






REVIEW

Deep learning for cybersecurity in smart grids: Review and perspectives

Jiaqi Ruan^{1,2}  | Gaoqi Liang^{1,2} | Junhua Zhao^{1,2}  | Huan Zhao¹  | Jing Qiu³  | Fushuan Wen⁴  | Zhao Yang Dong⁵

¹School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

²Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China

³School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia

⁴College of Electrical Engineering, Zhejiang University, Hangzhou, China

⁵School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Correspondence

Junhua Zhao, Gaoqi Liang and Jiaqi Ruan, School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 518172, Shenzhen, China.
Email: zhaojunhua@cuhk.edu.cn;
lianggaoqi@cuhk.edu.cn;
jiaqiruan@link.cuhk.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 72171206, 71931003, 72061147004; Shenzhen Institute of Artificial Intelligence and Robotics for Society; Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network, Grant/Award Number: ZDSYS2022060100601002; Guangdong Regional Joint Fund-Youth Fund, Grant/Award Number: 2021A1515110084

Abstract

Protecting cybersecurity is a non-negotiable task for smart grids (SG) and has garnered significant attention in recent years. The application of artificial intelligence, particularly deep learning (DL), holds great promise for enhancing the cybersecurity of SG. Nevertheless, previous surveys and review articles have failed to comprehensively investigate the intersection between DL and SG cybersecurity. To address this gap, this study presents a survey of the latest advancements in DL technology and their relevance to SG cybersecurity. First, the functional mechanisms and scope of application of common DL techniques are explored. Subsequently, SG cyberthreats are categorised into distinct types of cyberattacks that have not been systematically examined in previous surveys. Based on this, a thorough review of the application of DL techniques in addressing each cyberthreat along with recommendations and a generalised framework for enhancing cyberattack detection using DL is offered. Finally, insights are provided into the emerging challenges presented by DL applications in SG cybersecurity that are yet to be widely acknowledged, and potential research avenues are proposed to address or alleviate these challenges.

KEYWORDS

artificial intelligence, big data, cyberattack, cybersecurity, deep learning, smart grid

1 | INTRODUCTION

The electric power grid is the most fundamental and sophisticated industrial system in modern society [1]. By inheriting legacies and applying advanced information communication technology, traditional power systems have evolved into smart grids (SGs) with cyber-physical systems [2]. High-speed two-way communication in the cyber layer endows SG with the ability to integrate immense uncertain resources, such as renew-

ables and electric vehicles, thereby achieving efficient resource distribution and reliable operation [3]. Advanced metering infrastructure (AMI) is a key element for fulfilling the SG cyber infrastructure [4]. Data collected from various end-points, such as smart meters and phasor measurement units (PMUs), are communicated through the AMI to the supervisory control and data acquisition (SCADA) system for real-time analyses and controls, thereby ensuring the secure, reliable, and economical operation of SGs [5]. The communication

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Energy Conversion and Economics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and the State Grid Economic & Technological Research Institute Co., Ltd.

network of the SG cyber infrastructure comprises a wide area network (WAN), neighbour area network (NAN), and home area network (HAN) [6]. However, these networks are naturally vulnerable to cyberthreats, resulting in the instability of SGs after encountering cyberattacks, such as the Ukraine and Venezuela cyberattacks witnessed in the real world [7].

Guarding cybersecurity is of utmost significance in protecting SG cyber assets and physical operations. The cybersecurity of SGs can be classified into availability, integrity, accountability, and confidentiality (CIAA) [4][6]. Availability ensures timely and reliable access to information. Integrity protects information from unauthorised manipulation and destruction. Accountability ensures that the operations of SGs are trackable and that any actions are recordable. Confidentiality preserves the authorised restrictions on access to and disclosure of information. Essentially, any cyberattack is a breach of CIAA. Compromising the cybersecurity in SG may result in a series of accidents, such as operational failures, synchronisation loss, power supply interruption, high financial damages, social welfare damages, data theft, cascading failures, and complete blackouts [8].

Owing to the rapid development of technologies, AI has been widely applied in SG to reduce labour and manual costs, and it has also been considered an effective measure for enhancing cybersecurity [9]. AI was proposed in 1956 and is defined as “the science of making computers do things that human needs intelligence to do” [10]. Owing to the advent of big data and the dramatic increase in computing power, DL has been recognised as a promising AI technique. It can mine implicit relationships in big data, helping analyse and explore knowledge beyond the human domain [11]. To date, several studies have applied DL to improve SG cybersecurity.

However, the wide application of DL models in SGs also incurs novel cyberthreats, where the DL technique is used for load forecasting, stability assessment, fault detection, control strategies, demand response etc. Although the DL model achieves a highly accurate evaluation, it is extremely vulnerable to subtle orchestrated input disturbances that can lead to a complete deviation in the result [12]. Once illegitimate attackers leverage this vulnerability to launch adversarial cyberattacks, they can have destructive and catastrophic effects on SGs [13]. However, defending against such cyberattacks on SGs has not drawn special attention in the extant literature. Moreover, there is no article that systematically and extensively reviews the advantages and disadvantages of the DL technique applied to SG cybersecurity and the envisioned future.

Based on the aforementioned analyses and challenges, this study investigates the extant literature relevant to both DL and SG cybersecurity through a systematic and comprehensive review. Specifically, it makes the following contributions:

- We categorise prevalent SG cyberthreats according to distinct cyberattack modalities, addressing the omission of certain innovative threats in prior literature. Based on this, we present a comprehensive and systematic analysis of DL techniques applied to SG cybersecurity, scrutinizing the spe-

cific DL methodologies employed and identifying notable advancements and innovations within the relevant work.

- Given the potential of DL in identifying cyberattacks, we provide recommendations and introduce a generalized framework to assist researchers in developing high-performing DL models for cyberattack detection.
- The application of DL in SG also encounters novel challenges. We present an outlook to summarize these challenges that have not yet received significant attention but are critical for practical application.
- We provide insights that may be potential solutions in addressing or mitigating these novel challenges. We also envision DL and AI prospects, which may enhance future SG cybersecurity.

The remainder of this paper is organised as follows. Section 2 introduces DL techniques and describes the functionalities of common DL mechanisms to help readers easily understand these DL models in subsequent reviews. A comprehensive and extensive review is presented in Section 3 to investigate DL methods relevant to SG cybersecurity, and a generalised framework is provided to assist researchers in developing high-performance DL models for cyberattack detection. Section 4 elaborates on novel challenges and presents an outlook for potential mitigation directions. Finally, Section 5 summarises the study.

2 | DEEP LEARNING TECHNIQUES

2.1 | Development and structure of deep learning

AI is proposed to reduce manual labor and cost and make the machine as close to human performance as possible by endowing and setting the machine with expert knowledge. With the advent of big data, traditional machine learning methods have shown an insufficient ability to handle large-scale data, failing to learn complex functions in high-dimensional spaces. However, through innovative learning mechanisms, DL can extract critical characteristics from big data and has been developed as a promising direction for AI. The essence of DL is to stack and connect heterogeneous learning layers with different functionalities, deeply digging into the implicit relationships of data to accurately fit the mapping from input to output.

In terms of learning types, DL models are mainly used in three kinds, *i.e.* supervised learning, unsupervised learning, and reinforcement learning [14]. Supervised learning can only handle fully labelled samples and is commonly used for regression and classification. Conversely, unsupervised learning can handle unlabelled samples and is mainly used for clustering. Reinforcement learning can be based on intelligent agents learning the decision-making process under a given dynamic environment to maximise the accumulated rewards. Moreover, based on supervised and unsupervised learning, semi-supervised learning was developed to learn partially labelled samples, fitting the entire data for regression or classification.

2.2 | Deep learning mechanisms

A crucial part of designing a DL model is to select learning mechanisms because various learning mechanisms are intended at processing data with different structures.

2.2.1 | Fully-connected layer

The fully connected layer (FCL) is the most commonly used technique in DL. It maps features from one space to another using a linear transformation, where all the features in the previous layer connect to each new feature. In other words, each new feature is a weighted sum of the original features. The well-known multi-layer perceptron (MLP) model [15] uses stacked FCLs to extract useful features. FCL is generally used to extend or shrink the dimensions of data features.

2.2.2 | Convolutional neural network

The convolutional neural network (CNN) was developed by mimicking biological visual perception and is commonly used to process image data and extract features [16]. Convolutional and pooling layers are the two key components of a CNN [17]. The convolutional layer learns the visual characteristics and reduces hyperparameters using a receptive field and weight-shared mechanism. The pooling layer is inserted between convolutional layers. It can reduce the spatial size of the data to further compress the number of hyperparameters, which effectively avoids overfitting.

2.2.3 | Recurrent neural network

Human cognition is based on past experiences and memories. To employ such a mechanism, the recurrent neural network (RNN) [18] was developed to handle time-series data; it considers the latest and past information. The RNN structure is a recurrent process, wherein the output of the current time step of the time series is related to the outputs of the previous time steps. However, the training of RNN models may result in gradient explosions or vanishing problems when the input considers excessively long past information (also called the long-range dependence problem). To mitigate this challenge, long short-term memory (LSTM) [19] and gated recurrent unit (GRU) [20], which are two famous variants of RNN, were proposed. They implement gate-based approaches to overcome the long-range dependence problem in training and can effectively learn long-term time-series data.

2.2.4 | Autoencoder

Unlike other DL mechanisms, the autoencoder (AE) [21] is a neural network (NN) structure whose input and output have the same architectures. An AE can train the input to be represented

in a lower dimension using an unsupervised learning method without prior knowledge. The represented features can then be reconstructed into their original information without losing valuable information. A typical AE comprises three important components. The first is the encoding architecture, which consists of a series of layers with a progressively decreasing number of neurons. The second is the latent view representation, in which the input is compressed into the lowest dimension. However, this information is not lost, and the compressed features can still represent the original features. The last is the decoding architecture. It can be viewed as a mirror of the encoding architecture to reconstruct the compressed information, that is, the latent view representation, to the original information by gradually increasing the number of neurons in each layer. The denoising autoencoder (DAE) is an improved version of the AE that extracts and composes robust features [22]. By zeroing the input features with a given probability distribution (also called corrupted data), DAE can learn and eliminate useless features to reduce the gap between the training and test sets, further improving the robustness of the model.

2.2.5 | Restricted Boltzmann machine and deep belief network

The Restricted Boltzmann machine (RBM) is a stochastic 2-layers NN that aims to learn a probability distribution over its input [23]. It comprises a visible layer and a hidden layer, where the visible and hidden units are binary values that show the activation of the units, thus acting as a feature extractor. The RBM has a bi-directional symmetric connection between the two layers and has no intra-layer connection; however, it is fully connected between the visible and hidden layers. The RBM can calculate the energy in visible and hidden units using an energy-based model and learn the discrete probability distribution. However, it is a shallow model that may show worse accuracy over its learned distribution owing to its two layers. Consequently, the deep belief network (DBN) [24] was developed by stacking multiple RBMs, where contrastive divergence is applied to each sub-network. The DBN is a probabilistic generative model that can learn to reconstruct its input with a set of samples probabilistically. This composition results in a fast, layer-by-layer, unsupervised learning procedure.

2.2.6 | Residual network

Current deep neural networks (DNNs) exhibit degradation with an increase in the number of layers. On the premise that a DNN can converge, with an increase in network depth, the performance of the DNN gradually improves to saturation and then rapidly decreases. The degradation problem is not caused by overfitting but by the loss of input information. A residual network (ResNet) [25] can alleviate this problem. It allows the input to be propagated to higher layers from the bottom layers to spread the information forward and backward more smoothly. Some studies have shown that ResNet-based DNNs converge

faster and outperform same-layer DNNs without using the ResNet mechanism. Moreover, [26] revealed that ResNet can significantly preserve the spatial structure of gradients, thereby avoiding the shattering gradient problem. In the DL era, the appearance of ResNets has enabled the building of deeper DL models to learn knowledge from big data.

2.2.7 | Graph neural network

In reality, some data have graphical structures. There would be inevitable losses of graphical information if these data were directly processed by general DL mechanisms, such as FCL or CNN. To effectively handle graph data, graph neural network (GNN)-related learning mechanisms can be used [27]. Generally, a graph consists of nodes and edges. The nodes contain entity information, and the edges show the related information of the entities. Studies on GNNs have two motivations. The first motivation is for a CNN to extract local spatial features and combine them to build expressive representations. It derives a convolutional operation on graphs, namely, a graph convolutional network (GCN) [28]. The second motivation is graph embedding, which aims at nodes, edges, or subgraphs to learn vector representation in a lower dimension. The applications of GNN learning can be classified into graph- and node-focused fields. As the name suggests, the graph-focused field only concentrates on the regression or classification of a graph and has no tasks on nodes, whereas the node-focused field has only tasks on specific nodes. The advantage of GNN-based DL mechanisms is that they learn the information spread between nodes to extract more crucial features for representation learning [29].

2.2.8 | Generative adversarial network

A generative adversarial network (GAN) [30], which is a class of DL frameworks, comprises a generative model and a discriminative model. The generative model aims to generate realistic samples to deceive the discriminative model. Simultaneously, the discriminative model seeks to distinguish between the generated false and real samples. Under this conflict, the generative model can learn the distribution from the real samples and generate equally accurate fake samples. The discriminative model can gradually learn the difference between the generated and real samples and eventually discern fake and real samples. GAN-based DL models are commonly used to learn sample distributions and solve imbalanced problems in datasets.

2.2.9 | Attention mechanism

In cognitive science, humans selectively focus on certain information while ignoring other visible information owing to information processing bottlenecks. Inspired by this phenomenon, the attention mechanism (AM) [31] was built into DL to focus on valuably learnable parameters while giving lit-

tle attention to other less useful parameters. The essence of AM is a function (or scoring mechanism) that queries the key-value pair of a parameter with regard to other parameters to score these learnable parameters and assign the corresponding weights. However, using a single AM may result in inaccurate attention. Therefore, multi-head attention mechanisms [32] were proposed to learn attention in parallel and improve learning stability and robustness. Because AM can assign weights to each parameter, it is naturally explainable, which results in AMs also being used for interpretability. Previous studies have revealed that most AM-based DL models can generally achieve lower errors and higher accuracy.

2.3 | Deep learning in smart grid

DL-based models have been widely utilized in the SG environment. However, well-performing DL models are not built by blindly selecting DL mechanisms. The success of DL lies in its use of big data. Heterogeneous DL mechanisms can be blended according to the structures of available data and tasks to build tailored and fine-grained DL models. The overall framework for applying DL mechanisms to SG cybersecurity is illustrated in Figure 1. As cyberattacks threaten the data in SGs, DL mechanisms can be orchestrated to extract complex features from the data and detect abnormal data. SG data can be divided into two categories. The first type of data is defined spatiotemporally, wherein spatial and temporal data are handled by the GNN and RNN, respectively. The second category is 1D or non-1D data. The FCL can extract the features of the 1D data, and the non-1D data are processed using CNN. Other DL mechanisms have different functionalities for data processing, and they can be combined to improve the DL model performance.

3 | REVIEW OF DEEP LEARNING APPLICATIONS FOR CYBERSECURITY IN SMART GRIDS

Cybersecurity is a critical concern for SGs, which are complex and interconnected systems of electrical infrastructures that use digital technology to deliver electricity to consumers. SG technologies, such as AMI, SCADA systems, and distributed energy resources, have enabled utilities to improve the efficiency, reliability, and resilience of electric grids. However, these technologies have also increased the vulnerability of SGs to cyberthreats, including attacks that can cause physical damage or grid disruption.

Safeguarding cybersecurity is of paramount importance and is a non-negotiable task for SGs because cyberthreats can result in financial losses, large-scale blackouts, and casualties [52]. The DL technology presents both opportunities and challenges to SG cybersecurity. Although DL can be used to identify information closely related to cyberattacks from the massive data generated by SG, thereby facilitating the development of the corresponding defense mechanisms, adversaries can also exploit DL to explore SG's cyber vulnerabilities and

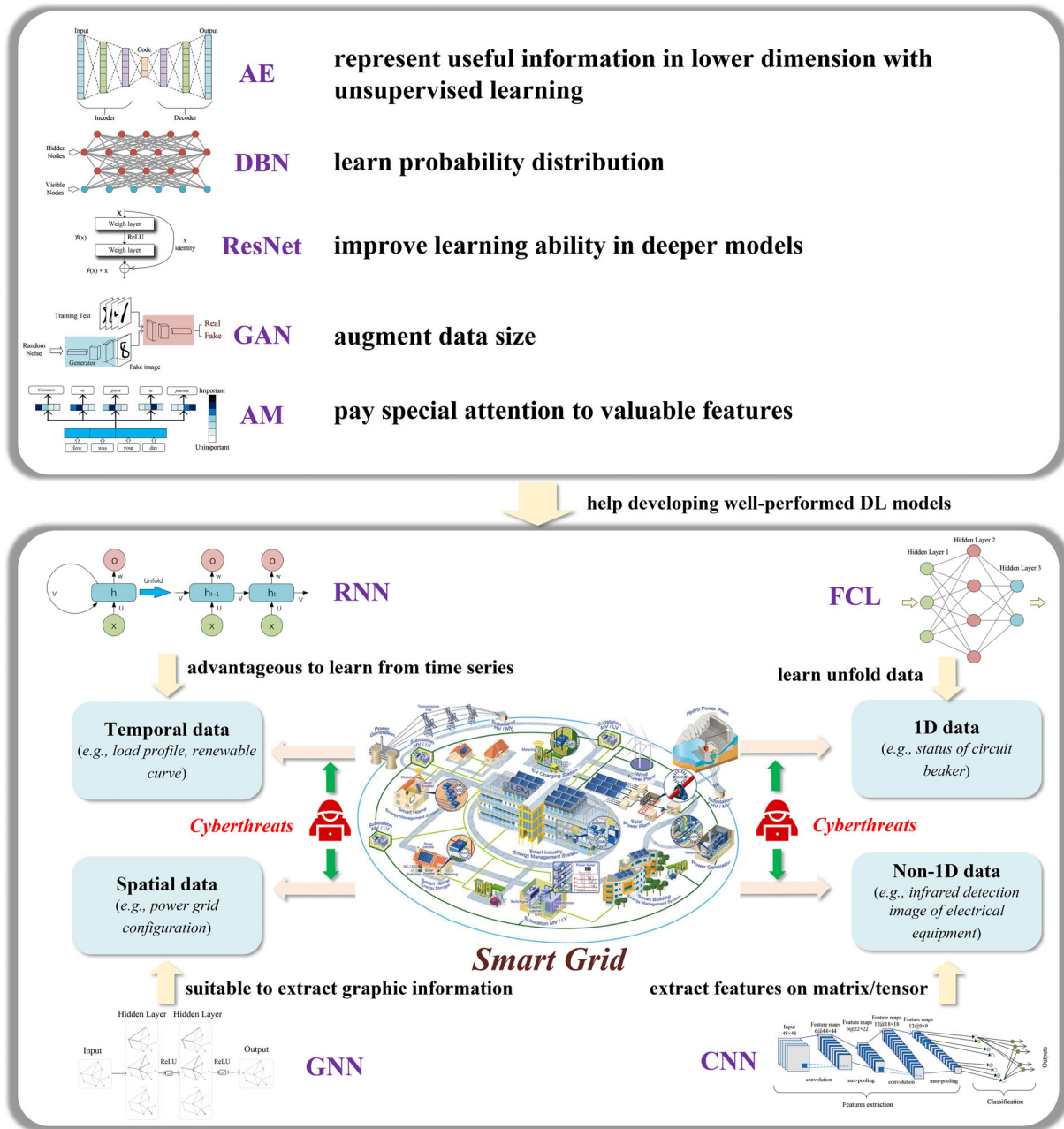


FIGURE 1 Overall framework of applying DL mechanisms in SG cybersecurity.

develop new cyberattack strategies. This section classifies the most prevalent cyberthreats and examines recent developments in DL applications for different types of cyberthreats in SGs.

The cyberthreats in SG can be categorised into five types based on the purpose or target of the attack: electricity theft attack (ETA), false data injection attack on state estimation (FDIA), false command attack (FCA), communication traffic attack (CTA), and adversarial learning attack (ALA). ETA is commonly performed by individual users who manipulate meter readings to reduce their electricity bills. FDIA targets the state estimation (SE) function in the SCADA system and deceives the bad data detector (BDD) by injecting well-designed

false data into normal measurements, thus falsifying the SE results and inducing incorrect dispatch schemes. FCA involves modifying the real-time data required to create control strategies or directly tampering with control commands, causing a wide range of real-time control issues. CTA disrupts available data, thereby breaking the availability and integrity of cyber assets and disabling SGs. ALA aims to compromise the DL models used in SGs, resulting in inaccurate results and improper decision-making.

To conduct a comprehensive review of relevant literature, specific bibliometric techniques were employed. First, several databases were utilized, including IEEE Xplore, IET, Elsevier, Springer Nature, Scopus, ACM Digital Library, Web of Science,

TABLE 1 Brief classification of literature in terms of cyberthreats: electricity theft attack (ETA), false data injection attack (FDIA), false command attack (FCA), communication traffic attack (CTA), and adversarial learning attack (ALA).

Cyberthreats	Literature	Brief description
ETA	[33][34–44]	Directly decreasing electricity consumption to reduce payments
	[45–47]	Reduction of payments with balanced community consumption
	[48]	Increasing renewable generation for illegal profits
FDIA	[49–51]	Development of false data injection attack strategies
	[52–83]	Development of false data injection attack detection methods
	[84]	Development of data recovery approaches
FCA	[85–90]	PMU-related false command attacks
	[91–93]	False command attacks on transmission protective relays
	[94, 95]	Others
CTA	[96–99]	Denial of service-related communication traffic attacks
	[100, 101]	Time delay attack-related communication traffic attacks
	[102]	Others
ALA	[12, 103–106]	Mitigation methods
	[107–109]	Adversarial learning attacks on prediction fields of SG
	[50, 51, 110–112]	Adversarial learning attacks on SG control strategies

and Google Scholar. Second, various combinations of keywords were explored, such as deep learning + smart grid/power system/energy system + cybersecurity/attack, to ensure that all pertinent studies were identified. In addition, to ensure that high-quality literature is included in the review, a rigorous evaluation process was applied. As journal papers are generally more valuable and significant than conference papers, nearly all journal papers are included as review articles, and less valuable and low-quality conference papers are excluded after careful evaluation. A brief classification of the literature is given in Table 1, and the statistics of the main DL mechanisms applied to each cyberthreat are shown in Table 2.

3.1 | Electricity theft attack

The act of ETA, also known as non-technical loss (NTL), is a pervasive issue that can have significant financial and safety implications. The perpetrators of ETA, which are typically individual users, alter meter readings to lower their electricity bills or gain illegal profits. This practice can cause enormous revenue losses for utilities and even pose real threats to public safety [33]. A survey revealed that the economic loss of utili-

TABLE 2 Statistics of main DL mechanisms applied in cyberthreats: electricity theft attack (ETA), false data injection attack (FDIA), false command attack (FCA), communication traffic attack (CTA), and adversarial learning attack (ALA).

	FCL	CNN	RNN	AE	DBN	ResNet	GNN	GAN	AM
ETA	[33, 39, 42, 44–47]	[33, 34, 36, 37, 42, 46–48]	[34, 39–48]	[38, 40, 43]	[35]	-	-	[46]	[40]
FDIA	[49, 53, 56, 59, 62, 66, 70, 75, 80, 83]	[56, 63, 74, 78, 81, 83]	[53–55, 57, 60, 62, 65, 66, 72–75, 81]	[52, 54, 55, 57, 59, 67, 68, 76]	-	[61, 83]	[69, 75, 77, 83]	[59, 68, 84]	[55, 64, 71, 73]
FCA	[92–94]	[90, 91, 94]	[88, 90, 94]	[85–87, 91]	[85]	-	-	[89]	-
CTA	[97, 99, 100]	[98, 102]	[97, 100, 101]	-	-	[98]	-	-	-
ALA	[50, 108]	[109, 111]	[109]	[60]	-	-	-	[107]	[107]

ties worldwide due to ETAs amounts to \$100 billion per year [113]. In addition to financial losses, ETA can cause physical damage to the SG infrastructure. Specifically, ETA can lead to power transformer overloading, voltage fluctuations, and power quality problems. Transformer overloading occurs because of an increased demand for power without an accompanying increase in supply, leading to failures and blackouts. Imbalanced power supply and demand can also cause voltage fluctuations, which can damage appliances and equipment connected to the grid. Additionally, ETA can result in power quality problems, such as voltage sags and swells, which can cause electronic device malfunctions or failures.

The relevant ETA studies fall into three categories. The first kind is concerned with directly modifying meter readings to decrease the electricity consumption sent to utilities, thereby reducing electricity payments. Ref. [34] is the first study that proposed the use of a DL model to identify the NTL problem in distributed systems. In this study, CNN and LSTM were orchestrated to extract spatial and temporal features from monthly energy consumption data to classify regular and irregular consumers. Ref. [35] aimed at transmission networks, proposing a conditional DBN-based model to extract features from geographically distributed sensors/meters. It helps reduce the complexity of the training and execution times. Ref. [33] proposed a wide and deep CNN model to analyze the ETA behavior. In the model, the wide component is an FCL to learn global knowledge from 1D electricity consumption data. The deep CNN part can identify the non-periodicity of ETA and the periodicity of normal electricity usage by processing the 1D data into 2D data with respect to weeks. The two components are then combined using a weighted sum of their output as hidden features feeding into a logistic loss function as a classifier to identify ETA behaviours. Ref. [36] proposed an energy theft detection scheme based on a combined CNN model with the Paillier homomorphic encryption algorithm to detect whether the metering data has an abnormal behavior to preserve energy privacy. To improve feature processing ability, [37] presented a hybrid CNN random forest (RF) model for automatic ETA detection. The CNN is similar to an automatic feature extractor in investigating massive and varying smart meter data to learn features between different hours of the day and different days. The RF is an output classifier based on the obtained features to detect electricity theft by the consumer. Similarly, [42] developed a CNN-GRU-MLP-based detector to identify malicious customers; [43] designed an AE and bidirectional GRU model to identify anomalies in electricity consumption patterns; [44] proposed an ensemble-based DL detector to improve accuracy and quickly detect false readings.

Several studies concerning the first kind offer distinct contributions to the literature. Given that most existing ETA data are unlabelled and labelling entails a significant labour cost, [38] proposed a semi-supervised DL approach for ETA detection. This method employs a DAE for unsupervised feature extraction, thereby addressing the scarcity of data resources pertaining to ETA detection in power distribution systems. Data poisoning is another prevalent issue in ETA datasets if utilities fail to identify ETA behaviour. Training on such datasets

can impair the detector's ability to distinguish between benign and malicious behaviours. To address this challenge, [40] introduced a robust sequential ensemble ETA detector based on a deep AE with AM and GRUs. The AE with AM structure consists of several LSTM layers and an AM layer to facilitate the differentiation between benign and malicious samples. GRU layers are efficient in capturing patterns of temporal correlation and sequential information in customers' electricity consumption time series data, subsequently enhancing the detector's robustness against data poisoning. In addition to energy consumption data, other information can be used for ETA identification. Ref. [39] considered geographic, contractual, economic, and smart meter information, proposing an end-to-end solution founded on a hybrid LSTM-MLP model. The LSTM network analyses the raw daily energy consumption history, whereas the MLP integrates non-sequential data, such as contracted power or geographical information. This model can self-learn features for detecting anomalies and fraud in smart meters and is also the first DL architecture for ETA detection that accommodates both sequential and non-sequential data. Considering that customers exhibit varying energy consumption behaviours and may employ unique ETA strategies, [41] developed a customer-specific DL model to identify benign and malicious customers. The customer-specific model utilises vector embedding to represent the energy consumption profiles. Subsequently, a deep GRU-based classifier was trained on these embeddings to detect the ETA behaviour for each customer. This reduces generalisation but improves detection accuracy.

In the second part of the ETA studies, the attack strategy is improved by redistributing electricity consumption in a community. The attackers decrease their electricity readings while augmenting the usage of other households, thereby ensuring that the aggregated bill for all customers in the community remains the same. This overcomes the disadvantage of the attack method employed in the first kind, wherein the SG operator inaccurately aggregates the system load level owing to the less received power usage, leading to imbalanced dispatching that deteriorates power quality and affects attackers' usage. In the second group, [46] designed a covert ETA strategy based on LSTM and GAN to mimic normal consumption patterns and concurrently tamper with neighbouring meters. Considering that these well-trained models may be ineffective against novel and covert attacks in real-world applications, [46] proposed a hybrid CNN model with two wide components and one deep component to detect ETAs. In the hybrid model, one wide component and one deep component are similar to [33], and the other wide component is an FCL used to learn representations from the correlation features. Similar to defending ETAs, [45] developed a three-stage ETA detection system based on MLP, LSTM, and GRU. Stage 1 is a multi-DL model forecasting system for the next 24 hours of energy consumption; stage 2 is a primary decision-making model to determine energy theft predictions; and stage 3 further checks for possible energy theft and improves detection accuracy. Moreover, [47] proposed a general multi-data-source deep hybrid learning-based detector to identify the ETA in a net-metering system. It is not applica-

ble for detecting individual users; however, it can effectively find anomalies in the net-metering system.

The third category is suitable for users with installed distributed renewable generators. The attack strategy reduces consumed electricity readings and falsely increases the renewable power generated. The attacker can charge illegal revenue by aggrandising the readings recorded for renewable power generation sold to the power grid. Within this category, [48] is the only study that investigates ETAs in the distributed generation (DG) domain and proposes a hybrid CNN-GRU DL model to detect cyberattacks of ETA in renewable energy-based DG units. The model can extract the temporal correlation between DG smart meter readings, irradiance data, and SCADA meter readings to enhance detection performance.

3.2 | False data injection attack on state estimation

FDIA has recently drawn significant attention because of its covert and destructive nature. It was first proposed in [114] by finding that measurement data can be potentially compromised through orchestrated designs, thereby deceiving the residual-based BDD and resulting in malicious falsification of the SE outcome. Accurate estimation of the power system's state variables, such as voltage, current, and power flow, is crucial for ensuring its stability and reliable operation. The injection of false data through FDIA can cause an incorrect SE, leading to erroneous control actions and consequently jeopardising the system's stability. In severe cases, FDIA can cause cascading failures that eventually culminate in large-scale blackouts. For instance, if false data are injected into the SE process that provides the impression that there is less demand on the system than there actually is, the system may reduce power generation to match the perceived demand, which can lead to a supply-demand imbalance and potentially cause blackouts. DL-related studies on FDIA can be divided into those exploring FDIA strategies and those developing defense mechanisms.

The exploration of FDIA strategies seeks to utilise DL to explore unknown vulnerabilities, develop novel attack strategies, or identify effortless attack scenarios. To date, only this kind of study has been explored in the literature. Ref. [49] proposed a coordinated cyber-physical topology attack strategy and used a deep reinforcement learning (DRL)-based approach to determine the minimal attack resources. The proposed attack method trips a physical transmission line and masks the outage signal by creating a fake one for another line in the cyber layer using FDIA. To achieve minimal attack resources, an MLP-based DRL model was developed to determine physical- and cyber-tripped lines by modelling the power system as the dynamic environment, tripped lines as the action, and grid topology information as the state. Moreover, this method can handle load uncertainty, thereby increasing the probability of successful attacks. Considering that massive DL-based methods have been developed for detecting FDIAs, the existing FDIA modelling may fail to deceive such novel DL-based detectors. To concurrently bypass BDDs and DL-based detectors, [50] and

TABLE 3 Classification of studies relevant to false data injection attack detection.

Classification	Literature
Using deep learning as auxiliaries	[52, 66, 72, 74]
Simply developing classifiers for detection	[53, 54, 69, 76, 77, 81]
Locating false data injection attacks	[55, 56, 60, 63, 68, 75]
Resorting to deep reinforcement learning for detection	[64, 65, 79, 95]
Detecting attacks with specific targets	[57, 70, 78, 80]
Addressing the problem of attack samples insufficiency	[58, 59, 67, 83]
Considering disturbances from renewable energy integration	[60, 61]
Handling the privacy problem in constructing detectors	[62, 71, 73]

[51] designed novel FDIA strategies by introducing adversarial samples (also called perturbation vectors) into FDIAs, thereby deceiving BDDs and DL-based detectors.

The development of defence mechanisms falls into two fields, that is, detection and recovery. Detection-related DL studies involve building FDIA detectors using the latest or historically available data. Researchers have made significant efforts to develop various FDIA detection approaches, resulting in the most relevant studies in this field. A brief description of FDIA detection is shown in Table 3. Some studies prefer to use DL as auxiliaries and combine it with other techniques, such as optimisation and statistics, to identify possible data manipulation caused by FDIAs. Specifically, [52] proposed a DL-based optimisation model to ensure the normal state bounds resulting from load forecasting uncertainties and consequently detect potential FDIAs, where an AE-based DNN for load forecasting was used to shorten the width of the state bounds and further improve the detection accuracy. Similarly, [72] used ensemble DL mechanisms for load forecasting to construct state bounds for FDIA detection. Moreover, [66] presented a real-time FDIA identification scheme by comparing the predicted states with the SE result, which demonstrates a scalable, real-time, and effective state forecasting approach with minimal error margin. In addition, [74] proposed a DL-based pseudo-measurement model to quantitatively describe the measurement uncertainty and further determine the state bounds for FDIA detection.

Developing DL-based classifiers has also been effective for FDIA detection, in which various difficulties have been addressed. Considering that the previous work for detecting FDIA generally adopted spatial data only at the latest time, [53] proposed a DNN-based detector with GRU and FCL. It can learn from temporal data correlation in consecutive states and extract temporal-spatial characteristics to distinguish FDIA from normal operation events. Because massive data and the high dimensionality of the original data of the power system will overburden the computational resources, [54] used an invertible AE to reduce the original data dimension. The data processed by the dimensionality reduction module were then input into an LSTM module to train a classifier for FDIA detection. This

method can reduce computational time while preserving high detection accuracy. Several solutions exist for FDIA detection. Ref. [69] proposed a GCN framework to analyze the graphical aspect of FDIAs by exploiting the graphical structures of the power network. A stacked AE was designed in [76] along with an extremely randomised tree classifier to address FDIA issues. Ref. [81] proposed a spatiotemporal DL network for FDIA detection in AC-model power grids. As previous works did not consider power grid topology changes, [77] considered the FDIA issue under topology dynamics using a gated GNN and graph attention network-based model for detection.

Some studies have considered identifying the locations of FDIAs for detection purposes. Ref. [55] employed an AE structure for locating FDIAs. Within the AE structure, the RNN and AM modules were federatively employed to promote the correlation extraction of time series and improve location accuracy. Ref. [56] proposed a DL-based locational detection architecture to detect the exact location of FDIAs, where CNN and FCL were used in the DL model. The CNN captures the inconsistencies and co-occurrence dependencies of measurements. The locational detection model aims to develop a multilabel classification for each metre, thus exhibiting possible intrusion. Similarly, [63] showcased a CNN-based model to determine the exact intrusion points in real time, and [68] designed an AE-GAN-based methodology for detecting and locating FDIA under a DC power flow and recovering the falsified measurement/state variables. Moreover, [75] proposed a localisation detection method that disaggregates primary data into graph-structured data and designed specialised GCN and GRU networks for FDIA location and detection.

Some researchers have also considered the use of DRL for FDIA detection. Ref. [65] proposed a DRL-based detection approach that combines the LSTM network to extract the system state features from previous time steps and consequently determine whether the system is currently being attacked. Similar to [64, 65] added an AM to the DRL-based detection algorithm, which calculated the attention distribution of all input information, and the SE results were dynamically weighted according to the attention distribution. In this manner, it can extract more representative and distinguishable state features as observations, which can better facilitate decision-making in DRL for FDIA detection. Ref. [79] focused on low latency detection of cyberattacks with DRL, which can minimize detection delay while ensuring high detection accuracy. Moreover, [95] proposed a resilient optimal defensive strategy with the distributed DRL approach to evaluate the extent of supply security or voltage stability of the microgrids affected by FDIA.

Detection of FDIAs with specific targets is also important. Aiming at FDIAs that can induce rescheduling and load shedding, [57] used feature vectors extracted from normally estimated states to train an LSTM-AE model. The well-tuned model can process the extracted spatial and spectral features to further calculate and update the deviation of the measurements, and the logistic regression classifier can identify FDIA from normal system operation events. To address the ever-increasing cyberattack challenge in SGs, a visualisation-based attack detec-

tion framework using DL techniques was developed in [78], which provides cybersecurity researchers with improved techniques for uncovering trends, identifying outliers, recognising correlations, and communicating their results. For transmission line overflow FDIAs, [70] developed a simple MLP method for detection, which has high detection performance and low computational complexity. Moreover, [80] developed a detection framework based on DNNs that used bad data generated from line overflow FDIAs for training purposes and detected this type of FDIAs.

DL-based FDIA detectors commonly encounter insufficient sample problems because labelled attack data are extremely scarce, and data distribution shifts may occur when loads, topology, or system dynamics change. These problems can lead to a bias in the training data and render the DL model intractable. Considering these concerns, [58] proposed a self-adaptive intrusion detection framework based on semi-supervised domain-adversarial training, which extracts novel features to unify data distributions across two domains and improves classifier robustness against the shift. This demonstrates that domain-adversarial training can be used to detect unknown cyberthreats. Similarly, [59] overcame the deficiency of labelled attack samples in distribution systems by using AE and GAN. In this approach, an AE consisting of FCLs was used for dimension reduction and feature extraction of measurements in three-phase unbalanced networks. Moreover, GAN generated training samples and was applied to the AE structure to detect abnormal measurements by capturing the unconformity between anomalies and secure measurements. Advantageously, this method requires only a few labelled data to train a well-performing detector. To eliminate label limitations, [67] proposed a self-supervised clustering model with a stacked AE network. It can achieve clean clustering of data into benign and compromised samples without labelled supervision. Ref. [83] also proposed a real-time FDIA detection scheme without using any attack samples. This was achieved using a spatiotemporal GNN to extract state features in both spatial and temporal dimensions and consequently output state quantiles as upper and lower detection bounds. Moreover, a CNN-ResNet-based super-resolution perception network was designed to reconstruct high-frequency SE results, thereby improving the temporal information for detection learning.

The ability of most detectors to identify FDIAs may deteriorate with the increasing penetration of renewable energy sources (RESs). This is because disturbances from renewables aggravate SG uncertainties, making it difficult for detectors to distinguish FDIA from normal operations. To overcome this problem, [60] designed an LSTM-embedded DNN to model the dynamic behaviours of SG influenced by RES or system reconfiguration for real-time FDIA detection. This detection framework can handle uncertainties raised by integrating the RES and locating the attacked buses. Similarly, [61] studied the effects of different RES penetration levels on the accuracy of previous FDIA detection methods and proposed the use of causal and dilated convolution and ResNet, instead of the RNN family widely employed in the extant literature, for FDIA detection.

Privacy is another problem associated with FDIA detection. Because most DL-based FDIA detectors require global information on the power grid, they may incur a privacy problem between sub-grids. Ref. [62] proposed a sub-grid-oriented framework for FDIA detection by collaboratively learning the relationship between a specific sub-grid and the remaining sub-grids. It uses MLP to model the spatial relationship between bus/line representations and LSTM to learn the temporal relationship from time-series measurement data. The measurement data collected by each sub-grid are not shared with the other sub-grids. Only the feature representation for each sub-grid is transmitted for collaborative training. This significantly reduces data latency and preserves the privacy of each sub-grid. Similar to overcoming the privacy problem, [71] is the first to leverage secure federated learning by combining transformer, federated learning, and the Paillier cryptosystem for detecting FDIA in the SG. Moreover, a deep federated learning-based decentralised FDIA detection method, which is capable of parallel computing and can reliably identify stealthy FDIA on all nodes simultaneously, was utilized in [73].

The second part of FDIA defence is recovery. In this field, researchers expect to recover the compromised data to the pre-attack value because it is crucial for the operator to make subsequent decisions and dispatches of the SG [7]. However, only a few studies have considered this problem. Ref. [84] is the first work to utilize DL for data recovery. This study proposed a GAN-based recovery model against FDIAs that can capture deviations from ideal measurements while generating non-tampered data to replace tampered data. Considering the computational burden of DL, a smooth training technique and an adaptive window were adopted to accelerate the GAN training process. The simulation results reveal that the recovered measurements are sufficiently close to the true measurements to maintain SE integrity. Similarly, [68] developed a pattern-matching algorithm for recovery. It designs an AE-GAN framework to generate various candidate sets of normal measurement vectors with various operational topologies and subsequently localises and recovers the falsified measurements and state variables by comparing the falsified measurement vectors with the normal measurement vectors in the candidate set.

3.3 | False command attack

Unlike FDIA, which maliciously modifies the SE result, FCA targets the control center of SG, thereby resulting in various real-time control issues. Given the critical importance of PMU measurements, as control centres may depend directly on them or on the inferences drawn from them to make crucial decisions [85], the manipulation of PMU data is a particular type of FCA that can obscure the real-time operational status of SG. This kind of attack can produce incorrect or unreliable data regarding power generation, consumption, or distribution, resulting in reduced grid efficiency, higher costs, and hazardous situations. Another mainstream of FCA is to directly send fraudulent commands, such as protective relays and line trip commands,

which can cause severe faults and bring the SG to a halt. For instance, an attacker can send false commands to a smart grid's distribution system, causing improper power distribution or equipment failure.

With regard to potential PMU-related FCAs, [86] presented a stacked AE framework for automatic and adaptive detection in transmission systems. The proposed framework leverages the automaticity of unsupervised learning in the stacked AE to reduce the reliance on system models and human expertise in complex security scenarios. It enables the detection of cyberattacks on data injection, remote tripping command injection, and relay-setting changes. Similarly, [87] proposed using a stacked DAE to reconstruct a noise-free input from noise-corrupted perturbations. Moreover, ensemble learning, which trains multiple learners and combines their decisions to solve problems, was used by [87] to improve detection robustness. Ref. [88] considered that fault current limiters might also be targets of FCAs; therefore, an ultrafast active response strategy based on an LSTM-enabled DRL model was proposed to generate real-time response action. Ref. [89] further considered that practical power systems cannot provide numerous faulty and FCA samples for DL model training; thus, a GAN-based semi-supervised method was designed to learn partially labeled data to diagnose PMU-related FCAs and faults. To detect the FCA anomaly of each PMU measurement, [85] built a distributed anomaly measurer based on an AE-DNN. The AE structure was unrolled from two RBMs. The probability mechanism of RBM is more adaptable for working in synergy with other sources of information to enhance detection performance. Similarly, [90] proposed a method for optimising CNN-LSTM with particle swarm optimization (PSO) to detect abnormal measurement values in PMU.

Aiming at transmission protective relays in substations, [91] proposed a detection system based on a 1D CNN-embedded AE structure, wherein current and voltage measurements representing various types of faults on transmission lines were used as the input. This detection system can distinguish FCAs from three-phase-to-ground, two-phase-to-ground, single-phase-to-ground, and phase-to-phase faults. For substation control authorities, [92] proposed a DRL-based recovery strategy for optimally reclosing tripped transmission lines using FCAs. The DRL endows the strategy with good adaptability under different FCA scenarios and the ability for real-time optimal or near-optimal decision-making. Regarding power distribution systems, [93] used MLP to design a fault and attack location and a classification system to classify and locate cyber and physical anomalies, including FCAs on protection devices, replay attacks on communication networks, and physical faults on distribution lines. This system takes as input the transient short-circuit current and voltage measured by protection relays, the relay command status, and the fault alarm from fault indicators. It was found that even without fault indicators, voltage and current can provide sufficient information to allow accurate classification between faults and attacks and the location of attacks.

Because every SG typically has limited attack examples and is unwilling to share such attack examples owing to highly sensitive information, [94] presented a federated DL scheme with

CNN, GRU, and FCL to detect command injection attacks, and a Paillier cryptosystem-based secure communication protocol was crafted to preserve the security and privacy of the model parameters. However, this scheme can only be applied to same-domain SGs. Moreover, a multi-agent DRL-based algorithm was proposed in [95] to automatically discover the vulnerable spots in the conventional index-based cyberattack detection schemes and generate coordinated stealthy destabilising FCAs on cyber-protected islanded DC microgrids.

3.4 | Communication traffic attack

CTA targets network communication channels with the aim of eavesdropping, intercepting, and modifying traffic packets. Unlike other cyberattacks that seek to steal data or gain unauthorised access to computer systems, CTA aims to disrupt the availability and integrity of cyber assets. The consequences of CTA include delayed response time, system instability, equipment damage, and reduced system efficiency. In SG control systems, CTA can cause delays in the response time, making it difficult to promptly detect and respond to system disturbances. This is particularly concerning given that SG requires real-time communication to maintain stability. CTA can also lead to communication delays and errors, which can result in instability and blackouts. Moreover, it can cause physical damage to the equipment connected to the SG, including communication devices and power generation systems, leading to power outages and system failures. Finally, CTA can reduce the overall efficiency of SG by increasing the amount of energy required to transmit and process data, resulting in increased costs and decreased performance.

The main reason for launching CTA is owing to encrypted data that attackers cannot decrypt. It hampers attackers from achieving the desired results, such as those from FDIA and FCA. Meanwhile, CTA is a straightforward way for launching a cyberattack. Denial-of-service (DoS) attacks and time delay attacks (TDAs) are common CTA strategies. They can harm the operation of SGs, such as the closed-loop control in power grids that depend critically on timely feedback to accurately adapt its operations in real time.

DoS attacks involve sending massive packets to jam the communication channel, rendering network resources unavailable for the SG [96]. They are difficult to detect and mitigate because they typically do not attempt to access private data. Ref. [97] considered that detecting DoS attacks would be more efficient by observing the interpacket patterns in the time domain and proposed an LSTM-MLP-based intrusion detection system to detect DoS attacks. However, thousands of users may be present on a communication channel. Thus, it is more difficult for defenders to distinguish malicious from legitimate users. This is an improved version of the DoS attack called a distributed denial-of-service (DDoS) attack. Therefore, a DL framework based on CNN and ResNet was designed in [98] to provide the early detection of DDoS attacks; an accuracy higher than 91% was achieved. Furthermore, instead of detecting DoS-related attacks, [99] considered improving a network's throughput and

communication efficiency to enhance its immunity from DoS attacks. Based on this, a DRL approach was proposed to fulfill the goal.

TDA can be launched by delaying the transmission of control command packets without altering the content [100]; therefore, it is difficult to detect when inspecting the sending packets. To handle TDA, leveraging the temporal information of packets is a promising approach. Ref. [100] proposed a DL model based on stacked bidirectional LSTM units and MLP to detect TDA on data traces from sensor measurements of temperature, pressure, and power generation. Similarly, a hierarchical LSTM model was designed in [101] to process relevant sensor outputs in closed-loop control systems of SGs to simultaneously detect and characterise TDA.

In addition to DoS attacks and TDAs, [102] considered inspecting network traffic anomalies and designed a network intrusion detection system using CNN to characterise the salient temporal patterns of SCADA traffic. It leverages the traffic monitoring capabilities of SCADA networking devices without interrupting SCADA system operation.

3.5 | Adversarial learning attack

Despite their high accuracies on various SG applications, DL models are extremely vulnerable to a special cyberattack called ALA, which has three main types. The first type is data poisoning [115]. DL models generally require numerous samples for training. In the training process, the knowledge extracted from the samples is continuously learned by the DL model via forward and backward propagation. However, if attackers maliciously pollute a part of the sample, the DL model learns incorrectly from the polluted samples. Consequently, it deteriorates the performance of the trained DL model as poisoned. The second type is the modification of network parameters. The network parameters of a well-trained DL model are orchestrated through learning. If certain parameters are modified by adversaries, the compromised DL model may produce completely different results. In the third type, the input data is changed for practical applications. Although the DL model remains intact, it can still make abnormal judgments if the input is falsified. A famous real-world case of this phenomenon was when hackers tricked a Tesla into veering into the wrong lane by placing a series of small stickers on the road. Therefore, certain well-trained DL models are vulnerable to subtle input disturbances.

The effects of ALAs on the performances of DL models in SG applications are of great concern. In the specific context of SG, these attacks can be utilised to manipulate the output of various DL models, such as energy demand or fault diagnosis data, thereby resulting in erroneous decisions. The potential physical repercussions of these attacks on power grids are significant. For instance, if an adversary manipulates demand prediction data such that the smart grid underestimates demand, it could lead to energy shortages and even blackouts. Similarly, manipulating fault diagnosis data can result in misguided maintenance decisions, leading to equipment failure and physical damage.

Because DL is recognised as the heart of AI, academia and industry communities have recently made a large influx of contributions to ALA. To defend ALA, [12] and [103] suggested three directions, that is, using modified training during learning or modified input during testing, modifying networks by adding more layers/sub-networks or changing loss/activation functions, and using external models as network add-ons when classifying unseen examples. In addition, [104] proposed the construction of robust ALA detection systems, such as using known attack samples for training and leveraging unsupervised learning approaches combined with statistical hypotheses to reject samples whose maximum posterior class probabilities are less than a given threshold. As DL technology relies on data, [105] proposed the establishment of a data-preserving AI system, that is, blockchain-based learning data environments, to verify the integrity of learning data. Thus, it can prevent data deterioration and defend against certain kinds of ALA. In addition, [106] presented both reactive and proactive approaches. Reactive methods, such as adversarial detection, input reconstruction, and network verification, are intended to block adversarial examples. Conversely, proactive methods build more robust models that resist adversarial examples, such as network distillation, adversarial training, and classifier ensembles.

DL technology has been widely exploited in the SG environment. One mainstream approach is load forecasting. Ref. [107] is the first paper to evaluate the cybersecurity of load forecasting under ALA. It revealed that an attacker, who does not require knowing the load forecasting model or the underlying power system parameters, can successfully launch ALA by only injecting malicious temperature data from online weather APIs. Consequently, the attacker can manipulate load forecasts in arbitrary directions and cause significant damage to SG operations. To improve the robustness of DL-based load forecasting models against ALA, [107] suggested that variational AE and GAN may be potential solutions. Variational AEs, such as DAE, can eliminate useless content in the input, whereas GAN can generate samples with perturbations and distinguish them in the discriminator. Moreover, [109] proposed a cyber-secure DL framework that can accurately predict the electric load while effectively defending against ALA. In addition, for the forecasting problem, [108] examined ALA on the prediction of solar PV based on irradiance, module temperature, and ambient temperature. It revealed that the accuracy of an MLP-based PV prediction model decreases by 15% on average when 20% of the training/testing samples are modified. Damage, such as blackouts, can be catastrophic if an attacker has complete knowledge of the PV profile and obtains access to the network. Additionally, [111] systematically investigated the attack requirements and credibility of six representative ALAs based on a voltage stability assessment application.

With the advantages of model-free and real-time computational capabilities, DRL has shed light on decision-making problems in SGs, such as frequency control and network reconfiguration. However, DRL models also encounter ALA, which leads to incorrect control actions that may induce hazardous situations. Therefore, [110] proposed a vulnerability assessment approach for DRL models in power system topology opti-

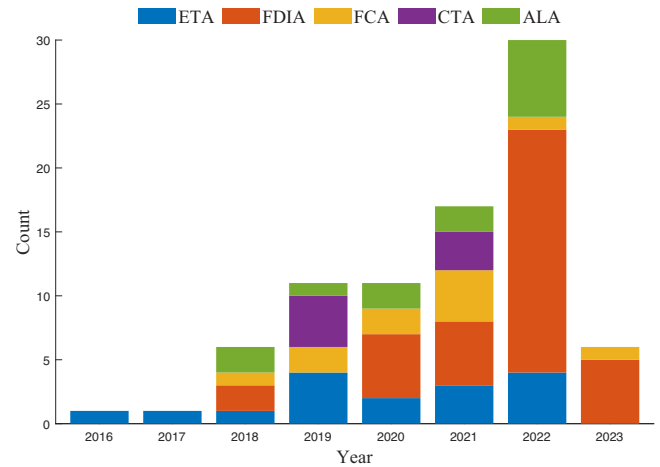


FIGURE 2 Research frequency of different cyberattacks since 2016.

misation, enabling power grid operators to identify or reduce security risks before practically applying DRL models. An ALA strategy called adversarial perturbation was also constructed in [110] to minimise DRL rewards by considering when-to-perturb and how-to-perturb. To address the DRL-based security-constrained optimal power flow (SCOPF) problem, [112] developed an adversarial example generation method to realise a targeted ALA considering nonlinear physical constraints in power systems via two main stages: constructor function design and unconstrained optimisation problem transformation.

Considering that FDIA can bypass traditional BDDs but may be detected by DL-based detectors, [50] proposed an ALA strategy by introducing perturbation vectors into FDIAs, which can deceive DL-based detectors while achieving targeted attack effects on SE results. Similarly, [51] introduced the joint adversarial example for FDIAs, which is performed by adding perturbations to state variables that can ensure a stealth attack to both BDDs and DL-based detectors.

3.6 | Research trends

To offer a comprehensive evaluation of the current research advancements in this domain, Figure 2 depicts the frequency of various cyberattacks that have been studied since 2016, whereas Figure 3 shows the usage counts of different DL mechanisms over the same timeframe.

As observed in 2, ETA has emerged as a pioneering research subject employing DL technology, and the number of relevant studies has remained relatively stable over the years. In contrast, FDIA has recently garnered considerable academic interest, witnessing a notable surge in relevant publications since 2018. Both FCA and CTA, characterised by their comparatively singular attack characteristics, have consistently demonstrated a limited presence in the literature. However, ALA, as an emergent attack type, has gradually attracted increasing attention and is expected to undergo a significant increase in future research.

The use of different DL mechanisms also exhibits certain trends, as shown in Figure 3. This reveals that FCL and RNN

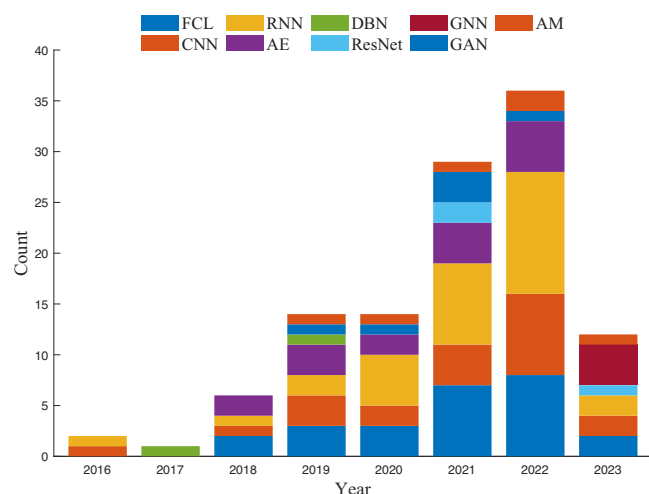


FIGURE 3 Counts of different DL mechanisms used since 2016.

have been widely used in recent years. This is because FCL, as the most fundamental DL mechanism, possesses a powerful ability to handle data. In contrast, RNN is proficient in processing time-series data, making it easier to identify temporal correlations that may be broken by cyberattacks. Additionally, CNN is skilled in local feature extraction, making it an effective tool for identifying anomalies, whereas AE can perform feature compression and eliminate irrelevant noise, making it another widely utilised mechanism. Looking forward, it is expected that GNN and AM will be adopted significantly. GNN can better extract relevant features for identifying attacks in a naturally graphical power grid system, whereas AM has an outstanding ability to focus on valuable features when identifying attack anomalies.

3.7 | Suggestions on threat prevention

As mentioned above, there are five main cyberthreats in SGs. Although some of them are handled well by developing DL techniques as countermeasures, there is currently no common guidance for developing well-performing DL models for cyberattack detection. This section provides suggestions and a general guide for exploiting DL mechanisms to cope with cyberthreats.

The RNN and GNN families are effective at processing temporal and spatial data. The five cyberthreats can first be classified according to their spatio-temporality. As discussed in Sections 3.1–3.5, each cyberthreat has specific targets. Individual malicious users launch the ETA to modify the meter readings. To detect ETAs, using the RNN family is beneficial for identifying inconsistencies in the time-series data. The FDIA aims to modify the SE results through the SCADA system. It is usually launched by compromising the measurements. Therefore, there are two ways to detect FDIAs, that is, by detecting the measurement data or the SE result. The RNN family is still suitable for examining temporal anomalies. However, the SE result has a spatial property that includes the grid structure

and power flow in the transmission lines. Hence, exploiting the GNN-RNN structure would be advantageous for FDIA detection based on the SE results. The FCA refers to a change in the control command, which includes the on/off status of lines and the active/reactive power control of units. As such, falsifying the control command violates the correlations of the command values in the time domain, and the RNN family can be utilised for FCAs detection. The CTA is achieved by contaminating the network communication channels. Because the data transmitted through the channel are inherently temporal correlations, it also suggests employing the RNN family to detect CTAs.

The ALA differs from other cyberthreats, which can be launched by poisoning the input data. Therefore, developing a model for detecting the input data and recovering compromised input data is an effective countermeasure. Doing so enables the use of the AE model. The AE refines useful information by compressing the data into a lower dimension and thus can help eliminate contaminated content. In fact, the AE is also conducive to extracting valuable features from data targeted by ETA, FDIA, FCA, and CTA. After compression, the data can further help detect these cyberthreats.

The remaining DL mechanisms can be combined with RNN, GNN, and AE techniques to improve learning ability and detection performance. The AM can be used to enhance the learning ability for temporal data. CNN can handle complex data and identify local anomalies. The DBN enables the exploration of the distribution of big data and efficiently uses hidden layers. If the network is extremely deep, the ResNet can be used to facilitate the information transfer. If the available data are insufficient for DL model training, the GAN can be used to augment the data size. Finally, the FCL can be used to change the feature dimensions and is typically used as the output layer. A general guide for building a DL-based cyberattack detection model is shown in Figure 4.

4 | OUTLOOK AND PERSPECTIVES

Although the DL technology has made significant contributions to SG cybersecurity, there are novel challenges that must be solved. We present an outlook to highlight these challenges that have not yet been explored and provide perspectives that may help address and mitigate them.

4.1 | Label imbalanced samples

DL is a powerful technique that requires massive and label-balanced datasets to extract and learn underlying knowledge. However, SGs are still in their early stages, and cyberattack cases are insufficient to create a large, label-balanced dataset for DL training. In addition, the detection systems in place may not be able to identify all cyberattacks, resulting in a shortage of samples. Although some historical data can be used to label cyberattacks, the process is labour-intensive and expensive. This problem has often been overlooked in the literature.

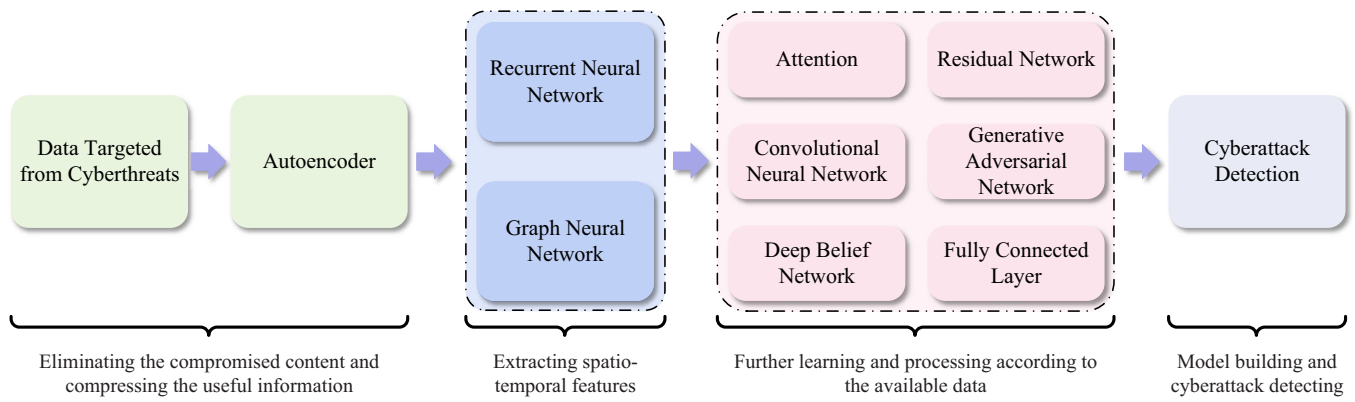


FIGURE 4 General guide for developing the DL model for cyberattack detection.

One solution is to generate artificial cyberattack samples. Dynamic simulators can be created to mimic the operation of practical SGs, allowing researchers to implement cyberattacks on the simulators and obtain sufficient samples. GANs can also be used to generate infrequent attack data by pitting the generator against the discriminator. However, researchers must determine the proportion of true cyberattack samples that should be fed into the GAN and how to train the model effectively in different SG environments.

Another solution is to use unsupervised or semi-supervised learning approaches. DAE structures, which are typical unsupervised learning structures, can be used to separate false data from the original data. Semi-supervised learning techniques can be used to learn from small-scale samples. This enables learning the characteristics of available labelled cyberattacks to explore unlabelled samples and build detectors to defend against ever-evolving cyberattacks.

Transfer and collaborative learning are also promising strategies. Transfer learning involves fine-tuning pre-trained models that have been trained on datasets from other domains with characteristics similar to SGs. By adapting pre-trained models to SG-specific data, transfer learning can effectively build cyberattack detectors for SGs. Collaborative learning involves multiple SGs sharing their data and models to collectively improve detector performance. This approach can address the problem of insufficient data by pooling resources from different SGs.

4.2 | Selection of deep learning models

The essence of DL is to extract useful information and learn knowledge from data. However, many types of DL mechanisms exist, of which the common types are described in Section 2.2. Because each DL mechanism is designed to handle different data types, it is important to select suitable learning mechanisms to build high-performance DL models. From the above review, it is clear that the present methods focus greatly on MLP, CNN, RNN and AE to extract features from either temporal or non-temporal data. However, power grids are graphical. Some information losses will be inevitable if the dynamic graph

information of a power grid is neglected. Therefore, employing GNN-related DL models would be a promising way to capture graphical changes and extract more features, particularly with a change in the power grid topology, to offset this challenge.

However, a great novelty in the current DL development is AM. Although RNN models can handle time series, it is still believed that AM can extract more relevant features and ignore insignificant content. Hence, AM should be considered in future DL model designs. Even small improvements resulting from AE can have significant impacts, such as defending against cyber attacks and preventing catastrophic disasters in the power grid.

4.3 | Data privacy

Protecting sensitive data from cyberattacks is critical in the operation of SGs. One solution is privacy-preserving data, which can defend against cyberattacks by keeping data confidential. Federated learning [116] is a technique used for machine learning model training that aims to preserve data privacy by not sharing data between entities but learning the data in a distributed manner. This innovative approach is a promising way to preserve the privacy of SG data. However, when used for DL model training, the performance of the model can be significantly reduced. Therefore, the performance of federated DL models must be improved to protect SG data privacy.

Moreover, although entities do not require sharing their data when participating in federated learning, which seems to ensure data privacy, there are still two potential attacks that pose security risks. The first one is external attacks, in which attackers can reconstruct the sensitive information of specific entities by stealing the gradients uploaded by all the entities during the federated learning training process. The second attack occurs on the central server, which can directly access all gradients uploaded by the entities, thereby directly reconstructing and obtaining specific sensitive information. To address these security risks, researchers may need to develop more robust algorithms for federated learning to better preserve data privacy. For example, mainstream algorithms typically require uploading the gradients of the parameters for federated learning. To

mitigate these two potential attacks, researchers may consider replacing the uploaded gradients with other prototypes, such as the outputs from the representation layers. Even if attackers can obtain these prototypes, inferring their original meanings is difficult, making it challenging to conduct the attacks.

Encryption is another alternative for data protection. In an encrypted environment, most cyberattacks are infeasible because it is challenging to decrypt the transmitted data and modify them for the desired results. The two mainstreams of state-of-the-art encryptions are multi-party computation [117] and homomorphic encryption [118]. Multi-party computation has the advantage of low computational costs but requires high communication conditions. In contrast, homomorphic encryption requires less communication but is burdened with computational power. The tradeoff between the two should be carefully assessed before considering their practical applications.

4.4 | Detection mechanisms

Most of the current DL-based cyberattack detectors are binary classifiers, which means they only indicate the presence or absence of cyberattacks or their probabilities. Although this can raise possible alarms regarding data manipulation, operators cannot objectively determine the reality of an attack and its location. To improve detection mechanisms, DL can be used to establish normal operational bounds for all SG indicators, such as voltage, current, power, meters, and sensors. As these indicators are the primary targets of cyberattacks, out-of-bound indicators can be viewed as abnormal behaviours, and operators can follow these clues to determine the existence of cyberattacks and locate them.

Detectors designed for specific cyberattacks are unsuitable for the SG concept. In addition to generating, transmitting, and smartly distributing electric power, SG requires smart measures with sufficient power to automatically detect cyberattacks. DL-related methods can help improve SG cybersecurity by considering more data to construct a highly generalised and robust cyberattack defense mechanism. DL can explore more implicit information with more data fed and defend against all types of discovered cyberattacks, distinguishing them from normal faults and thus can defend against ever-evolving cyberattacks. Although defending against these cyberattacks is difficult, it can be designed to be integrable and upgradeable to mix with more mature methods in the future. In addition, the superiority of DL in knowledge discovery, mining, and extraction can be further developed to explore unseen vulnerabilities and novel cyberattack types. Studying them and developing countermeasures can effectively prevent them, thereby enhancing SG resilience.

4.5 | High-risk deep learning application scenarios

The application of DL offers substantial benefits to SG. However, the opaque nature of DL models poses a threat to SG security. Unlike traditional models, DL models lack intermediate

examination during operation, leading to a lack of transparency that can result in improper decisions that compromise SG security, particularly when external interfaces are used to transmit input data that are vulnerable to attacks. A relevant example is the utilisation of DL-based renewable energy forecasting models. The high penetration of renewable energy in SGs can make them susceptible to blackouts during extreme weather conditions when the reserve capacity is unable to offset the gap between power supply and demand. This will induce voltage and frequency instability to affect the SG operation, further leading to cascading failures or blackouts if the security and stability control is inappropriate. Examples of this include the 2019 England Blackout [119] and the 2020 California Rolling Blackouts [120]. Attackers can compromise renewable energy forecasting performance by tampering with input data, such as meteorological data transmitted through external online weather forecast system interfaces. Such tampering can lead to inaccurate predictions for renewable energy, potentially resulting in blackouts if the remaining reserve capacity cannot compensate for power generation errors. Moreover, after a long-term utilisation of DL models, operators may become overly reliant on the DL model and find it more difficult to recognise or respond to such anomalies in the SG.

To address this threat, developing explainable AI (XAI) techniques to make the decision-making processes of DL models transparent and interpretable is crucial. By incorporating interpretability algorithms, such as local interpretable model-agnostic explanations (LIME) [121] and SHapley Additive exPlanations (SHAP) [122], the interpretability of DL models can be enhanced, enabling operators to identify vulnerabilities and potentially attacked data, and develop appropriate countermeasures. In addition, implementing robust security measures to protect against attacks that target input data transmitted through external interfaces is essential. Techniques, such as data validation, can be used to verify the integrity of input data before it is used in DL models, preventing attackers from tampering with the input data. Data watermarking and provenance can be used to validate the authenticity and integrity of input data. Moreover, DL models should be continuously monitored for anomalies and updated regularly to identify any changes in their behaviour and ensure that they operate as expected. Regular updates in these high-risk DL application scenarios can also address any vulnerabilities that may have been identified.

5 | CONCLUSION

The incorporation of advanced communication technologies has enabled the development of a digitized cyber layer for efficient interaction and management in SGs. However, this digitisation has resulted in an increased vulnerability of SG's cyber assets to cyberthreats. To address this challenge, DL has emerged as a promising direction in AI owing to its ability to learn and analyse large volumes of data generated by SG and make incontestable contributions to improve SG cybersecurity. This study presents a comprehensive and systematic review of the recent advances in DL-related meth-

ods relevant to SG cybersecurity. First, DL techniques and common DL mechanisms, including their functionalities, are investigated and elaborated. This study further classifies common SG cyberthreats based on different cyberattacks, such as ETA, FDIA, FCA, CTA, and ALA. Moreover, it examines the applications of DL techniques to each cyberthreat and discusses its novelty. Additionally, general guidelines and suggestions are provided to researchers to develop a well-performing DL model for cyberattack detection. The guidelines analyse the functions of different DL mechanisms, providing viewpoints on DL-based detector modelling according to various types of data and cyberattacks. Despite the widespread use of DL to defend against cyberattacks, several challenges remain. Therefore, an outlook is presented to highlight the current challenges encountered and provide insights into possible mitigation directions.

Future developments of DL have the potential to significantly enhance the cybersecurity of SG by providing more effective and robust defense mechanisms. By enabling better detection, prevention, and response capabilities, DL can help protect critical infrastructure, maintain a reliable power supply, and ensure the resilience of SG against evolving cyberthreats. However, SG operators, regulators, and technology providers must collaborate and invest in the development and deployment of DL-based cybersecurity solutions. As SGs become more interconnected and dependent on digital technologies, the importance of effective cybersecurity measures, including DL, continues to grow.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 72171206, Grant 71931003, and Grant 72061147004; in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society; in part by the Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network under Grant ZDSYS20220606100601002; and in part by the Guangdong Regional Joint Fund-Youth Fund under Grant 2021A1515110084.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable - no new data generated, or the article describes entirely theoretical research.

ORCID

Jiaqi Ruan  <https://orcid.org/0000-0003-2584-0738>

Junhua Zhao  <https://orcid.org/0000-0001-5446-2655>

Huan Zhao  <https://orcid.org/0000-0002-3133-3137>

Jing Qiu  <https://orcid.org/0000-0001-8507-0558>

Fushuan Wen  <https://orcid.org/0000-0002-6838-2602>

REFERENCES

1. Tan, S., De, D., Song, W.-Z., Yang, J., Das, S.K.: Survey of security advances in smart grid: A data driven approach. *IEEE Commun. Surv. Tutorials* 19, 397–422 (2016)
2. Wang, Q., Zhang, G., Wen, F.: A survey on policies, modelling and security of cyber-physical systems in smart grids. *Energy Convers* 2, 197–211 (2021)
3. Wang, H., Ruan, J., Ma, Z., Zhou, B., Fu, X., Cao, G.: Deep learning aided interval state prediction for improving cyber security in energy internet. *Energy* 174, 1292–1304 (2019)
4. Cui, L., Qu, Y., Gao, L., Xie, G., Yu, S.: Detecting false data attacks using machine learning techniques in smart grid: A survey. *J. Netw* 170, 102808 (2020)
5. Jakaria, A.H.M., Rahman, M.A., Gokhale, A.: Resiliency-aware deployment of SDN in smart grid SCADA: A formal synthesis model. *IEEE Trans. Netw. Serv. Manage.* 18, 1430–1444 (2021)
6. Mrabet, Z.E., Kaabouch, N., Ghazi, H.E., Ghazi, H.E.: Cyber-security in smart grid: Survey and challenges. *Comput. Electr. Eng.* 67, 469–482 (2018)
7. Ruan, J., Liang, G., Zhao, J., Qiu, J., Dong, Z.Y.: An inertia-based data recovery scheme for false data injection attack. *IEEE Trans. Ind. Inf.* 18(11), 7814–7823 (2022)
8. Hossain, E., Khan, I., Un-Noor, F., Sikander, S.S., Sunny, M.S.H.: Application of big data and machine learning in smart grid, and associated security concerns: a review. *IEEE Access* 7, 13960–13988 (2019)
9. Aslani, M., Hashemi-Dezaki, H., Ketabi, A.: Reliability evaluation of smart microgrids considering cyber failures and disturbances under various cyber network topologies and distributed generation's scenarios. *Sustainability* 13, 5695 (2021)
10. Kotsiopoulos, T., Sarigiannidis, P., Ioannidis, D., Tzovaras, D.: Machine learning and deep learning in smart manufacturing: The smart grid paradigm. *Comput. Sci. Rev.* 40, 100341 (2021)
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* 521, 436–444 (2015)
12. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* 6, 14410–14430 (2018)
13. Chen, Y., Tan, Y., Deka, D.: Is machine learning in power systems vulnerable? In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE, Piscataway (2018)
14. Bécue, A., Praça, I., Gama, J.: Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities. *Artif. Intell. Rev.* 54, 3849–3886 (2021)
15. Gardner, M.W., Dorling, S.R.: Artificial neural networks (the multi-layer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* 32, 2627–2636 (1998)
16. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. *Patt. Recogn.* 77, 354–377 (2018)
17. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. IET, Stevenage (2017)
18. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681 (1997)
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* 9, 1735–1780 (1997)
20. Dey, R., Salem, F.M.: Gate-variants of Gated Recurrent Unit (GRU) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1597–1600. IEEE, Piscataway (2017)
21. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* 313, 504–507 (2006)
22. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103. Association for Computing Machinery, New York (2008)
23. Le Roux, N., Bengio, Y.: Representational power of restricted boltzmann machines and deep belief networks. *Neural Comput.* 20, 1631–1649 (2008)
24. Hinton, G.E.: Deep belief networks. *Scholarpedia* 4, 5947 (2009)

25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE, Piscataway (2016)
26. Balduzzi, D., Fream, M., Leary, L., Lewis, J.P., Ma, K.W.D., McWilliams, B.: The shattered gradients problem: If resnets are the answer, then what is the question? In: *Proceedings of the 34th International Conference on Machine Learning*, pp. 342–350. PMLR (2017)
27. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. *IEEE Trans. Neural Networks* 20, 61–80 (2009)
28. Zhang, S., Tong, H., Xu, J., Maciejewski, R.: Graph convolutional networks: A comprehensive review. *Computat. Social Netw.* 6, 11 (2019)
29. Ruan, J., Liang, G., Zhao, J., Lei, S., He, B., Qiu, J., et al.: Graph deep learning-based retail dynamic pricing for demand response. *IEEE Trans. Smart Grid* (2023)
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.: Generative adversarial networks. *Commun. ACM* 63, 139–144 (2020)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is All you Need. In: *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., Red Hook (2017)
32. Li, J., Wang, X., Tu, Z., Lyu, M.R.: On the diversity of multi-head attention. *Neurocomputing* 454, 14–24 (2021)
33. Zheng, Z., Yang, Y., Niu, X., Dai, H.N., Zhou, Y.: Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Trans. Ind. Inf.* 14, 1606–1615 (2018)
34. Bhat, R.R., Trevizan, R.D., Sengupta, R., Li, X., Bretas, A.: Identifying nontechnical power loss via spatial and temporal deep learning. In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 272–279. IEEE, Piscataway (2016)
35. He, Y., Mendis, G.J., Wei, J.: Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism. *IEEE Trans. Smart Grid* 8, 2505–2516 (2017)
36. Yao, D., Wen, M., Liang, X., Fu, Z., Zhang, K., Yang, B.: Energy theft detection with energy privacy preservation in the smart grid. *IEEE Internet of Things J.* 6, 7659–7669 (2019)
37. Li, S., Han, Y., Yao, X., Yingchen, S., Wang, J., Zhao, Q.: Electricity theft detection in power grids with deep learning and random forests. *J. Electr. Comp. Eng.* 2019, e4136874 (2019)
38. Hu, T., Guo, Q., Shen, X., Sun, H., Wu, R., Xi, H.: Utilizing unlabeled data to detect electricity fraud in AMI: A semisupervised deep learning approach. *IEEE Trans. Neural Networks Learn. Syst.* 30, 3287–3299 (2019)
39. Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P., Gómez-Expósito, A.: Hybrid deep neural networks for detection of non-technical losses in electricity smart meters. *IEEE Trans. Power Syst.* 35, 1254–1263 (2020)
40. Takiddin, A., Ismail, M., Zafar, U., Serpedin, E.: Robust electricity theft detection against data poisoning attacks in smart grids. *IEEE Trans. Smart Grid* 12, 2675–2684 (2021)
41. Takiddin, A., Ismail, M., Nabil, M., Mahmoud, M.M.E.A., Serpedin, E.: Detecting electricity theft cyber-attacks in ami networks using deep vector embeddings. *IEEE Syst. J.* 15, 4189–4198 (2021)
42. Ibrahim, M.I., Mahmoud, M.M.E.A., Alsolami, F., Alasmay, W., AL-Ghamdi, A.S.A.M., Shen, X.: Electricity-theft detection for change-and-transmit advanced metering infrastructure. *IEEE Internet Things J.* 9, 25565–25580 (2022)
43. Pamir, Javaid, N., Qasim, U., Yahaya, A.S., Alkhamash, E.H., Hadjouni, M.: Non-technical losses detection using autoencoder and bidirectional gated recurrent unit to secure smart grids. *IEEE Access* 10, 56863–56875 (2022)
44. Abdulaal, M.J., Ibrahim, M.I., Mahmoud, M.M.E.A., Khalid, J., Aljohani, A.J., Milyani, A.H., et al.: Real-time detection of false readings in smart grid AMI using deep and ensemble learning. *IEEE Access* 10, 47541–47556 (2022)
45. Li, W., Logenthiran, T., Phan, V.T., Woo, W.L.: A Novel smart energy theft system (SETS) for IoT-based smart home. *IEEE Internet Things J.* 6, 5531–5539 (2019)
46. Cui, L., Guo, L., Gao, L., Cai, B., Qu, Y., Zhou, Y., et al.: A Covert electricity-theft cyber-attack against machine learning-based detection models. *IEEE Trans. Ind. Inf.* 18(11), 7824–7833 (2022)
47. Badr, M.M., Ibrahim, M.I., Mahmoud, M., Fouda, M.M., Alsolami, F., Alasmay, W.: Detection of false-reading attacks in smart grid net-metering system. *IEEE Internet Things J.* 9, 1386–1401 (2022)
48. Ismail, M., Shaaban, M.F., Naidu, M., Serpedin, E.: Deep learning detection of electricity theft cyber-attacks in renewable distributed generation. *IEEE Trans. Smart Grid* 11, 3428–3437 (2020)
49. Wang, Z., He, H., Wan, Z., Sun, Y.: Coordinated topology attacks in smart grid using deep reinforcement learning. *IEEE Trans. Ind. Inf.* 17, 1407–1415 (2021)
50. Tian, J., Wang, B., Li, J., Wang, Z., Ma, B., Ozay, M.: Exploring targeted and stealthy false data injection attacks via adversarial machine learning. *IEEE Internet Things J.* 9, 14116–14125 (2022)
51. Tian, J., Wang, B., Wang, Z., Cao, K., Li, J., Ozay, M.: Joint adversarial example and false data injection attacks for state estimation in power systems. *IEEE Trans. Cybern.* 52, 13699–13713 (2022)
52. Wang, H., Ruan, J., Wang, G., Zhou, B., Liu, Y., Fu, X., et al.: Deep learning-based interval state estimation of AC smart grids against sparse cyber attacks. *IEEE Trans. Ind. Inf.* 14, 4766–4778 (2018)
53. Yu, J.J.Q., Hou, Y., Li, V.O.K.: Online false data injection attack detection with wavelet transform and deep neural networks. *IEEE Trans. Ind. Inf.* 14, 3271–3280 (2018)
54. Li, Y., Huo, W., Qiu, R., Zeng, J.: Efficient detection of false data injection attack with invertible automatic encoder and long-short-term memory. *IET Cyber-Phys. Syst.: Theory Appl.* 5, 110–118 (2020)
55. Kundu, A., Sahu, A., Serpedin, E., Davis, K.: A3D: Attention-based auto-encoder anomaly detector for false data injection attacks. *Electr. Power Syst. Res.* 189, 106795 (2020)
56. Wang, S., Bi, S., Zhang, Y.J.A.: Locational detection of the false data injection attack in a smart grid: A multilabel classification approach. *IEEE Internet Things J.* 7, 8218–8227 (2020)
57. Yang, L., Zhai, Y., Li, Z.: Deep learning for online AC false data injection attack detection in smart grids: An approach using LSTM-autoencoder. *J. Netw. Comp. Appl.* 193, 103178 (2021)
58. Zhang, Y., Yan, J.: Semi-supervised domain-adversarial training for intrusion detection against false data injection in the smart grid. In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, Piscataway (2020)
59. Zhang, Y., Wang, J., Chen, B.: Detecting false data injection attacks in smart grids: a semi-supervised deep learning approach. *IEEE Trans. Smart Grid* 12, 623–634 (2021)
60. Mohammadpourfard, M., Khalili, A., Genc, I., Konstantinou, C.: Cyber-resilient smart cities: detection of malicious attacks in smart grids. *Sustain. Cities Soc.* 75, 103116 (2021)
61. Almutairi, F., Scekkic, L., Elmoudi, R., Wshah, S.: Accurate detection of false data injection attacks in renewable power systems using deep learning. *IEEE Access* 9, 135774–135789 (2021)
62. Yin, X., Zhu, Y., Hu, J.: A subgrid-oriented privacy-preserving microservice framework based on deep neural network for false data injection attack detection in smart grids. *IEEE Trans. Ind. Inf.* 18, 1957–1967 (2022)
63. Mukherjee, D.: A novel strategy for locational detection of false data injection attack. *Sustain. Energy Grids Netw.* 31, 100702 (2022)
64. Huang, R., Li, Y., Wang, X.: Attention-aware deep reinforcement learning for detecting false data injection attacks in smart grids. *Int. J. Electr. Power Energy Syst.* 147, 108815 (2023)
65. An, D., Zhang, F., Yang, Q., Zhang, C.: Data integrity attack in dynamic state estimation of smart grid: attack model and countermeasures. *IEEE Trans. Autom. Sci. Eng.* 19, 1631–1644 (2022)
66. Mukherjee, D., Chakraborty, S., Abdelaziz, A.Y., El-Shahat, A.: Deep learning-based identification of false data injection attacks on modern smart grids. *Energy Rep.* 8, 919–930 (2022)
67. Bhattacharjee, A., Mondal, A.K., Verma, A., Mishra, S., Saha, T.K.: Deep latent space clustering for detection of stealthy false data injection attacks against AC state estimation in power systems. *IEEE Trans. Smart Grid* 14(3), 2338–2351 (2022)

68. Huang, X., Qin, Z., Xie, M., Liu, H., Meng, L.: Defense of massive false data injection attack via sparse attack points considering uncertain topological changes. *J. Mod. Power Syst. Clean Energy* 10, 1588–1598 (2022)
69. Vincent, E., Korki, M., Seyedmahmoudian, M., Stojcevski, A., Mekhilef, S.: Detection of false data injection attacks in cyber–physical systems using graph convolutional network. *Electr. Power Syst. Res.* 217, 109118 (2023)
70. He, Z., Khazaei, J., Moazeni, F., Freihaut, J.D.: Detection of false data injection attacks leading to line congestions using Neural networks. *Sustain. Cities Soc.* 82, 103861 (2022)
71. Li, Y., Wei, X., Li, Y., Dong, Z., Shahidepour, M.: Detection of false data injection attacks in smart grid: A secure federated deep learning approach. *IEEE Trans. Smart Grid* 13, 4862–4872 (2022)
72. Lu, K.D., Wu, Z.G., Huang, T.: Differential evolution-based three stage dynamic cyber-attack of cyber-physical power systems. *IEEE/ASME Trans. Mechatron.* 28(2), 1137–1148 (2022)
73. Tahir, B., Jolfaei, A., Tariq, M.: Experience-driven attack design and federated-learning-based intrusion detection in industry 4.0. *IEEE Trans. Ind. Inf.* 18, 6398–6405 (2022)
74. Wang, Y., Xing, A., Qu, Z., Han, X., Dong, H., Georgievitch, P.M.: False data injection attack detection based on interval affine state estimation. *Electr. Power Syst. Res.* 210, 108100 (2022)
75. Han, Y., Feng, H., Li, K., Zhao, Q.: False data injection attacks detection with modified temporal multi-graph convolutional network in smart grids. *Comp. Sec.* 124, 103016 (2023)
76. Majidi, S.H., Hadayeghpars, S., Karimipour, H.: FDI attack detection using extra trees algorithm and deep learning algorithm-autoencoder in smart grid. *Int. J. Crit. Infrastruct. Prot.* 37, 100508 (2022)
77. Li, X., Wang, Y., Lu, Z.: Graph-based detection for false data injection attacks in power grid. *Energy* 263, 125865 (2023)
78. Moayyed, H., Mohammadpourfard, M., Konstantinou, C., Moradzadeh, A., Mohammadi-Ivatloo, B., Aguiar, A.P.: Image processing based approach for false data injection attacks detection in power systems. *IEEE Access* 10, 12412–12420 (2022)
79. Li, Y., Wu, J.: Low latency cyberattack detection in smart grids with deep reinforcement learning. *Int. J. Electr. Power Energy Syst.* 142, 108265 (2022)
80. Khazaei, J., Moazeni, F.: Neural networks-based detection of line overflow cyberattacks on AC state-estimation of smart grids. *IEEE Syst. J.* 17(2), 2399–2410 (2022)
81. Yin, X., Zhu, Y., Xie, Y., Hu, J.: PowerFDNet: Deep learning-based stealthy false data injection attack detection for AC-model transmission systems. *IEEE Open J. Comput. Soc.* 3, 149–161 (2022)
82. Zhang, H., Yue, D., Dou, C., Hancke, G.P.: Resilient optimal defensive strategy of micro-grids system via distributed deep reinforcement learning approach against FDI attack. *IEEE Trans. Neural Networks Learn. Syst.* 1–11 (2022)
83. Ruan, J., Fan, G., Zhu, Y., Liang, G., Zhao, J., Wen, F., et al.: Super-resolution perception assisted spatiotemporal graph deep learning against false data injection attacks in smart grid. *IEEE Trans. Smart Grid* 1–1 (2023)
84. Li, Y., Wang, Y., Hu, S.: Online generative adversary network based measurement recovery in false data injection attacks: A cyber-physical approach. *IEEE Trans. Ind. Inf.* 16, 2031–2043 (2020)
85. Wang, J., Shi, D., Li, Y., Chen, J., Ding, H., Duan, X.: Distributed framework for detecting PMU data manipulation attacks with deep autoencoders. *IEEE Trans. Smart Grid* 10, 4401–4410 (2019)
86. Wilson, D., Tang, Y., Yan, J., Lu, Z.: Deep learning-aided cyber-attack detection in power transmission systems. In: 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE, Piscataway (2018)
87. Hu, C., Yan, J., Wang, C.: Robust feature extraction and ensemble classification against cyber-physical attacks in the smart grid. In: 2019 IEEE Electrical Power and Energy Conference (EPEC), pp. 1–6. IEEE, Piscataway (2019)
88. Wei, F., Wan, Z., He, H., Lin, X.: Ultrafast active response strategy against malfunction attack on fault current limiter. *IEEE Trans. Smart Grid* 11, 2722–2733 (2020)
89. Farajzadeh-Zanjani, M., Hallaji, E., Razavi-Far, R., Saif, M., Parvania, M.: Adversarial semi-supervised learning for diagnosing faults and attacks in power grids. *IEEE Trans. Smart Grid* 12, 3468–3478 (2021)
90. Bitirgen, K., Filik, Ü.B.: A hybrid deep learning model for discrimination of physical disturbance and cyber-attack detection in smart grid. *Int. J. Crit. Infrastruct. Prot.* 40, 100582 (2023)
91. Khaw, Y.M., Abiri-Jahromi, A., Arani, M.F.M., Sanner, S., Kundur, D., Kassouf, M.: A deep learning-based cyberattack detection system for transmission protective relays. *IEEE Trans. Smart Grid* 12, 2554–2565 (2021)
92. Wei, F., Wan, Z., He, H.: Cyber-attack recovery strategy for smart grid based on deep reinforcement learning. *IEEE Trans. Smart Grid* 11, 2476–2486 (2020)
93. Ganjkhani, M., Gilanifar, M., Giraldo, J., Parvania, M.: Integrated cyber and physical anomaly location and classification in power distribution systems. *IEEE Trans. Ind. Inf.* 17, 7040–7049 (2021)
94. Li, B., Wu, Y., Song, J., Lu, R., Li, T., Zhao, L.: DeepFed: Federated deep learning for intrusion detection in industrial cyber–physical systems. *IEEE Trans. Ind. Inf.* 17, 5615–5624 (2021)
95. Abianeh, A.J., Wan, Y., Ferdowsi, F., Mijatovic, N., Dragičević, T.: Vulnerability identification and remediation of FDI attacks in islanded DC microgrids using multiagent reinforcement learning. *IEEE Trans. Power Electron.* 37, 6359–6370 (2022)
96. Paffenroth, R.C., Zhou, C.: Modern machine learning for cyber-defense and distributed denial-of-service attacks. *IEEE Eng. Manage. Rev.* 47, 80–85 (2019)
97. Gao, J., Gan, L., Buschendorf, F., Zhang, L., Liu, H., Li, P., et al.: Omni SCADA intrusion detection using deep learning algorithms. *IEEE Internet Things J.* 8, 951–961 (2021)
98. Hussain, B., Du, Q., Sun, B., Han, Z.: Deep Learning-based DDoS-attack detection for cyber–physical system over 5G network. *IEEE Trans. Ind. Inf.* 17, 860–870 (2021)
99. Qiu, C., Yu, F.R., Yao, H., Jiang, C., Xu, F., Zhao, C.: Blockchain-based software-defined industrial internet of things: a dueling deep QoS-learning approach. *IEEE Internet Things J.* 6, 4627–4639 (2019)
100. Lou, X., Tran, C., Yau, D.K.Y., Tan, R., Ng, H., Fu, T.Z., et al.: Learning-based time delay attack characterization for cyber-physical systems. In: 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE, Piscataway (2019)
101. Ganesh, P., Lou, X., Chen, Y., Tan, R., Yau, D.K.Y., Chen, D., et al.: Learning-based simultaneous detection and characterization of time delay attack in cyber-physical systems. *IEEE Trans. Smart Grid* 12, 3581–3593 (2021)
102. Yang, H., Cheng, L., Chuah, M.C.: Deep-learning-based network intrusion detection for SCADA systems. In: 2019 IEEE Conference on Communications and Network Security (CNS), pp. 1–7. IEEE, Piscataway (2019)
103. Li, J.H.: Cyber security meets artificial intelligence: A survey. *Front. Inf. Technol. Electron. Eng.* 19, 1462–1474 (2018)
104. Miller, D.J., Xiang, Z., Kesidis, G.: Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proc. IEEE* 108, 402–433 (2020)
105. Kim, J., Park, N.: Blockchain-based data-preserving AI learning environment model for AI cybersecurity systems in IoT service environments. *Appl. Sci.* 10, 4718 (2020)
106. Hao, J., Tao, Y.: Adversarial attacks on deep learning models in smart grids. *Energy Rep.* 8, 123–129 (2022)
107. Chen, Y., Tan, Y., Zhang, B.: Exploiting vulnerabilities of load forecasting through adversarial attacks. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems. E-Energy '19, pp. 1–11. Association for Computing Machinery, New York (2019)

108. Tufail, S., Batool, S., Sarwat, A.I.: False data injection impact analysis in AI-based smart grid. In: SoutheastCon 2021, pp. 01–07. IEEE, Piscataway (2021)
109. Moradzadeh, A., Mohammadpourfard, M., Konstantinou, C., Genc, I., Kim, T., Mohammadi-Ivatloo, B.: Electric load forecasting under False Data Injection Attacks using deep learning. *Energy Rep.* 8, 9933–9945 (2022)
110. Zheng, Y., Yan, Z., Chen, K., Sun, J., Xu, Y., Liu, Y.: Vulnerability assessment of deep reinforcement learning models for power system topology optimization. *IEEE Trans. Smart Grid* 12, 3613–3623 (2021)
111. Song, Q., Tan, R., Ren, C., Xu, Y., Lou, Y., Wang, J., et al.: On credibility of adversarial examples against learning-based grid voltage stability assessment. *IEEE Trans. Dependable Secure Comput.*, pp. 1–14. (2022)
112. Zeng, L., Sun, M., Wan, X., Zhang, Z., Deng, R., Xu, Y.: Physics-constrained vulnerability assessment of deep reinforcement learning-based SCOPF. *IEEE Trans. Power Syst.* 38(3), 2690–2704 (2022)
113. Northeast Group LLC.: Electricity theft and non-technical losses: Global markets, solutions, and vendors (2017). Available: <http://www.northeast-group.com>
114. Liu, Y., Ning, P., Reiter, M.K.: False data injection attacks against state estimation in electric power grids. *ACM Trans. Inf. Syst. Secur.* 14, 13:1–13:33 (2011)
115. Alfeld, S., Zhu, X., Barford, P.: Data poisoning attacks against autoregressive models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30. AAAI Press, Menlo Park, CA (2016)
116. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag.* 37, 50–60 (2020)
117. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing. STOC '09, pp. 169–178. Association for Computing Machinery, New York (2009)
118. Du, W., Atallah, M.J.: Secure multi-party computation problems and their applications: A review and open problems. In: Proceedings of the 2001 Workshop on New Security Paradigms. NSPW '01, pp. 13–22. Association for Computing Machinery, New York (2001)
119. Bialek, J.: What does the GB power outage on 9 August 2019 tell us about the current state of decarbonised power systems? *Energy Policy* 146, 111821 (2020)
120. Yang, J., Dong, Z.Y., Wen, F., Chen, Q., Liang, B.: Spot electricity market design for a power system characterized by high penetration of renewable energy generation. *Energy Conv. Econ.* 2, 67–78 (2021)
121. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM, New York (2016)
122. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., Red Hook (2017)

How to cite this article: Ruan, J., Liang, G., Zhao, J., Zhao, H., Qiu, J., Wen, F., Dong, Z.Y.: Deep learning for cybersecurity in smart grids: Review and perspectives. *Energy Convers. Econ.* 4, 233–251 (2023).
<https://doi.org/10.1049/enc2.12091>