Deep Learning Based Attack Detection for Cyber-Physical System Cybersecurity: A Survey

Jun Zhang, *Senior Member, IEEE*, Lei Pan, *Member, IEEE*, Qing-Long Han, *Fellow, IEEE*, Chao Chen, *Member, IEEE*, Sheng Wen, *Member, IEEE*, and Yang Xiang, *Fellow, IEEE*

***Abstract*—With the booming of cyber attacks and cyber criminals against cyber-physical systems (CPSs), detecting these attacks remains challenging. It might be the worst of times, but it might be the best of times because of opportunities brought by machine learning (ML), in particular deep learning (DL). In general, DL delivers superior performance to ML because of its layered setting and its effective algorithm for extract useful information from training data. DL models are adopted quickly to cyber attacks against CPS systems. In this survey, a holistic view of recently proposed DL solutions is provided to cyber attack detection in the CPS context. A six-step DL driven methodology is provided to summarize and analyze the surveyed literature for applying DL methods to detect cyber attacks against CPS systems. The methodology includes CPS scenario analysis, cyber attack identification, ML problem formulation, DL model customization, data acquisition for training, and performance evaluation. The reviewed works indicate great potential to detect cyber attacks against CPS through DL modules. Moreover, excellent performance is achieved partly because of several high-quality datasets that are readily available for public use. Furthermore, challenges, opportunities, and research trends are pointed out for future research.**

***Index Terms*—Cyber-physical system, cybersecurity, deep learning, intrusion detection, pattern classification.**

I. Introduction

CYBER-physical systems (CPSs) suffer from cyber attacks when they are increasingly connected to the cyber space. According to [1] published in 2017, more than 30 surveys were published to cover the cybersecurity issue in the CPSs. Cyber attacks have become increasingly sophisticated and prevalent as automated attacking tools, and professional hacking groups have started to get involved. A successful cyber attack against a CPS may be disastrous, catastrophic, or even fatal [2]–[6]. However, it is a challenge to defend against cyber attacks on CPSs. Many CPS systems lack cybersecurity mechanisms like message authentication, resulting in challenges to detect false data injection attacks. A lack of universal encryption, especially on the systems employing dated technologies, makes it challenging to defend against eavesdropping attacks. System states need to be referred to detect replay attacks. In addition, the use of dated technology in operation limits the choices of defenses to network traffic in most cases [7].

Deep learning (DL) [8], [9] delivers superior performance to traditional machine learning (ML) solutions. Whenever there is adequate data, DL models almost deliver excellent results. However, DL models have been slowly applied to solve the CPS cybersecurity issue compared with other fields such as NLP, image processing, software vulnerability [10], [11], and many more [12]–[17]. It is also observed that many DL models have been proposed in recent publications to detect CPS cyber attacks. A widely accepted view to explain the difficulty of detecting cyber attacks on CPSs was accredited to the degree of complexity when superposing cybersecurity over CPSs [2].

There exist a few short-length survey papers on CPS cybersecurity [1], [2], [18], [19]. Some papers investigated data-driven methods for detecting cyber attacks against CPS systems [18], [20]. However, there is no detailed discussion on applying DL methods to detect CPS cyber attacks. A short survey was provided in [18] with a four-step framework to apply DL methods on CPS issues, including cybersecurity, adaptability, recoverability, and many more, without a specific focus on cybersecurity. Furthermore, most of the cited works in [18] were published between 2012 and 2016, but this survey includes most papers between 2017 and 2021. A survey of surveys was presented in [1] without relevance to DL models. A comprehensive survey on the cyber attacks against CPSs was presented in [20] without investigating the DL models. Various methods of detecting cyber attacks in the CPSs were summarized in [2] without using DL methods. A comprehensive list of CPS attacks and challenges were provided in [19] but overlooking ML, or DL approaches. A cybersecurity analysis framework was proposed in [21] without utilizing the rich sources of available data. A recently published survey in [22] presents cybersecurity control and state estimation from active and passive defence perspectives.

We aim to review current research works on the advances of

Manuscript received June 14, 2021; revised July 23, 2021; accepted July 31, 2021. Recommended by Associate Editor Mengchu Zhou. *(Corresponding author: Qing-Long Han.)*

Citation: J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, and Y. Xiang, "Deep learning based attack detection for cyber-physical system cybersecurity: a survey," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 3, pp. 377–391, Mar. 2022.

J. Zhang, Q.-L. Han, S. Wen, and Y. Xiang are with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC 3122, Australia (e-mail: {junzhang; qhan; swen; yxiang}@swin.edu.au).

L. Pan is with the School of Information Technology, Deakin University, Geelong, VIC 3216, Australia (e-mail: l.pan@deakin.edu.au).

C. Chen is with the College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia (e-mail: chao.chen@jcu.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2021.1004261

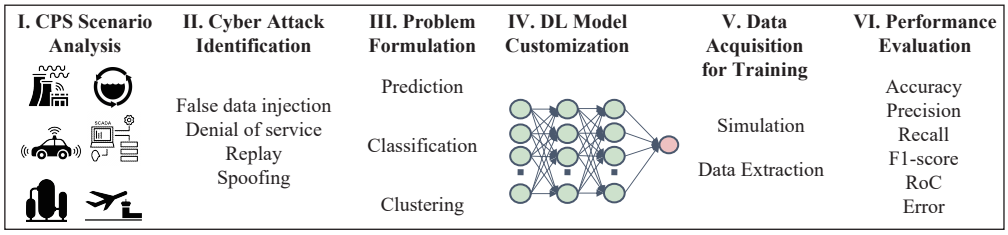

Fig. 1. The DL driven methodology for CPS cybersecurity considers the essential needs for training robust and usable DL models in the context of cyber attacks against the CPS systems.

DL driven solutions for detecting cyber attacks in the CPS domain. It provides an overview for readers to quickly understand and step into the field by following our six-step DL driven methodology. Our six-step methodology considers the complete cycle of DL application from broad scenarios to performance evaluation. This paper caters to researchers, practitioners, and students interested in building DL-based cybersecurity applications in CPSs. The key contributions of this survey are three-fold:

1) We conduct an up-to-date review of detecting cyber attacks in CPSs using DL models and propose a six-step methodology to position and analyze the surveyed works.

2) We provide an overview for the state-of-the-art solutions with preservation of technical details.

3) Based on the methodology, we discuss the challenges and future research directions.

The rest of this survey is organized as follows: Section II proposes a research methodology for deep learning driven CPS cybersecurity. Section III presents the reviews on state-of-the-art research. Section IV discusses the research challenges and future work. Finally, Section V concludes this survey.

II. Research Methodology

Our methodology represents a deep understanding of the surveyed papers. The process consists of six steps, including CPS scenario analysis, cyber attack identification, DL problem formulation, DL model construction, data acquisition, and performance evaluation. Fig. 1 shows a process of detecting cyber attacks in the context of a CPS by using DL models. For example, a smart grid may suffer from erroneous controls derived by electric load forecasts [20], [23]. Falsely injected messages containing maliciously crafted information need to be identified and eliminated before committing the prediction process. A stacked AutoEncoder (AE) proposed in [24] may serve as a reliable regressor to predict the energy load on the system. The chosen AutoEncoder was subsequently trained with sufficient simulation data. At last, the DL model delivered excellent prediction results with the mean absolute percentage error of 3.51% on annual predictions.

A. Step I: CPS Scenario Analysis

The normal operations of CPSs rely on several important factors, including dependability, real-time operation, fault tolerance, cybersecurity, and many more. We must consider these requirements holistically. Dependability consists of service availability and reliability to minimize the system downtime; real-time operation is a critical factor for maintaining the system operation when the inputs and environment rapidly change; fault tolerance requires that the critical components of the system have sufficient backups to prevent the system from shutting down; and cybersecurity requirements are becoming more and more prominent when many CPSs are connected to the cyber space to improve the quality of system control and the overall level of quality of service. According to Mitchell *et al*. [2], there are four primary categories of characteristics of CPS intrusion detection, including physical process monitoring, closed control loops, attack sophistication, and legacy technology.

Physical process monitoring: Physical properties of a CPS should be constantly monitored to identify any anomalies of the system because many physical processes of the CPS follow the laws of physics.

Closed control loops: CPS events are significantly more regular and predictable than user-triggered events because many CPS events are driven by the preset feedback-based controllers.

Attack sophistication: Sophisticated cyber attacks are increasingly popular in the CPS context because the potentially huge payoff for a successful cyber attack may bring sensitive information, valuable intelligence for military or finance operations, and many more.

Legacy technology: Legacy hardware commonly used in the CPSs cannot interact with software-defined control because of the existing mechanical and hydraulic control.

Analyzing the characteristics of a CPS scenario will help craft an appropriate cybersecurity problem. The involvement of physical signals enriches the input variables and complicates the design of any security solutions for CPSs. Although the behaviors of simplified proof-of-concept systems are relatively regular and predictable, real-world systems often operate in a noisy environment with unprecedented cyber threats.

B. Step II: Cyber Attack Identification

Upon completion of identifying the CPS scenario, we need to define a set of appropriate cyber attacks associated with CPS characteristics. For example, we will have more confidence to detect the falsely injected network packets if physical processes of the CPS components are properly monitored; cyber attacks like replay attacks may be detected on a CPS with a closed control loop; unknown attacks and sophisticated attacks like web attacks need to be considered if there is any concern of attack sophistication; denial of service (DoS) attacks and replay attacks are more prevalent in the

presence of legacy technology.

Based on the surveyed articles, we identify many common cyber attacks. Some frequent cyber attacks against the industrial control network include false data injection attacks, DoS attacks, replay attacks, and alike; and some frequent cyber attacks against the software-based controllers with a centralized server include brute force attacks, botnets, web attacks, heartbleed attacks, infiltration attacks and many more. Effective and efficient detection of these cyber attacks can be leveraged by using DL models, so we will need to translate the cybersecurity problem to the ML domain.

C. Step III: ML Problem Formulation

After aligning the cyber attacks to the CPS characteristics, the research problem can be translated to the ML/DL domain. ML is defined in [25] as "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." DL is referred to in [8] as solving a complex problem by using a hierarchy of more straightforward concepts without too much human intervention. The definition of ML is general, and we will implement an ML solution in multiple steps. In this step, we need to define the task *T*, including classification, clustering, regression, etc. A classification task requires that the trained model allocates its output to a pre-defined set of "classes" which could be the specific cyber attack categories; a clustering task often requires that the trained model allocates its output to a few "clusters" which could indicate normal traffic or attack traffic; a regression task is also known as a prediction task which requires the trained model to predict some numerical values. For example, a classification problem was found in [26] to differentiate cyber attack types; a clustering problem was found in [27] to separate covert messages from the normal messages; and a regression problem was set in [24] to predict the electric load in a smart grid. The choice of the ML tasks will impact the construction of the DL models.

D. Step IV: DL Model Customization

The DL model is constructed by selecting an architecture suitable for the research problem and optimizing parameters. The choice of DL models should be made according to actual needs. For example, autoencoders are good at translating the input data so that they are suitable for learning the representations of the data often required in prediction or regression tasks [24]; convolution networks (CNNs) and other models are usually used in classification tasks [28].

The configuration of the chosen DL model also depends on the available data. A DL model with a large number of neurons per layer will almost always require more data than a DL model with the same design but a few neurons per layer. Some trade-offs can also be made by stacking more hidden layers inside the DL model instead of expanding the layer size. The ways and insights of the customizing model can be explored based on a thorough understanding of DL algorithms and CPS cybersecurity data. Furthermore, we can achieve improvement at various levels by combining the choice of DL models with a specific research problem.

E. Step V: Data Acquisition for Training

Data acquisition is a critical step for training DL models. The quality and quantity of data determine the effectiveness of solving the research problem. Also, data can serve as the source for setting up ground truth and affect the prediction model's performance. One of the simplest methods to collect data is through simulation. This method is often used to generate datasets for power grids such as IEEE 9-bus, 14-bus, 30-bus, and 118-bus systems in Matlab. The other method relies on several existing datasets harvested by other researchers. These datasets include the SWaT dataset[1], the SCADA IDS dataset[2], the CICIDS2017 dataset[3], the UNSW-NB15 dataset[4], and the KDD99 Cup dataset[5].

Different cyber attacks were included in the datasets:

1) The SWaT dataset contained eleven days of network traffic collected from a scaled-down water treatment plant. And there were no attacks during the first seven days. It includes 36 types of cyber attacks that are most commonly seen in today's CPS systems.

2) The SCADA IDS dataset contained network traffic logs of a SCADA IDS system. It includes seven types of cyber attacks — injection random response packets, hide the real state of the controlled process, inject malicious state commands, inject malicious parameter commands, inject malicious function code commands, DoS attack, and recon attack.

3) The CICIDS2017 dataset [29] contained network traffic logs collected from an industrial control system. It includes six types of cyber attacks — brute force attacks, botnet, DoS attack, web attack, heartbleed attack, and infiltration attack.

4) The Bot-IoT dataset [30] contained network traffic logs collected from an IoT setup. It includes three types of cyber attacks — infiltration, DoS attack, and information theft.

5) The power system dataset [31] contained the network traffic collected from a power grid. It includes three cyber attacks — data injection, remote command injection, and replay attacks.

6) The UNSW-NB15 dataset [32] contained the network traffic information extracted form a 100 GB packet capture dump. It includes nine cyber attacks — DoS attacks, exploits, recon attacks, worms, fuzzers, web penetration attacks, backdoors, shellcode, and generic attacks against block ciphers.

7) The KDD99 Cup dataset [33] contained the network traffic information presented at the ACM SIGKDD conference in 1999. It includes four cyber attacks — DoS attacks, unauthorized accesses, privilege escalation, and probing attacks.

F. Step VI: Performance Evaluation

The last step is used to determine whether the DL model

[1] http://itrust.sutd.edu.sg/dataset/SWaT
[2] https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets
[3] https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets
[4] https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
[5] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

meets our expected objectives through performance evaluation. The performance is usually measured according to various metrics. We divide the performance metrics into two categories according to the tasks: 1) For prediction or regression tasks, a number of error metrics are used to measure the performance, including mean absolute error (MAE), mean relative error (MRE), root of mean squared error (RMSE), and mean absolute percentage error (MAPE). 2) For classification or clustering tasks, there are a few standard metrics, including accuracy, recall, precision, false positive rate (FPR), F1 score. And occasionally, graphical plots like receive operating characteristic (ROC) curves are used by plotting TPR as y-axis and FPR as x-axis to depict the trade-offs between benefits and costs. Finally, area under ROC curve (auROC) is used to indicate the cumulative strength of a particular ROC curve.

In many cases, FPR poses challenges for the DL models because the false alarms almost always result in excessive costs associated with manual verification. And it is always challenging to detect the rare or even unknown attacks as proven in [34] so that most of the surveyed literature aimed to maximize the TPR while minimizing the FPR. On the other hand, the error rate can be tolerated more generously in the regression task than in the classification task. By leveraging comprehensive evaluation metrics, we can decide whether the outputs of a specific DL model are satisfactory. Whenever there are unsatisfactory results, the process should be repeated with proper adjustments.

III. CPS Cybersecurity With Deep Neural Models

This Section surveys the relevant literature of detecting cyber attacks in the context of CPSs by following the research methodology described in Fig. 1. In particular, the body of the literature is divided into two parts according to the DL architectures, which will be elaborated below.

A. Representation Learning for Attack Detection

An AutoEncoder-based (AE) model was proposed in [26] to preserve privacy information in the context of smart power networks. Data privacy violations are becoming more and more popular in smart power networks. It is challenging to defend against inference attacks, because the smart power networks represent the CPS characteristics of physical process monitoring, closed control loops, attack sophistication, and legacy technology. The research problem of defending against inference attacks was translated into a classification problem in the ML domain. A Variational AutoEncoder (VAE) was proposed to provide transformed features for the ultimate classification task and transform raw data into an encoded format for preventing inference attacks. A VAE is a feedforward model used for encoding an input into new data codes using a set of weighted parameters. The VAE consisted of one input layer, four hidden layers, and one output layer. The transformed data from the output layer were written to the database for publication. Two datasets were used to evaluate the VAE, i.e., the power system dataset [31] and the UNSW-NB15 dataset [32]. The Power system dataset is a multi-class dataset involving 37 scenarios that include 8 natural events, 28 intrusive events, and 1 no event; and the UNSW-NB15 dataset includes a combination of current normal and attack records. 300,000 random samples of legitimate and attack observations were chosen from each dataset for assessing the performance of the proposed framework. Although the VAE was only employed as a part of the intrusion detection system, its strength was demonstrated while transforming complex data into a simple form. The VAE achieved 0.921 for accuracy and 0.005 for loss on the power system dataset, and 0.998 for accuracy and 0.0001 for loss on the UNSW-NB15 dataset.

An AE-based solution was proposed in [35] to detect various cyber attacks in the context of industrial control networks. There exist many kinds of cyber attacks when control networks are connected to the internet. The research problem generally reflects the CPS characteristics of attack sophistication. And it was translated to a classification problem in the ML domain. Hence, a 7-layer AE consisted of an input layer, four hidden layers, and an output layer. The input layer had 41 units corresponding to the feature space's dimension, and the output layer had five units corresponding to the five types of network traffic. In particular, the last hidden layer was a softmax layer to provide the stability of the model. The AE was trained using the NSL-KDD dataset [33]. As an early study, the proposed AE suffered low performance in detecting small classes like probe attack and remote attack. The stacked AE achieved 0.978 for accuracy over the five categories. The model achieved an F1 score of 0.9683.

An AE-based model was proposed in [24] to detect cyber attacks in the context of smart grids. One big challenge is a large number of control parameters. The smart power networks represent the CPS characteristics of physical process monitoring and legacy technology. The smart grid's essential controller is based on state estimation, so the lower and upper bounds of each state variable need to be predicted as accurately as possible. Hence, this research problem was translated into a regression problem in the ML domain. A stacked AE (SAE) was proposed to process the smart grid data. The SAE consisted of an input layer, three vanilla AEs, and a logistic regressor as the output layer. The SAE was trained with simulated data representing IEEE 9-bus, 14-bus, 30-bus, and 118-bus systems. Overall, the SAE in this study achieved excellent results in predicting the electric load forecast. The mean absolute percentage error (MAPE) was used to evaluate the SAE's accuracy. And the SAE achieved a MAPE of 3.51% on an annual prediction and outperformed the baseline models like SVM and BP. Despite the SAE model's simplicity, the empirical studies showed its applicability and consistency in performing load forecasts.

Another AE-based model was proposed in [36] to detect Phasor measurement unit data manipulation attacks (PDMAs) in smart grids. PDMAs are challenging to be detected because of the similarity between PDMAs and man-in-the-middle attacks with infiltration of communication networks. This problem represents the CPS characteristics of physical process monitoring and attack sophistication. The main idea was to detect anomalies based on the normal operation patterns from the data collected from the PMUs with PDMA-free measurements in a distributed manner. Hence, the research

problem was translated to a regression problem in the ML domain. A deep AE (DAE) was constructed by stacking four RBM models. Training the deep AE required multiple stages by fine-tuning the intermediate RBMs. The input layer took 108 numerical features, and the output layer is a regressor. The dataset was collected from a simulated IEEE 9-bus system and had 250 000 records. The deep AE was trained by using 200 000 benign records, 20 000 records were reserved as the validation dataset, and the testing dataset consisted of 30 000 samples with half from attack records. The studies showed that the deep AE outperformed the baseline models like OCSVM, C4.5, MLP, SVM, and kNN. The DAE achieved 0.941 for accuracy, 0.996 for precision, 0.886 for recall, and 0.9038 for F1 score. Despite the success of using the deep AE in this study, it is challenging to obtain benign data from real-world power networks, and new methods may need to be explored.

An AE-based model was proposed in [37] to detect attacks against physical measurements in the context of smart grids. This problem represents the CPS characteristics of physical process monitoring and legacy technology. It is challenging to derive useful features for intrusion detection in a noisy environment in a real-world factory. Hence, the research problem was translated into a classification problem in the ML domain. A stacked denoising AE (SDAE) was proposed to learn the advanced features from the input data. And the learned features were fed to an ELM for classification. Simulated data from gas turbines were collected and used to train the model. The proposed model achieved excellent results with an FPR of 0.000006, which was significantly below the required FPR of 0.01. The SDAE was used as a part of the IDS model but demonstrated its strength in extracting useful features to represent physical measurements from a noisy environment.

An AE-based model was proposed in [28] to detect cyber attacks in the context of the industrial control systems. This problem represents the CPS characteristics of physical process monitoring, attack sophistication, and legacy technology. Hence, the problem was translated into a classification problem in the ML domain. An AE was proposed to extract features for a 1D CNN classifier. The AE consisted of five layers, including an input layer, a corruption layer applying Gaussian noise to the input, a fully connected layer with an activation function, an encoder layer, a decoding layer as the output layer to generate the extracted feature. The SWaT dataset was used to train the model. In particular, the training time for the AE was less than half-second, which was significantly faster than the 1D CNN model. The AE achieved 0.890 for precision, 0.827 for recall, and 0.844 for F1 score. In summary, the AE model was validated as a powerful and efficient method to extract useful features.

An LSTM autoencoder architecture was proposed to detect cyber attacks in the context of the autonomous vehicles (AVs) [38]. AVs are linked together by using communication technologies, and thus are vulnerable to network attacks, such as Denial of Service, replay ans spoofing attacks. Such attacks can be inferred from network traffic. Authors designed an LSTM autoencoder to detect these cyber attacks. Statistical features from network traffic were extract to represent the activities of AVs. The designed neural network architecture was consisted of two types layers, LSTM and fully connected layer. A number of LSTM layers were used to encode the representation of the transformed likelihood stream. Then the reconstructed output was produced by the fully connected layer. Two datasets, i.e., Car Hacking dataset and UNSW-NB15 were used to evaluate the proposed scheme. In particular, the proposed LSTM based autoencoder achieved 0.99 for precision, 1.0 for recall, and 0.99 for F1 score in the Car Hacking dataset. While on UNSW-NB15 dataset, the proposed scheme achieved 0.1 for precision, 0.97 for recall, and 0.98 for F1 score. In a word, this work can successfully detect multiple types of attack vectors.

Remark 1: Research works employing AE-based architecture were summarized in Table I. Most of them focused on smart grids or power network systems. Due to the difficulties in smart grids' control systems, most AE models were used to learn the useful features of an intrusion detection system or predict the electric load as an indicator of cyber attacks. Moreover, the AE models were relatively small in size, so that they could be trained in a short amount of time.

B. Cyber Recognition with Deep Learning Methods

1) Cybersecurity Pattern Recognition with Deep Neural Networks (DNNs): A DNN-based model was proposed in [39] to learn the communication patterns between electronics control units (ECUs) in the context of in-vehicular network security. The security of communication messages among ECUs is vital because a group of ECUs can control and monitor a vehicle's status during a maneuver. It is challenging to ensure cybersecurity because most communications between ECUs are through the controller area network protocol, which has no support for authentication or integrity check. Specifically, fake packets injected into the open communication channel through the controller area network protocol pose severe cybersecurity risks. Detecting the fabricated or modified packets in the vehicular setup needs to meet the requirements of physical process monitoring and legacy technologies. This intrusion detection problem was translated into a binary classification problem in the ML domain. That is, statistical features were extracted from high-dimensional CAN packet data through a dimension reduction process to represent the normal and attack packets. A 5-layered DNN model was constructed based on a standard DBN model by adding a binary classification layer as the final output layer. The DBN's coefficient weights were determined through an unsupervised pre-training process, but the final DNN model was trained with a bottom-up supervised manner. During each simulation round, a total of 200 000 packets were generated by the Open Car Test-bed and Network Experiments (OCTANE) generator. A 70:30 split was made to divide training and testing sets. Many experiments were conducted by varying the layers of the DNN model from 5 to 11 to investigate the trade-offs between performance and efficiency. The empirical results demonstrated the effectiveness of the proposed DNN model while comparing it with ANN and SVM. The best performance was achieved as 0.978 for

TABLE I
RESEARCH WORKS EMPLOYING AUTOENCODERS (AES)

Reference	Scenario	Attack Targets	Cyber Attacks	AE Architecture	Dataset	Performance
[26]	Smart power networks	Sensor nodes and Communication links	Inference attacks against privacy data	VAE with 6 layers	the power system dataset and the UNSW-NB15 dataset	0.921 for accuracy and 0.005 for loss on the power system dataset, and 0.998 for accuracy and 0.0001 for loss on the UNSW-NB15 dataset
[35]	Industrial control networks	Controllers and Communication links	DoS attack, probe attack, remote attack, and unauthorized access attacks	AE with 7 layers	the NSL-KDD dataset	0.978 for accuracy and 0.9683 for F1 score
[24]	Smart grids	Controllers	Replay attacks	Stacked AE of 5 layers	Simulated IEEE 9-bus, 14-bus, 30-bus, and 118-bus systems in Matlab	3.51% for MAPE
[36]	Smart grids	Controllers and Communication links	Phasor measurement unit data manipulation attacks	Deep AE with 4 stacked RBM models	Simulated IEEE 9-bus with 250 000 records	0.941 for accuracy, 0.996 for precision, 0.886 for recall, and 0.9038 for F1 score
[37]	Smart grids	Sensor nodes, Controllers and Actuator nodes	Unspecified cyber attacks	Stacked denoising AE	Simulated data from gas turbines with 7 804 600 data samples	0.000006 for FPR
[28]	Water treatment plant	Sensor nodes, Controllers, and Actuator nodes	False data injection attacks	AE with 5 layers	the SWaT dataset	0.890 for precision, 0.827 for recall, and 0.844 for F1 score
[38]	Internet of Vehicles	Sensor nodes, Controllers, Actuator nodes, and Communication links	fuzzy attack, DoS attack, data spoofing attack, exploits attack, and reconnaissance attack	LSTM-based AE	Car hacking dataset and UNSW-NB15 dataset	0.99 accuracy on CAN dataset, 0.96 accuracy on UNSW-NB15 dataset

accuracy, 0.016 for false positive rate, and 0.028 for false negative rate. Given the detection ratio of over 99%, the proposed DNN model showed good potentials to detect fake packets on vehicular networks despite that the DNN models' efficiency with more than five layers needed to be improved to meet the real-time requirements.

Another DNN-based model was proposed in [40] to learn the network traffic patterns in the electric power grid context. The cybersecurity of an electric power grid largely depends on state estimation underpinning critical control processes for the grid. It is challenging to detect false data injection attacks against the state estimation because a skilled cyber attacker may disguise the injected data stealthily with the inside knowledge of system topology. Such successful attacks may blackout an entire region due to the falsely impacted state estimation because the injected data value is progressively added to the legitimate signal and the Gaussian noise values. Detecting the injected data fed to the state estimation model needs to meet the requirements of physical process monitoring and closed control loops in a power grid. Hence, this intrusion detection problem was translated into a binary classification problem with the objective function of simultaneously minimizing the number of false positives and false negatives. A series of measurement vectors were created for a specific time slot so that a compromised vector contains any injected component corresponding to a false data inject attack. Four variations of DNN models were constructed with different settings — 1 or 3 hidden layers, 100 or 150 neurons per hidden layer, whether to use L1 regularization. The DNN models were trained with the standard stochastic gradient descent approach using the back-propagation method. The activation function was tanh. The most accurate DNN model was also the most complex among all the four DNN models. In terms of accuracy, the DNN model of 3 hidden layers with 150 neurons on each layer without L1 regularization outperformed generalized linear models, gradient boosting machines, a distributed random forest classifier, and the other three DNN models. The best performance was 0.9802 for precision, 0.9895 for recall, 0.9852 for F1 score, and a low false alarm rate of 0.1840. The proposed DNN model demonstrated the effectiveness of a simulated IEEE 14-bus power grid without testing realistic datasets generated by Real-time Digital Simulation (RTDS) and physical testbeds.

A DNN-based model was proposed in [41], [42] to detect the anomalies in the context of secure water treatment (SWaT). The SWaT system represents many challenges of securing CPSs, including physical process monitoring, closed control loops, attack sophistication, and legacy technologies. A divide-and-conquer strategy was employed in this scenario by separating different sets of sensors and actuators into groups according to their functionalities. Thereafter, the high dimension and complexity of detecting anomalies were mitigated. Data points with a low probability were regarded as outliers because common processes in the system usually generated the normal data points. The proposed DNN model included an LSTM layer and 100 intermediate layers. The LSTM layer predicted the actuator's position based on its historical positions. Each hidden layer was fully connected to an output layer with a bi-linear function. The cost function was defined by the cross-entropy of the real probability distribution and the predicted probability distribution. The dataset was the log entries collected from 51 sensors and actuators of a testbed for 11 days. According to [41], [42], the DNN detected 13 of the total 36 scenarios, and the DNN model achieved 0.98295 for precision, 0.67847 for recall, and 0.80281 for F1 score. However, it took two weeks to train the DNN model and 8 hours to complete testing the data. Due to its inefficiency, DNN has limited use for real-world

applications.

A DNN-based model was proposed in [27] to detect covert message transmission in the context of a chemical process plant. The covert messages containing critical control information exfiltrated from the actuators can be used to detect anomaly operations on hardware devices. Because the covert channel was established without any modification to the system, conducting data analytics on the covert channel had to involve CPS characteristics, including physical process monitoring and closed control loops. It is challenging for simplistic solutions to detect the messages transmitted via a covert channel when analog emission from the physical instrument was used to disguise the existence of the messages. Hence, the problem of detecting covert messages was translated to a clustering problem. A 10-layers DNN received the inputs from the digitized audio samples and produced binary outputs to indicate whether there was a covert message transmission or not. The ten layers included two dropout layers, four linear layers, three ReLU activation functions, and a tanh activation function. The DNN model was trained using Adam optimizer to maintain the learning quality. The dataset consisted of 9 minutes of audio recordings recorded by a Hardware-In-The-Loop (HITL) simulator. Due to the nature of analog signals, the performance of the DNN model was measured by an accuracy of 0.95 on testing data in online operation. The successful application of the DNN model opened many opportunities to develop and deploy real-time monitoring mechanisms over critical actuators using audio sampling and covert channels.

2) Cybersecurity Pattern Recognition with Convolutional Neural Networks (CNNs): A CNN-based model was proposed in [43] to detect cyber attacks in the context of an industrial water treatment plant. Because the SWaT dataset was used in this study, it inherited many CPS characteristics, including physical process monitoring, closed control loops, attack sophistication, and legacy technologies. An anomaly detection approach was used for improving the detection rate of the cyber attacks on the SWaT dataset. A 1D CNN model was constructed by stacking a set of 1D convolution layer, a ReLU function, and a 1D max-pooling layer before applying to flatten, dropout, and a final fully connected layer. Among all the 36 different attack scenarios, the CNN model successfully detected 31. The CNN model achieved 0.912 for precision, 0.861 for recall, and 0.886 for F1 score. When using 8 layers, the 1D CNN model reached 0.967 as the AUC value with only 15 535 parameters. This model's training time and testing time for one epoch were 88 seconds and 47 seconds, respectively. Based on the positive results of this empirical study, the proposed CNN model achieved excellent results in detecting cyber attacks on the SWaT dataset and demonstrated good potentials in real-world deployment.

A CNN-based model was proposed in [44] to detect message injection attacks in the vehicular networks. Due to the lack of security protections to the controller area network (CAN), network packets with malicious contents can be easily injected to the CAN bus resulting in gaining control of the vehicle. The detection of malicious packets on a physical vehicle needs to be conducted according to the CPS requirements. Two Raspberry Pi devices acting as a listener and an attacker were connected to an operational passenger vehicle via the OBD-II port to validate the model's effectiveness. The attacker had four attacking modes — DoS attack, fuzzy attack, drive gear spoofing attack, and engine RPM gauge spoofing attack. It is challenging to effectively and efficiently detect all of the attacks in real-time. A deep CNN model named Inception-ResNet was employed with two blocks of convolution layers, pooling layers, a fully connected layer, and a softmax layer before generating the final output. The activation function was set to the ReLU function. The whole dataset was split into four sub-datasets according to the attack scenarios. The CNN model outperformed other classifiers, including LSTM, ANN, SVM, kNN, NB, and Decision Trees. Each attack lasted for 3 to 5 seconds during the four recording sessions of forty minutes each. Thus, each dataset contained 300 injection attack instances. The number of messages in the four attacking scenarios are 3 078 250 (DoS normal), 587 521 (DoS attack), 3 347 013 (fuzzy normal), 491 847 (fuzzy attack), 2 766 522 (gear normal), 597 252 (gear spoofing), 2 290 185 (RPM normal), and 654 897 (RPM spoofing). False negative rate (FNR) and error rate (ER) were used to measure the CNN model's performance. In terms of FNR, the CNN model achieved 0.06, 0.07, 0.10, and 0.24 for gear spoofing, RPM spoofing, DoS attack, and fuzzy attack, respectively; in terms of ER, the CNN model achieved 0.03, 0.04, 0.05, and 0.18 for DoS attack, RPM spoofing, gear spoofing, and fuzzy attack, respectively. The empirical studies showed that the CNN model was a promising technique for detecting false message injection attacks to vehicular networks.

A federated deep learning scheme (DeepFed) in [45], to detect cyber threats targeting industrial CPSs. Li et al. firstly designed a CNN-GRU based intrusion detection model. Second, they took account into the federated learning scenario, and built a framework to allow multiple industrial CPSs to build an intrusion detection model together. Lastly, they applied a secure communication protocol based on Paillier cryptosystem to preserve the privacy of model parameters.

Industrial CPSs are not only targeted by traditional cyber threats, such as DoS attacks, but also by some cyber threats which are customized to industrial systems, such as response injection attacks. Furthermore, in the federation based CPS cyber threats detection framework, eavesdropping attacks on data sources and model parameters are emerging. Thus, the authors proposed DeepFed, which was based on CNN-GRU to detect such threats. The model architecture was consisted of a CNN module, a GRU module, an MLP module, and a softmax layer. To evaluate the performance of DeepFed, they ran experiments on a real-world dataset collected from a gas pipelining system, with 80% splited for training and the rest 20% for testing. The performance varied when different number of local agents (K) were in consideration. If K = 3, DeepFed could achieve an accuracy, precision, recall, F-score of 99.20%, 98.86%, 97.34%, and 98.08%. Overall, all metrics could reach over 97%. The experiments demonstrated the effectiveness of DeepFed to detect different types of cyber

threats to industrial CPSs.

3) Cybersecurity Pattern Recognition with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Models: An LSTM-based model was proposed in [46] to identify anomaly sensor behaviors in the water treatment plant context. Detecting the cyber attacks in a real plant is challenging because of the complex setups of the sensors and actuators representing many CPS characteristics, including physical process monitoring, closed control loops, attack sophistication, and legacy technologies. Specifically, the anomaly behaviors and the sensors under attack needed to be identified at the same time. The LSTM model using a cumulative sum was proposed to learn the sensors' behaviors to achieve a low false positive rate. The LSTM model consisted of three LSTM stacks with 100 hidden units each. The cumulative sum of the sequence predictions was introduced to reduce false positives because of their capabilities of indicating small deviations over time. And its loss function was a mean-square loss function. Ten attack scenarios of the SWaT dataset was used for training the LSTM model. The attack scenarios were made up of six single-stage-single-point attacks, two single-stage-multi-point attacks, one multi-stage-single-point attack, and one multi-stage-multi-point attack. It took about 24 hours to train the LSTM model, which is a big drawback. Nine out of ten attack scenarios were detected without listing detailed results for specific performance metrics. This work was an early study of applying DL models on the SWaT dataset with only limited success. Nevertheless, it showed good potentials for applying DL models for detecting cyber attacks against CPSs.

An LSTM-based model was proposed in [47] to learn the traffic pattern generated by unknown or zero-day attacks in the context of a control system for a gas pipeline. Detecting unknown or zero-day attacks is challenging because most intrusion detectors on industrial control systems are manually configured for specific protocols and systems. A hybrid strategy of combining packet contents and temporal dependencies and a fast signature-based Bloom filter were used to triage anomaly traffic based on the network packets' contents. A subsequent and slow LSTM model was dedicated to capturing the unknown or zero-day attacks based on the time-series information. The proposed LSTM model consisted of two LSTM layers and a softmax layer. Each LSTM layer had 256 fully connected neuron cells, and the output softmax layer had 613 bits to match the length of signatures used by the Bloom filter. A dataset of network logs collected from the gas pipeline was used to train the model, where there were seven types of cyber attacks with a total of 214 580 normal network packets and 60 048 attack packets. Combining a fast Bloom filter and a slow but powerful LSTM model demonstrated its success in detecting cyber attacks in a small SCADA system. Regarding the detection ratio (recall), the hybrid model outperformed the other six classifiers for six out of seven attacks, including Bloom filter, Bayesian network, SVDD, isolation forest, PCA-SVD, and Gaussian mixture model. The proposed model also achieved 0.92 for accuracy, 0.94 for precision, 0.78 for recall, and 0.85 for F1 score. This study showed the promising application of the LSTM model to detect time-series anomalies supplemented by an efficient and lean model like Bloom filter.

An RNN-based model was proposed in [48] to detect various cyber attacks in the context of smart grids. There are various cyber attacks against smart grids, such as DoS attacks, data infiltration, and so on. Detecting all the attacks on a large scale smart grid network is challenging because of the CPS characteristics in terms of attack sophistication and legacy technology. Hence, a vanilla RNN model was trained by using the truncated backpropagation through time (BPTT) algorithm. Three datasets were fed to the RNN model to cover a wide range of cyber attacks. In particular, the CICIDS2017 dataset [29] included brute force attacks, botnets, DoS attacks, web attacks, heartbleed attacks, and infiltration attacks; the Bot-IoT dataset [30] consisted of infiltration, DoS attack, and information theft; and the power system dataset [31] included data injection, remote command injection, were used to replay attacks. The CICIDS2017 dataset contains 2 830 743 records, the Bot-IoT dataset 73 360 900 records, and the power system dataset 78 404 records. Experiments were conducted on the three datasets separately, and the results showed that the RNN model outperformed the benchmark classifiers like SVM, random forest, and NB in terms of mean false positive rate — 0.00986 for the CICIDS2017 dataset, 0.01281 for the Bot-IoT dataset, and 0.03986 for the power system dataset. In terms of accuracy, the RNN model achieved 0.98941 for the CICIDS2017 dataset, 0.99912 for the Bot-IoT dataset, and 0.96882 for the power system dataset, respectively. Although the proposed vanilla RNN performed well across the dataset, its integration with the blockchain component of the DeepCoin system for detecting fraudulent transactions remains unclear but may inspire further research works to use blockchain technologies together with DL models.

An LSTM-based model was proposed in [49] to detect anomalous automatic dependent surveillance-broadcast (ADS-B) messages transmitted between airplanes and control towers. A typical ADS-B message includes an aircraft's flight information, such as flight number, speed, GPS coordinates, altitude, and many more. Due to historical design and implementation, the ADS-B system is an open messaging system without authentication or encryption. Therefore, it is challenging to defend ADS-B systems against malicious message injection attacks in the aviation industry because the ADS-B messages are transmitted very frequently at an average rate of 4.2 messages per second. Detecting the spoofed ADS-B messages represents the CPS characteristics of legacy technology. Hence, the problem was translated into an unsupervised anomaly detection problem in the ML domain. An LSTM model was proposed to capture the message sequences to detect anomalies according to estimated credibility scores. The LSTM model consisted of two vanilla RNN models stacked as an encoder-decoder paradigm. The LSTM was trained to reconstruct sequences of benign messages with minimal errors, where the reconstruction error score was calculated by using the cosine similarity. A large-scale flight tracking dataset named *Flightradar24* was used to train the LSTM model using flight data collected from 13 international airports. The LSTM achieved 1.00 as recall, 0.03

as FPR, and 0.99 as TPR. The LSTM outperformed the baseline anomaly detection methods, including HMM-GMM, one-class SVM (OCSVM), local outlier factor (LOF), isolation forest (IF), and DBSTREAM. In particular, the spoof messages through RND and ROUTE attacks were detected almost instantaneously, but the SHIFT Down and SHIFT Up attacks were detected with some delay. The LSTM model achieved excellent results in detecting the spoofing messages.

An LSTM-based model was proposed in [50] to predict multimedia data requests transmitted in the cyber-physical industrial networks. Multimedia data as the carrier for many cyber attacks are manifold, including image files, audio files, and many more. The prediction of multimedia data through a cache resource allocation system became a new proposal to defend the underlying network, representing the CPS characteristics of attack sophistication. Hence, this research problem was translated as a regression problem in the ML domain. A vanilla LSTM model was proposed to capture the spatio-temporal relations of the multimedia contents. Specifically, the LSTM consisted of 3 layers where the input layer's size was equal to the number of multimedia types, and the hidden layer's size was half of the input size. A two-year-long network traffic logs collected from a factory-intensive area in Tianjin, China, were considered the normal traffic of the dataset; all cyber attacks were manually added to the dataset. And various ratios of training data and testing data were used to train the LSTM model. The empirical studies showed that the LSTM accurately predicted the multimedia contents. The LSTM outperformed three baseline models: SVM, ANN, and RNN. The performance was measured in MAE, MRE, and RMSE, where the LSTM achieved 3.9857 for MAE, 0.0553 for MRE, and 4.6166 for RMSE. Despite the LSTM's excellent performance, predicting multimedia contents in the industrial network remains open.

4) Cybersecurity Pattern Recognition with Deep Belief Network (DBN): A DBN-based model was proposed in [51] to detect false data injection attacks in the context of energy internet. It is challenging to detect stealthy cyber attacks due to many control signals and meter reading data on the energy internet. And this represents the CPS characteristics of physical process monitoring, closed control loops, and legacy technology. In this study, the detection problem was formulated as a dual bi-level programming problem with the upper and lower bounds because the variations of the electric loads can be predicted. And the prediction problem is translated as a regression problem in the ML domain. A DBN model was trained to forecast electric load. The DBN was made of three stacked RBM models forming six layers — an input layer, four hidden layers, and a logistic regression layer as the output layer. The DBN was trained with the data collected from simulated IEEE 14- and 118-bus systems. An overall error rate was used to measure the DBN model's performance, and the DBN achieved a 2.73% error rate that was almost 3% lower than the benchmark model SVM. In this study, the DBN used only as a forecasting component of the intrusion detection model is a good example of trading off between the DL model and other programming solutions.

A DBN-based model was proposed in [52] to detect false data injection attacks in electric power networks. The detection of false data injection attacks represents the CPS characteristics of physical process monitoring, closed control loops, and legacy technology. The research problem was translated into a classification problem in the ML domain. The detection accuracy was investigated through the features automatically generated by DL models. Specifically, a DBN model was proposed as a baseline approach to compare with an RNN model and a graph neural network (GNN) model. The DBN was created by using the default settings of Tensorflow package without optimization. The data was collected from two simulated systems, IEEE 30-bus and 118-bus. The DBN achieved 0.9939 for precision and 0.98231 for recall on the IEEE 30-bus, which was almost identical to the GNN model and slightly higher than the RNN model. However, on the IEEE 118-bus, the DBN model's performance was consistently better than the RNN model and comparable to the graph neuron network model. Overall, the DBN model shows consistency, and the GNN model shows some promising future.

Remark 2: Table II lists thirteen different DL models for detecting various cyber attacks across various CPS scenarios. Among the models, four were DNN-based, two CNN-based, five RNN-based, and two DBN-based. It took longer time and more computational resources to train some models than others because of the nature of the architectures and parameter settings. In general, CNNs, DNNs, and DBNs were faster to train than RNNs and LSTMs. But the slower models like LSTMs or even an embedded LSTM layer had their merits in forecasting electric power load soon. On the contrary, DNNs, CNNs, and DBNs provided more powerful capabilities to detect anomalies. Last but not least, DNNs earned their popularity partly because of their simple setup and various hyperparameter fine-tuning techniques.

IV. Challenges and Future Opportunities

Six potential areas are depicted in Fig. 2 where challenges and new research directions may arise. These six areas correspond to the six steps of our research methodology, as shown in Fig. 1. Our research methodology helps provide an overview of the research literature and extract important elements for comparative analysis. The analysis results underpin research challenges and opportunities in the near future.

A. New CPS Cybersecurity Scenarios

The papers studied the communication networks in CPS scenarios. The majority of the surveyed publications investigated the CPS scenarios in water treatment plants or smart grids, which accounted for thirteen out of twenty surveyed papers. Among the remaining six papers, two studied vehicle networks, one on generic industrial control networks, one on chemical process plant, one on aviation communication networks, and one on gas pipeline controllers. The imbalanced topics suggest that CPS scenarios were underrepresented.

Applying DL to the 21st-century manufacturing industry is an emerging topic. DL has been used to detect flaws and

TABLE II
CYBERSECURITY PATTERN RECOGNITION WITH DEEP LEARNING

Reference	CPS Scenarios	Attack Targets	Cyber Attacks	DNN Model	Dataset	Performance
[39]	Vehicular network	Sensor nodes and Controllers	False data injection attack	DNN with 5 layers	Simulated data using OCTANE	0.978 for accuracy, 0.016 for FPR, and 0.028 for FNR
[40]	Power grid networks	Sensor nodes and Controllers	False data injection attacks	DNN with 5 layers	Simulated IEEE 14-bus system in Matlab	0.9802 for precision, 0.9895 for recall, 0.9852 for F1 score, and 0.1840 for FPR
[41] & [42]	Water treatment plant	Sensor nodes, Controllers, Actuator nodes, and Communication links	DoS attacks, false data injection attacks, replay attacks	DNN with 102 layers	SWaT dataset	0.98295 for precision, 0.67847 for recall, and 0.80281 for F1 score
[27]	Chemical process plant	Communication links	Covert message transmission	DNN with 10 layers	9 minutes audio recorded by HITL	0.95 for accuracy
[43]	Water treatment plant	Sensor nodes, Controllers, and Actuator nodes	DoS attacks, false data injection attacks, replay attacks	CNN with 6 layers	SWaT dataset	0.912 for precision, 0.861 for recall, and 0.886 for F1 score
[44]	Vehicular network	Sensor nodes, Controllers, Actuator nodes, and Communication links	False data injection attacks, DoS attacks, fuzzy attacks	CNN with 7 layers	Real vehicle data packets collected by Raspberry Pi	0.12 for FNR and 0.075 for error rate
[45]	Industrial CPS	Controllers and Communication links	Reconnaissance attack, Response injection attack, Command injection attack and DoS attack	CNN-GRU	real data collected from a gas pipelining system	0.992 for accuracy, 0.9885 for precision, 0.9747 for recall, 0.9814 for f-score
[46]	Water treatment plant	Sensor nodes	Anomaly sensor behaviors	LSTM with 6 layers	SWaT dataset	No specific performance metrics provided
[47]	Gas pipeline controller	Controllers and Communication links	DoS attacks, web attacks, infiltration, port scanning, bots	LSTM with 3 layers	CICIDS2017 dataset	0.92 for accuracy, 0.94 for precision, 0.78 for recall, and 0.85 for F1 score
[48]	Smart grids	Controllers and Communication links	DoS attacks, web attacks, infiltration, port scanning, bots	RNN with 1 layer	CICIDS2017 dataset, Bot-IoT dataset and power system dataset	0.0208 for FPR and 0.986 for accuracy
[49]	Aviation traffic network	Communication links	False data injection attacks	LSTM with 2 layers	Flightradar24 dataset	1.0 for recall, 0.03 for FPR, and 0.99 for TPR
[50]	Industrial factory network	Communication links	Multimedia carrying exploits	LSTM with 3 layers	Network traffic collected from factories	3.9857 for MAE, 0.0553 for MRE, and 4.6166 for RMSE
[51]	Smart grids	Sensor nodes and Controllers	False data injection attacks	DBN with 6 layers	Simulated IEEE 14-bus and 118-bus systems in Matlab	0.0273 for error rate
[52]	Smart grids	Sensor nodes and Controllers	False data injection attacks	DGN with 3 layers	Simulated IEEE 30-bus and 118-bus systems in Matlab	0.9939 for precision and 0.98231 for recall

defects during the manufacturing process of complicated items such as semiconductor [53], and 3D printing [54]. They were excluded because no cybersecurity issues were covered. A hybrid model was proposed in [53] to detect wafer defect patterns during the semiconductor fabrication process, and an AE-based model was proposed in [54] to detect defects in the 3D printed objects. These two works primarily considered the causes of defects as a production issue or a modeling issue without the presence of cyber attacks. However, cyber attacks and threats may exist in the manufacturer's network or in the cloud server where the design models are stored. Moreover, blockchain technologies have been studied in two of the surveyed papers [26], [48] in smart grids for providing additional cybersecurity and privacy guarantee. We anticipate that blockchain will be studied much more widely together with CPS. And the diversity and development of CPS scenarios will require intensive studies on cybersecurity.

B. Identification of New Cyber Threats

Almost all the surveyed papers studied false data injection attacks. The detection of stealthy false data injection attacks is challenging because of the large amount of noise produced in the CPS and the lack of cybersecurity mechanisms to authenticate devices and messages transmitted across the network. There are a few types of false injection attacks, depending on the attacker's information and goals. For example, no advanced information is required for launching DoS attacks; only recorded packets are required for replay attacks; scanning tools are required for probing attacks; automated tools are required for fuzzy attacks. The effective and efficient analysis of the network traffic is crucial for defending against these attacks [55]–[57].

Furthermore, CPS cybersecurity is much broader than cyber attacks against CPS systems. Detecting the cyber attacks that originated in the cyber space and penetrated the physical domain remains a great challenge. For example, the Stuxnet attack [58] utilized many cyber attack elements like system vulnerabilities, network sniffing, and many more, to exploit the kinetic system. Such attacks may cause significant damage

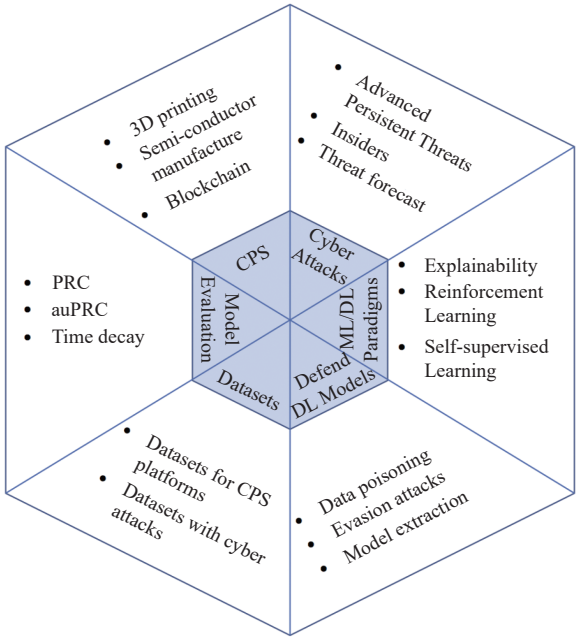

Fig. 2. Research directions for DL driven CPS cybersecurity.

and loss to the physical infrastructure. There are proposals to mitigate these sophisticated attacks by monitoring the CPS system's cyber component and the physics component. However, it remains an open problem to correlate the inter-dependent monitoring mechanism in a distributed and real-time manner.

Considering the characteristics of CPS cybersecurity, it needs to explore three topics — advanced persistent threats, insider attacks, and cyber incident forecasts. Advanced persistent threats (APTs) quickly appear due to the availability of attack tools and techniques and are difficult to defend against because of their frequently updated signatures and evolving behaviors based on reconnaissance results [59]. Insiders are very challenging to defend against because their specialized knowledge can be utilized to circumvent security checkpoints deployed on the perimeters of the critical infrastructure [60]. Despite its potential high false positives, cyber incident forecasting is valuable to prepare early for fortifying strong defense and improve the overall cyber resilience [61]. Due to available big data, many aspects of these three topics can be significantly advanced and improved by adopting ML and DL models. In particular, new insights and knowledge can be obtained via visualization, and deep analysis. We expect that emerging cyber attacks will precede the defense mechanisms, but the risk can be mitigated through the data-driven approach.

C. Adopting New ML/DL Paradigms

All the surveyed papers followed traditional ML paradigms, including supervised and unsupervised learning. Apart from four papers that examined regression problems and three papers on clustering problems, the remaining papers studied classification problems. The dominating use of the supervised learning paradigm reflected the importance of the well-labeled data. In particular, network packets were labeled as normal or attack traffic, and the attack types often were differentiated. Such reliance on labeled data restricted the wide adoption of ML or DL methods. We advocate for researchers and practitioners to try new ML/DL paradigms. These new paradigms include reinforcement learning and self-supervised learning, together with improving the model's explainability. Instead of relying on learning the records, reinforcement learning focuses on experience [62]. It is very suitable in isolated environments where many CPSs are currently deployed. For example, a reinforcement learning model was constructed in [63] to predict the normal driving behaviors. Deep reinforcement learning was applied to build human-level intelligence [64]. Deep reinforcement learning may have huge potentials in fighting against cyber attacks on CPSs. Self-supervised learning is a relatively new paradigm to generate data labels automatically. Self-supervised learning is a special kind of supervised learning without the need for the manual labeling process. In theory, self-supervised learning is ideal for changing environments where zero-day attacks and unknown attacks exist because the relationship between input data is deeply analyzed. A popular and successful model is BERT [65] for deriving deep bidirectional representations from unlabeled data. Because of BERT's huge success in NLP and many other domains, we anticipate that self-supervised learning will prosper in the CPS domain because DL models generally suffer from poor explainability [66]. We know how well the DL model works but do not know why the model works. Explainability is important for many CPS problems when we need to improve the system. Fortunately, tools like LEMNA [67] are available to explain DL-based cybersecurity applications. LEMNA can identify the critical features for a trained DL model like RNN and LSTM. We are optimistic to predict that the DL models will be more and more explainable when new tools and techniques are invented and used.

D. Defending the Trained DL Models

No surveyed works are considered to defend the trained DL models against various attacks. However, we would like to highlight the importance of defending the trained DL models because of the computational expenses to train the DL models, the important roles of the trained models, and potential dangers if the DL models are compromised. Due to the hunger for training samples, the DL models are sometimes trained with data from untrustworthy sources. Thus, adversarial attacks are prevalent because of linear behavior in high-dimensional space [68]. For example, Android HIV was proposed in [69] to automatically generate adversarial Android malware that the existing detectors failed to detect.

Adversarial attacks are categorized into two types, including evasion attacks and poisoning attacks [70]. Evasion attacks aim to evade a trained classifier by altering input data, and poisoning attacks aim to affect the trained classifier by injecting poisoning samples to the training dataset. These attacks differ by their time of occurrence. Poisoning attacks usually occur during the training phase, but evasion attacks occur during the test phase. Much research efforts have been devoted to fighting against data poisoning attacks, including robust optimization, adding some adversarial samples to

training data, blind-spot removal, examining decision boundaries, and many more. There are many strategies to counter poisoning attacks, such as outlier removal, data sensitization, frequent model retraining, classifier ensembles, randomizing the classifier's output, and many more. Last but not least, model stealing attacks were proven in [71] as a big threat to the trained model. The attackers can obtain sufficient information to mimic an ML or DL model by launching many queries and aggregating the results. The extracted information can be used to build a mimicking model for the attacker to find possible evasion attacks. We strongly advocate that cyber defense should be conducted as soon as possible due to the unawareness of adversarial attacks in the CPS scenarios.

E. Enriching CPS Cybersecurity Datasets

Among the surveyed papers, datasets collected in the field dominated the simulation with a 14:6 ratio. Simulated data were investigated in the two CPS scenarios — smart grids and vehicular networks. In smart grids, Matlab was the only choice for simulating electric load in five papers; and the OCTANE simulator was used in one paper on vehicular networks. However, there is a significant risk of solely relying on proprietary products like Matlab because the availability of such products may be discontinued unprecedently. On the other hand, field data is independent of the simulation platform and offers researchers good flexibility. Among the papers using field data, five papers chose the SWaT dataset, two papers the CICIDS2017 dataset, and the rest seven papers different datasets. The SWaT dataset dominated the field data category for a few reasons: 1) The network traffic data were continuously collected for 11 days from the control networks and from the sensors of a physical testbed, 2) the traffic with and without attack were chronologically separated for easy use, and 3) there were 36 attack scenarios against different components of the testbed. To our surprise, the NSL-KDD dataset [33], also known as KDD99-cup, was studied in one surveyed paper despite its age of 20+ years. Using such a dated dataset may cause people to draw biased conclusions because many cyber attacks were not included in the NSL-KDD dataset. Therefore, we recommend the researchers to use datasets like the UNSW-NB15 dataset [32] where recent cyber attacks were present.

Moreover, new datasets will always be valuable and appreciated. Ideally, the new datasets are open-sourced field data collected from physical testbeds. Several CPS testbeds are proposed to facilitate detecting cyber attacks, such as [72]–[74]. The recent trend of increasing interest in building CPS testbeds may benefit researchers to collect high-quality attack and defense data. The new datasets should be large enough to exploit DL models' power, and both new and old cyber attacks should be included because cyber attacks evolve quickly. If labeling data is challenging, then chronologically separating the attacks from the normal traffic is a feasible idea. Artificially blending the data entries representing attacks into a set of normal traffic records should be avoided because the simple data augmentation method does not consider feasibility, attacking sequences, and possible changes of correlations. To benefit the advancement of research and knowledge, we strongly encourage more and more high-quality datasets to be made available to the community.

F. Improving the Model Evaluation

Standard performance metrics were used in most of the surveyed papers. False positives were investigated, along with accuracy and error rates. It is proven in [34] that it is significantly more difficult to detect the rarely occurred attacks than the common ones derived by the Bayesian laws. In real-world CPSs, cyber attacks may rarely occur, so the DL models trained in the lab settings may be invalid. Since most papers did not investigate the impact of the imbalanced data between normal and attack traffic, the empirical results may be substantially biased or inflated. Cross comparisons in [75] showed that the precision-recall curve (PRC) and the area under precision-recall curve (auPRC) were more resilient to imbalanced data than ROC and auROC. Therefore, new studies should consider reporting PRC and auPRC.

Furthermore, time decay should be considered in future studies because each trained ML or DL model's performance will inevitably degrade over time. When the cyber attacks rapidly evolve, the models trained with old data will struggle with detecting new attacks. A time decay metric was proposed in [75] to evaluate a trained model's performance loss. By studying the time decay, we will be able to decide when the model needs to be retrained. We strongly hope to see future work similar to [75] in the context of CPS and cyber attacks. Once the in-depth knowledge is developed and gained, we may expect to mitigate the risk of CPS cyber attacks.

V. CONCLUSION

This survey provides a current view of detecting cyber attacks in the CPSs. A six-step DL driven methodology is proposed to summarize and analyze the twenty recently published papers in this survey. Specifically, a panoramic view is obtained through inspecting the CPS scenarios, identifying cybersecurity problems, translating the research problem to the ML/DL domain, constructing the DL model, preparing datasets, and finally evaluating the model. Cyber attacks persist as an ongoing and prominent threat to the security and safety of the CPSs. The reviewed works show great potential to exploit CPS cyber data through DL models because of their promising performances. The excellent performance is achieved partly because of several high-quality datasets that are readily available for public use. In addition to following the success of current research, we also identified promising research topics, including integration with blockchain, detection of advanced persistent threats, adopting new ML and DL paradigms, prevention of adversarial and model extraction attacks, enriching datasets, and applications of additional performance metrics. We are optimistic and confident that the research in this field will flourish.

REFERENCES

[1] J. Giraldo, E. Sarkar, A. A. Cardenas, M. Maniatakos, and M. Kantarcioglu, "Security and privacy in cyberphysical systems: A survey of surveys," *IEEE Design & Test*, vol. 34, no. 4, pp. 7–17, 2017.

[2] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques

for cyber-physical systems," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–29, 2014.

[3] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Automatic control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[4] X. Ge, Q.-L. Han, M. Zhong, and Z.-M. Zhang, "Distributed Krein space-based attack detection over sensor networks under deception attacks," *Automatica*, vol. 109, Article 108557, Nov. 2019.

[5] W. He, Z. Mo, Q.-L. Han, and F. Qian, "Secure impulsive synchronization in Lipschitz-type multi-agent systems subject to deception attacks," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 5, pp. 1326–1334, 2020.

[6] X. Yang, L. Shu, J. Chen, M. A. Ferrag, J. Wu, E. Nurellari, and K. Huang, "A survey on smart agriculture: Development modes, technologies, and security and privacy challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 2, pp. 273–302, 2020.

[7] X.-M. Zhang, Q.-L. Han, X. Ge, and L. Ding, "Resilient control design based on a sampled-data model for a class of networked control systems under denial-of-service attacks," *IEEE Trans. Cybernetics*, vol. 50, no. 8, pp. 3616–3626, 2020.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for emg-based human-machine interaction: A review," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 3, pp. 512–533, 2021.

[10] G. Lin, S. Wen, Q.-L. Han, J. Zhang, and Y. Xiang, "Software vulnerability detection using deep neural networks: A survey," *Proc. the IEEE*, vol. 108, no. 10, pp. 1825–1848, 2020.

[11] S. Liu, G. Lin, Q.-L. Han, S. Wen, J. Zhang, and Y. Xiang, "Deepbalance: Deep-learning and fuzzy oversampling for vulnerability detection," *IEEE Trans. Fuzzy Systems*, vol. 28, no. 7, pp. 1329–1343, 2020.

[12] R. Coulter, Q.-L. Han, L. Pan, J. Zhang, and Y. Xiang, "Code analysis for intelligent cyber systems: A datadriven approach," *Information Sciences*, vol. 524, pp. 46–58, 2020.

[13] J. Qiu, J. Zhang, L. Pan, W. Luo, S. Nepal, and Y. Xiang, "A survey of android malware detection with deep neural models," *ACM Computing Survey*, vol. 53, no. 6, pp. 126:1–126:36, 2021.

[14] M. Wang, T. Zhu, T. Zhang, J. Zhang, S. Yu, and W. Zhou, "Security and privacy in 6G networks: New areas and new challenges," *Digital Communications and Networks*, vol. 6, no. 3, pp. 281–291, 2020.

[15] Y. Miao, C. Chen, L. Pan, Q.-L. Han, J. Zhang, and Y. Xiang, "Machine learning based cyber attacks targeting on controlled information: A survey," *ACM Computing Surveys*, vol. 54, no. 7, Article No. 139, pp. 1–36, 2022.

[16] J. Shen, T. Zhou, D. He, Y. Zhang, X. Sun, and Y. Xiang, "Block design-based key agreement for group data sharing in cloud computing," *IEEE Trans. Dependable and Secure Computing*, vol. 16, no. 6, pp. 996–1010, 2019.

[17] Z. Liu, B. Li, Y. Huang, J. Li, Y. Xiang, and W. Pedrycz, "Newmcos: towards a practical multi-cloud oblivious storage scheme," *IEEE Trans. Knowledge and Data Engineering*, vol. 32, no. 4, pp. 714–727, 2019.

[18] C. S. Wickramasinghe, D. L. Marino, K. Amarasinghe, and M. Manic, "Generalization of deep learning for cyber-physical system security: A survey," in *Proc. the 44th Annual Conf. of the IEEE Industrial Electronics Society*, 2018, pp. 745–751.

[19] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security-A survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802–1831, 2017.

[20] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[21] Y. Ashibani and Q. H. Mahmoud, "Cyber physical systems security: Analysis, challenges and solutions," *Computers & Security*, vol. 68, pp. 81–97, 2017.

[22] D. Ding, Q.-L. Han, X. Ge, and J. Wang, "Secure state estimation and control of cyber-physical systems: A survey," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 176–190, 2021.

[23] L. Ding, Q.-L. Han, B. Ning, and D. Yue, "Distributed resilient finite-time secondary control for heterogeneous battery energy storage systems under denial-of-service attacks," *IEEE Trans. Industrial Informatics*, vol. 16, no. 7, pp. 4909–4919, 2020.

[24] H. Wang, J. Ruan, G. Wang, B. Zhou, Y. Liu, X. Fu, and J. Peng, "Deep learning-based interval state estimation of ac smart grids against sparse cyber attacks," *IEEE Trans. Industrial Informatics*, vol. 14, no. 11, pp. 4766–4778, 2018.

[25] T. M. Mitchell, *Machine learning*. McGraw-Hill, Inc., New York, NY, USA, 1997.

[26] M. Keshk, B. Turnbull, N. Moustafa, D. Vatsalan, and K.-K. R. Choo, "A privacy-preserving framework based blockchain and deep learning for protecting smart power networks," *IEEE Trans. Industrial Informatics*, vol. 16, no. 8, 2020.

[27] P. Krishnamurthy, F. Khorrami, R. Karri, D. Paul-Pena, and H. Salehghaffari, "Process-aware covert channels using physical instrumentation in cyber-physical systems," *IEEE Trans. Information Forensics and Security*, vol. 13, no. 11, pp. 2761–2771, 2018.

[28] M. Kravchik and A. Shabtai, "Efficient cyber attacks detection in industrial control systems using lightweight neural networks," *IEEE Trans. Dependable and Secure Computing*, 2021. DOI: 10.1109/TDSC.2021.3050101

[29] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in *Proc. the 4th Int. Conf. on Information System Security Privacy*, 2018, pp. 108–116.

[30] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[31] S. Pan, T. Morris, and U. Adhikari, "Developing a hybrid intrusion detection system using data mining for power systems," *IEEE Trans. Smart Grid*, vol. 6, no. 6, pp. 3104–3113, 2015.

[32] N. Moustafa and J. Slay, "The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems," in *Proc. the 4th Int. Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 2015, pp. 25–31.

[33] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in *Proc. the 1999 IEEE Symposium on Security and Privacy*, 1999, pp. 120–132.

[34] S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection," *ACM Trans. Information and System Security*, vol. 3, no. 3, pp. 186–205, 2000.

[35] S. Potluri, N. F. Henry, and C. Diedrich, "Evaluation of hybrid deep learning techniques for ensuring security in networked control systems," in *Proc. the 22nd IEEE Int. Conf. on Emerging Technologies and Factory Automation*, 2017, pp. 1–8.

[36] J. Wang, D. Shi, Y. Li, J. Chen, H. Ding, and X. Duan, "Distributed framework for detecting pmu data manipulation attacks with deep autoencoders," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4401–4410, 2018.

[37] W. Yan, L. K. Mestha, and M. Abbaszadeh, "Attack detection for securing cyber physical systems," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8471–8481, 2019.

[38] J. Ashraf, A. D. Bakhshi, N. Moustafa, H. Khurshid, A. Javed, and A. Beheshti, "Novel deep learning-enabled lstm autoencoder architecture for discovering anomalous events from intelligent transportation systems," *IEEE Trans. Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4507–4518, 2021.

[39] M.-J. Kang and J.-W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *PloS One*, vol. 11, no. 6, pp. 1–17, 2016.

[40] M. Ashrafuzzaman, Y. Chakhchoukh, A. A. Jillepalli, P. T. Tosic, D. C. de Leon, F. T. Sheldon, and B. K. Johnson, "Detecting stealthy false data injection attacks in power grids using deep learning," in *Proc. the 14th Int. Wireless Communications & Mobile Computing Conf.*, 2018, pp. 219–225.

[41] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, and J. Sun, "Anomaly detection for a water treatment system using unsupervised machine

learning," in *Proc. the 2017 IEEE Int. Conf. on Data Mining Workshops*, 2017, pp. 1058–1065.

[42] Q. Lin, S. Adepu, S. Verwer, and A. Mathur, "Tabor: A graphical model-based approach for anomaly detection in industrial control systems," in *Proc. the 2018 Asia Conf. on Computer and Communications Security*, 2018, pp. 525–536.

[43] M. Kravchik and A. Shabtai, "Detecting cyber attacks in industrial control systems using convolutional neural networks," in *Proc. the 2018 Workshop on Cyber-Physical Systems Security and Privacy*, 2018, pp. 72–83.

[44] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Vehicular Communications*, vol. 21, Article No. 100198, 2020.

[45] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "Deepfed: Federated deep learning for intrusion detection in industrial cyber-physical systems," *IEEE Trans. Industrial Informatics*, vol. 17, no. 8, pp. 5615–5624, 2021.

[46] J. Goh, S. Adepu, M. Tan, and Z. S. Lee, "Anomaly detection in cyber physical systems using recurrent neural networks," in *Proc. the 18th IEEE Int. Symposium on High Assurance Systems Engineering*, 2017, pp. 140–145.

[47] C. Feng, T. Li, and D. Chana, "Multi-level anomaly detection in industrial control systems via package signatures and lstm networks," in *Proc. the 47th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks*, 2017, pp. 261–272.

[48] M. A. Ferrag and L. Maglaras, "Deepcoin: A novel deep learning and blockchain-based energy exchange framework for smart grids," *IEEE Trans. Engineering Management*, vol. 67, no. 4, pp. 1285–1297, 2020.

[49] E. Habler and A. Shabtai, "Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages," *Computers & Security*, vol. 78, pp. 155–173, 2018.

[50] B. Jiang, J. Yang, G. Ding, and H. Wang, "Cyber-physical security design in multimedia data cache resource allocation for industrial networks," *IEEE Trans. Industrial Informatics*, vol. 15, no. 12, pp. 6472–6480, 2019.

[51] H. Wang, J. Ruan, Z. Ma, B. Zhou, X. Fu, and G. Cao, "Deep learning aided interval state prediction for improving cyber security in energy internet," *Energy*, vol. 174, pp. 1292–1304, 2019.

[52] Y. Li and Y. Wang, "Developing graphical detection techniques for maintaining state estimation integrity against false data injection attack in integrated electric cyberphysical system," *Journal of Systems Architecture*, vol. 105, Article No. 101705, 2020.

[53] G. Tello, O. Y. Al-Jarrah, P. D. Yoo, Y. Al-Hammadi, S. Muhaidat, and U. Lee, "Deep-structured machine learning model for the recognition of mixed-defect patterns in semiconductor fabrication processes," *IEEE Trans. Semiconductor Manufacturing*, vol. 31, no. 2, pp. 315–322, 2018.

[54] Z. Shen, X. Shang, M. Zhao, X. Dong, G. Xiong, and F.- Y. Wang, "A learning-based framework for error compensation in 3D printing," *IEEE Trans. Cybernetics*, vol. 49, no. 11, pp. 4042–4050, 2019.

[55] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, and Y. Guan, "Network traffic classification using correlation information," *IEEE Trans. Parallel and Distributed systems*, vol. 24, no. 1, pp. 104–117, 2013.

[56] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu, "Robust network traffic classification," *IEEE/ACM Trans. Networking*, vol. 23, no. 4, pp. 1257–1270, 2015.

[57] R. Coulter, Q.-L. Han, L. Pan, J. Zhang, and Y. Xiang, "Data-driven cyber security in perspective-intelligent traffic analysis," *IEEE Trans. Cybernetics*, vol. 50, no. 7, pp. 3081–3093, 2020.

[58] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[59] A. Alshamrani, S. Myneni, A. Chowdhary, and D. Huang, "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1851–1877, 2019.

[60] L. Liu, O. De Vel, Q.-L. Han, J. Zhang, and Y. Xiang, "Detecting and preventing cyber insider threats: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1397–1417, 2018.

[61] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xiang, "Data-driven cybersecurity incident prediction: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1744–1772, 2019.

[62] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[63] T. Akazaki, S. Liu, Y. Yamagata, Y. Duan, and J. Hao, "Falsification of cyber-physical systems using deep reinforcement learning," in *Proc. the 2018 Int. Symposium on Formal Methods*, Springer, 2018, pp. 456–465.

[64] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv*: *1810.04805*, 2018.

[66] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proc. the 2020 Conf. on Fairness, Accountability*, *and Transparency*, 2020, pp. 648–657.

[67] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "Lemna: Explaining deep learning based security applications," in *Proc. the 2018 ACM SIGSAC Conf. on Computer and Communications Security*, 2018, pp. 364–379.

[68] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. the Int. Conf. on Learning Representations*, 2015, pp. 1–11.

[69] X. Chen, C. Li, D. Wang, S. Wen, J. Zhang, S. Nepal, Y. Xiang, and K. Ren, "Android HIV: A study of repackaging malware for evading machine-learning detection," *IEEE Trans. Information Forensics and Security*, vol. 15, pp. 987–1001, 2020.

[70] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[71] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. the 25th USENIX Security Symposium*, 2016, pp. 601–618.

[72] H.-K. Shin, W. Lee, J.-H. Yun, and H. Kim, "Implementation of programmable CPS testbed for anomaly detection," in *Proc. the 12th USENIX Workshop on Cyber Security Experimentation and Test, 2019* (*https://www.usenix.org/system/files/cset19-papershin.pdf*).

[73] S. Choi, J. Choi, J.-H. Yun, B.-G. Min, and H. Kim, "Expansion of ics testbed for security validation based on mitre attack techniques," in *Proc. the 13th USENIX Workshop on Cyber Security Experimentation and Test*, 2020 (*https://www.usenix.org/system/files/cset20-paper-choi.pdf*).

[74] M. Conti, D. Donadel, and F. Turrin, "A survey on industrial control system testbeds and datasets for security research," *arXiv preprint arXiv*: *2102.05631*, 2021.

[75] F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, "Tesseract: Eliminating experimental bias in malware classification across space and time," in *Proc. the 28th USENIX Security Symposium*, 2019, pp. 729–746.

Jun Zhang (M'12–SM'18) received the Ph.D. degree in Computer Science from the University of Wollongong, Wollongong, NSW, Australia, in 2011. He is the Co-founder and Director of the Cybersecurity Lab, Swinburne University of Technology, Australia. His research focuses on cybersecurity. He is the Chief Investigator of several projects in cybersecurity, funded by the Australian Research Council (ARC). He has published more than 100 research papers in reputed international journals and conferences, such as the IEEE CST, PIEEE, TIFS, TDSC, CSUR, CCS, PETS. Two of his papers were selected as the featured articles in IEEE TDSC and IT Professional. He has been recognised in The Australians Top Researchers special edition publication (09/2020) as the leading researcher in the field of Computer Security & Cryptography. He is leading Swinburne Cybersecurity Lab to deliver value and impact to our

community, including significant contribution to Australian P-TECH education program and penetration test service to the industry partners.

Lei Pan (M'11) received the Ph.D. degree in Computer Science from Deakin University, Geelong, VIC, Australia. He is currently a Senior Lecturer with the Strategic Centre for Cyber Security Research and Innovation, School of Information Technology, Deakin University. His research interests cover broad topics in cybersecurity and privacy. He has authored 70 research papers in refereed international journals and conferences, such as *IEEE Transactions on Information Forensics* and *Security*, *IEEE Transactions on Dependable and Security Computing*, *IEEE Transactions on Industrial Informatics*, and many more. He is a regular reviewer for prestigious journals like the *IEEE Transactions on Information Forensics and Security*, *the Internet of Things journal*. He also regularly serves as a technical panel member for international conferences like Trustcom, ATIS, NSS, AWMLC.

Qing-Long Han (M'09-SM'13-F'19) received the B.Sc. degree in Mathematics from Shandong Normal University, Jinan, China, in 1983, and the M.Sc. and Ph.D. degrees in Control Engineering from East China University of Science and Technology, Shanghai, China, in 1992 and 1997, respectively.

Professor Han is Pro Vice-Chancellor (Research Quality) and a Distinguished Professor at Swinburne University of Technology, Melbourne, Australia. He held various academic and management positions at Griffith University and Central Queensland University, Australia. His research interests include networked control systems, multi-agent systems, time-delay systems, smart grids, unmanned surface vehicles, and neural networks.

Professor Han was a Highly Cited Researcher in both Engineering and Computer Science (Clarivate Analytics, 2019-2020). He was one of Australia' s Top 5 Lifetime Achievers (Research Superstars) in Engineering and Computer Science (The Australian's Research Magazine, 2019-2020). He was the recipient of The 2021 M. A. Sargent Medal (the Highest Award of the Electrical College Board of Engineers Australia), The 2020 IEEE Systems, Man, and Cybernetics (SMC) Society Andrew P. Sage Best Transactions Paper Award, The 2020 IEEE Transactions on Industrial Informatics Outstanding Paper Award, and The 2019 IEEE SMC Society Andrew P. Sage Best Transactions Paper Award.

Professor Han is a Member of the Academia Europaea (The Academy of Europe) and a Fellow of The Institution of Engineers Australia. He has served as an AdCom Member of IEEE Industrial Electronics Society (IES), a Member of IEEE IES Fellow Committee, and Chair of IEEE IES Technical Committee on Networked Control Systems. He is Co-Editor-in-Chief of IEEE Transactions on Industrial Informatics, Co-Editor of Australian Journal of Electrical and Electronic Engineering, an Associate Editor for 12 international journals, including the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE INDUSTRIAL ELECTRONICS MAGAZINE, the IEEE/CAA JOURNAL OF AUTOMATICA SINICA, Control Engineering Practice, and Information Sciences and Computing, and a Guest Editor for 13 Special Issues.

Chao Chen (M'21) received the Ph.D. degree in Computer Science from Deakin University, Geelong, VIC, Australia, in 2017. He is currently a Senior Lecturer in Cybersecurity at College of Science and Engineering, James Cook University. From 2016 to 2018, he worked as a Data Scientist at Telstra to create customer value from huge and heterogeneous data sources using advanced analytics and big data techniques. He then worked at Swinburne University of Technology as a Research Scientist from 2018 to 2020. He is currently conducting interdisciplinary research between cybersecurity and artificial intelligence (AI), including AI for cybersecurity and security & privacy issues in AI models.

Sheng Wen (M'15) received the Ph.D. degree in Computer Science from the School of Information Technology, Deakin University, Australia, in 2015. He is currently a Senior Lecturer at Swinburne University of Technology. His focus is on modeling of virus spread, information dissemination, and defense strategies for the Internet threats. He is also interested in the techniques of identifying information sources in networks.

Yang Xiang (M'07-SM'12-F'20) received the Ph.D. degree in Computer Science from Deakin University, Geelong, VIC, Australia. He is currently a full Professor and the Dean of Digital Research & Innovation Capability Platform, Swinburne University of Technology, Australia. His research interests include cyber security, which covers network and system security, data analytics, distributed systems, and networking. He is also leading the Blockchain initiatives at Swinburne. He has published more than 300 research papers in many international journals and conferences. He is the Editor-in-Chief of the *SpringerBriefs on Cyber Security Systems and Networks*. He serves as the Associate Editor of *IEEE Transactions on Dependable and Secure Computing*, *IEEE Internet of Things Journal*, and *ACM Computing Surveys*. He served as the Associate Editor of *IEEE Transactions on Computers* and *IEEE Transactions on Parallel and Distributed Systems*. He is the Coordinator, Asia for IEEE Computer Society Technical Committee on Distributed Processing (TCDP).