IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 May 2023, accepted 1 June 2023, date of publication 9 June 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284681

RESEARCH ARTICLE

Electricity Theft Detection Using Deep Reinforcement Learning in Smart Power Grids

AHMED T. EL-TOUKHY[1,2], MAHMOUD M. BADR[3], MOHAMED M. E. A. MAHMOUD[1], (Senior Member, IEEE), GAUTAM SRIVASTAVA[4,5,6], (Senior Member, IEEE), MOSTAFA M. FOUDA[7,8], (Senior Member, IEEE), AND MAAZEN ALSABAAN[9]

[1]Department of Electrical and Computer Engineering, Tennessee Tech University, Cookeville, TN 38505, USA
[2]Department of Electrical Engineering, College of Engineering, Al-Azhar University, Cairo 11884, Egypt
[3]Department of Networks and Computer Security, College of Engineering, State University of New York (SUNY) Polytechnic Institute, Utica, NY 12201, USA
[4]Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada
[5]Research Centre for Interneural Computing, China Medical University, Taichung 404, Taiwan
[6]Department of Computer Science and Mathematics, Lebanese American University, Beirut 1102-2801, Lebanon
[7]Department of Electrical and Computer Engineering, College of Science and Engineering, Idaho State University, Pocatello, ID 83209, USA
[8]Center for Advanced Energy Studies (CAES), Idaho Falls, ID 83401, USA
[9]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Mohamed M. E. A. Mahmoud (mmahmoud@tntech.edu)

This work was supported by the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Project RSPD2023R636.

ABSTRACT In smart power grids, smart meters (SMs) are deployed at the end side of customers to report fine-grained power consumption readings periodically to the utility for energy management and load monitoring. However, electricity theft cyber-attacks can be launched by fraudulent customers through compromising their SMs to report false readings to pay less for their electricity usage. These attacks harmfully affect the power sector since they cause substantial financial loss and degrade the grid performance because the readings are used for energy management. Supervised machine learning approaches have been used in the literature to detect the attacks, but to the best of our knowledge, the use of reinforcement learning (RL) has not been investigated yet. RL can be better than the existing approaches because it can adapt more efficiently with the dynamic nature of cyber-attacks and consumption patterns due to its capability to learn by exploration and exploitation mechanisms and deciding optimal actions. In this article, a deep reinforcement learning (DRL) approach is proposed as a promising solution to the electricity theft problem. The samples of real dataset are employed as an environment and rewards are given based on detection errors made during training. In particular, the proposed approach is presented in four different scenarios. First, a global detection model is constructed using a deep Q network (DQN) and a double deep Q network (DDQN) with different architectures of deep neural networks. Second, the global detector is used to build a customized detection model for new customers to achieve high detection accuracy while preventing zero-day attacks. Third, changing the consumption pattern of the existing customers is taken into consideration in the third scenario. Fourth, the challenges of defending against newly launched cyber-attacks are addressed in the fourth scenario. Extensive experiments have been conducted, and the results demonstrate that the proposed DRL approach can boost the detection of electricity theft cyberattacks, and it can efficiently learn new consumption patterns, changes in the consumption patterns of existing customers, and newly launched cyber-attacks.

INDEX TERMS Security, electricity theft, false reading attacks, reinforcement learning, zero-day attacks, smart power grids.

The associate editor coordinating the review of this manuscript and approving it for publication was Neetesh Saxena.

I. INTRODUCTION

The smart grid (SG) is a new vision for the traditional power grid that aims to regulate and optimize grid operation,

facilitate reliable delivery of electricity, and keep track of the performance of all the system components. The architecture of SG includes different components, including electricity production stations, advanced metering infrastructure (AMI) network, system operator (SO), and transmission and distribution systems as shown in FIGURE 1 [1], [2]. The function of the AMI is to facilitate efficient bidirectional communications between the smart meters (SMs), installed at the customers' homes, and the SO [3], [4]. In contrast to the traditional monthly billing method of electricity consumption, in SG, fine-grained electricity consumption readings, e.g., every few minutes, are measured periodically by SMs and forwarded to the SO through AMI. Consequently, the SO can utilize these readings for demand and response management, load monitoring and forecasting purposes, calculating the consumption bill using a dynamic pricing approach, and managing the power generation efficiently [5], [6], [7], [8].

In traditional power grids, electricity theft can be conducted by tampering with mechanical meters physically, e.g., by line hooking. In SGs, fraudulent customers can launch cyber-attacks by hacking their SMs to manipulate the electricity consumption readings and report false data. Comparing to the traditional grids, the electricity theft is more severe in the case of SG because the attacks not only cause hefty financial losses but also may degrade the power grid's performance since the reported electricity consumption data are utilized for the grid management [9], [10], [11], [12]. In worldwide, electricity theft has a negative financial impact on both developing and developed countries. According to the World Bank's report, the annual global loss caused by illegal electricity usage is estimated at approximately $89.3 billion [13], [14], [15]. For developed countries, the loss is estimated at $6 billion, $173 million, and $100 million per year in the United States, United Kingdom, and Canada, respectively [1], [2], [16], [17]. On the other hand, electricity theft is worse in developing countries. For instance, India loses $17 billion per year while Brazil and China lose around 16% and 6% of their total electricity production, respectively [1], [15], [17].

Artificial intelligence (AI) has been incorporated in a broad range of applications in the power industry to address real-world challenges [18], [19]. AI is particularly useful in the integration of more renewable energy generators in the smart grid, as it can optimize electricity pricing and make it adaptive to fluctuations in the power generation due to unpredictable weather conditions. It can also be used to detect equipment failures to enhance the reliability of the grid, and forecast electricity demand and generation.

In recent years, machine learning (ML) has been used to eliminate the harmful consequences of electricity theft cyber-attacks [20], [21], [22]. Both supervised and unsupervised ML approaches, such as DL approaches, have been employed for electricity theft detection in SGs [3], [6], [11], [23]. However, these approaches have the following limitations. First, DL models are trained on a fixed dataset and may overfit the training data. Consequently, they learn to recognize specific patterns and features rather than generalizing to a broader range of patterns. Second, it is inefficient to adapt to changes in consumption patterns and new cyberattacks, requiring retraining the models on old and new data. This process is time-consuming and computationally extensive, especially for large datasets.

Over the last few years, AlphaGO and AlphaGO Zero, introduced by Google, paid much attention to the application of AI to solve difficult problems [24], [25]. Both AlphaGO and AlphaGO Zero demonstrated the fact that reinforcement learning (RL) is an emerging type of ML that has similar features to human learning due to its ability to adapt to the surrounding environment and learn by exploration and exploitation mechanisms [26], [27]. Also, it can model an agent and make optimal decisions regardless of the limited available knowledge about the surrounding environment. In other words, RL demonstrates outstanding decision-making ability, and it has several merits [28]. Firstly, RL seeks optimal decisions through direct interaction with the surrounding environment, similar to the way the human brain learns. Secondly, RL is adaptive and can make optimal decisions autonomously. Thirdly, compared to the traditional optimization methods, RL is flexible and can be used for real-world applications. Furthermore, DL has been integrated with RL approaches to address a wide range of complicated problems [29] such as cyber-attack detection. This integration is a research direction initiated and pioneered by Google DeepMind [30].

Consequently, RL can be a promising solution to overcome the aforementioned limitations of DL approaches since it has the capability of efficiently incrementing the learning of the model by retraining the model using new consumption patterns or newly traced attacks, without forgetting what has already been learned or requiring complete retraining of the model from scratch [28], [31]. The nature of the RL allows for a simple update to the model parameters, and adaptation for new consumption patterns and cyber-attacks.

In this paper, we investigate the use of RL to detect electricity theft cyber-attacks by considering four different scenarios (or cases). In the first scenario, an initial global detection model is investigated using a deep Q network (DQN) and a double deep Q network (DDQN) of different DL architectures. These architectures include feedforward neural network (FFNN), gated recurrent unit neural network (GRU), convolutional neural network (CNN), and hybrid architecture consisting of CNN and GRU (CNN+GRU). The model is global in the sense that it is trained on the data of a large number of consumers, and it can be used for detecting the false data of all customers. In the second scenario, a customized detection model is built for new customers by utilizing RL's ability to adapt to new data. This is accomplished by retraining the initial global model on the consumption readings of the new customer. This enables the model to adjust to the consumption patterns of the new customer to boost the detection accuracy and prevent zero-day attacks by

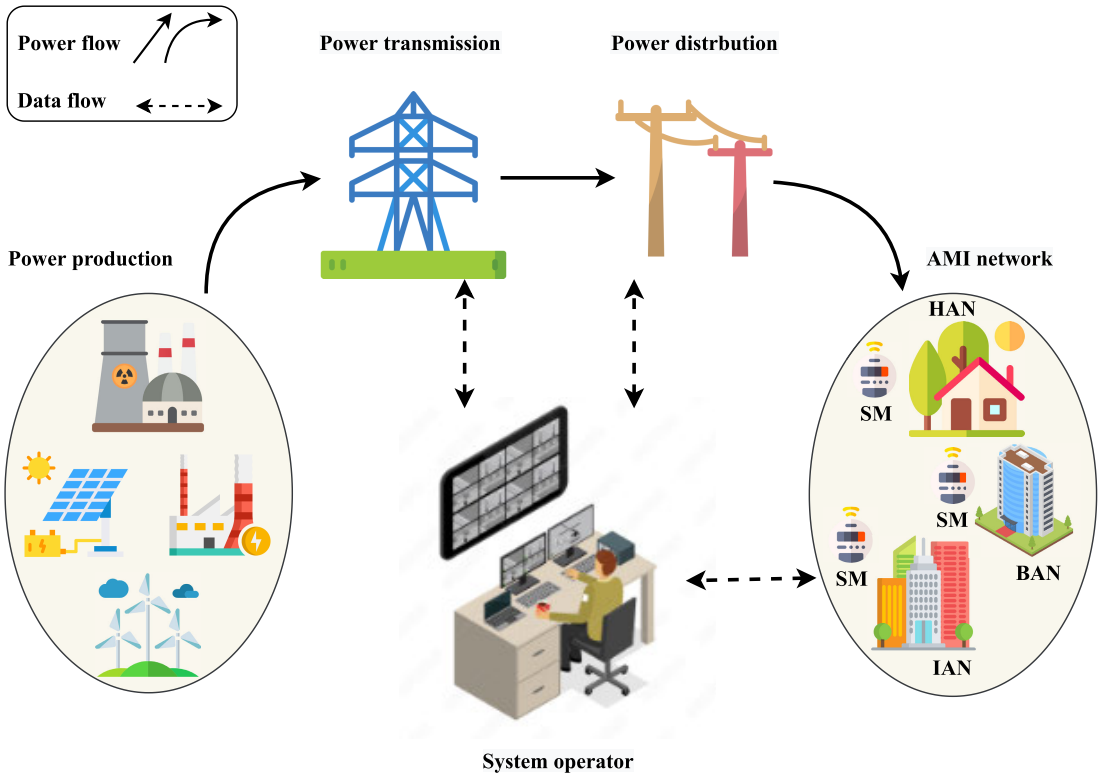


FIGURE 1. Smart grid model architecture [1].

detecting the consumers that send false data from day one. In the third scenario, we investigate the ability of our detector to learn changes in the consumption patterns that may occur due to various factors, such as changing the life-style of a customer, change in the number of occupants, and the purchase of new appliances. Finally, in the fourth scenario, we focus on the problem of newly launched cyber-attacks and investigate the ability of our detection model to learn new attacks. The experimental results demonstrate that the proposed DRL approach boosts the detection accuracy and achieves higher performance than the existing DL techniques. Also, it can adapt efficiently to new consumption patterns and cyber-attacks.

To the best of our knowledge, RL has not been used before for electricity theft cyber-attack detection, and in general, despite its attractive features, its use in cyber-security has received little attention. The key contributions of this paper are outlined as follows:

- RL-based DQN and DDQN detectors are proposed to detect electricity theft cyber-attacks.
- The ability of the RL detector to adapt to changes in the consumption patterns is investigated.
- The challenges of defending against newly launched cyber-attacks are addressed.

The remaining of this paper is structured as follows. Section II presents the related works in the literature that investigate electricity theft in SGs and our motivation for this work. Section III introduces the preliminaries used in the development of our detector. Dataset preparation for training the detectors is discussed in Section IV. Section V discusses the proposed DRL detection models. The detection performance is evaluated and analyzed in Section VI. Finally, Section VII concludes the paper.

II. RELATED WORKS AND MOTIVATION

In this section, we first review the related research studies that investigate electricity theft detection. Then, we provide a comparison between our proposed approach and the current ML approaches present in the literature. The objective of this comparison is presenting our motivation and addressing the limitations inherent in the existing ML approaches.

A. RELATED WORKS

To detect electricity theft, multiple methods have been proposed for the detection of electricity theft. These methods can be categorized as hardware-based methods, statistical and game theory methods, and data-driven methods.

1) HARDWARE-BASED METHODS

One of the methods to thwart electricity theft attacks is by using hardware tamper proof modules in the SMs to prevent hacking the meters to modify them to send false data [32]. However, these methods have several limitations. Specifically, these modules are costly and require full trust which

cannot be guaranteed in reality. That is why most of the proposals in the literature prefer data-driven methods over hardware-based methods [20], [22], [32].

2) STATISTICAL AND GAME THEORY METHODS

Different electricity theft detectors have been proposed using game theory [33], [34], [35], data mining, and statistical methods including state estimation [36], clustering, local outlier factor (LOF), and principal component analysis (PCA). For instance, the *k*-means clustering algorithm has been utilized in [37] to cluster customers by analyzing their electricity consumption readings. Then, LOF has been employed to identify the outlier candidates whose consumption readings are significantly different from the centers of their respective clusters. Furthermore, LOF is used to compute the anomaly score for each of the identified outlier candidates. Also, a PCA-based detector is proposed by Singh et al. [38]. The detector calculates an anomaly score for the data and compares to a predefined threshold value to classify it. Zheng et al. [39] proposed a novel approach for improving the detection of electricity theft. The approach combines the maximum information coefficient (MIC) data mining technique with the fast search and find of density peaks (CFSFDP) clustering technique. However, the statistical and game theory methods do not give good accuracy because they cannot capture the temporal aspect and complex patterns of the data [40].

3) DATA-DRIVEN METHODS

In order to detect false power consumption readings reported by malicious SMs, different ML-based detectors have been proposed in the literature. Some of these detectors utilize shallow ML detection algorithms [20], [21], [22] such as decision trees (DTs), logistic regression (LR), and support vector machine (SVM), while others employ DL detection algorithms [32], [41], [42]. While shallow ML detection algorithms require explicit feature extraction for providing good performance, DL detection algorithms possess the capability of automatically identifying and extracting the important features of the raw data using their deep layers. The given results in the literature confirm that DL is a promising approach that achieves better performance than shallow ML algorithms [32], [41], [42], [43], [44], [45], [46], [47].

Jokar et al. [20] have proposed customized electricity theft detectors, which are trained using real benign readings from the Irish dataset [48]. This paper introduced a series of six cyber-attacks to synthetically generate malicious samples. Also, for each customer, two SVM-based electricity theft detectors have been trained. One detector is a single-class SVM that is trained exclusively on benign data samples, while the other detector is a multi-class SVM that is trained on both benign and malicious samples. The experimental results indicate that the second detector has better performance in terms of detection rate and false alarm than the former detector.

A hybrid electricity theft detector has been proposed by Li et al. in [49]. The detector employs a hybrid architecture of CNN and random forest (CNN-RF) to identify electricity theft via analyzing the daily consumption readings. The CNN captures the features of electricity consumption readings, and the RF classifies the consumption readings samples. The results of the experiments confirmed that the hybrid model outperforms other detection models, such as Gradient-Boosted Decision Trees (GBDT), RF, LR, and SVM. Also, Hasan et al. [50] have proposed a hybrid CNN-LSTM model for electricity theft detection. The results of the model indicate a promising performance by achieving 89% classification accuracy. Another DL-based electricity theft detection model has been proposed by Zheng et al. in [32]. The detector employs CNN and MLP to detect fraudulent behaviors by analyzing the electricity consumption readings weekly. The detector has been trained on the state grid cooperation of China (SGCC) dataset [51] that includes malicious and benign samples, where the malicious samples represent 9% of the total samples. The experimental results demonstrated that the proposed detector outperforms other detectors such as LR, RF, SVM, and CNN.

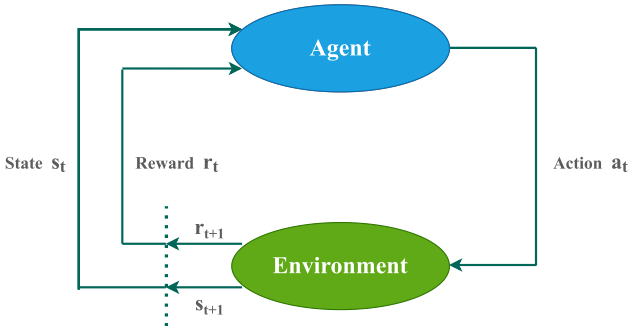


FIGURE 2. The RL model block diagram.

Most of the current ML-based methods for detecting electricity theft rely on fine-grained electricity consumption data, which can reveal sensitive information about the habits and activities of smart grid consumers. This information poses a threat to privacy and could be exploited for criminal purposes such as burglary. To address this issue, various privacy-preserving [1], [9], [52], secure communication solutions [53], [54], and secure federated learning approaches [55] have been proposed in the literature. However, our focus in this paper is solely on using RL techniques for electricity theft detection, and we do not address privacy and secure communication concerns. Nonetheless, the existing privacy-preserving and security methods can be incorporated into our proposed approach.

B. MOTIVATION

Referring to the above discussion, the existing works mainly focus on developing supervised/unsupervised ML techniques to detect electricity theft, and none of them investigated using RL for electricity theft detection. Since RL has the ability to efficiently adapt to changes in consumption patterns and new cyber-attacks, it is a promising solution for

electricity theft detection. Thus, this paper aims to investigate the use of RL for electricity theft detection in smart power grids. The core idea and working principle of the proposed approach lies in the use of DRL-DDQN model, which seeks to learn the electricity consumption patterns of the consumers and construct an optimal policy. This policy empowers the model to make well-informed decisions in the face of anomalous deviation in consumption patterns, which could be an indicator for an electricity theft attack. Consequently, the model can make the best actions to mitigate such attacks. We propose electricity theft detectors in four different architectures of neural networks while taking into consideration learning new consumption patterns and attacks in addition to addressing the zero-day cyber-attacks problem.

Specifically, this work aims to address the following limitations in the existing DL approaches in the literature.

- *Inefficient to adapt to changes in consumption patterns and attacks*: DL models are often trained on a fixed dataset and may not be efficient to adapt to changes in consumption patterns and attacks, requiring retraining on a new dataset. This process is time-consuming and computationally extensive, especially for large datasets. In contrast, RL can efficiently adapt to these changes by learning from past experiences and optimizing a policy that considers the new patterns. As a result, RL can lead to better performance and more efficient use of computational resources.
- *Lack of exploration*: DL models do not typically have a built-in exploration mechanism, and they rely solely on the training data to minimize a loss function. This makes the DL model less flexible and less capable of handling new situations. In contrast, RL can explore the space of possible actions by optimizing a policy that considers maximizing the expected rewards for each action. The exploration mechanism enables RL to retrain the agent using newly collected data that may contain information about previously unseen patterns. This exploration can help the agent to improve its policy and better handle the changes in the environment that were not encountered during the initial training. As a result, RL can provide improved performance.
- *Difficulty of generalization*: DL models may overfit the training data and thus, they learn to recognize specific patterns and features rather than generalizing to learn a broader range of patterns. As a result, they may not perform well on new or unseen data that was not present in the training set. This is a significant limitation in many classification tasks, especially in changing environments where the distribution of data may shift over time. In contrast, RL can learn policies that generalize to new environments based on past experiences. This is because the agent learns to optimize a policy based on maximizing the expected reward. Therefore, RL can lead to more reliable and accurate classification, even in new environments.

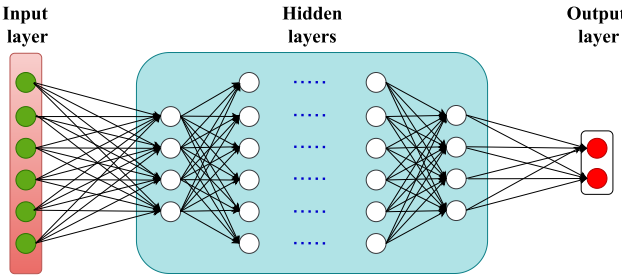


FIGURE 3. The typical architecture of FFNN [5].

III. PRELIMINARIES

A. REINFORCEMENT LEARNING (RL)

Apart from supervised and unsupervised learning, RL is considered the third type of ML that is characterized by its capability of self-learning and the development of behaviors through trial-and-error. The main architecture of an RL model consists of two key components, including an agent and an environment, as shown in FIGURE 2. The interactions between these components are described using three different concepts, state (s), action (a), and reward (r) [25], [56]. The RL process starts through a direct interaction between the agent and the environment sequentially in different time steps as shown in FIGURE 2. The agent takes an action a_t at a time step t and sends it to the environment. Accordingly, the environment's state changes from s_t at time t to a new state s_{t+1} at time $t+1$. Then, the agent receives from the environment a reward/penalty value represented by r_t that reflects how good/bad the action is. Generally, RL aims to maximize the total accumulated reward and establish a policy that maps states to actions. The total accumulated reward is expressed as follows.

$$R_t = \sum_{l=0}^{\infty} \gamma_l r_{t+l}, \tag{1}$$

where $\gamma \in [0, 1]$ is the discount factor that reflects the contribution of the future reward to the expected return. If $\gamma = 0$, it means the agent lacks foresight and is looking forward to maximizing the current rewards only. On the other hand, if γ gets closer to 1, the agent aims to be foresighted to the future rewards [56]. r_{t+l}, in Eq. 1, is the reward of the future time step.

RL introduces the value function $V^{\pi}(s)$, described in Eq. 2, to show how good the agent is in the state s and it is identified as the expectation at state s. Also, it depends on the policy π that maps the actions and states. When the optimal policy, which maximizes the action value achievable for state s, is adopted, an optimal value function that represents the highest value is obtained. The optimal policy and optimal value function are denoted as π^* and $V^*(s)$, respectively [57], and represented by Eqs. 3 and 4, respectively.

$$V^{\pi}(s) = E(R_t \mid S_t = s) \tag{2}$$

$$\pi^* = \arg\max_{\pi} V^{\pi}(s) \tag{3}$$

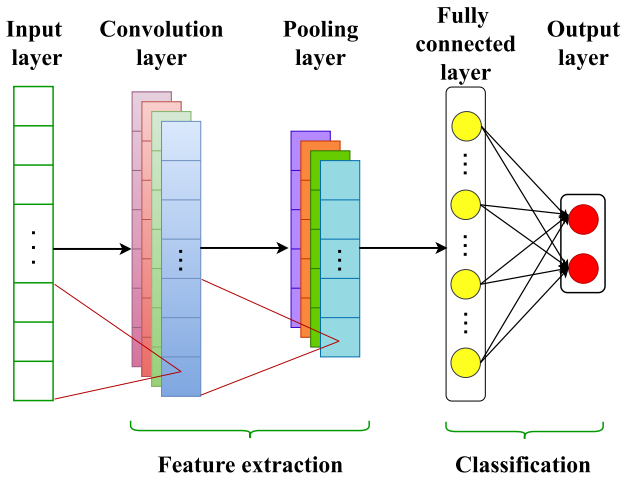


FIGURE 4. The typical architecture of CNN [5].

$$V^*(s) = \max_{\pi} V^{\pi}(s) \tag{4}$$

Similarly, a Q function is formulated to estimate a Q value using input pair (state, action) and outputs the reward. Consequently, the optimal policy π^* and the optimal Q value $Q^*(s, a)$ are represented as follows.

$$\pi^* = \arg\max_{a} Q^*(s, a). \tag{5}$$

$$Q^*(s, a) = R(s, a) + \gamma E_{s'}\left[V^*\left(s'\right)\right], \tag{6}$$

where $R(s, a)$ stands for the immediate reward gained by the agent after executing an action a and the transition from state s to another state s', $E_{s'}\left[V^*\left(s'\right)\right]$ stands for the expected future reward of state s' after transitioning from state s to s'.

In order to understand how the optimal policy is computed in the RL, it is important to discuss the exploration and exploitation mechanisms. Exploration means that the agent must explore and evaluate a variety of available actions to determine the best action selection in the future. Meanwhile, exploitation means that the agent must utilize the current knowledge to modify the action policy, thus maximizing the total rewards. The mathematicians investigate the exploration and exploitation using an ϵ-greedy policy. The agent can either explore the actions by randomly selecting an action from the set of available actions at state s with exploration rate ϵ or exploit a certain action with exploitation rate $1 - \epsilon$ considering the maximum Q value of this action. The exploration rate $\epsilon \in [0, 1]$ should start from the maximum value 1 and gradually decrease with the progress of the learning process [57], [58]. Eventually, the agent decides the optimal action using the exploitation mechanism and the current knowledge when the training model becomes more mature.

A distinguishable Q learning algorithm has been proposed in [59] as a model-free RL. It encourages the agent of the RL model in the Markovian domain to learn and behave optimally via practicing different actions sequentially. The goal of using the Q learning algorithm is to maximize the total accumulated reward using the Bellman equation:

$$Q^{new}(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha\left(r_t + \gamma \max_{a} Q(s_{t+1}, a_{t+1})\right), \tag{7}$$

where $(r_t + \gamma \max_a Q(s_{t+1}, a_{t+1}))$ is the update target and $\alpha \in [0, 1]$ is the learning rate. α indicates to what level the updated Q-value overrides the previous Q-value. If $\alpha = 0$, this indicates that the agent utilizes the prior knowledge only and did not learn anything from the new interaction. If $\alpha = 1$, this indicates that the agent ignores the prior knowledge and focuses only on exploring the available actions.

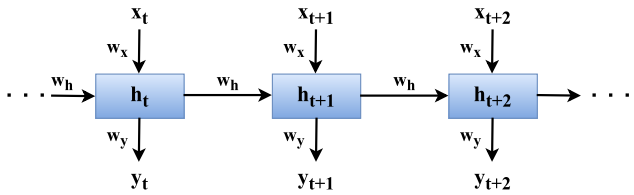
Q function is presented to estimate this reward maximization when the action is executed in a state. The iterative updating process is considered the main concept of the Q learning algorithm to continually update and learn the Q value giving the learning rate [25], [56], [57]. This updating process is repeated until the next state s_{t+1} becomes a terminal state at the end of the episode. Finally, the current and the last updated Q values are compared to check the convergence occurrence. If the convergence is not realized, the agent has to repeat this process in the next iteration. In the Q-learning algorithm, a Q-table is constructed to record Q-values of pairs (state, action). The architecture of the Q-table is formulated using rows that correspond to the set of the states and columns corresponding to the set of the available actions [29], [57]. With increasing the number of actions and states, a large memory is required to cope with the continuous increase of state-action space. Hence, it is inefficient to use a Q-table to handle this increasing space, especially in complex real problems. Fortunately, DL has been integrated with RL and initiated deep Q network (DQN) as a promising and powerful solution to this problem due to the interesting DL properties [29].

B. DEEP LEARNING (DL)

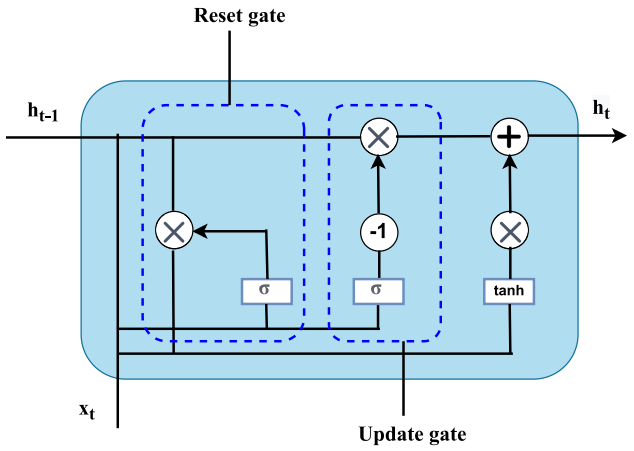
DL is a powerful technique of neural networks comprised of several hidden layers. The general structure of neural networks is constructed from input, hidden, and output layers. Recently, many applications, such as voice identification and face recognition, employed DL techniques to enhance the application performance due to the accuracy and flexibility of the DL. In this manuscript, a classification problem is studied to classify electricity consumption readings, hence, detecting the electricity theft attacks using various DL architectures. In particular, feed-forward neural networks (FFNNs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs) are used in the proposed RL approach. In the training phase of a DL model, an optimizer, an objective function, and labeled data sample are used to compute the optimum values of parameters for the model, including biases and weights.

1) FEED-FORWARD NEURAL NETWORK (FFNN)

FFNN is one of the most common architectures of DL used to solve a wide variety of problems, such as voice identification,



(a) RNN.



(b) GRU.

FIGURE 5. The typical architecture of (a) RNN and (b) GRU.

TABLE 1. Irish dataset characteristics.

Description of data	Value
Data consumption time frame	536 days
No. of customers	3600+
Fine-grained interval	30 Min
No. of readings per day (R)	48
No. of the considered customers	130

TABLE 2. Cyber attack functions.

Attack No	Attack function
1^{st}	$f_1(x_i(t)) = \beta x_i(t)$
2^{nd}	$f_2(x_i(t)) = \beta_t x_i(t)$
3^{rd}	$f_3(x_i(t)) = mean(x_i)$
4^{th}	$f_4(x_i(t)) = \beta_t mean(x_i(t))$
5^{th}	$f_5(x_i(t)) = \begin{cases} 0 & t \in [t_s, t_e] \\ x_i(t) & t \notin [t_s, t_e] \end{cases}$
6^{th}	$f_6(x_i(t)) = x_i(R - t)$

face recognition, and prediction systems. Multi-layer perceptron (MLP) is another name for FFNN. The typical architecture of an FFNN is shown in FIGURE. 3.

2) CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is introduced as a class of DL networks frequently used in a variety of applications such as speech processing, image processing, and natural language processing (NLP). CNN is characterized by the ability to capture intricate patterns and extract distinct features from the input data. A typical CNN structure comprises of an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer as depicted in FIGURE 4.

3) GATED RECURRENT UNIT (GRU) NEURAL NETWORK

GRU is a class of RNNs, used in different applications such as speech recognition and text generation. RNN is a combination of some hidden states and internal connections between the internal states as shown in FIGURE 5. The input data to an RNN is processed time step by time step. At each time step t, the current hidden state of the network h_t is updated using a transition function with two inputs, the previous hidden state h_{t-1} and the current time information x_t for, as follows.

$$h_t = F\left(x_t, h_{t-1}\right), \tag{8}$$

where F is a nonlinear activation function. Hence, the previous inputs can be stored and memorized using h_{t-1}, which indicates that GRU is considered a memory for input patterns.

IV. DATASET PREPARATION

In this section, we discuss the details of the preparation phase of the dataset, used for training and evaluating the proposed electricity theft detection model.

A. BENIGN SAMPLES

In this paper, the Irish smart energy trails [48], a real public dataset, are used to train and evaluate the proposed approach. This dataset was released by the Electric Ireland and Sustainable Energy Authority at the onset of 2012. The main characteristics of this dataset are presented in TABLE 1. In our experiments, we depend only on the SMs' readings of 130 randomly selected customers from the Irish dataset, yielding a total of 69,680 samples, where each sample represents the fine-grained readings (i.e., 48 readings) in one single day for one customer. All these samples are benign samples.

B. MALICIOUS SAMPLES

The proposed detector needs to be trained on both classes of data (benign and malicious samples). However, the malicious samples are not publicly available. Therefore, a set of electricity theft cyber-attacks, which are introduced in [20], are utilized in this work to imitate the electricity theft cyber-attacks. The proposed cyber-attacks are used to generate malicious samples via modifying the benign samples. Six attacks are summarized in TABLE 2, where, $x_i(t)$ denotes the true value of the electricity consumption reading of a customer i at time step t and $f(.)$ gives the reduced electricity consumption value.

The target of the 1^{st} attack is reducing the true consumption value by a random reduction factor β, where $0 < \beta < 1$. The

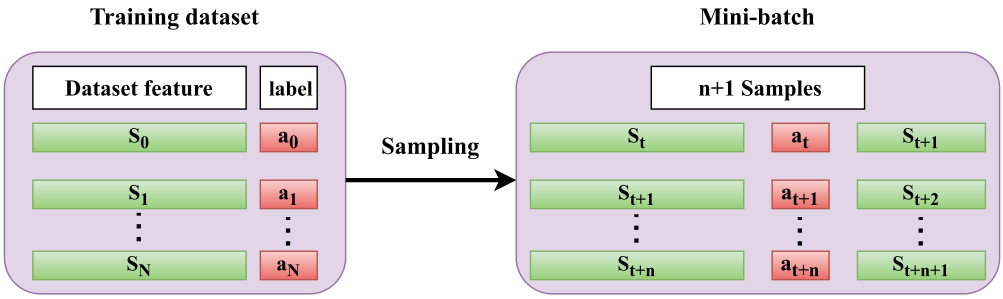


FIGURE 6. Mini-batch preparation for DQN and DDQN training [58].

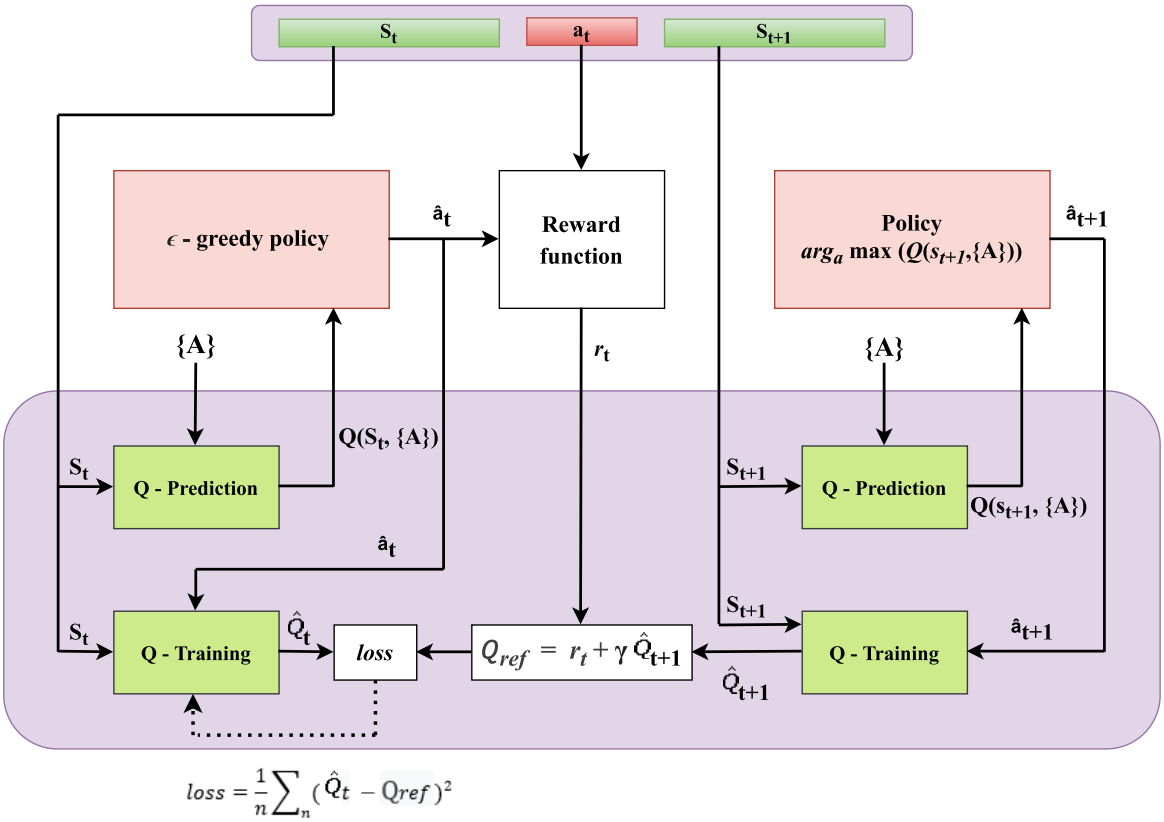


FIGURE 7. The DQN training scheme [58].

2^{nd} attack dynamically reduces the true consumption value by a time-varying reduction factor β_t, where $0 < \beta_t < 1$. The 3^{rd} attack reports the mean value of the electricity consumption readings over the day and the 4^{th} attack reduces the mean of consumption readings by the time-varying reduction factor. The 5^{th} attack enables the attacker to report zero consumption readings during a certain period of time defined as $[t_s, t_e]$; otherwise, the attacker reports the true consumption readings, where t_s and t_e identify the start and end time of electricity theft period, respectively. Finally, the 6^{th} attack reports the actual higher consumption readings during the low price periods. This attack is launched in case of using dynamic pricing for electricity consumption where the prices of the electricity vary during the day to reduce the load during peak hours.

C. DATA PREPROCESSING

Given the attack functions discussed in the previous subsection, to create the malicious samples, the parameters β and β_t must be set. These parameters are uniformly distributed random variables within the range of [0.1, 0.4] in the attack functions $f_1(.)$, $f_2(.)$, and $f_4(.)$. In attack function $f_5(.)$, t_s is a uniformly distributed random variable in the range of [0, 19] and t_e is set to 48. Further, after employing the proposed attack functions on the benign samples to create the malicious dataset, the corresponding daily benign and malicious

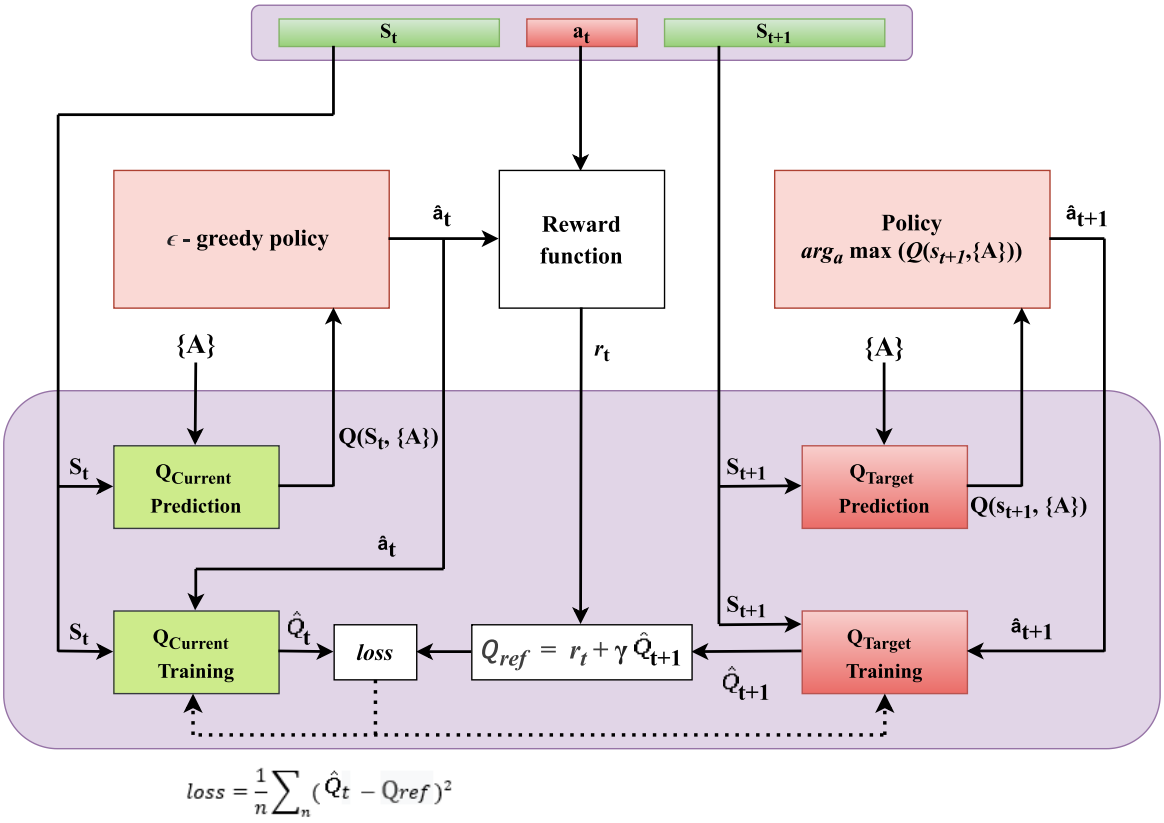


FIGURE 8. The DDQN training scheme [58].

samples for each customer are 536 and 3,216 (536*6 attacks), respectively, with a total of 3,752 samples. Consequently, the dataset is imbalanced since the number of benign samples is less than the number of malicious samples where the ratio between the two classes of samples is (1:6).

Utilizing the imbalanced dataset for training the electricity theft detector causes the detector to be biased toward the malicious samples' class because it is the majority class. Therefore, an adaptive synthetic (ADASYN) over-sampling technique is applied to avoid this issue by over-sampling the minority benign samples' class for each SM. Thus, the total daily samples of each customer is 6,432 representing the balanced benign and malicious samples, where each sample consists of 48 readings. Consequently, given the benign and malicious samples of 130 customers over 536 days, the dataset size is 836,160 (6,432*130) samples. The dataset is partitioned into two sets, one for training and the other for testing, in a 2:1 ratio. The number of samples in the training set is 557,440, and the number of samples in the testing set is 278,720.

V. THE PROPOSED REINFORCEMENT LEARNING MODELS

In the proposed models, it is required to set up the training process by creating the mini-batches of the training dataset for each model. The generic structure of the training dataset consists of a number of samples where each sample contains features and the corresponding label. In order to assimilate the basic concepts of DRL with the training dataset, the dataset features represent the state of the environment while the corresponding label represents the action. Thus, the general form for each sample of a mini-batch consists of the state of the current time slot s_t, the action of the current time slot a_t, and the state of the next time slot s_{t+1}. The generic structure of the mini-batch is illustrated in FIGURE 6, where a subset of samples called mini-batch is created by sampling the training dataset randomly. The generation process of the mini-batches is performed before the training model starts such that each mini-batch is constructed by $n+1$ sequential samples ranked by the index t. In this section, we discuss the training process of our RL-based DQN and DDQN models.

A. DEEP Q NETWORK (DQN)

In this work, a DQN model is employed to compute the value of the Q function. Given a certain state and action, a Q function value is obtained to represent the maximum expected reward for this specific pair (state, action). Consequently, depending on different Q function values for the different states, it is easy to formulate the policy function that maps states to actions, i.e., a certain action is executed for a certain state. The optimal policy π^* represents the maximum Q-values of the pairs (state, action), and it is computed from the Q function as explained in Eq. 5. The training process of

Algorithm 1 DDQN Training Algorithm

Input: Training epochs T, batch size B, exploration rate ϵ, discount factor γ, and learning rate α.
Output: The optimal action a^*.

```
Initialize the action value function Q(s, a) arbitrarily.
Initialize the state s by sampling the training dataset randomly.
for i = 0, 1, 2, ..., T do
    for each state s in i. do
        Input the state s_t and the actions set A in the current network in order to predict Q(s, A) for all actions.
        Use the ϵ-greedy policy to select the action â_t.
        Given s_t and â_t, obtain Q(s_t, â_t).
        Calculate the reward r_t.
        Input the next state s_{t+1} and the actions set A in the target network in order to predict Q(s_{t+1}, A) for all actions.
        Use arg max_a Q(s_{t+1}, A) policy to select â_{t+1}.
        Given s_{t+1} and â_{t+1}, obtain Q̂_{t+1}(s_{t+1}, â_{t+1}).
        Using Q̂_{t+1}, r_t, and γ, Obtain Q_ref.
        Calculate the loss function.
        Update the Q-value Q(s_t, a_t).
        Repeat until s_{t+1} is terminal.
    end for
    Repeat until getting to epoch T.
end for
Compute the optimal policy π* and optimal action a*.
Execute the optimal action a*_t at current time slot t.
```

DQN is initiated using the generic structure of the individual sample that contains the triple (s_t, a_t, s_{t+1}), where s_t is the current state, a_t is the true label, and s_{t+1} is the next state as represented in FIGURE. 7.

The training process in this work is performed using four different architectures of deep neural networks, including FFNN, CNN, GRU, and hybrid architecture (CNN+GRU). Also, the mean square error loss is applied between Q_t and Q_{ref}, where Q_t and Q_{ref} are the estimated Q-value of the current state and Q-reference value, respectively. Q_{ref} is calculated by summing the current state reward r_t to the result of the multiplication of the discount factor γ by the Q-value of the next state Q_{t+1}. The reward value r_t of a certain state is estimated depending on the comparison between the ground truth label a_t and the predicted label $\hat{a}_t$ at that certain state. When the two values are equal, i.e., the prediction is correct, the reward value is one, $r_t = 1$. Otherwise, the reward value is zero, $r_t = 0$, as indicated in the following equation.

$$r_t = \begin{cases} 1 & \hat{a}_t = a_t. \\ 0 & \hat{a}_t \neq a_t. \end{cases} \tag{9}$$

Furthermore, a prediction process of the Q function using all the combinations of a current state s_t and the set of different available labels A is performed to get the predicted label of the current state $\hat{a}_t$. This prediction process of $Q(s_t, A)$ is illustrated in FIGURE. 7, where $Q(s_t, A) = [Q(s_t, a_0), Q(s_t, a_1), ..Q(s_t, a_p)]$ and p is the total number of available labels. After that, an ϵ-greedy algorithm is employed to select the action from the input $Q(s_t, A)$ where the action is selected either using the exploitation process with probability ϵ or using the random exploration process with probability $1 - \epsilon$. On the other hand, an arg_a max(.) policy is employed to predict the action of the next state $\hat{a}_{t+1}$ as illustrated in the right-hand side of FIGURE 7. Both the next state s_{t+1} and the predicted action a_{t+1} are used to estimate the Q-value of the next state $\hat{Q}_{t+1}$ using the maximum policy formula $\hat{Q}_{t+1} = max_a Q(s_{t+1}, A)$. Once the model training phase is successfully completed, the proposed DRL model is employed for the action prediction by selecting the action which provides the maximum value of the Q function.

TABLE 3. Parameters of DRL schemes.

Parameter	Value
No. of training epochs (T)	10
Batch size (B)	128
Exploration rate (ϵ)	0.6
Discount factor (γ)	0.001
Learning rate (α)	0.00001

B. DOUBLE DEEP Q NETWORK (DDQN)

The generic structure of DDQN has the same elements as the DQN structure. However, there is only one difference in the next state prediction process. DDQN has two deep neural networks; the first network is called the current network and used to predict the Q function of the current state $\hat{Q}_t$, while the second network is called the target network and used to predict the Q function of the next state $\hat{Q}_{t+1}$. Although both the target network and the current network have the same architecture, the target network has a time delay synchronization. Therefore, the target network's parameters must be updated with those of the current network on a regular basis. The main cause for employing the second network (target network) is avoiding the moving target effect during gradient decent calculation over $(\hat{Q}_t - Q_{ref})^2$. Otherwise, the training and prediction mechanisms of the DDQN model are the same as explained in the DQN model. By updating the parameters of the target network, the DDQN is able to determine the optimal Q-value for the optimal action by minimizing the prediction loss between $\hat{Q}_t$ and Q_{ref}. This is accomplished through an increase in the accumulated reward used to calculate Q_{ref}. The typical architecture of the DDQN is presented in FIGURE 8 and the DDQN training process is presented in Algorithm 1.

TABLE 4. The parameters of the FFNN detection model.

Architecture	Parameters		
	Layer	Number of units	AF
FFNN	Input	48	Linear
	Dense	512	Relu
	Dense	700	Relu
	Dense	850	Relu
	Dense	1024	Relu
	Dense	512	Relu
	Dense	256	Relu
	Dense	200	Relu
	Dense	50	Relu
	Output	2	Softmax

TABLE 5. The parameters of the CNN detection model.

Architecture	Parameters		
	Layer	Number of units	AF
CNN	Input	48	Linear
	Conv1D	32	Relu
	Conv1D	64	Relu
	Conv1D	128	Relu
	Dense	64	Relu
	Dense	128	Relu
	Dense	256	Relu
	Dense	256	Relu
	Dense	512	Relu
	Dense	2	Softmax

TABLE 6. The parameters of the GRU detection model.

Architecture	Parameters		
	Layer	Number of units	AF
GRU	Input	48	Linear
	GRU	64	Sigmoid
	GRU	64	Tanh
	Dense	64	Relu
	Dense	128	Relu
	Dense	2	Softmax

VI. PERFORMANCE EVALUATION

In this section, the setup of our experiments and the performance evaluation metrics are discussed. After that, the experimental results of four scenarios conducted in this work are presented to assess the performance of the proposed detection approaches. The first scenario presents a global RL model that utilizes DQN and DDQN to detect false power consumption readings using various neural network architectures such as FFNN, CNN, GRU, and hybrid (CNN+GRU).

TABLE 7. The parameters of the hybrid (CNN+GRU) detection model.

Architecture	Parameters		
	Layer	Number of units	AF
CNN+GRU	Input	48	Linear
	Conv1D	64	Relu
	Conv1D	64	Relu
	Conv1D	128	Relu
	GRU	64	Tanh
	GRU	64	Tanh
	GRU	64	Tanh
	Dense	64	Relu
	Dense	128	Relu
	Dense	2	Softmax

In the second scenario, a customized model based on DDQN is developed for new customers. Firstly, the global model is utilized for detecting the electricity theft cyberattacks, and then, utilizing the distinct features of RL, the global model is retrained on the consumption readings of this customer to obtain the customized model. In this way, the detection accuracy can be high due to customizing the global model and zero-day attacks are thwarted due to using the global model at the beginning.

The third scenario focuses on changes in the consumption patterns of current customers. The customized model is updated by retraining it on the latest consumption data to learn the new patterns. Meanwhile, in the fourth scenario, we tackle the challenge of learning new cyber-attacks. The global model is initially trained to identify three types of attacks and when a new attack is discovered, the model is retrained using the newly collected consumption data to enhance its detection and defense capabilities.

The parameters of the proposed DRL detection schemes are adjusted and given in TABLE 3. In our experiments, we have used Python 3 libraries, including Numpy, Scikit-learn, Tensorflow, Pandas, Matplotlib, and Keras. Finally, all experiments have been run on the Google Colab platform which provides the ability to write and execute Python codes directly in the browser.

A. PERFORMANCE EVALUATION METRICS

The proposed detection approaches are assessed in terms of different evaluation metrics, including accuracy, precision, recall, false alarm, false negative rate, highest difference, and F-1 score. These metrics are computed using the true positive (TP), true negative (TN), false positive (FP), and false negative (FN), which are the core elements of the confusion matrix and are defined as follows:

- TP: It is the number of samples that are correctly classified as malicious.
- TN: It is the number of samples that are correctly classified as benign.

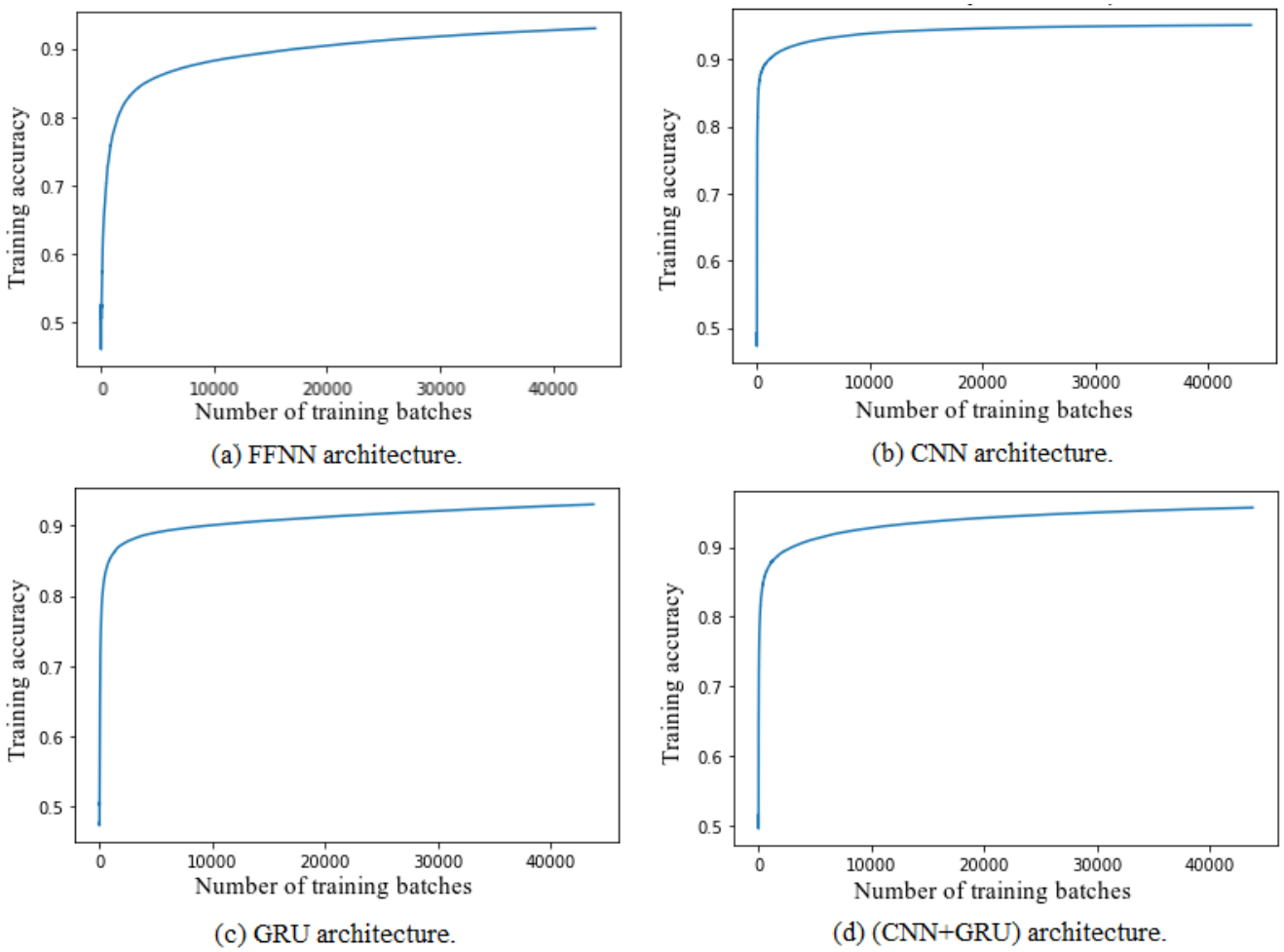


FIGURE 9. Training accuracy of different architectures of DDQN-based global model.

- FP: It is the number of samples that are misclassified as malicious.
- FN: It is the number of samples that are misclassified as benign.

The above evaluation metrics are explained as follows.

1) ACCURACY (ACC)

It computes the ratio of the number of correctly classified samples to the total number of evaluated samples. It is determined as follows:

$$ACC(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100. \quad (10)$$

2) PRECISION

It computes the ratio of the number of true positive samples to the sum of true positive samples and false positive samples. It is determined as follows:

$$Precision(\%) = \frac{TP}{TP + FP} \times 100. \quad (11)$$

3) RECALL

It computes the ratio of the number of true positive samples to the number of positive samples. It is determined as follows:

$$Recall(\%) = \frac{TP}{TP + FN} \times 100. \quad (12)$$

4) FALSE ALARM (FA)

It computes the ratio of the number of false positive samples to the total number of negative samples. It is determined as follows:

$$FA(\%) = \frac{FP}{FP + TN} \times 100. \quad (13)$$

5) FALSE NEGATIVE RATE (FNR)

It computes the ratio of the number of false negative samples to the total number of positive samples. It is determined as follows:

$$FNR(\%) = \frac{FN}{FN + TP} \times 100. \quad (14)$$

TABLE 8. Comparison between the performance of DL, DQN-RL, and DDQN-RL global models.

Metrics	DL				DQN-RL				DDQN-RL			
	FFNN	CNN	GRU	CNN+GRU	FFNN	CNN	GRU	CNN+GRU	FFNN	CNN	GRU	CNN+GRU
ACC (%)	92.42	93.14	91.10	94.71	95.02	95.84	95.84	96.84	94.63	95.22	95.82	97.33
Precision (%)	92.41	92.78	91.67	93.68	95.10	95.93	95.90	96.89	94.86	95.50	95.94	97.38
Recall (%)	92.40	93.52	90.38	95.84	95.02	95.84	95.84	96.84	94.63	95.22	95.82	97.33
FA (%)	7.56	7.23	8.17	6.42	4.47	2.99	3.75	2.68	2.43	1.37	2.48	2.06
FNR (%)	7.59	6.47	9.62	4.15	5.48	5.32	4.57	3.64	8.30	8.16	5.86	3.27
HD (%)	84.85	86.30	82.21	89.43	90.55	92.86	92.09	94.16	92.20	93.86	93.35	95.27
F1 (%)	92.40	93.15	91.02	94.75	95.06	95.89	95.87	96.86	94.74	95.36	95.88	97.35

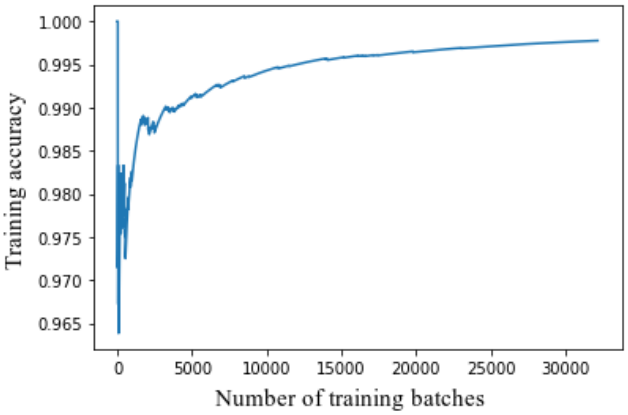


FIGURE 10. Training accuracy of DDQN-based hybrid(CNN+GRU) customized model for customer 14.

TABLE 9. Comparison between the performance of the global and customized models for different customers.

Metrics	Global model	Customized models		
		Customer 14	Customer 30	Customer 48
ACC (%)	97.332	99.439	98.102	99.502
Precision (%)	97.388	99.441	98.119	99.507
Recall (%)	97.332	99.439	98.102	99.502
FA (%)	2.06	0.25	0.93	0.14
FNR (%)	3.27	0.86	2.88	1.00
HD (%)	95.27	99.19	97.17	99.36
F1 (%)	97.355	99.440	98.110	99.504

6) HIGHEST DIFFERENCE (HD)

It computes the difference between recall and false alarm (*FA*), and it is determined as follows:

$$HD(\%) = Recall(\%) - FA(\%). \tag{15}$$

7) F-1 SCORE (F1)

It computes the harmonic mean between precision and recall. It is determined as follows:

$$F1(\%) = \frac{2 * Precision * Recall}{Precision + Recall} \times 100. \tag{16}$$

B. EXPERIMENTAL RESULTS OF SCENARIO 1

In the first scenario, a global RL detection model is constructed using a DQN and a DDQN of different DL architectures, including FFNN, CNN, GRU, and hybrid architecture (CNN+GRU). Different DL models, including FFNN, CNN, GRU, and hybrid (CNN+GRU) are utilized as benchmarks to evaluate the performance of the different RL detection models. The dataset discussed in section IV is used for training all the different architectures of DL and RL detection models. The parameters of the aforementioned detectors are provided in TABLES 4-7. Also, the training accuracy of different architectures of DDQN-based global model is presented in FIGURE 9. The figure shows that the training accuracy increases with increasing the number of training batches. Both the CNN and hybrid (CNN+GRU) architectures exhibit superior performance and faster convergence compared to other architectures such as FFNN and GRU. Also, it is worth noting that the FFNN architecture exhibits the slowest convergence rate compared to the other architectures.

The performance of the different detectors in scenario 1 is evaluated in terms of ACC, precision, recall, FA, FNR, HD, and F1, and the results are given in TABLE 8. This table compares the performance of DL-based, DQN-RL-based, and DDQN-RL-based global detectors. Firstly, we can observe that both CNN and GRU detectors achieve higher performance compared to FFNN detectors due to the distinguishable characteristics of CNN and GRU. CNN provides the detectors with the ability to extract features successfully while GRU enables the detectors to capture the correlation among the inputs. Secondly, the hybrid architecture of CNN and GRU provides the best performance compared to the other detectors as a result of the effective combination of their distinguishable characteristics. Thirdly, all the aforementioned detectors provide better performance in the case of using RL compared to using DL because RL has the capability to find the optimal actions using the reward concept during the training process.

For instance, the HD increases from 89.43% in the hybrid (CNN+GRU) DL-based detector to 95.27% in the corresponding DDQN-based detector (i.e., about 5.84% increase). Moreover, the FA decreases from 6.42% in the hybrid (CNN+GRU) DL-based detector to 2.06% in the

TABLE 10. The performance of old and updated customized models for different customers due to changing their consumption behavior.

Metrics	Old customized model			Updated customized models		
	Customer 14	Customer 30	Customer 48	Customer 14	Customer 30	Customer 48
ACC (%)	78.095	79.595	79.9	99.222	99.626	99.248
Precision (%)	79.865	79.717	80.519	99.233	99.234	99.230
Recall (%)	78.095	79.595	79.9	99.222	99.626	99.248
FA (%)	33.56	19.14	12.58	0.11	0.37	0.12
FNR (%)	10.0	15.06	27.84	1.57	0.38	1.45
HD (%)	44.54	60.455	67.32	99.11	99.26	99.10
F1 (%)	78.970	79.655	80.208	99.228	99.429	99.238

TABLE 11. Comparison between the performance of the global and customized models for newly launched attacks for different customers.

Metrics	Global model				Newly launched attacks customized models			
	Customer 5	Customer 20	Customer 25	Customer 35	Customer 5	Customer 20	Customer 25	Customer 35
ACC (%)	77.753	81.704	77.069	79.060	99.782	98.444	98.630	99.751
Precision (%)	80.271	83.648	81.416	85.170	99.387	98.481	98.648	99.752
Recall (%)	77.753	81.704	77.069	79.060	99.782	98.444	98.630	99.751
FA (%)	8.04	6.20	4.45	3.01	0.65	0.19	0.57	0.42
FNR (%)	36.21	30.48	41.14	42.68	0.44	2.90	2.16	0.49
HD (%)	69.72	75.5	72.62	82.16	99.13	98.26	98.07	99.33
F1 (%)	78.992	82.665	79.182	82.001	99.584	98.462	98.639	99.751

corresponding DDQN-based detector (i.e., about 4.36% decrease) and the F1 score increases from 94.75% to 97.35% (i.e., about 2.6% increase). Overall, the results indicate that the performance of DQN-RL- and DDQN-RL-based detectors are better than the performance of the DL-based detectors. Furthermore, the performance of the DDQN-RL-based detectors is better than the performance of the DQN-RL-based detectors due to the continuous updating of the target network parameters and avoiding the moving target effect. Finally, the results indicate that the DDQN-RL-based hybrid (CNN+GRU) detector outperforms all other detectors. Therefore, the hybrid structure of CNN+GRU using DDQN-based RL is chosen to design the detectors in the following scenarios.

C. EXPERIMENTAL RESULTS OF SCENARIO 2

In this scenario, a DDQN-based customized detection model is built for a new customer. In particular, if a new customer joins the smart grid, the readings of this customer are utilized for retraining a copy of the global detection model to create a customized model which has higher accuracy, while preventing zero-day attacks. The retraining process is performed sample by sample to obtain the new DDQN-based customized detection model. Also, the training accuracy of the DDQN-based hybrid (CNN+GRU) customized model for customer 14 is presented in FIGURE 10. The figure shows that the training accuracy increases with increasing the number of the training batches. TABLE 9 shows the comparison between the performance of the global detection model and the performance of the customized detection models of three randomly selected new customers. For a fair comparison, all the DDQN-based global and customized detection models have the same hybrid (CNN+GRU) structure. The results given in TABLE 9 indicate that all the customized detection models have better performance compared to the global detection model. For the different customers, the ACC, precision, recall, HD, and F1-score are higher, while the FA and FR are lower.

D. EXPERIMENTAL RESULTS OF SCENARIO 3

In this scenario, changing the consumption pattern of an existing customer is taken into consideration. The consumption pattern of a customer may change for several reasons, including changing the number of dwellers in the home and purchasing new electric appliances. In case of changing the consumption pattern of a customer, the DDQN-based customized detection model of this customer is retrained using the new consumption readings. The retraining process is performed sample by sample to obtain the updated DDQN-based customized detection model. TABLE 10 gives the results of the old and updated customized detection models for different customers due to changes in their consumption patterns. All the DDQN-based customized detection models have the same hybrid (CNN+GRU) structure. As observed from TABLE 10, the performance results of the updated customized detection models, i.e., after the retraining process, are better than the

performance results of the old customized detection model, i.e., before the retraining process, for the different customers. On one hand, the ACC and HD increase by up to 20% and 55%, respectively. On the other hand, FA and FNR decrease by up to 33% and 26%, respectively. Furthermore, comparing the performance results of the updated customized detection models to the original models in TABLE 9 indicates that there is a closer performance match for different customers.

E. EXPERIMENTAL RESULTS OF SCENARIO 4

In this scenario, retraining the model on newly discovered cyber-attacks is investigated. In particular, a global detection model is trained on the 1^{st}, 2^{nd}, and 4^{th} attacks presented in TABLE 2. However, the customers launch, the 3^{rd}, 5^{th}, and 6^{th} attacks presented in TABLE 2, using the new attacks, new malicious samples are computed and utilized for retraining the global detection model. The retraining process is performed sample by sample to obtain the new DDQN-based customized detection model. TABLE 11 gives the results of the global models, i.e., before the retraining process, and customized detection models of new cyber-attacks for different customers. For a fair comparison, all the DDQN-based global and customized detection models are constructed using a hybrid (CNN+GRU) structure. The results illustrate that the customized detection model can learn the newly launched cyber-attacks and provides a better capability for electricity theft detection. For the different customers, the ACC and HD increase by up to 21% and 30%, respectively. On the other hand, the FA and FNR decrease by up to 7.4% and 42%, respectively.

VII. CONCLUSION

In this paper, the use of RL for identifying the electricity theft cyber-attacks in smart power grids has been investigated. In particular, a series of cyber-attacks have been employed to create the malicious reading samples from benign readings of a real power consumption dataset. Then, deep RL detectors have been proposed for detecting electricity theft cyber-attacks. Specifically, we consider four scenarios. In the first scenario, the results indicate that the global detectors of RL-based DQN and DDQN achieve better performance compared to the performance of DL-based detectors. The results indicated that the RL-based detectors provide lower FA and higher HD. Moreover, the hybrid architecture of (CNN+GRU) provides the best performance compared to the other detectors as a direct result of the effective combination of their distinguishable characteristics. In the second scenario, a DDQN-based customized detector has been built for new customers. The results indicate that the customized detectors have better performance compared to the global detection model. In the third scenario, changing the consumption pattern of existing customers is investigated. The results indicate that the updated detection model achieves a comparable performance compared to the original detection model before the consumption patterns change. In the fourth scenario, training the model on new cyber-attacks is investigated and the results illustrate that the detection model can learn the new cyber-attacks and provide high accuracy, recall, and HD in addition to lower FA.

REFERENCES

[1] M. I. Ibrahem, M. Nabil, M. M. Fouda, M. M. E. A. Mahmoud, W. Alasmary, and F. Alsolami, "Efficient privacy-preserving electricity theft detection with dynamic billing and load monitoring for AMI networks," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1243–1258, Jan. 2021.

[2] A. Takiddin, M. Ismail, M. Nabil, M. M. E. A. Mahmoud, and E. Serpedin, "Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings," *IEEE Syst. J.*, vol. 15, no. 3, pp. 4189–4198, Sep. 2021.

[3] M. I. Ibrahem, M. M. Badr, M. M. Fouda, M. Mahmoud, W. Alasmary, and Z. Md. Fadlullah, "PMBFE: Efficient and privacy-preserving monitoring and billing using functional encryption for AMI networks," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2020, pp. 1–7.

[4] M. I. Ibrahem, M. M. Badr, M. Mahmoud, M. M. Fouda, and W. Alasmary, "Countering presence privacy attack in efficient AMI networks using interactive deep-learning," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–7.

[5] M. M. Badr, M. I. Ibrahem, M. Mahmoud, M. M. Fouda, F. Alsolami, and W. Alasmary, "Detection of false-reading attacks in smart grid net-metering system," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1386–1401, Jan. 2022.

[6] M. M. Badr, M. I. Ibrahem, M. Mahmoud, W. Alasmary, M. M. Fouda, K. H. Almotairi, and Z. M. Fadlullah, "Privacy-preserving federated-learning-based net-energy forecasting," in *Proc. SoutheastCon*, Mar. 2022, pp. 133–139.

[7] L. J. Lepolesa, S. Achari, and L. Cheng, "Electricity theft detection in smart grids based on deep neural network," *IEEE Access*, vol. 10, pp. 39638–39655, 2022.

[8] M. M. Badr, M. M. E. A. Mahmoud, Y. Fang, M. Abdulaal, A. J. Aljohani, W. Alasmary, and M. I. Ibrahem, "Privacy-preserving and communication-efficient energy prediction scheme based on federated learning for smart grids," *IEEE Internet Things J.*, vol. 10, no. 9, pp. 7719–7736, May 2023.

[9] M. Nabil, M. Ismail, M. M. E. A. Mahmoud, W. Alasmary, and E. Serpedin, "PPETD: Privacy-preserving electricity theft detection scheme with load monitoring and billing for AMI networks," *IEEE Access*, vol. 7, pp. 96334–96348, 2019.

[10] D. Gu, Y. Gao, K. Chen, J. Shi, Y. Li, and Y. Cao, "Electricity theft detection in AMI with low false positive rate based on deep learning and evolutionary algorithm," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4568–4578, Nov. 2022.

[11] M. M. Badr, M. I. Ibrahem, M. Baza, M. Mahmoud, and W. Alasmary, "Detecting electricity fraud in the net-metering system using deep learning," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–6.

[12] M. I. Ibrahem, S. Abdelfattah, M. Mahmoud, and W. Alasmary, "Detecting electricity theft cyber-attacks in CAT AMI system using machine learning," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–6.

[13] M. M. Badr, M. Mahmoud, M. Abdulaal, A. J. Aljohani, F. Alsolami, and A. Balamsh, "A novel evasion attack against global electricity theft detectors and a countermeasure," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 11038–11053, Jun. 2023.

[14] PR Newswire. (2014). *World Loses $89.3 Billion to Electricity Theft Annually, $58.7 Billion in Emerging Markets*. Accessed: Oct. 2020. [Online]. Available: https://www.prnewswire.com/news-releases/world-loses-893-billion-to-electricity-theft-annually-587-billion-in-emerging-markets-300006515.html

[15] W. Bank, "Reducing technical and non-technical losses in the power sector," Tech. Rep. 92639, 2009. [Online]. Available: http://documents.worldbank.org/curated/en/2009/01/20382190/reducing-technical-non-technical-losses-power-sector

[16] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Secur. Privacy Mag.*, vol. 7, no. 3, pp. 75–77, May 2009.

[17] N. Javaid, "A PLSTM, AlexNet and ESNN based ensemble learning model for detecting electricity theft in smart grids," *IEEE Access*, vol. 9, pp. 162935–162950, 2021.

[18] T. Ahmad, H. Zhu, D. Zhang, R. Tariq, A. Bassam, F. Ullah, A. S. AlGhamdi, and S. S. Alshamrani, "Energetics systems and artificial intelligence: Applications of industry 4.0," *Energy Rep.*, vol. 8, pp. 334–361, Nov. 2022.

[19] B. Ibrahim, L. Rabelo, E. Gutierrez-Franco, and N. Clavijo-Buritica, "Machine learning for short-term load forecasting in smart grids," *Energies*, vol. 15, no. 21, p. 8079, Oct. 2022.

[20] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.

[21] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2661–2670, May 2019.

[22] V. Ford, A. Siraj, and W. Eberle, "Smart grid energy fraud detection using artificial neural networks," in *Proc. IEEE Symp. Comput. Intell. Appl. Smart Grid (CIASG)*, Dec. 2014, pp. 1–6.

[23] M. Badr, "Security and privacy preservation for smart grid AMI using machine learning and cryptography," Ph.D. thesis, Tennessee Technol. Univ., Cookeville, TN, USA, 2022.

[24] M. Lapan, *Deep Reinforcement Learning Hands-On: Apply Modern RL Methods, With Deep Q-Networks, Value Iteration, Policy Gradients, TRPO, AlphaGo Zero and More*. Birmingham, U.K.: Packt Publishing, 2018.

[25] A. Kumari and S. Tanwar, "A reinforcement-learning-based secure demand response scheme for smart grid system," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2180–2191, Feb. 2022.

[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[27] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.

[28] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, Jul. 2020.

[29] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 1, 2021, doi: 10.1109/TNNLS.2021.3121870.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[31] O. Bouhamed, O. Bouachir, M. Aloqaily, and I. A. Ridhawi, "Lightweight IDS for UAV networks: A periodic deep reinforcement learning-based approach," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, May 2021, pp. 1032–1037.

[32] Z. Zheng, Y. Yang, X. Niu, H. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.

[33] S. Amin, G. A. Schwartz, A. A. Cardenas, and S. S. Sastry, "Game-theoretic models of electricity theft detection in smart utility networks: Providing new capabilities with advanced metering infrastructure," *IEEE Control Syst. Mag.*, vol. 35, no. 1, pp. 66–81, Feb. 2015.

[34] C.-H. Lin, S.-J. Chen, C.-L. Kuo, and J.-L. Chen, "Non-cooperative game model applied to an advanced metering infrastructure for non-technical loss screening in micro-distribution systems," *IEEE Trans. Smart Grid*, vol. 5, no. 5, pp. 2468–2469, Sep. 2014.

[35] T. Zhan, S. Chen, C. Kao, C. Kuo, J. Chen, and C. Lin, "Non-technical loss and power blackout detection under advanced metering infrastructure using a cooperative game based inference mechanism," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 4, pp. 873–882, Mar. 2016.

[36] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Sci. Technol.*, vol. 19, no. 2, pp. 105–120, Apr. 2014.

[37] Y. Peng, Y. Yang, Y. Xu, Y. Xue, R. Song, J. Kang, and H. Zhao, "Electricity theft detection in AMI based on clustering and local outlier factor," *IEEE Access*, vol. 9, pp. 107250–107259, 2021.

[38] S. K. Singh, R. Bose, and A. Joshi, "PCA based electricity theft detection in advanced metering infrastructure," in *Proc. 7th Int. Conf. Power Syst. (ICPS)*, Dec. 2017, pp. 441–445.

[39] K. Zheng, Q. Chen, Y. Wang, C. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1809–1819, Mar. 2019.

[40] A. Takiddin, M. Ismail, and E. Serpedin, "Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 663–676, Jan. 2023.

[41] M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1254–1263, Mar. 2020.

[42] R. R. Bhat, R. D. Trevizan, R. Sengupta, X. Li, and A. Bretas, "Identifying nontechnical power loss via spatial and temporal deep learning," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 272–279.

[43] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proc. IEEE*, vol. 109, no. 3, pp. 204–246, Mar. 2021.

[44] D. Wu, C. Wang, Y. Wu, Q.-C. Wang, and D. Huang, "Attention deep model with multi-scale deep supervision for person re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 1, pp. 70–78, Feb. 2021.

[45] M. J. Abdulaal, M. I. Ibrahem, M. M. E. A. Mahmoud, J. Khalid, A. J. Aljohani, A. H. Milyani, and A. M. Abusorrah, "Real-time detection of false readings in smart grid AMI using deep and ensemble learning," *IEEE Access*, vol. 10, pp. 47541–47556, 2022.

[46] T. Talaei Khoei and N. Kaabouch, "A comparative analysis of supervised and unsupervised models for detecting attacks on the intrusion detection systems," *Information*, vol. 14, no. 2, p. 103, Feb. 2023.

[47] T. T. Khoei and N. Kaabouch, "Densely connected neural networks for detecting denial of service attacks on smart grid network," in *Proc. IEEE 13th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2022, pp. 0207–0211.

[48] *Irish Social Science Data Archive*. Accessed: Sep. 2020. [Online]. Available: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[49] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," *J. Electr. Comput. Eng.*, vol. 2019, pp. 1–12, Oct. 2019.

[50] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. M. Islam, and J.-M. Kim, "Electricity theft detection in smart grid systems: A CNN-LSTM based approach," *Energies*, vol. 12, no. 17, p. 3310, Aug. 2019.

[51] *State Grid Corporation of China*. Accessed: Sep. 2020. [Online]. Available: http://www.sgcc.com.cn/

[52] M. S. Abdalzaher, M. M. Fouda, and M. I. Ibrahem, "Data privacy preservation and security in smart metering systems," *Energies*, vol. 15, no. 19, p. 7419, Oct. 2022.

[53] J. Luo, S. Yao, J. Zhang, W. Xu, Y. He, and M. Zhang, "A secure and anonymous communication scheme for charging information in vehicle-to-grid," *IEEE Access*, vol. 8, pp. 126733–126742, 2020.

[54] J. Luo, J. Liao, C. Zhang, Z. Wang, Y. Zhang, J. Xu, and Z. Huang, "Fine-grained bandwidth estimation for smart grid communication network," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 1225–1239, 2022.

[55] Y. Li, X. Wei, Y. Li, Z. Dong, and M. Shahidehpour, "Detection of false data injection attacks in smart grid: A secure federated deep learning approach," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4862–4872, Nov. 2022.

[56] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6629–6639, Nov. 2019.

[57] S. Dong, Y. Xia, and T. Peng, "Network abnormal traffic detection model based on semi-supervised deep reinforcement learning," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 4, pp. 4197–4212, Dec. 2021.

[58] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Application of deep reinforcement learning to intrusion detection for supervised problems," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112963.

[59] C. J. C. H. Watkins and P. Dayan, "*Q*-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

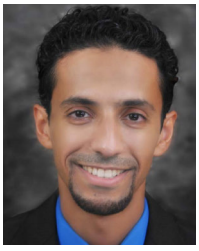
AHMED T. EL-TOUKHY received the B.Sc. and M.Sc. degrees in electrical engineering from Al-Azhar University, Cairo, Egypt, in 2011 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Tennessee Tech University, Cookeville, TN, USA. He is a Lecturer Assistant with the Faculty of Engineering, Al-Azhar University. His research interests include machine learning, reinforcement learning, cryptography, network security, 5G networks, and cognitive radio.

MAHMOUD M. BADR received the B.S. and M.S. degrees in electrical engineering (electronics and communications) from Benha University, Cairo, Egypt, in 2013 and 2018, respectively, and the Ph.D. degree in electrical and computer engineering from Tennessee Tech University, Cookeville, TN, USA, in 2022. He is currently an Assistant Professor with the Networks and Computer Security: Cybersecurity Department, College of Engineering, State University of New York (SUNY) Polytechnic Institute, USA. He is also a Lecturer Assistant with the Faculty of Engineering at Shoubra, Benha University, Egypt. His research interests include machine learning, blockchain, cryptography, 5G networks, and smart grids. He has been selected as a Poster Winner in Tennessee Tech University's Annual Research and Creative Inquiry Day, in 2021.

MOHAMED M. E. A. MAHMOUD (Senior Member, IEEE) received the Ph.D. degree from the University of Waterloo, in April 2011. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Tennessee Tech University, USA. He is the author of more than 100 papers published in IEEE conferences and journals. His research interests include security and privacy-preserving schemes for smart grids, e-health, and intelligent transportation systems. He has received the NSERC-PDF Award. He received the Best Paper Award from the IEEE International Conference on Communications (ICC 2009), Dresden, Germany, in 2009. He served as a technical program committee member for several IEEE conferences. He serves as an Associate Editor for IEEE Internet of Things Journal and *Peer-to-Peer Networking and Applications* (Springer).

GAUTAM SRIVASTAVA (Senior Member, IEEE) received the B.Sc. degree from Briar Cliff University, USA, in 2004, and the M.Sc. and Ph.D. degrees from the University of Victoria, Victoria, BC, Canada, in 2006 and 2011, respectively. He then taught for three years with the Department of Computer Science, University of Victoria, where he was regarded as one of the top undergraduate professors in computer science course instruction. In 2014, he joined a tenure-track position with Brandon University, Brandon, MB, Canada, where he is currently active in various professional and scholarly activities. He was promoted to the rank of Associate Professor, in January 2018. He has active research projects with other academics in Taiwan, Singapore, Canada, Czech Republic, Poland, and USA. He is constantly looking for collaboration opportunities with foreign professors and students. He is popularly known and is active in research in the field of data mining and big data. In his eight-year academic career, he has published a total of 43 papers in high-impact conferences in many countries and high-status journals (SCI and SCIE) and has also delivered guest lectures on big data, cloud computing, the Internet of Things, and cryptography at many Taiwanese and Czech universities. He received the Best Oral Presenter Award from FSDM 2017 which was held at the National Dong Hwa University (NDHU), Shoufeng, Hualien, Taiwan, in November 2017. He is the editor of several international scientific research journals.

MOSTAFA M. FOUDA (Senior Member, IEEE) received the B.S. (as the valedictorian) and M.S. degrees in electrical engineering from Benha University, Egypt, in 2002 and 2007, respectively, and the Ph.D. degree in information sciences from Tohoku University, Japan, in 2011. He was an Assistant Professor with Tohoku University and a Postdoctoral Research Associate with Tennessee Tech University, Cookeville, TN, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID, USA. He is also a Full Professor with Benha University. He has received several research grants, including the NSF Japan-U.S. Network Opportunity 3 (JUNO3). He has (co)authored more than 160 technical publications. His current research interests include cybersecurity, communication networks, signal processing, wireless mobile communications, smart healthcare, smart grids, AI, and the IoT. He has guest-edited a number of special issues covering various emerging topics in communications, networking, and health analytics. He is also serving on the editorial board for IEEE Transactions on Vehicular Technology and IEEE Access.



MAAZEN ALSABAAN received the B.S. degree in electrical engineering from King Saud University (KSU), Saudi Arabia, in 2004, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Canada, in 2007 and 2013, respectively. From 2015 to 2018, he was the Chairperson of the Department of Computer Engineering, KSU. He is currently an Associate Professor with the Department of Computer Engineering, KSU. He serves as a consultant for different agencies and received many grants from KSU and King Abdulaziz City for Science and Technology (KACST). His current research interests include wireless communications and networking, surveillance systems, vehicular networks, green communications, intelligent transportation systems, and cybersecurity.