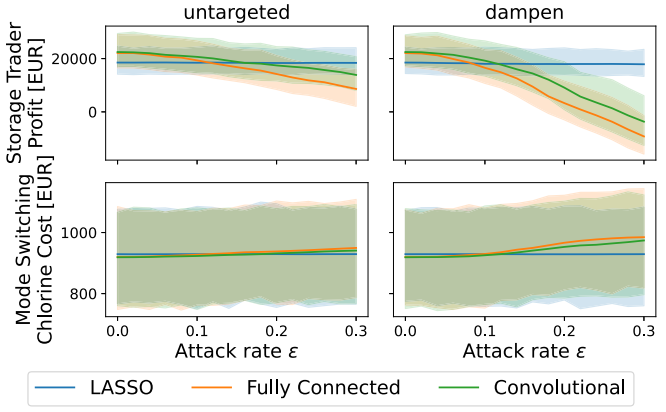

Fig. 6. White-box attacks: Profits [EUR] and costs [EUR] obtained in the Storage Trader (Problem (ST-DSM)) and Mode Switching Chlorine (Brée et al., 2019) DSM problems, respectively, for different attack rates ϵ. The figure shows average and distribution intervals estimated from profits in the test month of September 2019. Visualization using *seaborn* (Waskom, 2021).

achieved using the LASSO forecasts do not regress with increasing attack rates. The Mode Switching Chlorine case study shows small increases in costs for the untargeted attack and moderate increases in cost for the dampen attack in the case of both neural networks. Again, there are no visible changes for the attacks on the linear LASSO regression. Notably, the Storage Trader case study shows that the actual profits for the linear LASSO regression without attacks are lower than the two neural networks. The differences between the results for the different forecasting models confirm the observations in Section 5.1, where the LASSO forecasts do not change significantly with the attacked inputs. Furthermore, this confirms the observation that there is a trade-off between prediction accuracy and robustness towards adversarial attacks.

Of the two attack heuristics, the dampen heuristic leads to significantly more severe losses. For instance, the dampen attack leads to a strong decrease in profits in the Storage Trader case study and even negative profits for attack rates $\epsilon \geq 0.2$. Such negative profits indicate that the optimizer decided to buy during high-price hours and sell during low-price hours. As the attacks take a single gradient ascent step, a high attack rate may lead to an overshoot of the flat profile target and instead lead to a reversed prices profile. Meanwhile, the untargeted attack does not achieve negative profits for the considered attack rates. The actual cost obtained for the Mode Switching Chlorine case study shows similar differences between the two attack heuristics, where the dampen attack leads to a stronger increase in cost compared to the untargeted attack heuristic. For the Mode Switching Chlorine case study, the average losses lie within the typical range of cost values. It appears that the integer decisions made by the optimizer do not change significantly with the attack. Note that other MILP problems may react differently to adversarial attacks.

As intended by the intuition, the dampen attack pushes the decisions to a flat profile. Thus, the optimizer results in constant decisions that contradict the actual market and its fluctuations. As a result, the losses are comparatively high. Meanwhile, the untargeted attack heuristic forces the predictions away from the true value. Here, the direction of these perturbations is random due to the white noise that is added to the true predictions (see Eq. (3)). Hence, the effects of the untargeted attack on the scheduling decisions are difficult to estimate. Overall, the losses induced by the untargeted attack are comparatively low.

6.2. Black-box attacks

This Section applies the black-box attack proposed in Section 4.3 to attack the two DSM case studies. This black-box study assumes that the attacker has access to the input features and the trading decisions made by the DSM optimizer. For this work, we assume that the attacker has observed the trading decisions over the years of 2019 and 2020 and has collected both the residual load forecasts and the trading decisions over time. Then, the ASM is trained on the recorded residual load forecasts and the trading decisions. The trained ASM is then used to generate adversarial noise that is subsequently added to the input features of the actual decision-making process of forecasting and DSM optimization. Again, September 2019 is set aside as a test set.

Fig. 7 shows the actual profits and the actual costs in the case of the black-box attacks obtained for the Storage Trader and the Mode Switching Chlorine case studies, respectively. The columns of Fig. 7 show the results for the untargeted and the dampen attack, respectively. The attack rates are increased between 0 and 0.3 with equidistant steps

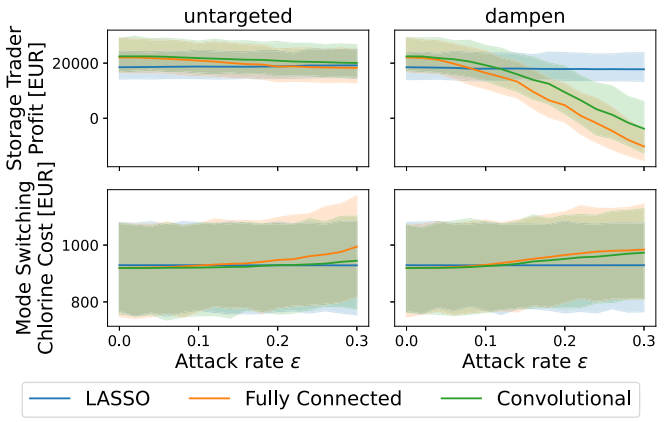

Fig. 7. Black-box attacks: Profits [EUR] and costs [EUR] obtained in the Storage Trader (Problem (ST-DSM)) and Mode Switching Chlorine (Brée et al., 2019) DSM problems, respectively, for different attack rates ϵ. The figure shows average and distribution intervals estimated from profits in the test month of September 2019. Visualization using *seaborn* (Waskom, 2021).

of 0.02. The profits and cost shown in Fig. 7 highlight that the black-box attack proposed in Section 4.3 can induce significant damage to the considered scheduling problems. In particular, the black-box dampen attack leads to similar damages as the dampen attack in the white-box case leading to negative profits for the Storage Trader with attack rates as small as 0.2. Meanwhile, the untargeted black-box attack leads to comparatively small changes to profits and costs in the Storage Trader and the Mode Switching Chlorine case studies, respectively. The losses observed in Fig. 7 mirror the results for the white-box attack shown in Fig. 6. Again, the dampen attack can flatten the profile of the scheduling decisions leading to significantly higher losses compared to the untargeted attack heuristic. The similarity between the effects of the white-box and the black-box attacks shows that the ASM is able to emulate the decision-making process sufficiently well to design attacks on the real system. Please note that Zhuo et al. (2023) have observed similar results in their study, where they found their black-box attacks to be as potent as white-box attacks. The ASM is used to design the adversarial noise via the FGSM algorithm in Eq. (2), where the ASM only provides gradient information of the decision-making process. Thus, the results for the black-box attack show that the ASM emulates the behavior of the EPF and DSM decision-making process to an extent that suffices to design potent attacks. Note that the FGSM applies the *sign* function to the gradients. Thus, the results only show that the gradients computed via the ASM yield attack designs that are competitive with white-box attacks. However, the results do not necessarily indicate that the ASM emulates the decision-making process perfectly in the sense that it could be used to replace it.

6.3. Discussion

The results in Figs. 6 and 7 show that adversarial attacks can lead to significant deterioration of the decisions made in a scheduling optimization. Of the two attack heuristics, the dampen attack shows the higher potential as it leads to losses in both case studies as well as both white-box and black-box cases. It appears that targeting a dampening of the trading decisions directly is a highly potent attack strategy that leads to higher financial losses than the untargeted attack. The dampen attack cancels the peaks in the price time series alleviating the option to plan for those peaks. Considering that the error metrics in Fig. 4 show roughly equal potency for both attack heuristics, error metrics like MSE and MAPE do not indicate the attack's potential to change the decisions of a downstream optimizer.

In all considered attacks on the EPF model alone and on the combined decision-making process, the LASSO model shows low susceptibility to adversarial attacks. This robustness is a result of the regularization of the scale parameters in the LASSO concept. The result is a forecasting model that is dominated by its bias term which is unaffected by the attacks. While this reliance on the bias term supports adversarial robustness, the LASSO model is also the least accurate EPF model and leads to the lowest profits or rather cost in DSM. In conclusion, the regularization of scaling parameters appears to enhance adversarial robustness. This is expected as large scaling factors can amplify minor changes in the input data.

From the attacker's perspective, the black-box ASM suffices to induce flat decision profiles that do not fit the profile of the day-ahead prices. In particular, there is no difference between losses induced by the white-box and the black-box dampen attacks. For real-world attack scenarios, this implies that attackers do not need to infiltrate a company network to implement adversarial attacks. Thus, the security and protection of the EPF model and the scheduling optimization model are insufficient to protect a company against adversarial attacks. Instead, the EPF model inputs, i.e., information typically obtained from outside sources, are the most vulnerable part of the decision-making process.

The black-box ASM is trained on almost two years of data which is a long time for the attackers to observe. With less data available, the emulation of the ASM will be worse. Whether an ASM trained on less data leads to less potent attacks is difficult to say and an investigation of different observation horizons is beyond the scope of this work.

7. Conclusions and open questions

This work highlights the potential of adversarial attacks to deteriorate the decisions made in the combined decision-making process of EPF and DSM. In particular, we propose a black-box attack strategy paired with heuristic attack targets that aim to deteriorate the decisions made by the optimizer. In two case studies of a grid-scale electricity storage and a chlorine production plant, the quantitative analysis shows how small modifications of input data can lead to significant losses. Notably, the results show that attackers do not need to know the EPF model or the scheduling optimization model to design effective attacks that lead to significant financial losses. In fact, attackers do not need access to the network of the company under attack. Instead, sampling historical data of residual load forecasts, i.e., the input data, and subsequent trading decisions by the optimizer, i.e., the outputs of the decision-making processes, is sufficient to fit an adversarial surrogate model that can be used to design adversarial noise for data manipulation.

For industrial applications, the success of the black-box attack presents a substantial risk. Effective attacks can be designed by using historical data of day-ahead residual load forecasts and the trading decisions by the company, i.e., using data that is stored in third-party databases outside of company networks. Furthermore, the tactic of evasion attacks is difficult to detect as they only marginally change the data, and no software or virus has to infiltrate the company network. Thus, established protection against cyber-attacks needs to be augmented by anomaly detection algorithms and other advanced security measures.

The investigation in this work presents a first step to discovering the potential of adversarial attacks on DSM and other scheduling problems. Notably, the discussed attack approach is just one of possibly numerous other attacks that could interfere with decision-making tasks like DSM. The possibilities for adversarial attacks in chemical engineering are far from understood and there remains a substantial risk for production companies that are oblivious to the issue.

Future investigations should consider alternative attack heuristics and utilize advanced methods to compute adversarial noise. White-box attacks could be used to engineer optimal attacks that are specifically

designed to worsen the objective instead of relying on heuristics like the dampen attack. Furthermore, the proposed black-box attack strategy relies on fitting a surrogate model for the unknown decision-making processes. Artificial neural networks like the one used in this work need a lot of data, and data availability is often limited. Alternative surrogate models like Gaussian Processes could yield comparable performance using significantly fewer samples.

Another critical avenue of future investigations is a defensive strategy to protect against adversarial attacks. Possible defensive strategies include regularization, adversarial training (Bai et al., 2021), and output constraining (Tan and Wu, 2024) of the EPF models to enforce robustness towards modified samples. An alternative or complementary approach is fault detection, where the predicted electricity prices and the DSM decisions are evaluated against historical data to decide if they follow typical patterns. Data that breaks with established patterns is likely to be compromised.

In summary, this work opens numerous questions in adversarial attacks and defense for the safe operation of industrial processes. The results shown in this work open many directions for further investigations in machine learning in chemical engineering, process operation, and safety considerations.

CRediT authorship contribution statement

Eike Cramer: Visualization, Writing – original draft, Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation. **Ji Gao:** Conceptualization, Formal analysis, Investigation, Software, Validation, Writing – review & editing.

Declaration of competing interest

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

Data availability

The training data for the electricity price forecasting models is publicly available on the Entso-E Transparency Platform (https://transparency.entsoe.eu/, ENTSO-E Transparency Platform (2022)). This work uses the day-ahead residual load forecasts, i.e., wind power generation, solar power generation, and load, and the day-ahead electricity prices for the BZN|DE-LU bidding zone in 2019 and 2020. Other research data or code will be made available upon request.

Acknowledgments

We gratefully acknowledge the financial support of the Kopernikus project SynErgie 3 by the Federal Ministry of Education and Research (BMBF) and the project supervision by the project management organization Projektträger Jülich, as well as from the Helmholtz Association of German Research Centers as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). The authors would like to thank Danimir Doncevic and Alexander Mitsos for their feedback on the manuscript.

Appendix A. Training of multi-period electricity prices forecasting models

The multi-period forecasting scheme is shown in Fig. A.8.

The three models used for electricity price forecasting are a linear LASSO regression, a fully connected neural network, and a convolutional neural network. The parameter for the LASSO regularization is set to 0.02. Tables A.2 and A.3 list the model structures for the two neural networks, respectively. The attributes of layers shown in Tables A.2 and A.3 are Linear (fully connected): Number of nodes, and Conv2D (2-dimensional convolution): number of filters, and filter size.

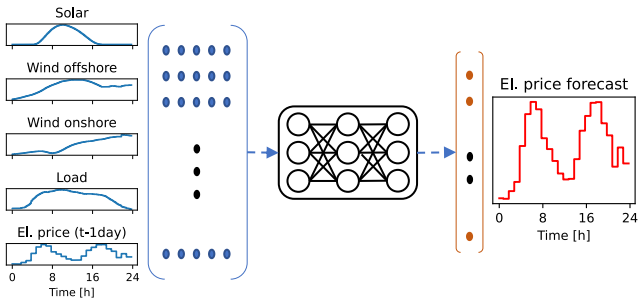

Fig. A.8. Sketch of the input/output structure of the multi-period forecasting scheme.

Table A.2
Structure of fully connected neural network.

Layer	Attributes	Activation
Linear	144	ReLU
Linear	144	ReLU
Linear	24	–

Table A.3
Structure of convolutional neural network.

Layer	Attributes	Activation
Conv2D	16, 3×3	ReLU
Conv2D	16, 3×3	ReLU
Linear	24	–

Table A.4
MSE for test sets in 5-fold cross-validation of the three forecasting models.

	LASSO	Fully connected	Convolutional
1	0.46	0.06	0.06
2	0.37	0.05	0.08
3	0.40	0.07	0.07
4	0.40	0.07	0.08
5	0.44	0.07	0.08

Table B.5
Structure of adversarial surrogate model **ASM(x)**.

Layer	Attributes	Activation
Linear	64	ReLU
Linear	64	ReLU
Linear	32	ReLU
Linear	24	–

All models are trained using the mean-squared-error (MSE) loss function over 100 epochs using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and a batch size of 32 with shuffled data. The input features are standardized, and the electricity prices, i.e., the labels, are scaled using the variance stabilizing transformation (Uniejewski et al., 2017). Scaling and other preprocessing is performed using the python-based machine learning library *scikit-learn* (Pedregosa et al., 2011).

Table A.4 shows the test loss values for the MSE in a 5-fold cross-validation. All models give consistent results, which indicates no bias through the selection of training and test sets. The LASSO regression shows about one order of magnitude higher loss values.

Appendix B. Adversarial surrogate model

Table B.5 shows the model structure for the adversarial surrogate model used for the black-box attacks in the main manuscript. The same structure is used for all forecasting model and DSM problem

combinations. The adversarial surrogate model is trained over 200 epochs, with a batch size of 32 after shuffling, and a learning rate of 0.001. The adversarial surrogate uses *l2*-regularization for the hidden layers. The model primarily aims to provide gradient information for the attacks. Thus, there is no separate test set for evaluation except for the test month used in the evaluation in the main text.

Nomenclature

Symbol	Description	Symbol	Description
δ	Small number	ASM($\mathbf{x}$)	Adversarial surrogate model
$\Delta\mathbf{x}$	Adversarial noise	$\mathbf{P}^{\mathrm{DA}}$	Day-ahead electricity prices
ϵ	White noise	$\mathbf{P}^{\mathrm{DA}}_{t-1\mathrm{day}}$	Day-ahead price on previous day
σ^2	White noise variance	P^{DA}_t	Day-ahead price at t-th hour
ε	Attack rate	$\mathbf{W}^{\mathrm{DA}}_{PV}$	Day-ahead PV forecast
$\mathbf{T}$	Machine learning model/EPF model	$\mathbf{W}^{\mathrm{DA}}_{Wind,on}$	Day-ahead wind onshore forecast
$\mathbf{x}$	Inputs	$\mathbf{W}^{\mathrm{DA}}_{Wind,off}$	Day-ahead wind offshore forecast
$\mathbf{y}$	Outputs	$\mathbf{W}^{\mathrm{DA}}_{Load}$	Day-ahead load forecast
y_i	ith dimension of outputs	$\mathbf{W}^{\mathrm{grid}}$	Grid interaction
$\mathbf{y}^*$	True realization/label	W^{in}_t	Charging rate
$\hat{\mathbf{y}}$	Target outputs	W^{out}_t	Discharging rate
$\tilde{\mathbf{y}}$	Perturbed outputs	W^{max}	Maximum (dis-)charging rate
$\tilde{\mathbf{T}}$	Modified inputs	SOC_t	Battery state of charge
$\mathcal{L}(\mathbf{x},\mathbf{y})$	Loss function	SOC^{max}	Maximum battery state of charge
$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\hat{\mathbf{y}})$	Loss gradient w.r.t. inputs	Δt	Time interval
$\mu_{\mathbf{y}}$	Mean of historical realizations	η	(Dis-)charging efficiency

References

Abadi, M., Agarwal, A., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/. URL https://www.tensorflow.org/. (Accessed on 08 August 2022).

Addis, B., Castel, C., Macali, A., Misener, R., Piccialli, V., 2023. Data augmentation driven by optimization for membrane separation process synthesis. Comput. Chem. Eng. 177, 108342. http://dx.doi.org/10.1016/j.compchemeng.2023.108342.

Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q., 2021. Recent advances in adversarial training for adversarial robustness. http://dx.doi.org/10.48550/ARXIV.2102.01356, arXiv URL https://arxiv.org/abs/2102.01356.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F., 2013. Evasion attacks against machine learning at test time. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13. Springer, pp. 387–402. http://dx.doi.org/10.1007/978-3-642-40994-3_25.

Bor, M.C., Marnerides, A.K., Molineux, A., Wattam, S., Roedig, U., 2019. Adversarial machine learning in smart energy systems. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems. pp. 413–415. http://dx.doi.org/10.1145/3307772.3330171.

Bozlak, u.B., Yaşar, C.F., 2024. An optimized deep learning approach for forecasting day-ahead electricity prices. Electr. Power Syst. Res. 229, 110129. http://dx.doi.org/10.1016/j.epsr.2024.110129.

Brée, L.C., Perrey, K., Bulan, A., Mitsos, A., 2019. Demand side management and operational mode switching in chlorine production. AIChE J. 65, e16352.

Chen, Y., Tan, Y., Zhang, B., 2019. Exploiting vulnerabilities of load forecasting through adversarial attacks. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–11. http://dx.doi.org/10.1145/3307772.3328314.

Colthorpe, A., 2021. Manufacturer reveals involvement in world's biggest battery energy storage system so far. https://www.energy-storage.news/. URL https://www.energy-storage.news/manufacturer-reveals-involvement-in-worlds-biggest-battery-energy-storage-system-so-far/.

Cramer, E., Witthaut, D., Mitsos, A., Dahmen, M., 2023. Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows. Appl. Energy 346, 121370. http://dx.doi.org/10.1016/j.apenergy.2023.121370.

Cui, L., Qu, Y., Gao, L., Xie, G., Yu, S., 2020. Detecting false data attacks using machine learning techniques in smart grid: A survey. J. Netw. Comput. Appl. 170, 102808. http://dx.doi.org/10.1016/j.jnca.2020.102808.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., Roli, F., 2019. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th USENIX Security Symposium. USENIX Security 19, USENIX Association, Santa Clara, CA, pp. 321–338. http://dx.doi.org/10.5555/3361338.3361361, URL https://www.usenix.org/conference/usenixsecurity19/presentation/demontis.

Ehsani, B., Pineau, P.O., Charlin, L., 2024. Price forecasting in the ontario electricity market via triconvgru hybrid model: Univariate vs. multivariate frameworks. Appl. Energy 359, 122649. http://dx.doi.org/10.1016/j.apenergy.2024.122649.

ENTSO-E Transparency Platform, 2022. ENTSO-E transparency platform. https://transparency.entsoe.eu/. URL https://transparency.entsoe.eu/. (Accessed: 01 March 2022).

European Power Exchange, 2021. EPEX SPOT documentation. http://www.epexspot.com/en/extras/download-center/documentation. (Accessed: 30 May 2021).

Fazlyab, M., Morari, M., Pappas, G.J., 2022. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. IEEE Trans. Automat. Control 67, 1–15. http://dx.doi.org/10.1109/TAC.2020.3046193.

Gomez, A.L.P., Maimo, L.F., Clemente, F.J., Morales, J.A.M., Celdran, A.H., Bovet, G., 2022. A methodology for evaluating the robustness of anomaly detectors to adversarial attacks in industrial scenarios. IEEE Access 10, 124582–124594. http://dx.doi.org/10.1109/ACCESS.2022.3224930.

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations. pp. 1–11, URL http://arxiv.org/abs/1412.6572.

Gurobi Optimization, LLC, 2023. Gurobi optimizer reference manual. URL https://www.gurobi.com.

Hao, J., Tao, Y., 2022. Adversarial attacks on deep learning models in smart grids. Energy Rep. 8, 123–129.

Hart, W.E., Laird, C.D., Watson, J.P., Woodruff, D.L., Hackebeil, G.A., Nicholson, B.L., Siirola, J.D., et al., 2017. Pyomo-Optimization Modeling in Python, vol. 67, Springer.

Heinrich, R., Scholz, C., Vogt, S., Lehna, M., 2023. Targeted adversarial attacks on wind power forecasts. Mach. Learn. 113, 863–889. http://dx.doi.org/10.1007/s10994-023-06396-9.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B., 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy. SP, IEEE, pp. 19–35. http://dx.doi.org/10.1109/SP.2018.00057.

Jedrzejewski, A., Lago, J., Marcjasz, G., Weron, R., 2022. Electricity price forecasting: The dawn of machine learning. IEEE Power Energy Mag. 20, 24–31.

Kapoor, G., Wichitaksorn, N., 2023. Electricity price forecasting in New Zealand: A comparative analysis of statistical and machine learning models with feature selection. Appl. Energy 347, 121446. http://dx.doi.org/10.1016/j.apenergy.2023.121446, URL https://www.sciencedirect.com/science/article/pii/S0306261923008103.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. pp. 1–11, URL http://arxiv.org/abs/1412.6980.

Koay, A.M., Ko, R.K., Hettema, H., Radke, K., 2023. Machine learning in Industrial Control System (ICS) security: Current landscape, opportunities and challenges. J. Intell. Inform. Syst. 60, 377–405. http://dx.doi.org/10.1007/s10844-022-00753-1.

Kurakin, A., Goodfellow, I.J., Bengio, S., 2018. Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security. Chapman and Hall/CRC, pp. 99–112. http://dx.doi.org/10.48550/arXiv.1607.02533.

Lago, J., Marcjasz, G., De Schutter, B., Weron, R., 2021. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. Appl. Energy 293, 116983. http://dx.doi.org/10.1016/j.apenergy.2021.116983, URL https://www.sciencedirect.com/science/article/pii/S0306261921004529.

Lee, J.H., Shin, J., Realff, M.J., 2018. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng. 114, 111–121. http://dx.doi.org/10.1016/j.compchemeng.2017.10.008.fOCAPO/CPC2017, URL https://www.sciencedirect.com/science/article/pii/S0098135417303538.

Maiti, R.R., Yoong, C.H., Palleti, V.R., Silva, A., Poskitt, C.M., 2023. Mitigating adversarial attacks on data-driven invariant checkers for cyber–physical systems. IEEE Trans. Dependable Secure Comput. 20, 3378–3391. http://dx.doi.org/10.1109/TDSC.2022.3194089.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Schäfer, P., Westerholt, H.G., Schweidtmann, A.M., Ilieva, S., Mitsos, A., 2019. Model-based bidding strategies on the primary balancing market for energy-intense processes. Comput. Chem. Eng. 120, 4–14. http://dx.doi.org/10.1016/j.compchemeng.2018.09.026.

Schweidtmann, A.M., Esche, E., Fischer, A., Kloft, M., Repke, J.U., Sager, S., Mitsos, A., 2021. Machine learning in chemical engineering: A perspective. Chem. Ing. Tech. 93, 2029–2039. http://dx.doi.org/10.1002/cite.202100083.

Shen, X., Liu, H., Qiu, G., Liu, Y., Liu, J., Fan, S., 2024. Interpretable interval prediction-based outlier-adaptive day-ahead electricity price forecasting involving cross-market features. IEEE Trans. Ind. Inform. 1–14. http://dx.doi.org/10.1109/tii.2024.3355105.

Şuvak, Z., Anjos, M.F., Brotcorne, L., Cattaruzza, D., 2022. Design of Poisoning Attacks on Linear Regression using Bilevel Optimization. Technical Report, University of Edinburgh.

Tan, W.G.Y., Wu, Z., 2024. Robust machine learning modeling for predictive control using Lipschitz-constrained neural networks. Comput. Chem. Eng. 180, 108466. http://dx.doi.org/10.1016/j.compchemeng.2023.108466.

Tang, N., Mao, S., Nelms, R.M., 2021. Adversarial attacks to solar power forecast. In: 2021 IEEE Global Communications Conference. GLOBECOM, IEEE, pp. 1–6.

Trebbien, J., Gorjão, L.R., Praktiknjo, A., Schäfer, B., Witthaut, D., 2023a. Understanding electricity prices beyond the merit order principle using explainable AI. Energy AI 13, 100250. http://dx.doi.org/10.1016/j.egyai.2023.100250.

Trebbien, J., Pütz, S., Schäfer, B., Nygård, H.S., Rydin Gorjão, D., 2023b. Probabilistic forecasting of day-ahead electricity prices and their volatility with LSTMs. In: 2023 IEEE PES Innovative Smart Grid Technologies Europe. ISGT EUROPE, pp. 1–5. http://dx.doi.org/10.1109/ISGTEUROPE56780.2023.10407112.

Uniejewski, B., Weron, R., Ziel, F., 2017. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans. Power Syst. 33, 2219–2229. http://dx.doi.org/10.1109/TPWRS.2017.2734563.

Wang, J., Srikantha, P., 2021. Stealthy black-box attacks on deep learning non-intrusive load monitoring models. IEEE Trans. Smart Grid 12, 3479–3492.

Waskom, M.L., 2021. Seaborn: statistical data visualization. J. Open Source Softw. 6, 3021.

Wen, H., Amin, M.T., Khan, F., Ahmed, S., Imtiaz, S., Pistikopoulos, E., 2023. Assessment of situation awareness conflict risk between human and AI in process system operation. Ind. Eng. Chem. Res. 62, 4028–4038. http://dx.doi.org/10.1021/acs.iecr.2c04310.

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. Int. J. Forecast. 30, 1030–1081.

Weron, R., Ziel, F., 2019. Electricity price forecasting. In: Routledge Handbook of Energy Economics. pp. 506–521. http://dx.doi.org/10.4324/9781315459653-36.

Wolff, G., Feuerriegel, S., 2017. Short-term dynamics of day-ahead and intraday electricity prices. Int. J. Energy Sector Manag. 11, 557–573. http://dx.doi.org/10.1108/IJESM-05-2016-0009.

Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K., 2020. Adversarial attacks and defenses in images, graphs and text: A review. Int. J. Autom. Comput. 17, 151–178. http://dx.doi.org/10.1007/s11633-019-1211-x.

Zeng, L., Qiu, D., Sun, M., 2022. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. Appl. Energy 324, 119688.

Zhang, Q., Grossmann, I.E., 2016. Enterprise-wide optimization for industrial demand side management: Fundamentals, advances, and perspectives. Chem. Eng. Res. Des. 116, 114–131, Process Systems Engineering - A Celebration in Professor Roger Sargent's 90th Year.

Zhuo, Y., Yin, Z., Ge, Z., 2023. Attack and defense: Adversarial security of data-driven fdc systems. IEEE Trans. Ind. Inform. 19, 5–19. http://dx.doi.org/10.1109/TII.2022.3197190.

Ziel, F., Steinert, R., Husmann, S., 2015. Efficient modeling and forecasting of electricity spot prices. Energy Econ. 47, 98–111. http://dx.doi.org/10.1016/j.eneco.2014.10.012.

Ziel, F., Weron, R., 2018. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. Energy Econ. 70, 396–420.