Vulnerability and Impact of Machine Learning-based Inertia Forecasting Under Cost-Oriented Data Integrity Attack

Yan Chen, Mingyang Sun, Zhongda Chu, Simon Camal, Georges Kariniotakis, Fei Teng

Vulnerability and Impact of Machine Learning-based Inertia Forecasting Under Cost-Oriented Data Integrity Attack

Yan Chen, *Student Member, IEEE,* Mingyang Sun, *Member, IEEE,* Zhongda Chu, *Member, IEEE,* Simon Camal, George Kariniotakis, *Senior Member, IEEE,* Fei Teng, *Senior Member, IEEE*

***Abstract*—With the increasing penetration of renewables, the power system is facing unprecedented challenges of low-inertia levels. The inherent ability of the system to defense disturbance and power imbalance through inertia response is degraded, and thus, system operators need to make faster and more efficient scheduling operations. As one of the most promising solutions, machine learning (ML) methods have been investigated and employed to realize effective inertia forecasting with considerable accuracy. Nevertheless, it is yet to understand its vulnerability with the growing threat of cyberattacks. To this end, this paper proposes a methodological framework to explore the vulnerability of ML-based inertia forecasting models, with a special focus on data integrity attacks. In particular, a cost-oriented false data injection attack is proposed, for the first time, with the primary objective to significantly increase the system operation cost while retaining the stealthiness of the attack via minimizing the differences between the pre-perturbed and after-perturbed inertia forecasts. Moreover, we propose four vulnerability assessment metrics for the ML-based inertia forecasting models. Case studies on the GB power system demonstrate the vulnerability and impact of the ML-based inertia forecasting models, as well as the stealthiness and transferability of the proposed cost-oriented data integrity attacks.**

***Index Terms*—Inertia forecasting, power system operation, machine learning, cyber security, power system economics.**

I. Introduction

DECARBONISATION agenda significantly increases the penetration of RES, which drives the power system towards a low inertia system as most of these technologies are interfaced via converters that do not supply rotational inertia to the system [1]. By definition, system inertia refers to the ability of a system to oppose changes in frequency due to the resistance provided by the kinetic energy stored in rotating masses synchronously connected to the power system [2]. As disturbance or supply/demand imbalances occur, the inertial response is first released, combined with the subsequent primary control, to determine the maximum frequency offset. Therefore, it is of great importance to conduct accurate inertia forecasting to determine appropriate frequency responses and reserves in advance for system stability enhancement, and cost-effective system operation [3].

In the literature, the prevailing inertia forecasting methods mainly focus on the contribution from synchronous equipment on the generation side. The authors in [3] forecast system inertia using the current operation plan submitted by each synchronous generator in the system, which contains the short-term estimation of the unit state and operating limit. In [4] authors use generation prediction and corresponding inertia constants to realize short-term inertia prediction. As the system inertia is evolving from a relatively controllable variable to a time-varying variable affected by external uncertainties, more accurate inertia forecasting methods, need to be developed with the consideration of recessive-related data such as date, weather, and renewable generation information [5]. In recent years, data-driven methods such as ML have been applied to power system inertia forecasting[5], [6], [7], [8], [9]. The authors in [5] designed a power system inertia forecasting tool based on artificial neural network (ANN) under a scenario with high penetration of wind generation. It comprehensively considers the power production from synchronous generators (SGs), renewable generation, and coupled motor loads, achieving an effective and accurate estimation of regional inertia. In [7], authors proposed a decomposable time series model for inertia forecasting, which takes the trend, seasonal and irregular components of the historical inertia value into consideration.

Although existing studies have demonstrated the superior performance of machine learning algorithms in terms of accuracy of the forecasting task, with the growing threat of cyberattacks, it is imperative to explore the vulnerability and impact of machine learning-based forecasting models before implementing them in practice. In particular, machine learning-based forecasting models highly rely on other forecast information, such as power generation output from various sources, system load, weather, etc. Compared with conventional forecasting models that are trained based on real measurements, the forecasted data used in inertia forecasting are derived from other modules and transmitted through API,

This work was supported in part by the National Natural Science Foundation of China under Grants 52161135201, U20A20159, 62103371, by the EPSRC under Grant EP/W028662/1 and by European Union Horizon 2020 under Smart4RES Project (No. 864337). (*Corresponding authors: Mingyang Sun and Fei Teng*)

Y. Chen, M. Sun are with the State Key Laboratory of Industrial Control Technology and the College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China, e-mail: chenyan16@zju.edu.cn, mingyangsun@zju.edu.cn;

Z. Chu, F. Teng are with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K., e-mail: zc4915@ic.ac.uk; f.teng@imperial.ac.uk;

S.Camal, G.Kariniotakis are with the MINES Paris, PSL Research University, PERSEE-Center for Processes, Renewable Energies and Energy Systems, Sophia-Antipolis 06904, France, e-mail:simon.camal@minesparis.psl.eu; georges.kariniotakis@minesparis.psl.eu.

which becomes vulnerable to cyberattacks. In the machine learning community, the adversarial attack against neutral networks is first proposed in [10] and studied extensively. The attack exploits the over-fitting property of the model and can achieve error outputs with high confidence by adding small perturbations to the test samples. On the other hand, poisoning attacks utilize the data-driven models' innate trust of training data and implement false data injection to reshape the spatial distribution of training set and hence indirectly affect model performance [11]. In addition, a backdoor attack is an advanced form of poisoning attack, which implants backdoor into the model so that the model can identify the special samples that are flagged by triggers and perform as the design of the attacker [12].

All aforementioned approaches mainly focus on classification problems, which aim to design perturbations that can make the model misclassify the data input. However, very few studies have been conducted to deal with regression problems [13], especially in terms of subpopulation attack[14], which aims to compromise the performance of particular subpopulation samples rather than the global samples. In a classification problem, it is easy to attack based on classes defined by labels and attackers can design complex and flexible attacks by the classifier's consensus decision to samples with the same label. For example, [15] can use malicious samples with targeted labels to surround specific benign samples and induce the model to misclassify the surrounded benign samples to the targeted label. However, in the regression problem, the model output continuous values based on a unified decision function, it is difficult to observe intuitionally the decision influence between samples and construct an attack based on decision influence between samples. In the power system community, recent studies have demonstrated the vulnerability of present machine learning-based energy forecasting models against data integrity attacks [16]. However, it only utilizes a simple unconstrained attack method without considering the potential countermeasures from the defender. In particular, anomaly detection algorithms [17] have been widely applied, and therefore significant and consistent attacks may be easily detected and filtered. In this context, it becomes crucial to investigate the feasibility of attacks that simultaneously achieve destructiveness and stealthiness.

Furthermore, existing attack methods can be classified as targeted attacks and untargeted attacks. For a classification task, the untargeted attack aims to misclassify the input data to any other classes without a specific target, whereas the targeted attack needs to set a target class pre-defined by attacker [18]. For a regression problem, the objective of the untargeted attack is to maximize the forecast error, while the targeted attack aims to minimize the distance between the forecasts and the pre-determined output targets. It is of great importance to note that, for power system applications, arbitrarily tampering with the data to maximize the forecasting error renders the possibility of being detected by defense mechanisms. In other words, the attacks need to be designed as stealthy attacks by injecting adversarial data with elaborately designed attack targets to influence the system stability and operational cost. In particular, as illustrated in [19], asymmetric system operation costs are observed for load or RES output forecasting errors with the same magnitude but over different periods or with opposite signs and thus, the targeted attacks under full system knowledge may result in higher operation cost with the same level of injected perturbations.

For inertia forecasting, to ensure the frequency stability of power systems, system operators need to maintain an adequate level of system inertia, whereas inaccurate inertia forecasts will lead to extra or improper scheduling and increase the operation cost. Inertia forecasting errors will drive the cost increment, but the exact economic consequence may vary in different system conditions, time periods, and even the signs of error [20]. In other words, higher inertia forecasting errors do not necessarily mean higher operation cost. On this basis, it is imperative to investigate the vulnerability and impact of machine learning-based inertia forecasting models.

To this end, we propose a cost-oriented data integrity attack that aims to stealthily increase the system operation cost by implementing well-designed inertia misprediction at predetermined cost-sensitive time periods, via injecting perturbations to the training or test stages. It is important to note that the proposed cost-oriented attack is designed as a subpopulation attack, whose goal is to realize inertia misprediction in cost-sensitive periods and maintain the benign prediction in the rest periods. To the best of our knowledge, it is the first attempt to develop a subpopulation attack to a regression problem. The main original contribution of this paper can be summarized as the following:

1) This paper proposes the concept of cost-oriented attack for machine learning models under the application of power system inertia forecasting. More specifically, cost-oriented data poisoning attacks and cost-oriented adversarial attacks are developed to increase the power system operation cost via cost-informed attack period selection and minimum perturbation injection calculated by gradient-based approaches.

2) The proposed attack design process explicitly considers the stealthiness of the attack. In particular, the proposed attack algorithms focus on a specific set of target points within cost-sensitive periods, while maintaining normal accurate forecasting for the rest intervals so that the overall forecasting error can be closely retained before and after the attack.

3) The effectiveness and impacts of the proposed methodological framework are validated based on the GB power system. Furthermore, the transferability of the proposed cost-oriented attacks is revealed across different machine learning-based inertia forecasting approaches.

4) A vulnerability assessment framework is developed for machine learning-based inertia forecasting that, for the first time, explores the potential risks of data integrity attacks in the offline model training and online test stages, respectively. Assessment metrics are proposed to assess the impacts of the considered threats in terms of forecasting accuracy and power system operation cost, respectively.

The rest of the paper is organized as follows. Section II formulates the ML-based inertia forecasting models, the attacker's objective and capacity constraints, and then introduces existing data integrity attack methods. The proposed cost-oriented data integrity attacks are illustrated in Section III.

Section IV presents the proposed vulnerability analysis framework for ML-based inertia forecasting models with developed assessment metrics. Section V conducts comprehensive numerical experiments to evaluate the effectiveness and impacts of the proposed methodological framework. Section V draws the conclusions.

II. Problem Statement

A. Machine Learning-based Inertia Forecasting Model

This paper aims to investigate the vulnerability of a family of ML-based inertia forecasting models that capture the relationship between system inertia and selected correlated features, including the power system operation data. Mathematically, let $S = \{(x_t, y_{t+k})\}_{t=1}^{T}$ denote the data set of the power system historical data constructed for inertia forecasting, here y_{t+k} is the value of system inertia, k is the model's forecast horizon which is set as one day. x_t represents the input feature vector, composed of the system variables available at timestamp t:

$$x_t := (x_t^{\text{forecast}}, x_t^{\text{IFs}}, x_t^{\text{state}}, y_t) \tag{1}$$

where x_t^{forecast} includes the features of the day-ahead forecasted synchronous generation $\hat{P}^{\text{gen}}$ and loads $\hat{P}^{\text{load}}$, as well as the forecasted renewable generation outputs (solar $\hat{P}^{\text{solar}}$, onshore wind $\hat{P}^{\text{onshore}}$ and offshore wind $\hat{P}^{\text{offshore}}$) in the region; x_t^{IFs} represents the power flow with surrounding region systems; x_t^{state} contains environment state information such as dates, temperature and weather; y_t is the historical ground truth of inertia.

To fully explore the vulnerabilities of various machine learning-based inertia forecasting models, a series of representative models, including Linear Regression (LR), Feedforward Neural Network (FNN), Recurrent Neural Network (RNN), and Long-short-term Memory Network (LSTM), are considered in this paper. The objective of training these models is to obtain a function parameterized by θ that captures the relationship between the input selected features and the output predicted system inertia, expressed as follows:

$$f_\theta(x_t^{\text{forecast}}, x_t^{\text{IFs}}, x_t^{\text{state}}, y_t) = \hat{y}_{t+k} \tag{2}$$

where f_θ is a linear or a non-linear function depending on the selected ML models.

B. Attackers' Objective

Power system inertia is closely related to frequency stability. The operators schedule generation or topology changes in advance based on inertia forecasts. In this paper, the objective of the attacker is assumed to influence the operation of power systems and increase the operation cost by manipulating the predicted value of system inertia. Specifically, over-prediction in system inertia during the forecasting stage would lead to insufficient inertia in real-time, and high-cost SGs with fast responding capability have to be dispatched online for inertia provision, causing significantly increased system operation cost or risk of inertia shortage. On the contrary, inertia under-prediction results in more SGs being prepared during the day-ahead scheduling period, leading to less efficient part-loading operation.

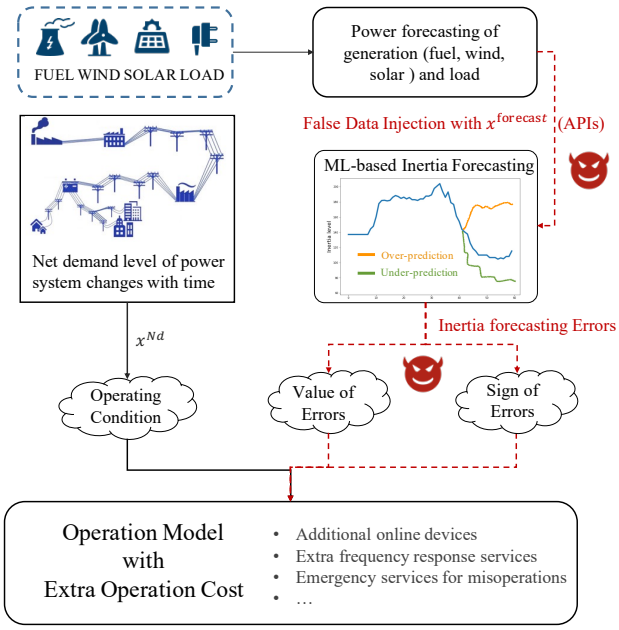


Fig. 1. General framework of cost-oriented attack.

The specific attack strategy is shown in Fig. 1. We find that the impact intensity of inertia forecasting errors on the scheduling model varies with the error values, error signs, and the operation conditions when the error occurs. Therefore, based on the knowledge of scheduling model, the attackers aims to design the optimal attack targets for the expected inertia forecasts under different system net demand x^{Nd} levels, which balances the cost impact and stealthiness. The attacker achieves the manipulation of inertia forecasts through false data injection against the externally forecasted system load and generation outputs.

C. Attackers' Knowledge and Capability

To investigate the impact of attacker's knowledge, we consider two kinds of attack scenarios: white-box attacks and black-box attacks. In white-box attack, as the worst case, we assume that the attacker has the full knowledge of training set D_{tr}, test set D_{test}, model structures, and even model parameters θ. Since this strong assumption may not always hold, we also investigate black-box attack where the attacker does not know the model information but can self-train an alternative model $f_\theta^{'}$ based on the training set D_{tr} and use the alternative model $f_\theta^{'}$ to generate the attack. In both attack scenarios, attackers just have the ability to modify one and only one kind of data set (D_{tr} or D_{test}), which differs from the backdoor attacks that require modification of both D_{tr} and D_{test}. In addition, black-box attacks can be used to evaluate the transferability of attacks between different models with more details given in Sec.IV and Sec.V.

The characteristic variables contained in x_t^{forecast} are predicted values of various system power conditions, which have

less physical constraints compared with measured values. Furthermore, these kinds of variables are imported from other system modules via APIs, which are more vulnerable to be tampered externally [16]. Therefore, considering the feasibility of the attack, we assume the attacker can only tamper with variables in x_t^{forecast} for a data integrity attack. Note that attackers strictly follow attack limits, keeping ground truth y clean during data poisoning in the training phase, which is different from current poisoning attacks, such as [21], [22]. Further, in order to avoid excessive data tampering being detected by the system operator, the following constraint is enforced to the disturbance added by the attacker:

$$||x_t^j - x_t^{j^*}|| \leq \xi ||x_t^j||, \quad \forall x_t^j \in x_t^{\text{forecast}} \tag{3}$$

where ξ represents the constraints of disturbance magnitude.

D. Existing Attack Methods

In the ML community, extensive studies have been conducted on the security of machine learning algorithms. In this subsection, we introduce two types of well-developed data integrity attack methods.

1) Data Poisoning: Data poisoning occurs at the training stage. The attackers contaminate the training set D_{tr} in order to change the parameters of the target model and reconstruct it [23]. Current data poisoning attacks focus on classification tasks, which have specific attack targets defined by labels (misclassification or assigned classification). According to the discrete label characteristic, flexible strategies can be constructed, such as inducing test data to be classified as target labels by deploying poison samples with the same label around them [15].

Nevertheless, it is difficult to directly employ the above methods to regression problems because the outputs of the model are continuous values. Therefore, [21], the problem of data poisoning attacks in regression is formulated as the following continuous space bilevel optimization problem:

$$\arg\max_{D_{tr}} \ W(f_\theta, D_{val}) \tag{4a}$$

$$s.t. \ \theta \in \arg\min L(D_{tr}, f_\theta) \tag{4b}$$

where L represents the loss function for training, by training on contaminated training set D_{tr} the attacked model is manipulated. D_{val} represents an untrained identically distributed validation set, as attackers are restricted to have no access to the test set, the attack effect is evaluated by model performance on D_{val}. As the output at a single timestamp lacks specific meaning in regression problems, the current works design the attack effect W with overall performance, such as maximizing the MSE (Mean Square Error) of the output [22], [21].

2) Adversarial Attacks: Regarding the test stage, the adversarial attack was first proposed in [10], where they found that the mapping of input and output of deep neural networks is mainly discontinuous. This is shown by the fact that adding small imperceptible disturbance to the input samples can lead to image misclassification with high confidence. Thus, adversarial attacks manipulate the input of samples during the test phase without changing or polluting the existing model.

In the literature, the adversarial attack has been studied as a great threat to ML algorithms in various application fields. In the power systems community, the vulnerabilities arising from adversarial attacks have been explored and shown in a series of key applications, including the MLP-based false data injection detection [24], RNN-based energy theft detection [25], and CNN-based voltage stability assessment [26]. For a regression problem, the authors in [16] proposed a black-box adversarial attack against the deep learning-based load forecasting models by tampering with the temperature features transmitted by the APIs. The attack aims to cause a consistent upper or lower deviation on load forecast, with the disturbance added to the temperature features as little as possible. It can be mathematically expressed as the following optimization problem [16]:

$$\arg\min_{\delta} \ \ \kappa f_\theta(x_t + \delta x_t) + \beta \Phi(\delta). \tag{5}$$

where δ represents the disturbance, β and Φ are used to constrain the magnitude of disturbance, and κ sets the direction of deviation (higher or lower).

Overall, for both data poisoning and adversarial attacks, most of the existing methods in the literature only aim to degrade the overall forecasting performance (e.g., maximizing the Mean Square Error (MSE) [22]). However, significantly decreased forecasting performance can be easily aware by the system operator or detected via deployed detection mechanisms. In other words, stealthiness can not be guaranteed if the attack target only considers its effectiveness.

III. Cost-oriented Data Integrity Attacks

This section introduces the proposed cost-oriented data integrity attack to investigate the vulnerability and impact of ML-based inertia forecasting models. In particular, cost-oriented data poisoning attacks and cost-oriented adversarial attacks will be defined and illustrated, respectively.

A. Cost-oriented Target Selection

To design an effective and stealthy cost-oriented attack, the first step is to determine the attack target for inertia forecasting via understanding its impacts on system operational costs under different operating conditions. In other words, this step aims to determine the cost-sensitive periods and the targeted sign (i.e., positive or negative) of difference between the original and the manipulated inertia forecasting outputs.

Let x_t and $\hat{y}_{t+k}$ denote the system operation condition at timestamp t and the corresponding forecasted system inertia in the future, respectively, the system operation cost $C(\hat{y_{t+k}}|x_t)$ can be calculated based on the simulation model. More details of the employed stochastic optimization model and the decisions made at different stages are illustrated in [27]. Then

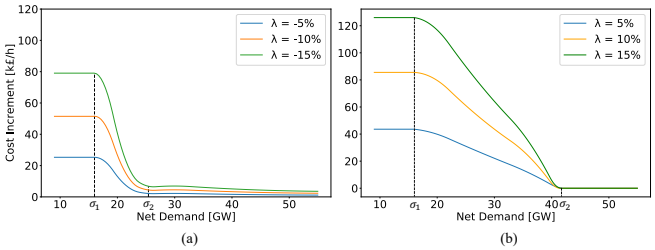


Fig. 2. Average system operation cost increment under inertia misprediction with $\lambda < 0$ (a) and $\lambda > 0$ (b) .

the proposed cost-oriented attack for inertia forecasting can be described as the following problem:

$$\underset{\mathcal{A}_{\text{Trigger}}}{\arg\max} \quad \sum_{t=1}^{T} C((1+\lambda_{t+k})\hat{y_{t+k}}|x_t) - C(\hat{y_{t+k}}|x_t) \tag{6a}$$

$$s.t. \quad |\lambda_{t+k}| < \text{Max}_\lambda, \forall t \in [1,T] \tag{6b}$$

$$\sum_{t=1}^{T} sgn(|\lambda_{t+k}|) = \text{Num}_\lambda \tag{6c}$$

where λ_{t+k} represents the change ratio before and after attack at timestamp $t+k$. In particular, $\lambda_{t+k} \neq 0$ and $\lambda_{t+k} = 0$ indicate that the value of system inertia forecast at timestamp $t+k$ is being or not being manipulated, respectively. If $\lambda_{t+k} \neq 0$, the sign of λ_{t+k} represents the direction of targeted forecast value (i.e., > 0 increase or < 0 decrease) , which is determined based on the calculated operation cost; $\mathcal{A}_{\text{Trigger}} = \{\lambda_{t+k}|\lambda_{t+k} \in \mathbb{R}\}_{t=1}^{T}$ is the trigger set, aiming at maximizing the difference of system operation costs over all the time stamps T before and after the attack via obtaining the optimal manipulated system inertia forecasts $(1+\lambda_{t+k})\hat{y}_{t+k}$; Max_λ represents the maximum value of λ_{t+k}; Num_λ indicates the restricted number of timestamps can be manipulated. Note that the smaller number of Num_λ means that the attack is more stealthy.

To achieve the aforementioned objective, the first step is to investigate the influence of inertia forecasting errors under different levels of net demand quantified via the incremental system operation cost compared with the no attack case, as depicted in Fig. 2. For the case of positive λ (Fig. 2-(b)), the system inertia is overpredicted during the forecasting process leading to less CCGTs (slower SGs with lower marginal cost) to be prepared. As a result, when the actual value of total inertia is available to system operators at the real-time stage, they can only turn on additional OCGTs (faster SGs with higher marginal cost) to compensate for the forecasting error, thus inducing more operational cost. Understandably, a larger magnitude of the attack causes higher cost increment. In the extreme case, where no sufficient OCGTs are available in real-time, the system may risk destabilizing in the event of a large outage.

Furthermore, this cost increment increases approximately in a linear manner as the net demand decreases from around $\sigma_2 = 42\,(\text{GW})$ since the increased wind power makes the inertia from SGs more critical in maintaining the frequency constraints and attacks with the same magnitude would require more OCGTs to compensate shorted system inertia, hence, the larger cost increment. This increasing trend becomes saturated below the net demand of $\sigma_1 = 17\,(\text{GW})$, as the additional wind power, in this case, cannot be utilized due to the frequency constraints. Therefore, the system operating conditions and the cost are not influenced.

On the other hand, the attacks with $\lambda < 0$ (Fig. 2-(a)) also induce positive cost increment as more CCGTs are prepared to be online during the forecasting process to compensate for the inertia shortage due to the attacks. However, this increment is negligible at high net demand since the discounted system inertia, in this case, is enough after the attack due to the large amount of online SGs. Similarly to the previous case, the cost increment almost remains a constant below the net demand of $\sigma_1 = 17\,(\text{GW})$ for the same reason, and the cost increment is approximately proportional to the magnitudes of the attacks.

Overall, it can be observed that the attack impact on the system operation cost is asymmetrical in terms of sign of injected attack vector, and vary regularly with the system operating condition. To this end, the attack target for cost-oriented approaches can be determined as follows:

$$sign(\lambda_{t+k}) = \begin{cases} 1, & (x_t^{Nd} < \sigma_o) \text{ AND } (\Delta_{x_t} C(\lambda_{t+k}) > 0) \\ -1, & (x_t^{Nd} < \sigma_o) \text{ AND } (\Delta_{x_t} C(\lambda_{t+k}) < 0) \\ 0, & other \end{cases} \tag{7a}$$

$$\Delta_{x_t} C(\lambda_{t+k}) = C((1+\lambda_{t+k})\hat{y_{t+k}}|x_t) - C((1-\lambda_{t+k})\hat{y_{t+k}}|x_t) \tag{7b}$$

where $x^{Nd} = \hat{P}^{\text{load}} - \hat{P}^{\text{solar}} - \hat{P}^{\text{onshore}} - \hat{P}^{\text{offshore}}$ represents the day-ahead forecast value of net demand level, $\sigma_o \in (\sigma_1, \sigma_2)$,

$$\sigma_1 \leftarrow \frac{\mathrm{d}C((1+\lambda)y|x^{Nd})}{\mathrm{d}x^{Nd}} = 0, \ \frac{\mathrm{d}^2 C((1+\lambda)y|x^{Nd})}{\mathrm{d}x^{Nd2}} < 0 \tag{8a}$$

$$\sigma_2 \leftarrow \frac{\mathrm{d}C((1+\lambda)y|x^{Nd})}{\mathrm{d}x^{Nd}} = 0, \ \frac{\mathrm{d}^2 C((1+\lambda)y|x^{Nd})}{\mathrm{d}x^{Nd2}} > 0 \tag{8b}$$

First, the sign of attack target λ_{t+k} need to be determined, corresponding to the selection of attack intervals ($\lambda_{t+k} \neq 0$) and attack directions (over-prediction or under-prediction). As shown in Fig. 2, the system can be divided into three cost-sensitive levels by demarcation points σ_1 and σ_2, which are the extreme points of cost increment-net demand curve. To make the attack intervals preferably cover the high-sensitive intervals and exclude the low-sensitive intervals, the attack interval can be finally defined as $x^{Nd} < \sigma_o$, $\sigma_o \in (\sigma_1, \sigma_2)$.

For the points within attack intervals ($\lambda_{t+k} \neq 0$), the next step is to further determine whether over-prediction ($\lambda_{t+k} > 0$) or under-prediction ($\lambda_{t+k} < 0$) can lead to higher cost increment with Eq. 7b. It is worth mentioning that by modifying $\sigma_o \in (\sigma_1, \sigma_2)$, we can control the size of attack intervals and balance the effectiveness and concealment of the cost-oriented attack. More specifically, the attack can be more stealthy by making σ_o closer to σ_1 or increase operation cost as much as possible by making σ_o closer to σ_2.

Furthermore, for the magnitude of λ_{t+k}, it can be designed large enough to significantly enhance the attack effect as the cost increment is proportional to the magnitude of λ_{t+k} under the same net demand level. However, the impacts of attack will be restricted by the attackers' capability as illustrated in Eq. 3 as well as the upper limits that how much the decision boundary of the ML model can be changed.

After determining the cost-oriented target $\mathcal{A}_{\text{Trigger}} = \{\lambda_{t+k}\}_{t=1}^{T}$, the general attack vector generation problem can be expressed as follows:

$$\underset{\theta or D_{test}}{\arg\min} \quad \frac{1}{N_{test}} \sum_{t=1}^{N_{test}} (f_\theta(x_t) - (1+\lambda_{t+k}) y_{t+k})^2 \tag{9a}$$

$$s.t. \quad x_t \in D_{test}, \quad t = 1, 2, ..., N_{test} \tag{9b}$$

where x_t are optimized to minimize the objective function that measures the proximity of the inertia forecasts to the attack target; D_{test} represents the test set.

To achieve this target, we propose Cost-Oriented Data Poisoning (CODP) and Cost-Oriented Adversarial Attack (COAA) for the offline training and online test stages, respectively, considering the pre-determined cost-sensitive periods and the attack targets. In particular, the CODP generates the Trojan model by poisoning the training set, whereas the COAA keeps the established model and modifies the test set to generate adversarial samples. The details of the design are given in the following parts.

Algorithm 1 Cost-oriented Data Poisoning

Initialization: Training set $(D_c \cup D_p^{(0)})$, the initial poisoning set $D_p^{(0)} = (x_t, y_{t+k})_{t=1}^{Np}$, validation set D_{val}, Searching step size α_0, convergence coefficient β
Output: D_p^{best}

$\theta^{(0)} \leftarrow argmin_\theta L(D_c \cup D_p^{(0)})$
$\hat{y} \leftarrow f_{\theta^{(0)}}(x),\ x \in D_{val}$
$\mathcal{A}_{\text{Trigger}} \leftarrow argmax C((1+\lambda)\hat{y}|x) - C(\hat{y}|x),\ x \in D_{val}$ ▷ *Identify attack target*
$w^{(0)} \leftarrow W(D_{val}, \mathcal{A}_{\text{Trigger}}, \theta^{(0)})$
for $i = 0, 1, ..., N_{update} - 1$ **do** ▷ *Loops for update θ_p*
  for $t = 1, 2, ..., N_p$ **do** ▷ *Loops for update D_p*
    $x_t^{(i+1)} \leftarrow L_{search}(x_t^{(i)}, \nabla_{x_t} W(D_{val}, \mathcal{A}_{\text{Trigger}}, \theta^{(i)}), w^{(i)})$
    $D_p^{(i+1)} \leftarrow x_t^{(i+1)}$
  end for
  $\theta^{(i+1)} \leftarrow argmin_\theta L(D_c \cup D_p^{(i+1)})$
  $w^{(i+1)} \leftarrow W(D_{val}, \mathcal{A}_{\text{Trigger}}, \theta^{(i+1)})$
  if $||w^{(i+1)} - w^{(i)}|| < \epsilon_{stop}$ **then**
    break ▷ *satisfy condition and early termination*
  end if
end for
$D_p^{best} \leftarrow D_p^{(i+1)}$

Def: $L_{search}(x_t, \nabla_{x_t} W, w)$**:**
Output: x_t^{best}

$x_t^{(0)} \leftarrow x_t$, $\alpha^{(0)} \leftarrow \alpha_0$, $w^{(0)} \leftarrow w$
for $j = 0, 1, ..., N_{search} - 1$ **do**
  $x_t^{\text{forecast}(j+1)} = x_t^{\text{forecast}(j)} - \alpha^{(j)} \cdot \nabla_{x_t^{\text{forecast}}} W$ ▷ *Inject disturbance on specific features $x^{forecast}$*
  $\theta^{(j+1)} \leftarrow argmin_\theta L(D_c \cup D_p(x_t^{\text{forecast}(j+1)}))$
  $w^{(j+1)} \leftarrow W(D_{val}, \mathcal{A}_{\text{Trigger}}, \theta^{(j+1)})$
  if $w^{(j+1)} > w^{(j)}$ **then**
    $x_i^{best} \leftarrow x_t^{\text{forecast}(j)}$
    break
  end if
  $\alpha^{(j+1)} \leftarrow \beta \cdot \alpha^{(j)}$
  $x_i^{best} \leftarrow x_i^{\text{forecast}(j+1)}$
end for

B. Cost-oriented Data Poisoning

For the proposed CODP approach, let $D_{tr} = D_p \cup D_c$ denotes the training set, we randomly select the samples in the training set with proportion α_p as poisoning set D_p, while the rest of samples in D_c will not be tampered. By adding disturbance to D_p, we can change the spatial distribution of the whole training set and reconstruct the mapping relationship between input features and predicted inertia value with the model parameterized by θ_p. In this way, the poisoned model will work abnormally (i.e., stealthily increase the system operation cost with targeted inertia forecasts). Note that the validation set D_{val} is applied in CODP to evaluate attack effect W and generate attack, as the data poisoning attackers are restricted from obtaining knowledge of test set. However it should be emphasized that the generated attack will be finally applied to test set so that to achieve a complete CODP attack, as shown in Fig. 3.

The ML-based inertia forecasting model is first initialized based on clean data set $(D_c \cup D_p^0)$, and obtains the output forecasts $\hat{y}$ on validation set. On this basis, as illustrated in Sec. III-A, the cost-oriented target $\lambda \in \mathcal{A}_{\text{Trigger}}$ can be determined with Eq. 7 before executing the poisoning attack. Under the attack target λ, the proposed CODP attack can be expressed with a bilevel optimization problem [21], [28]:

$$\underset{D_p}{\arg\min} \quad W = \frac{1}{N_{val}} \sum_{t=1}^{N_{val}} (f_{\theta_p}(x_t) - (1+\lambda_{t+k}) y_{t+k})^2 \tag{10a}$$

$$s.t. \quad \theta_p \in \underset{\theta_p}{\arg\min} \ L(D_p \cup D_c, f_\theta) \tag{10b}$$

$$x_t \in D_{val}, \quad t = 1, 2, ..., N_{val} \tag{10c}$$

where the upper-level objective function W is defined to minimize the gap between inertia forecasts and cost-oriented target, and the decision vector is poisoning samples $x_p \in D_p$; the lower-level objective function $L(D_p \cup D_c, f_\theta)$ represents the loss function of ML-based inertia forecasting model, corresponding to the process of training to update the poisoned model θ_p. It can be found that the calculation of W depends on the lower-level decision vector poisoned model θ_p, meanwhile, the update of θ_p also considers the change of upper-level decision vector D_p.

To solve the above bilevel optimization problem, the Gradient Decent Method is employed in this work. The poisoning samples $x_p \in D_p$ will be modified with gradient $\nabla_{x_p} W$ to optimize the upper-level objective function

$W(D_{val}, \mathcal{A}_{\text{Trigger}}, \theta_p)$. Considering the dependency of θ_p on x_p in the lower-level objective function, the gradient can be made a decomposition with the chain rule [21], [28]:

$$\nabla_{x_p} W = \nabla_{x_p} \theta_p(x_p)^T \cdot \nabla_{\theta_p} W. \tag{11}$$

where $\nabla_{x_p} \theta_p(x_p)^T$ involves the training process of inertia forecasting model, and the influence of x_p on model parameters θ_p is also reflected. For the neural network model, $\nabla_{x_p} \theta_p(x_p)^T$ is a non-convex problem and it is difficult to obtain an accurate numerical solution. Therefore, we choose the linear regression model as the forecasting model in the CODP and demonstrate its transferability to other ML-based models [22]. For linear regression model $f(x, \theta) = \omega^T x + b$, the loss function L can be expressed as:

$$L = \frac{1}{N_{tr}} \sum_{t=1}^{N_{tr}} (\omega^T x_t + b - y_{t+k})^2 + \eta \cdot \Omega(\omega) \tag{12}$$

where $\Omega(\omega)$ is the regularization term for feature selection. Under different regularization strategies, it is defined as $\Omega(\omega) = 0$ (Least Squares), $\Omega(\omega) = \|\omega\|_1$ (LASSO) and $\Omega(\omega) = \frac{1}{2}\|\omega\|_2{}^2$ (Ridge). Thus the lower-level optimization representing the training process can be translated with Karush-Kuhn-Tucker (KKT) equilibrium condition: $\nabla_\theta L(D_c \cup D_p) = 0$. In order to ensure the validity of the above condition when updating x_p [21], [28], we set its derivative with respect to x_p to be zero, which is finally expressed as:

$$\nabla_{x_p}(\nabla_\theta L(D_c \cup D_p)) = 0 \tag{13a}$$

$$\nabla_{x_p} \nabla_\theta L + \nabla_{x_p} \theta^T \cdot \nabla_\theta^2 L = 0 \tag{13b}$$

$$\nabla_{x_p} \theta_p(x_p)^T = -\nabla_{x_p} \nabla_\theta L \cdot (\nabla_\theta^2 L)^{-1} \tag{13c}$$

Due to the regularization term, the L could be not differentiable (e.g., $\Omega(\omega) = \|\omega\|_1$), and this is solved with subgradients referring to [28]. Specifically, the $\nabla_{x_p} \theta_p(x_p)^T$ can be obtained from Eq. 13 as:

$$\nabla_{x_p} \theta_p(x_p)^T = -\frac{2}{N_{tr}} \begin{bmatrix} M & \omega \end{bmatrix} \begin{bmatrix} \Sigma + \eta v & \mu \\ \mu^T & 1 \end{bmatrix}^{-1} \tag{14}$$

where $\Sigma = \frac{1}{N_{tr}} \sum_t x_t x_t^T$, $\mu = \frac{1}{N_{tr}} \sum_t x_t$, and $M = x_p \omega^T + (f(x_p) - y_p)\mathbb{I}$. The term v is zero in Least Squares and LASSO, and the identity matrix $\mathbb{I}$ in Ridge.

Meanwhile, the gradient $\nabla_{\theta_p} W$ can be expanded as:

$$\nabla_{\theta_p} W = \begin{bmatrix} \frac{2}{N_{val}} \sum_{t=1}^{N_{val}} (f_\theta(x_t) - (1 + \lambda_{t+k}) y_{t+k}) \cdot x_t \\ \frac{2}{N_{val}} \sum_{t=1}^{N_{val}} (f_\theta(x_t) - (1 + \lambda_{t+k}) \cdot y_{t+k}) \end{bmatrix} \tag{15}$$

On this basis, each $x_p \in D_p$ can be updated through line search, where we conduct multi-step search and shrink step size after each step, until object function stops decreasing or reach the searching times. While the whole poisoning set D_p is updated with line search, we will retrain the model θ_p with the updated poisoning set and test the poisoning effect with objection function $W(\theta_p)$. The above processes are iterated alternately, and the poisoning set $D_p^{(i+1)}$ is constructed based on the model updated in the previous round $\theta_p^{(i)}$ in each iteration. The algorithm will end up after finite iterations or meet the stop condition $||w^{i+1} - w^i|| < \epsilon_{\text{stop}}$ in advance. Finally, the poisoned model θ_p^{best} and poisoning set D_p^{best} can be obtained. Overall, the complete algorithm of the proposed CODP is presented in Algorithm 1.

Algorithm 2 Cost-oriented Adversarial Attack

Initialization: Existing forecasting model f_θ, test set $(x_t, y_{t+k}) \in D_{test}, t = 1, 2, ...N_{test}$
Output: ${x_t}^* \in D_{test}^*$

$\hat{y} \leftarrow f_\theta(x),\ x \in D_{test}$
$\mathcal{A}_{\text{Trigger}} \leftarrow C((1+\lambda)\hat{y}|x) - C(\hat{y}|x)$ ▷ *Identify attack target*
for $t = 1, 2, ..., N_{test}$ **do**
  if $\lambda_t \neq 0$ **then** ▷ *Locate cost-sensitive periods*
    $x_t^* = \text{FGSM}(x_t, f_\theta, \lambda_{t+k})$
    $D_{test}^* \leftarrow x_t^*$
  end if
end for
return D_{test}^*

Def: FGSM$(x_t, f_\theta, \lambda_{t+k})$:

for $i = 1, 2, ..., N_{search}$ **do** ▷ *Loops for searching the gradient*
  for $j \in x^{forecast}$ **do** ▷ *Features can be attacked*
    $\text{W}(x_t, \lambda_{t+k}, \theta) = (f_\theta(x_t) - (1 + \lambda_{t+k}) y_{t+k})^2$
    $x_t^{j\,(i+1)} = x_t^{j\,(i)} - \alpha \cdot sign(\nabla_{x_t^{j\,(i)}} \text{W}(x_t, \lambda_{t+k}, \theta))$
  end for
end for

C. Cost-oriented Adversarial Attack

To achieve the cost-oriented target, another way to influence the performance of the ML-based inertia forecasting model is to directly inject false data into the test set. Therefore, we propose the Cost-oriented Adversarial Attack (COAA) to generate adversarial samples for the test set and keep the original training set and forecasting model. According to Eq. 7, we first determine the cost-sensitive periods $\lambda_{t+k} \neq 0$ for the online test stage so that the inertia forecasts in these intervals are expected to perform abnormally as designed.

Then the next step is to generate adversarial samples only within the cost-sensitive periods of the online test stage so as to precisely achieve the different effects between the sensitive and non-sensitive periods. In particular, the adversarial example at timestamp t can be obtained via solving the following problem:

$$\underset{x_t^{\text{forecast}*}}{\arg\min} \quad (f_\theta(x_t) - (1 + \lambda_{t+k}) y_{t+k})^2 \tag{16a}$$

$$s.t. \quad x_t \in D_{test} \wedge \lambda_{t+k} \neq 0, \tag{16b}$$

Given the white-box model f_θ, the gradient between object function $W(x_t, \lambda_{t+k}, \theta) = (f_\theta(x_t) - (1 + \lambda_{t+k}) y_{t+k})^2$ and input features x_t can be computed efficiently using backpropagation. Therefore, we use the Fast Gradient Method (FGSM) [29] to solve the optimization problem. We search for the optimal adversarial sample with multiple iterations, in each iteration i, x_i is updated depending on the sign of the gradient:

$$x_t^{(i+1)} = x_t^{(i)} - \alpha \cdot sign(\nabla_{x_t^{(i)}} (W(x_t, \lambda_{t+k}, \theta)). \tag{17}$$

Note that we only manipulate the features in $x_t^{\text{forecast}} = (\hat{P}_t^{\text{load}}, \hat{P}_t^{\text{gen}}, \hat{P}_t^{\text{solar}}, \hat{P}_t^{\text{onshore}}, \hat{P}_t^{\text{offshore}})$ considering the limitation of attackers' capability given in Eq. 3. The detailed information is elaborated in Algorithm 2.

IV. Vulnerability Analysis Framework

This section proposes the vulnerability analysis framework for the ML-based inertia forecasting models, which is designed to provide a reference from the perspective of security for the selection and improvement of the ML-based inertia forecasting models. In Sec. III, we propose two cost-oriented attacks from the perspective of power systems operation cost. The generation processes of the proposed attacks can be summarized in Fig. 3 in the white-box case. Then the generated attack vectors can be used to evaluate the vulnerability of different tested ML-based inertia forecasting models in the black-box case due to the characteristic of transferability. In particular, the proposed CODP and COAA methods can generate a tampered malicious training set $D_c \cup D_p$ or test set D_{test}^*, respectively, which are the carrier of attacks. Thus we can explore the vulnerabilities in the offline training and online test stages by applying the tampered malicious data set to various forecasting models, as shown in Fig. 3.

Based on the output of influenced inertia forecast, we design the following metrics to quantify the changes in the forecasting performance of the model before and after the attack and assess the vulnerability:

$$\text{MAPE}_{\text{pre}} = \frac{100}{N_t} \sum_t \left| \frac{y_{true,t} - y_{pre,t}}{y_{true,t}} \right| \tag{18a}$$

$$\text{MAPE}_{\text{atk}} = \frac{100}{N_t} \sum_t \left| \frac{y_{true,t} - y_{atk,t}}{y_{true,t}} \right| \tag{18b}$$

$$\text{Rate} = \frac{\text{MAPE}_{\text{atk}} - \text{MAPE}_{\text{pre}}}{\text{MAPE}_{\text{pre}}} \tag{18c}$$

where $t \in \mathcal{A}_b$ or $\mathcal{A}_p$ or $\{1, ..., N_{test}\}$, the corresponding capacity of which is N_t.

$$\text{Suc} = \frac{N_s}{N_p} * 100\%; \tag{19}$$

$$\text{Rag} = \frac{100\%}{N_s} \sum_{t \in \mathcal{A}_s} \left| \frac{y_{pre,t} - y_{atk,t}}{y_{pre,t}} \right| \tag{20}$$

$$\text{R}_{err} = \frac{N_b \sum_{t \in \mathcal{A}_p} \left| \frac{y_{true,t} - y_{atk,t}}{y_{true,t}} \right|}{N_p \sum_{t \in \mathcal{A}_b} \left| \frac{y_{true,t} - y_{atk,t}}{y_{true,t}} \right|} \tag{21}$$

where $\mathcal{A}_b = \{t \in \mathbb{Z} \mid \lambda_t = 0\}$, $\mathcal{A}_p = \{t \in \mathbb{Z} \mid \lambda_t \neq 0\}$, $\mathcal{A}_s = \{t \in \mathcal{A}_p \mid \lambda_t (y_{atk,t} - y_{pre,t}) > 0\}$; the y_{true}, y_{pre} and y_{atk} are the true value and the predicted value before and after the attack of inertia, respectively; the $\mathcal{A}_p$ and $\mathcal{A}_b$ represent the cost-sensitive and non-sensitive periods in test set; within the cost-sensitive periods, we defined the area where the predicted value changes in the desired direction (e.g., over-prediction when $\lambda_t > 0$) after the attack as successful attack intervals, which was represented by $\mathcal{A}_s$; the N_b, N_p and N_s denote the number of samples in their corresponding intervals.

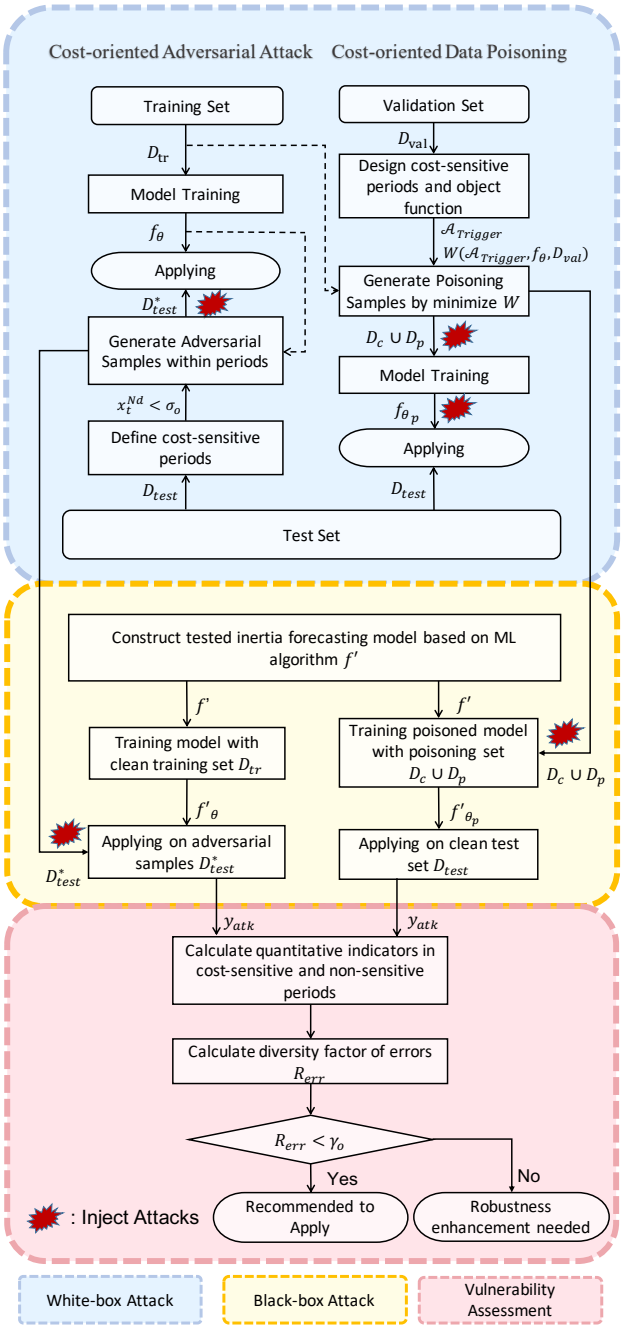


Fig. 3. The proposed vulnerability assessment framework for ML-based inertia forecasting model under cost-oriented attacks.

As the cost-oriented attacks are designed to have different effects inside and outside cost-sensitive periods, the quantitative indicators will be calculated respectively. First, as the forecasting error is an important index to evaluate the performance of forecasting models, we can observe the attack effect on model performance based on error variation before and after attack. We calculate the forecasting errors before (MAPE_{pre}) and after (MAPE_{atk}) attacks by regions, as well as the proportion of change Rate between them. In non-sensitive periods, the closer between MAPE_{pre} and MAPE_{atk}, the less impact the attack has in these periods. In other words, the attack is more secluded and the model

is less robust. Second, as illustrated in Eq. 7, the attack target is designed with a specified misprediction direction, thus we compare the inertia forecast before and after attack to evaluate whether the attack accurately achieves an over-prediction or under-prediction effect. $\mathrm{Suc} \in [0, 100\%]$ is the success rate, i.e., the proportion of samples whose predicted values misestimate in the desired direction ($\lambda_t > 0$ or $\lambda_t < 0$) after attack within this periods. And Rag is the average value of the variation rangeability of predicted values before and after the attack of the above samples. A higher Suc and Rag indicate lower robustness of the model.

Comprehensively, the model with lower error in non-sensitive periods and higher error in cost-sensitive periods indicates lower robustness, which means a more significant impact difference between different periods. It will generate higher operation cost increment under the same global error level, proved in the experiments in Sec. V. Therefore, we propose a comprehensive index R_{err} to evaluate the vulnerability of ML-based inertia forecasting models under cost-oriented attacks, which can quantify the diversity factor of error levels inside and outside cost-sensitive periods. As shown in Fig. 3, after calculating the diversity factor of error, based on a user-defined threshold γ_o, the models with $\mathrm{R}_{err} < \gamma_o$ indicate satisfied robustness and thus can be recommended to be used in practice; while those models having high diversity factor (i.e., $\mathrm{R}_{err} > \gamma_o$) are suggested to improve their robustness further (e.g., via adversarial training approach).

V. Case Study

A. Data Set and Model Construction

We evaluate the performance of the proposed approaches based on a three-year power system data set of Great Britain from 2016 to 2018, which is composed by public data from Elexon's BM reports ENTSO-E's data transparency platform National Grid's data explorer [U+FF0C] and the Nordpool website . These data are combined under a consistent time zone with $30mins$ time granularity, and the finished features are summarized in Tab. I, containing short-term prediction values of system power, interconnector flows with outside, environmental indicators such as weather characteristics, seasonal factors, and time factors, as well as system inertia. The details about the construction of the data set can be obtained in [9].

TABLE I
Selected features for inertia forecasting model.

Feature type	**Feature name**
Day-ahead forecast capacity	load, coal, solar, onshore, offshore
Interconnector flows	French, Dutch, Irish, East-West
Environment state	season, data, weekday or weekend, hour, temperature
Inertia	total system inertia (TARGET)

Regarding the ML-based day-ahead inertia forecasting model, the first two-year data from 1st January 2016 to 31st December 2017 are used as the training set in the COAA, whilst it is divided into the training set and verification set with a proportion of 3:1 in the CODP. Furthermore, the rest of the one-year data in 2018 is used as the test set to evaluate the model performance under different attack methods consistently. In terms of the employed forecasting model, we select four representative algorithms [30]: LR, FNN, RNN, and LSTM to assess their vulnerability. Meanwhile, we need to emphasize that the vulnerability assessment framework is universal and applicable to all kinds of machine learning models. For the COAA, we assume that the attackers only have access to querying and tampering with the test set. On the other hand, for the CODP, the attackers are assumed to have access to querying and tampering with the training set, hence the validation set is utilized to evaluate the attack effect in the test phase.

B. Impact Analysis: Forecasting Performance (White-Box)

The objective of this part lies in investigating the impacts of proposed cost-oriented approaches in terms of the inertia forecasting performance under the assumption of white-box. Furthermore, the final target of the generated white-box attack is to influence a black-box inertia forecasting model based on transferability, as illustrated in Sec. IV. Therefore, the CODP is constructed on an LR forecasting model, which can satisfy the convex optimization condition for gradient calculation, and the COAA is constructed on an LSTM forecasting model for deep learning. According to the cost impact analysis of the data set in Sec. V-A on the operation simulation model, as shown in Fig. 2. The division index of cost-sensitive periods is $x_t^{Nd} < \sigma_o$ (under normalization $\sigma_o = 0.35$) and the cost-oriented target is set as $\lambda_{t+k} > 0$ for over-prediction effect with higher cost increment. The data tampering is limited within features in x^{forecast}. For the CODP, the proportion of poisoning samples α_p is set to 0.5.

First, to visualize the effects of proposed approaches, Fig. 4 illustrates the inertia forecasts under CODP and COAA attacks over the same period, respectively. Note that the cost-sensitive periods are represented by gray shade, and the true value, predicted value before and after the attack of system inertia are indicated by the curves in green, blue, and orange (or red), respectively. As can be seen, the predicted inertia forecasts are evidently increased during the cost-sensitive periods, whereas the inertia values of non-sensitive periods are generally kept at the same levels compared with the non-attack case, demonstrating the effectiveness and the stealthiness of the proposed methods. It should be emphasized that, for the proposed CODP, the poisoning samples are randomly injected into the training dataset without the limitation of injection periods. It works by modifying the learned regression rules of the poisoned model in order to guarantee stealthiness via distinguishing the forecasting performance within and outside the targeted periods.

Furthermore, to quantitatively investigate the impacts of the proposed methods across different levels of disturbance constraints ξ, the results of both COAA and CODP approaches are given in Tab. II and Tab. III with $\xi = 5\%$, $\xi = 10\%$, $\xi = 20\%$ and $\xi = 30\%$, respectively. First, it can be seen that both of the proposed COAA and CODP methods

TABLE II
EFFECT OF WHITE-BOX CODP ATTACKS WITH DIFFERENT LEVELS OF DISTURBANCE CONSTRAINTS (ξ).

ξ	Non-sensitive Periods			Cost-sensitive Periods					Global Periods		
	MAPE_{pre}	MAPE_{atk}	Rate	MAPE_{pre}	MAPE_{atk}	Rate	Suc	Rag	MAPE_{pre}	MAPE_{atk}	Rate
5%	9.337	9.554	2.32%	11.653	13.035	11.86%	99.19%	2.81%	9.959	10.488	5.31%
10%		9.563	2.42%		13.591	16.63%	99.85%	3.81%		10.644	6.88%
20%		9.573	2.52%		13.818	18.58%	99.89%	4.16%		10.712	7.56%
30%		9.624	3.07%		14.078	20.80%	100.00%	4.62%		10.819	8.64%

TABLE III
EFFECT OF WHITE-BOX COAA ATTACKS WITH DIFFERENT LEVELS OF DISTURBANCE CONSTRAINTS (ξ).

ξ	Non-sensitive Periods			Cost-sensitive Periods					Global Periods		
	MAPE_{pre}	MAPE_{atk}	Rate	MAPE_{pre}	MAPE_{atk}	Rate	Suc	Rag	MAPE_{pre}	MAPE_{atk}	Rate
5%	9.383	9.383	0.00%	10.908	11.717	7.42%	100.00%	3.83%	9.792	10.059	2.73%
10%					12.596	15.48%	100.00%	6.08%		10.316	5.35%
20%					12.742	16.82%	100.00%	6.47%		10.358	5.78%
30%					12.759	16.98%	100.00%	6.51%		10.362	5.83%

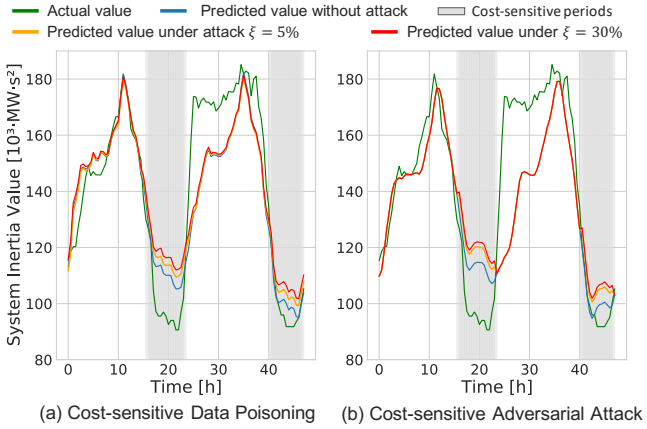


Fig. 4. Inertia forecasts under white-box CODP (a) and COAA (b) with different levels of disturbance constraints ($\xi = 5\%, 30\%$).

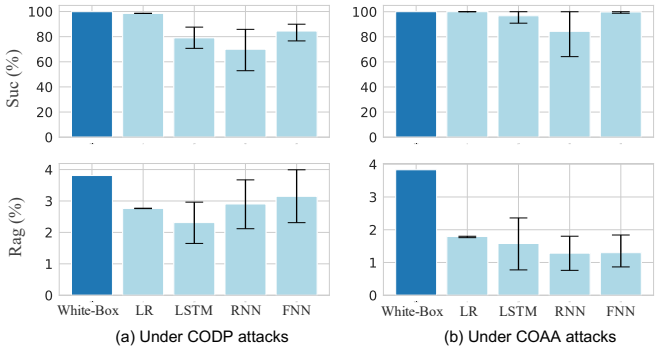


Fig. 5. Transferability analysis of black-box CODP (a) and COAA (b) based on various ML-based inertia forecasting models

can achieve the pre-determined attack targets indicated by the overpredicted inertia forecasts at most timestamps (i.e., $Suc > 99\%$) and the increment of forecasting error between 7% to 20% during the cost-sensitive periods. Moreover, under the same level of ξ, the proposed CODP can result in more severe impacts on the forecasting performance than that of the COAA, as shown by the 7.4% to 59.8% higher metric values of Rate of errors within cost-sensitive periods; while the COAA can influence the forecasting performance closer to the target with higher metric values of Suc and Rag. On the other hand, with the increasing value of ξ, the constraint of injection magnitude is relaxed, and thus higher Rag and $Rate$ can be obtained for both CODP and COAA approaches.

Additionally, the stealthiness of the proposed approaches can be quantitatively demonstrated via the following results: 1) both CODP and COAA approaches can result in higher forecasting errors during the cost-sensitive periods (i.e., $Rate$ is up to 20.8%) while keeping normal forecasting performance during the non-sensitive periods (i.e., $Rate < 3.1\%$); 2) the overall forecasting performance is almost retained at the same level before and after the injection of attacks with the metric values of $Rate$ between approximately 2.73% to 8.64%. In particular, the MAPE within non-sensitive periods remains the same under each COAA attack thanks to the point-by-point attack strategy of the cost-oriented adversarial attack, which performs more stealthy than CODP attack without considering the data integrity attack injection path (training or test phase).

C. Impact Analysis: Forecasting Performance (Black-Box)

This subsection aims to explore the transferability of the proposed methods on different ML-based models and then make a comparison among them. As illustrated in Sec. IV, we implement black-box attacks based on the malicious data set (D_p and D_t^*) generated in the above white-box attacks. To fairly analyze the transferability of the proposed methods, we select CODP($\xi = 10\%$) and COAA($\xi = 5\%$), which exhibit similar forecasting performance influence (e.g., Rag and Suc).

The result is shown in Fig. 5, which focuses on the effect of targeted attack (over-prediction) within the cost-sensitive periods. We carried out multiple black-box attacks for each model. The average value and distribution range (extreme

values) of the metrics Rag and Suc are represented in the histograms with error bars, respectively.

As can be seen, the pre-determined attack effects still exist when applying the proposed cost-oriented attacks to the black-box models, which can be demonstrated with the metrics of $Suc > 69\%$ and $Rag > 1.2\%$. Meanwhile, compared with the white-box attacks, the attack effects in the black-box cases are diminished to some extent: the success rate Suc of CODP within cost-sensitive periods exhibits a significant decrease between 1.4% to 29.9%; while the COAA attack maintains a high Suc but the metric value of Rag is decreased by $53.2\% - 66.5\%$.

D. Impact Analysis: System Operation Cost

In this subsection, we aim to explore further the impact of the proposed approaches in terms of the system operation cost based on the system operation model introduced in Sec. III-A. Here we compare the cost-oriented attacks with existing data integrity attacks, including the global no-target data poisoning (G-DP) in [21] and the global target adversarial attack (G-AA) in [16], which were developed to diminish the overall forecasting performance.

To quantify the impacts on system operation costs, the increased rate of system operation cost under conventional global attack and the proposed cost-oriented attack (i.e., the ratio of change in system operation cost before and after the injection of attack) are calculated and presented in Tab.IV. To ensure fairness, we make the above attacks have nearly the same global errors under white-box assumption.

As can be seen, the cost increment induced by the proposed COAA and CODP in both white-box and black-box cases is more than twice the costs of the corresponding global attacks, respectively. Furthermore, it can be found that cost-oriented attacks, CODP and COAA, exhibit higher transferability than G-DP and G-AA regarding the impacts of operation cost, as shown by the fact that the average cost increments under black-box attacks are 51.58% (CODP) and 46.1% (COAA) of those under white-box attacks, and the corresponding metric values for G-DP and G-AA are only 30.39% and 36.00%, respectively.

TABLE IV
INCREASE RATE OF OPERATION COST UNDER GLOBAL ATTACK AND COST-ORIENTED ATTACK.

Type	Name	White box	Black-box			
			LR	LSTM	RNN	FNN
Data Poisoning	G-DP[21]	2.81%	0.29%	0.73%	1.30%	1.72%
	CODP	**4.97%**	**1.61%**	**2.29%**	**3.00%**	**4.03%**
Adversarial Attack	G-AA[16]	2.30%	1.61%	0.76%	0.91%	0.30%
	COAA	**5.20%**	**3.79%**	**2.03%**	**2.48%**	**1.15%**

Moreover, the significant impacts of the injection periods selection for the proposed cost-oriented attacks are further investigated via comparing the system operation costs between the cases with randomly selected periods (i.e., random adversarial attack) and cost-oriented periods with low net demand level $x^{Nd} < \lambda_o$, as shown in Fig. 6. Note that

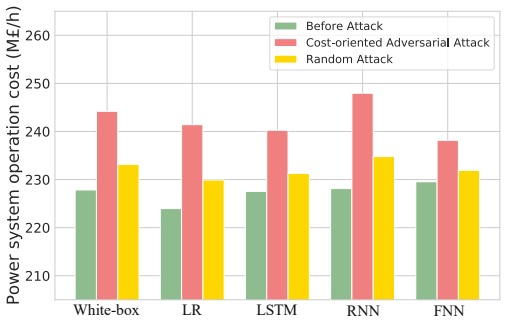


Fig. 6. System operation cost under COAA and RAA ($\xi = 20\%$)

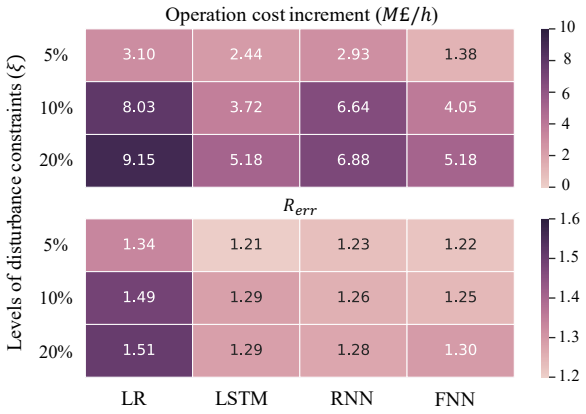


Fig. 7. Operation cost increment (M£/h) and R_{err} under COAA across different levels of disturbance constraint (ξ).

the total injection periods of both attacks are kept the same. For the white-box case, it can be seen that the average cost increment induced via cost-oriented injection is significantly higher (i.e., approximately 207%) than random attacks, and this impact difference is even more significant with about 3.2 times higher cost increments in the black-box case. Therefore, the investigated ML-based inertia forecasting model is more vulnerable during the cost-sensitive periods, and more protection resources should be allocated there.

Additionally, a sensitivity analysis across different levels of disturbance constraints (ξ) is also provided in Fig. 7, quantified via the increment of system operation cost (M£/h) and R_{err}. It is understandable that, with the increasing level of injected disturbance, the model is more vulnerable, and higher operation costs will be obtained. For example, the average cost increment for all ML models increases from 2.42 M£/h ($\xi = 5\%$) to 6.64 M£/h ($\xi = 20\%$) along with the decrements of forecasting accuracy. This provides an idea of mitigating such attacks via applying outlier detection during the cost-sensitive periods to identify the data with high disturbance. Finally, the consistency between R_{err} and operation cost also demonstrates the effectiveness of the proposed metric to quantify the vulnerability of ML-based inertia forecasting models under various potential threats.

VI. Conclusion

This paper proposes a novel vulnerability assessment framework for ML-based inertia forecasting models, considering the economic impact on system operation when the inertia forecasting models suffer potential attacks. In this framework, we propose cost-oriented data integrity attacks, which aim to stealthily maximize the system operation cost by locating the cost-sensitive periods and injecting disturbances following the pre-defined cost-oriented target. In particular, the CODP and COAA approaches are designed for the training and test processes, respectively, along with a series of vulnerability assessment metrics. Case studies based on real data collected from the GB power systems are carried out to analyze the vulnerability under the proposed attacks in both white-box and black-box cases based on multiple state-of-the-art forecasting models. The results demonstrate the effectiveness and stealthiness of the proposed approaches in terms of both forecasting accuracy and power system cost increment. At the same level of inertia forecasting errors, the operation cost increment under the proposed cost-oriented attacks is up to approximately 226% of that under the existing untargeted data integrity attacks.

Future work will be conducted to develop the corresponding detection and defense mechanisms to build a robust ML-based inertia forecasting model, which may include conducting adversarial training, adding a cost-sensitive term into the loss function, or developing anomaly detection techniques considering the probability distribution of input data. Moreover, it is also feasible and valuable to extend the concept of cost-oriented attack to other forecasting tasks in power systems, such as load or renewable energy forecasting.

References

[1] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbič, "Foundations and challenges of low-inertia systems," in *2018 Power Systems Computation Conference (PSCC)*. IEEE, 2018, pp. 1–25.

[2] Y. Bian, H. Wyman-Pain, F. Li, R. Bhakar, S. Mishra, and N. P. Padhy, "Demand side contributions for system inertia in the gb power system," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 3521–3530, 2017.

[3] P. Du and J. Matevosyan, "Forecast system inertia condition and its impact to integrate more renewables," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1531–1533, 2017.

[4] Entso-E, "Fast frequency reserve – solution to the nordic inertia challenge," Tech. Rep., 2019.

[5] E. R. Paidi, H. Marzooghi, J. Yu, and V. Terzija, "Development and validation of artificial neural network-based tools for forecasting of power system inertia with wind farms penetration," *IEEE Systems Journal*, vol. 14, no. 4, pp. 4978–4989, 2020.

[6] D. Wilson, J. Warichet, M. Eves, and N. Al-Ashwal, "D2.3: Lessons learned from monitoring forecasting kpis on impact of pe penetration." H2020 MIGRATE project, Tech. Rep., 2018.

[7] F. Gonzalez-Longatt, M. Acosta, H. Chamorro, and D. Topic, "Short-term kinetic energy forecast using a structural time series model: study case of nordic power system," in *2020 international conference on smart systems and technologies (SST)*. IEEE, 2020, pp. 173–178.

[8] J. Graham, E. Heylen, Y. Bian, and F. Teng, "Benchmarking explanatory models for inertia forecasting using public data of the nordic area," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2022, pp. 1–6.

[9] E. Heylen, J. Browell, and F. Teng, "Probabilistic day-ahead inertia forecasting," *IEEE Transactions on Power Systems*, 2021.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[11] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 1467–1474.

[12] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[13] N. Müller, D. Kowatsch, and K. Böttinger, "Data poisoning attacks on regression learning and corresponding defenses," in *2020 IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2020, pp. 80–89.

[14] M. Jagielski, G. Severi, N. Pousette Harger, and A. Oprea, "Subpopulation data poisoning attacks," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 3104–3122.

[15] C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7614–7623.

[16] Y. Chen, Y. Tan, and B. Zhang, "Exploiting vulnerabilities of load forecasting through adversarial attacks," in *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, 2019, pp. 1–11.

[17] Y. Chakhchoukh, P. Panciatici, and L. Mili, "Electric load forecasting based on statistical robust methods," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 982–991, 2010.

[18] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2484–2493.

[19] G. Li and H.-D. Chiang, "Toward cost-oriented forecasting of wind power generation," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2508–2517, 2018.

[20] J. Zhang, Y. Wang, and G. Hug, "Cost-oriented load forecasting," *arXiv preprint arXiv:2107.01861*, 2021.

[21] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 19–35.

[22] Y. Liang, D. He, and D. Chen, "Poisoning attack on load forecasting," in *2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2019, pp. 1230–1235.

[23] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack method against neural networks," *arXiv preprint arXiv:1703.01340*, 2017.

[24] A. Sayghe, J. Zhao, and C. Konstantinou, "Evasion attacks with adversarial deep learning against power system state estimation," in *2020 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2020, pp. 1–5.

[25] J. Li, Y. Yang, and J. S. Sun, "Searchfromfree: Adversarial measurements for machine learning-based energy theft detection," in *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2020, pp. 1–6.

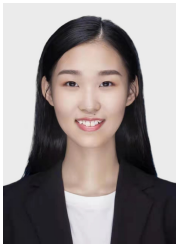
[26] Q. Song, R. Tan, C. Ren, and Y. Xu, "Understanding credibility of adversarial examples against smart grid: A case study for voltage stability assessment," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, 2021, pp. 95–106.

[27] Z. Chu, U. Markovic, G. Hug, and F. Teng, "Towards optimal system scheduling with synthetic inertia provision from wind turbines," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4056–4066, 2020.

[28] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, "Is feature selection secure against training data poisoning?" in *international conference on machine learning*. PMLR, 2015, pp. 1689–1698.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

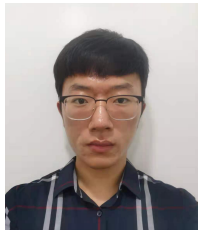
[30] A. Almalaq and G. Edwards, "A review of deep learning methods applied on load forecasting," in *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2017, pp. 511–516.



Yan Chen (Student Member, IEEE) received the B.Eng. degree from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. She is currently pursuing the MA.Eng. degree with the Department of Polytechnic, Zhejiang University, Hangzhou, China. Her research interests include artificial intelligence security and machine learning in power system.

Mingyang Sun (Member, IEEE) received the Ph.D. degree from the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., in 2017. From 2017 to 2019, he was a Research Associate and a DSI Affiliate Fellow with Imperial College London.

He is currently a Professor of Control Science and Engineering under the Hundred Talents Program with Zhejiang University, Hangzhou, China. Also, he is an Honorary Lecturer with Imperial College London. His research interests include AI in energy systems and cyber-physical energy system security and control.



Zhongda Chu (Member, IEEE) received the M.Sc. degree in electrical engineering and information technology from the Swiss Federal Institute of Technology, Zürich, Switzerland, in 2018 and the Ph.D. degree in electrical engineering from Imperial College London, London, U.K., in 2022. He is currently a research associate with the Department of Electrical and Electronic Engineering, Imperial College London. His research interests include control and optimization of power systems with high power electronics penetration.

Simon Camal received his Eng. degree in Energy and Environmental Engineering from Mines Nancy, France in 2010 and his European Master of Science in Renewable Energy from Loughborough University, UK in 2011. He obtained his PhD at MINES ParisTech - PSL University in 2020 on forecasting and optimization of ancillary service provision by renewable energy power plants. He is currently Project Manager of the Horizon2020 Smart4RES Project, working at MINES ParisTech Center for Processes, Renewable Energies and Energy Systems (PERSEE) in Sophia Antipolis, France.

George Kariniotakis (Senior Member, IEEE) was born in Athens, Greece. He received the Eng. and M.Sc. degrees from Greece, in 1990 and 1992, respectively, and the Ph.D. degree from Ecole des Mines de Paris, Paris, France, in 1996. Currently he is Professor at MINES ParisTech. He is head of the Renewable Energies and Smart Grids Group at PERSEE Centre. He has authored more than 300 scientific publications in journals and conferences. He has been involved as participant or coordinator in more than 45 RD projects in the fields of renewable energies and distributed generation. Among them, he was the coordinator of some major EU projects in the field of wind power forecasting such as Anemos, Anemos.plus and SafeWind projects. Currently he is the coordinator of the H2020 Smart4RES project. His scientific interests include among others timeseries forecasting, decision making under uncertainty, modeling, management and planning of power systems.

Fei Teng (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from Beihang University, China, in 2009, and the M.Sc. and Ph.D. degrees in electrical engineering from Imperial College London, U.K., in 2010 and 2015, respectively, where he is currently a Senior Lecturer with the Department of Electrical and Electronic Engineering. His research focuses on the power system operation with high penetration of Inverter-Based Resources (IBRs) and the Cyber-resilient and Privacy-preserving cyber-physical power grid.