Classification and Analysis of Adversarial Machine Learning Attacks in IoT: a Label Flipping Attack Case Study

Mahdi Abrishami
University of New Brunswick (UNB)
Fredericton, Canada
mahdi.abrishami@unb.ca

Sajjad Dadkhah
University of New Brunswick (UNB)
Fredericton, Canada
sdadkhah@unb.ca

Euclides Carlos Pinto Neto
University of New Brunswick (UNB)
Fredericton, Canada
e.neto@unb.ca

Pulei Xiong
Cybersecurity, National
Research Council Canada
pulei.xiong@nrc-cnrc.gc.ca

Shahrear Iqbal
National Research Council
Fredericton, New Brunswick, Canada
shahrear.iqbal@nrc-cnrc.gc.ca

Suprio Ray
University of New Brunswick (UNB)
Fredericton, Canada
sray@unb.ca

Ali A. Ghorbani
University of New Brunswick (UNB)
Fredericton, Canada
ghorbani@unb.ca

***Abstract*—With the increased usage of Internet of Things (IoT) devices in recent years, various Machine Learning (ML) algorithms have also developed dramatically for attack detection in this domain. However, the ML models are exposed to different classes of adversarial attacks that aim to fool a model into making an incorrect prediction. For instance, label manipulation or label flipping is an adversarial attack where the adversary attempts to manipulate the label of training data that causes the trained model biased and/or with decreased performance. However, the number of samples to be flipped in this category of attack can be restricted, giving the attacker a limited target selection. Due to the great significance of securing ML models against Adversarial Machine Learning (AML) attacks particularly in the IoT domain, this research presents an extensive review of AML in IoT. Then, a classification of AML attacks is presented based on the literature which sheds light on the future research in this domain. Next, this paper investigates the negative impact levels of applying the malicious label-flipping attacks on IoT data. We devise label-flipping scenarios for training a Support Vector Machine (SVM) model. The experiments demonstrate that the label flipping attacks impact the performance of ML models. These results can lead to designing more effective and powerful attack and defense mechanisms in adversarial settings. Finally, we show the weaknesses of the K-NN defense method against the random label flipping attack.**

I. Introduction

Internet of Things (IoT) environment is designated as a system of connected devices embedded with sensors to collect and exchange data and execute complex tasks [1]. Over the past years, there has been an upsurge growth in the usage of IoT gadgets. One of the major reasons for increasing the use of IoT devices is because they require less power consumption, provide more effortless connectivity, and are more convenient to use [2]. Because of this, the internet has accelerated the spread of IoT devices and has built a strong connection by providing service applications to different sectors such as industries, healthcare, smart cities, smart home, etc. [3]. IoT has become an extension of the internet that provides a relation between the physical and digital world where sensors and actuators are integrated to provide connectivity [4].

Researchers have proposed to use various ML models based on popular datasets such as NSL-KDD Cup [5] and UNSW-NB15 [6] that learn the patterns from captured data and provide the prediction of whether an input sample is benign or malicious. However, the solutions provided using many of the available datasets are losing their relevance due to new attack variants and new protocols developed according to the changing requirements. Moreover, with the rise of new IoT technologies, IoT devices' existing approaches to secure these limited resource usage are becoming obsolete. Recently, some new datasets, including BoT-IoT [7] which is generated through designing a natural network environment and CIC IoT Dataset 2022 [8] are presented, which investigate different IoT behaviors in different scenarios.

With changing the requirements, the pattern of attacks is also changing. Nowadays, adversarial cyber attackers have started exploiting the models rather than targeting particular IoT devices. These attacks, which are called Adversarial Machine Learning (AML) attacks, vary in type. Two important categories of these attacks include evasion and poisoning. In evasion attacks, malicious test samples are adopted by the adversary [9]. Following this approach, the attacker forces the implemented model to classify the data incorrectly and thus, making the system fail [10]. Poisoning attacks are aimed to target the data or model in ML training [11]. Data

manipulation is an essential category of poisoning attacks that further falls into two categories: manipulating the data in the training stage or the label of the training samples [11].

Considering the growing importance of IoT in the recent years and the new advances in adversarial attacks in this domain, a comprehensive review of the recent works related to this topic is required. However, to the best of our knowledge, no research work has specifically investigated adversarial attacks in IoT. Moreover, There are just a few works addressing the label flipping attacks as an important category of poisoning attacks.

Thereupon, the main aim of this research is to present an extensive review of the recent works regarding the adversarial machine learning in IoT. Moreover, several experiments are conducted regarding the label flipping attacks to shed light on different aspects of this attack category as an important issue in the IoT domain.

Following are the main contributions of this research:

- Presenting an extensive review of Adversarial Machine Learning (AML) in the IoT domain.
- Presenting a classification of AML attacks in the IoT domain.
- Investigating different scenarios for the label flipping attacks.
- Examining the effectiveness of the K-NN defense method against the random label flipping attack.

II. Background

Malware detection, network intrusion detection, and spam detection are just a few of the many areas where machine learning (ML) is important. It is typical to presume that a machine learning model will be used in a benign environment. In other words, it is assumed that no adversarial element will influence how well ML models function. This presumption, however, is not necessarily true [12]. Tricking ML models into producing inaccurate predictions is called adversarial machine learning (AML). In recent years, as the volume of data generated in different domains has increased significantly, poisoning attacks are considered to be an important category of threat, particularly where the data is collected from users (for IoT environments and sensor networks) or where the data labeling is crowdsourced. As an example, label flipping attacks are one of the major types of poisoning attacks resulting in significant performance degradation [13]. With respect to the growing interest in applications of ML in IoT, investigating the potential threats impacting the effectiveness and performance of ML models in this domain poses great importance.

A. Machine Learning (ML) in the IoT Domain

Considering the various vulnerabilities in the IoT domain, ML algorithms are widely being used to tackle the potential issues. Based on the requirements, data analysis can be performed in IoT devices or the cloud. Cloud refers to remote data servers or edge servers that bring the computation close to the IoT devices. As the data analysis may be done in IoT devices, considering the limited processing power, using lightweight ML models is preferred. Moreover, more processing may be performed on the data in cloud [14]. Some of the use cases of ML in IoT include outlier and intrusion detection [15], signal authentication and device identification [16], [17], spectrum sensing [18], and smart grids [19]. Therefore, with respect to the broad use of ML methods in IoT, investigating the threats against deployed models seems to pose great importance.

B. General Overview of Adversarial Machine Learning (AML)

The training and testing phases of the ML pipeline have different adversarial techniques. *Poisoning* is an important technique applied to the training phase and includes modifying training data [20]. In Indirect Poisoning, the adversarial modification of data is done before the preprocessing stage. Direct poisoning includes data injection, data manipulation, or logic corruption. In data injection, the goal is to change the distribution of data in a training set and is achieved through injecting *"adversarial inputs"* to the training set. However, the features and labels of existing data samples are not changed. The decision boundaries of an ML model can be shifted through this attack. In data manipulation, labels or features of data instances in the training set are altered [11].

The aim of attacks corresponding to the test phase is not to alter the data for training or decision boundary, but to generate samples such that they can fool a model in the testing phase [21]. Some of the important gradient-based techniques of evasion include [22] Fast Gradient Sign Method (FGSM) [23], Jacobian-based Saliency Map (JSMA) [24], and Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [25]. In gradient-free attacks, adversarial perturbations are applied with no background information of the model that is targeted. In other words, the target model is used as an oracle to which adversarial samples are sent and the output is observed [26].

C. Potential Cases of Adversarial Attacks in the IoT Domain

Based on [27], in an IoT/WSN (Wireless Sensor Networks) outlier detection setting, the collected data is sent to the gateway. Then, information is sent from the gateway to a server. In this step, the corresponding error, event, or malicious activity is detected to notify the end user. According to this paper, the ML methods that can be applied for outlier detection include statistical-based methods (parametric and non-parametric), supervised learning-based methods (such as SVM, Bayesian learning methods, K-NN, and Neural Networks), and unsupervised learning-based methods (including K-means clustering and PCA). However, supervised learning is used more widely in this domain. There are some limitations to the adoption of ML algorithms in the IoT domain. One of the most major constraints is the lack of computational power. In other words, the complex ML algorithms are hard to be deployed on the resource constraint devices [28].

Distributed approaches between sensing devices can also be adopted for the purpose of outlier detection [29], [30]. However, frequent communications are required in this approach [31]. The task of outlier detection can be performed on the individual sensors without the need to communicate with other

sensors. For instance, [31] has proposed a solution for outlier detection in WSN where autoencoder instances are run on sensors. The sensors send the input and output of autoencoders to the cloud via a gateway. In other words, the training process takes place in the cloud. Then, the model parameters (that are updated) are sent back from the cloud to the sensors. Nevertheless, outlier detection using ML algorithms is likely to be targeted by adversaries.

Signal authentication is one of the major tasks of IoT where adversarial attacks are likely to happen. Signals are sent from IoT devices (in the perceptual layer) to a gateway to control the operation of IoT devices. However, this signal is likely to be manipulated. Dynamic data injection attacks may happen in this stage. The gateway has to authenticate the signals sent from IoT devices. However, as the gateway is resource-constraint, it should optimally choose signals for authentication [16]. Traditional solutions such as ML-based Radio Frequency fingerprint identification can help with device identification. However, these solutions are facing some challenges such as huge amounts of training data. Deep learning methods are used widely for device identification and signal recognition. But, they are susceptible to adversarial attacks. The attacks may be targeted or non-targeted [32].

Malware detection in IoT is also vulnerable to attacks conducted by adversaries. In terms of combating malware in IoT, there are two signature-based solutions. The first solution is host-based where the detection system is installed on devices. However, these solutions seem to be inadequate with respect to the developments in malware attacks. More importantly, resource-constrained devices cannot benefit from these solutions, particularly as these solutions are signature-based. Another solution is to put the detection mechanisms in the cloud [33].

Tampering with sensors' measurements is one of the potential adversarial attacks [34]. As an example, to investigate the effect of gases emitted by a factory, the government may decide to measure the emissions of the corresponding factory. However, the factory manager may tamper with the sensors' measurements by releasing chemicals near the sensors when air quality is good. Therefore, it is hard to derive that bad air quality is related to the emissions of this factory [35].

The data fusion and aggregation steps are vulnerable to attacks as well [36]. Data collection from different sensors is an important task where the data noise is filtered. This data collection should be context-aware, privacy-preserving, reliable, and real-time. However, there is the potential that several devices sending data to the fusion center are controlled by an adversary and therefore, the decision is compromised [37].

Wireless communications that are used widely in the IoT domain are at the risk of over-the-air wireless attacks as they are broadcasted. However, the use of ML in this domain lacks security. There are several techniques for adversarial attacks including exploratory, evasion, and poisoning where the aim is to get an understanding of the target model, evade the model in the test phase to make wrong decisions, and provide the model with wrong training data to affect the decisions respectively. These attack approaches can be mapped to the wireless domain to act as the jamming, spectrum poisoning, and priority violation attacks [38].

III. Literature review

The summary of several research works related to adversarial machine learning in IoT is presented in Tables I and XI. These works are compared in terms of the investigated attack(s), related attack categories, used models, and dataset.

Figure 1 shows the proposed classification of Adversarial Machine Learning (AML) attacks in the IoT domain according to the literature review section. Machine learning models can be deployed in different layers in edge devices or sensors, gateways, and the cloud. In the figure, they are connected to the layers with dashed lines. However, there are threats against the models in each layer. Over-the-air wireless attacks that include priority violation, jamming, and spectrum poisoning can be deployed against the transmitters. Moreover, tampering with sensor measurements is another potential issue in this layer. IoT devices are likely to be controlled by adversaries in this layer which can result in inaccurate output results.

ML models for device identification and authentication can be deployed in gateways. However, adversaries are capable of fooling these models into making wrong decisions. If intrusion detection systems are used in gateways or the cloud, they have the potential of being targeted for adversarial attacks. The process of data fusion in the fog layer is also vulnerable to threats. The data fusing process can be performed in the cloud level as well [39], making this layer vulnerable. If the classification of smart home IoT devices is performed in the cloud, the utilized ML algorithms may be attacked by adversaries. Based on [40], data processing is possible to be performed in smart gateways in smart homes. Therefore, smart gateways are vulnerable to adversarial attacks. There is also the possibility of attacks against data operation retrieval in a Security Operation Center (SOC).

IV. Label flipping attacks

Label flipping is a major subcategory of data poisoning aiming at manipulating data labels to impact the ML model's performance adversely [13]. These attacks cause major problems for the ML-based systems, particularly in noisy or uncertain environments like complex networks and IoT [41]. In a label flipping attack, the attacker can control the label of a limited proportion of samples. Poisoning attacks are shown to be effective in impacting the performance of ML algorithms such as neural networks, deep learning systems, Support Vector Machines (SVM), and embedded feature selection methods. Although deep learning systems have shown great performance when dealing with samples with clean labels, their effectiveness is degraded in the case of existing samples with flipped labels [41], [42].

TABLE I. SUMMARY OF THE RESEARCH WORKS REVIEWED IN THE LITERATURE REVIEW SECTION

Author	Investigated Attack(s)	Related Categories	Used Models	Dataset	Description
Baracaldo et al. (2018) [35]	Tampering with sensors' measurement	poisoning	SVM (for evaluation)	Dataset used in [43] and MNIST	Provenance meta-data is used for defense against poisoning attacks.
Luo et al. (2020) [37]	Controlling a small portion of IoT devices sending data to the fusion center	Poisoning	5-layer neural network - SVM	10000 samples collected from two Guassian distributions	Data is collected from manipulated IoT devices and an attack model is learnt.
sagduyu et al. (2019) [38]	Jamming, Spectrum Poisoning, Priority violation	Exploratory, Evasion, and Poisoning attacks	FNN	Spectrum sensing data in the experimental setting	• An exploratory attack followed by a poisoning or evasion attack in the wireless communication domain are suggested and launched. • A defense mechanism based on the stackelberg game is proposed.
Shi et al. (2018) [44]	Spectrum data poisoning	Exploratory, Evasion (and Poisoning against adversary if the defense is performed)	FNN	Spectrum sensing data in the experimental setting	The transmitter's behavior is inferred and the spectrum sensing data is effected by the adversary to manipulate the transmitter's decisions.
Shi et al. (2018) [45]	Jamming	Exploratory and Poisoning (in the defense against adversary)	FNN	Sensing results of the transmitter and adversary in the experimental setting	• The transmission decisions of the transmitter are captured by the adversary and then, a model is trained to predict the future decisions and jam them. • A defense based on some deliberate wrong actions made by the transmitter is proposed.
Erpek et al. (2018) [46]	Jamming	Exploratory and Poisoning (in the defense against adversary)	FNN	Spectrum sensing data in the experimental setting	• The transmitter's decisions are collected by the adversary and used to build a model to predict and jam the future transmissions. • GAN is used to reduce the training time process for the adversary. • A defense based on a few wrong actions made deliberately by the transmitter is proposed.
Kim et al. (2020) [47]	Attack against signal modulation	Evasion	DNN (VT-CNN2)	GNU radio ML dataset RML2016.10a	A number of adversarial attacks against modulation classifiers in the wireless communications domain are proposed and evaluated.
Singh and Sikdar (2022) [48]	Attack against appliance classification in smart home environment	Evasion	DNN	UK-DALE, and REFIT	Proposed a new gradient ascent-based white-box adversarial attack
Bao et al. (2021) [32]	Attacks against device identification	Evasion	CNN	Generated signals by the authors	• Several attacks are examined against the CNN-based device identification method. • Combined evaluation of indicators is proposed to enhance the evaluation.
Sadeghi and Larsson (2019) [49]	Attacks against end-to-end communication systems	Evasion	MLP and CNN-based autoencoders	Simulated signals	Physical black-box adversarial attacks are investigated and methods to perform these attacks are suggested.
Ferdowsi and Saad (2019) [16]. [46]	Manipulating signals between IoTDs and the gateway	Poisoning (as the data injection attack in the signal transmission phase in investigated)	LSTM	For simulations, a real dataset from an accelerometer is used.	• A watermarking approach is proposed for IoT device authentication. • A game theoretic approach is proposed to predict vulnerable devices when the computational resources are limited. • A reinforcement learning algorithm is presented for the case where information is incomplete.
Sharaf-Dabbagh and Saad (2016) [50]	Object emulation	Poisoning	Infinite Gaussian Mixture model - transfer learning	Generated fingerprints	Proposed a mechanism for IoT object authentication based on device's fingerprints.

Due to the importance of label flipping attacks, different attack experiments against Support Vector Machines (SVM) are investigated. As SVM is determined to be robust in adversarial settings [43], it is a suitable option to be utilized for the experiments.

Fig. 2 shows the framework of experiments conducted regarding label flipping. The dataset contains samples of the DoS, DDoS, Reconnaissance, and theft attack categories. Then, the data is pre-processed. Afterward, three experiments are conducted using the SVM ML model to investigate the adversarial effects of the label flipping attack. In the first experiment, samples with small or large margins to the hyperplane are flipped using the RBF and linear SVM kernels. In the second experiment, the effect of flipping samples of specific attack categories on the performance of an SVM model is investigated. Finally, the effect of flipping specific attack categories on misclassified samples is explained. In the experiments' results, the percentage of samples with flipped labels is shown with $Noise(\%)$.

A. Preliminaries

1) ***Support Vector Machine (SVM)****:* We have performed our studies by varying kernel, i.e., using linear [51] and Radial Basis Function (RBF) [52]. SVM discriminates two classes of data by drawing an optimal hyperplane that maximizes the distance between data classes. Considering the data to have two classes, the Support Vector Machine (SVM) solves the following optimization problem [53]:

$$\begin{aligned} min_{w,\xi}\ & \frac{1}{2}W^TW + C\sum_{i=1}^{l}\xi_i \\ \text{Subject to } & y_i(W^T\phi(X_i)+b) \geq 1-\xi_i \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

where X_i, y_i, ξ_i, W, and C are input data, label vector, parameters associated with optimization, weight vector, and penalty parameter of error term respectively. According to [54], the RBF kernel seems to be the first option while using the Support Vector Machines (SVM). It can handle nonlinearity when the labels and attributes have a nonlinear relation. Actually, the samples are mapped to a higher dimension in a nonlinear manner. The kernel function for RBF is:

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0 \tag{2}$$

where γ is a kernel parameter. The function named linear discriminant is used to separate two different class samples at the time of training of the classifier [55]. The hyperplane discriminant function $df(k)$ is defined in Eq. (3).

$$df(k) = sgn(\sum_{k=1}^{n}\alpha_k y_k\,(i_k * i - b)) \tag{3}$$

where n is the number of training records ($k = 0, ..., n$), y_k, which is either -1 or $+1$, is the label label of training data, i_k are the support vectors and $0 \leq \alpha_k \leq C(constantC > 0)$, and b is a bias value.

2) ***Dataset Description****:* The BoT-IoT dataset [56] has been selected to conduct the experiments in this paper. The dataset is released by the Cyber Range Lab of UNSW Canberra and created in a realistic network environment, including normal and malicious traffic. The original dataset includes more than 72 million records with 46 features. However, a reduced version in which there are about 3,000,000 samples (5% of the original dataset) including best features has also been presented. The BoT-IoT dataset is imbalanced, i.e., the number of benign samples is significantly less than attacks in this dataset. Therefore, for the experiments, we have chosen to work on a smaller balanced dataset. In this way, the performance in terms of the detection rate is not significantly affected for two classes of data. The number of samples chosen for experiments is 20000 including 6481 DoS, 6482 DDoS, 6481 Reconnaissance, 79 Theft, and 477 benign samples. The target dataset is generated through sampling of the 5% dataset (mentioned earlier). As the 5% dataset is divided into training and testing sets, we have extracted samples from both sets while considering Theft/benign samples. On the other hand, just the training set (from the 5% dataset) has been considered for extracting DDoS, DoS, and Reconnaissance samples.

3) ***Data Pre-processing****:* For data pre-processing, 19 Features were selected. However, we have omitted six features as they can identify samples uniquely [57]. Further, MinMaxScaler() [58] is used to normalize the value of features between 0 and 1 as data scaling can affect the performance of ML models [59].

B. Applying Label Flipping on Samples with Shortest or Largest Distances to the Hyperplane of SVM

The percentage of samples selected to flip their labels range from 5% to 50% of the dataset's size. The increase is 5% for each turn. Two different SVM kernels, including linear and RBF, have been examined. To have a relative distance to the hyperplane, the decision_function(X) is used [60] (a model is trained by the training set first, then the training set is fed to the decision_function(X) of the trained model in order to acquire the relative distance). After measuring the samples' distances to the hyperplane, they are ordered based on the absolute value of the distance [57]. After acquiring the samples (with short or large distances to the hyperplane), their label is flipped and the model is trained.

The first experiment is conducted for two cases:

- Select samples having shortest distances to the hyperplane and flip their labels for the linear and RBF kernels.
- Select samples having the largest distances to the hyperplane and flip their labels for the linear and RBF kernels.

As discussed in [61], applying label flipping on samples with a large distance to the hyperplane should affect the model more significantly. However, no considerable difference in the model's performance was observed while flipping for two cases. But, applying label manipulation on samples with largest distances to the hyperplane has relatively more negative impacts, which can be seen in Fig. 3, confirming [61].

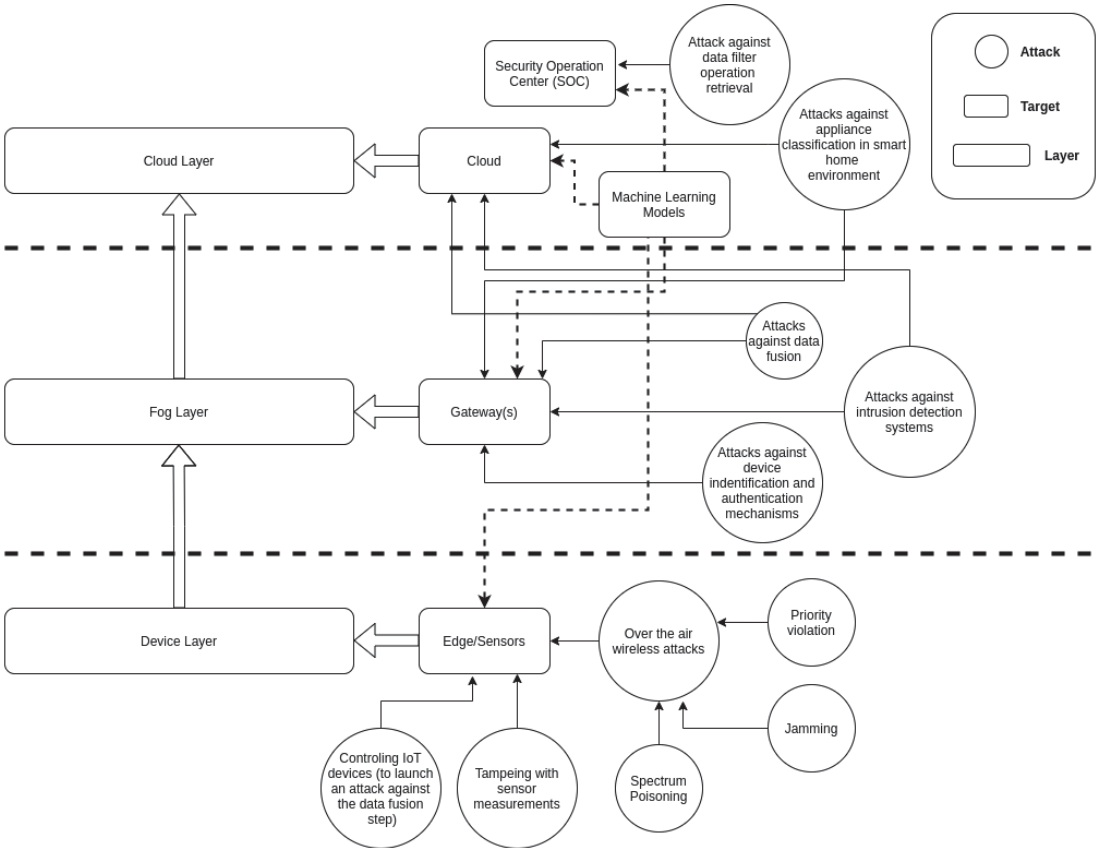

Fig. 1. A proposed classification of Adversarial Machine Learning (AML) in IoT

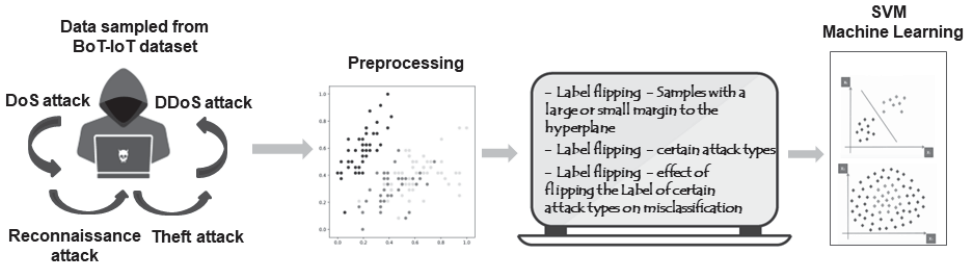

Fig. 2. The proposed framework of the study

Another study [62] emphasizes that SVM with RBF kernel is more robust against adversarial examples for the case of input data manipulation poisoning. While considering the results for the linear Kernel, the label noise applied on samples having the largest distances to the hyperplane has caused more adverse effects compared to RBF in general (there are some noise levels where the adverse effect is more for the RBF kernel). However, overall, the difference is not considerable. It can be assumed that the inconsistency may be related to the distribution of data or parameter settings. We also believe that the similarity between results in cases of samples with shortest distances and largest distances to the hyperplane is due to the fact that samples are very close in the feature space. In other words, the distance between the closest samples and furthest samples is negligible. The results corresponding to the first experiment are depicted in Fig. 3.

Tables II to V show the results for accuracy, precision, recall, and F1-score. In contrast with the accuracy, recall, and F1-score metrics, it can be seen that there is no decrease in the values of precision. As the number of false positives increases, the value of precision decreases. In the dataset used for this study, one label is in minority against the other one (same as most of the real-world intrusion detection datasets in which attack samples are fewer than benign samples). However, in this dataset, the number of benign samples are considerably fewer than attacks. The benign and attack categories are considered negative and positive types respectively. Therefore, as the number of benign samples is significantly fewer than attack

instances, the algorithm makes more errors when predicting the label of attack samples resulting in an increase in the number of false negatives. Moreover, the model does not have difficulties in detecting benign samples (which results in a reduction in the number of false positives). Therefore, as the recall and precision metrics depend on the number of false negatives and false positives respectively, the values of recall decrease whereas the values of precision remain relatively unchanged when applying label flipping.

C. Applying Label Flipping on Samples of a Specific Attack Category

In this attack scenario, the aim is to analyze the effect of label flipping applied on samples belonging to a specific attack category in a dataset that includes different attack types, i.e., instead of manipulating the label of just attack or benign samples or samples from different categories together, samples of specific attack types are targeted, and results are analyzed. In other words, samples are first chosen (with respect to the desired label noise percentage) having the largest distances to the hyperplane with RBF kernel and then, only samples from a specific category are flipped in turn. There is also the "Theft" attack category. However, as the number of samples for this attack are too few, it is not considered for this experiment.

Fig. 4 shows the achieved results for this experiment. In the case of DDoS, a small performance degradation is witnessed. While flipping just DoS samples caused a more major negative effect comparing to DDoS, the results demon-strate a significant change in performance for the case of Reconnaissance when more than 25% label noise is applied. Tables VI to VIII show the results for accuracy, precision, recall, and F1-score. As discussed in section IV-B, the values of precision remain unchanged as the number false positive are negligible.

The difference in results is related to the different positions of attack samples in the feature space. For example, it can be concluded that as DoS samples are located further from the hyperplane compared to the DDoS samples (we have ordered all the samples in the dataset based on their relative distance to the hyperplane, samples with the largest distances at the first of the list), more of them are chosen for label flipping and therefore, the model's performance is decreased more.

D. Investigating the Effect of Label Flipping Applied on Samples Having a Specific Attack Category on Type of the Misclassified Samples

In this experiment, the effect of label flipping on misclassified samples is analyzed. The aim is to identify which samples (from which category) are misclassified more when label flipping is applied just on a specific attack type. To this end, the attack categories of misclassified samples are recorded following the steps mentioned in the previous experiment. Then, they are compared with the target attack category for which labels are flipped. It is worth mentioning that 5-fold cross validation is used for all the experiments. For each fold, the noise is applied on the training set and the average is recorded for each noise level (after completion of five folds). Moreover, the parameters of SVM are default.

TABLE II. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EX-PERIMENT I, FLIPPING SAMPLES WITH SHORTEST DISTANCES TO THE HYPERPLANE WITH THE LINEAR KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	98	99	99	99
5%	95	98	96	97
10%	94	99	94	97
15%	89	99	89	94
20%	82	100	82	90
25%	79	100	78	88
30%	76	100	76	86
35%	76	100	75	85
40%	58	100	57	73
45%	55	100	54	70
50%	53	100	52	69

TABLE III. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EX-PERIMENT I, FLIPPING SAMPLES WITH LARGEST DISTANCES TO THE HYPERPLANE WITH THE LINEAR KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	98	99	99	99
5%	97	97	100	98
10%	89	97	91	94
15%	82	97	83	90
20%	76	96	78	86
25%	72	96	73	83
30%	66	96	68	80
35%	61	96	63	76
40%	54	95	55	70
45%	49	95	50	66
50%	46	95	47	63

For analysis, the total number of samples from each attack category whose labels are flipped, the number of misclassified samples from the same attack category as the chosen attack, and the number of misclassified samples from other attack categories for all five folds are counted. The presented numbers are the summation for all folds (the red, green, and blue lines). As depicted in Fig. 5, the results obtained using RBF kernel highlight that when more label noise is applied on samples of any attack category, the number of misclassified samples of the same attack type increases as well.

The results can lead to designing more effective attack scenarios. In other words, when the priority of an adversary is to reduce the performance of an intrusion detection system in terms of detecting a specific attack type, he can apply label flipping on samples of the same attack category in the training set. On the other hand, when designing defense strategies, if detecting a specific attack category poses greater importance, the security team can put more emphasis on protecting the samples of the same attack category in the training set.

V. DEMONSTRATING THE WEAKNESSES OF THE K-NN METHOD TO DETECT AND CORRECT LABEL NOISE IN DATASETS

Many defense mechanisms have been proposed in the literature to tackle the label flipping attack. Several data cleaning approaches are based on K-NN. However, these methods may not work well for all cases. In this section, we evaluate the robustness of a defense method that utilizes the K-NN model

TABLE IV. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EXPERIMENT I, FLIPPING SAMPLES WITH SHORTEST DISTANCES TO THE HYPERPLANE WITH THE RBF KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	99	99	99	99
5%	97	99	98	98
10%	92	98	93	96
15%	85	98	86	92
20%	81	98	82	89
25%	76	98	77	86
30%	71	98	72	83
35%	67	97	67	80
40%	61	97	61	75
45%	56	97	56	71
50%	51	96	52	67

TABLE V. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EX-PERIMENT I, FLIPPING SAMPLES WITH LARGEST DISTANCES TO THE HYPERPLANE WITH THE RBF KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	99	99	99	99
5%	95	99	95	97
10%	89	98	89	94
15%	84	99	85	91
20%	80	99	81	89
25%	73	98	73	84
30%	65	98	65	78
35%	63	99	63	77
40%	58	98	58	73
45%	54	98	53	69
50%	48	98	47	64

TABLE VI. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EXPERIMENT II, FLIPPING JUST DDOS SAMPLES HAVING LARGEST DISTANCES TO THE HYPERPLANE WITH THE RBF KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	99	99	99	99
5%	97	99	98	98
10%	97	99	97	98
15%	97	99	97	98
20%	97	99	97	98
25%	97	99	97	98
30%	96	99	97	98
35%	96	99	97	98
40%	96	99	96	98
45%	95	99	96	97
50%	93	99	93	96

TABLE VII. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EXPERIMENT II, FLIPPING JUST DOS SAMPLES HAVING LARGEST DISTANCES TO THE HYPERPLANE WITH THE RBF KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	99	99	99	99
5%	97	99	97	98
10%	96	99	96	98
15%	92	99	92	96
20%	91	99	91	95
25%	91	99	91	95
30%	91	99	91	95
35%	90	99	91	95
40%	88	99	88	93
45%	85	99	85	91
50%	81	99	81	89

TABLE VIII. ACCURACY, PRECISION, RECALL, AND F1-SCORE - EXPERIMENT II, FLIPPING JUST RECONNAISSANCE SAMPLES HAV-ING LARGEST DISTANCES TO THE HYPERPLANE WITH THE RBF KERNEL

Noise(%)	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
0%	99	99	99	99
5%	99	99	99	99
10%	99	99	99	99
15%	96	99	97	98
20%	95	99	96	97
25%	91	99	91	95
30%	78	99	78	87
35%	77	99	77	87
40%	76	99	76	86
45%	74	99	74	85
50%	74	99	74	84

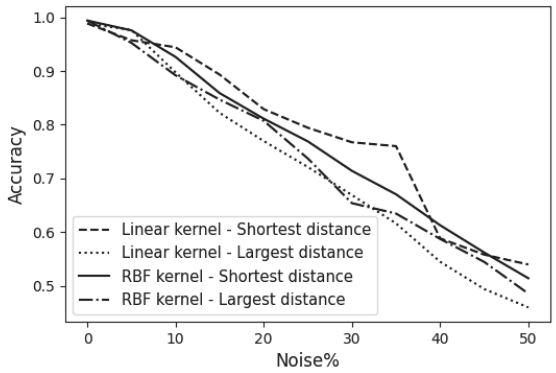

Fig. 3. The results for scenario I including label flipping applied on samples with large or small margins to the hyperplane

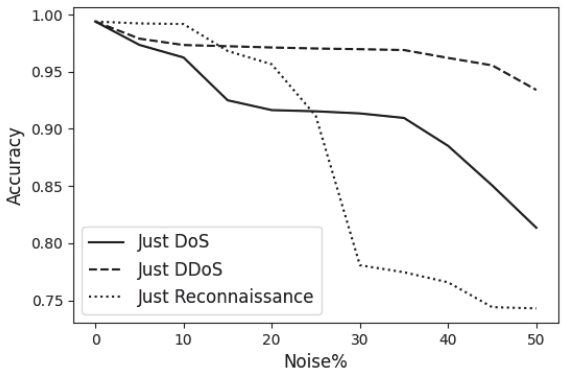

Fig. 4. Using RBF kernel while flipping label of samples from specific attack categories having large distances to the hyperplane

against the random label flipping attack. While applying the label flipping attack on a training set, many poisoning points will be far from the true samples (non-poisoning data points) having the same label. Using K-NN to mitigate the effect of label flipping attacks by relabelling malicious points is suggested in [13]. The procedure is as follows: for each sample in the training set, the K nearest neighbors are found based on the Euclidean distance. If most of the neighbors (according

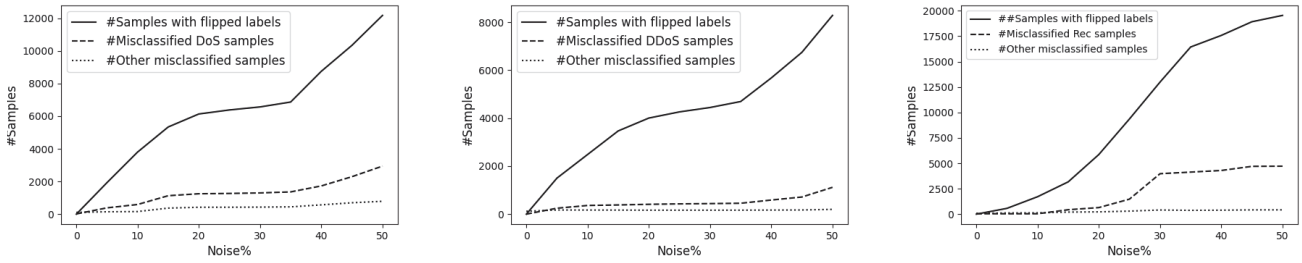

Fig. 5. The results for scenario III that include investigating the effect of flipping samples from specific categories on misclassified samples

to a threshold, ranging from 0.5 to 1) have a different label, the label of the sample is flipped. According to this paper, large and small values of k show better performance when the number of poisoning points is large and small respectively. For larger thresholds, the defense is less effective as fewer points are relabeled. This defense is designed for the case where specific data points are affected by label noise. However, The proposed approach has two drawbacks when noise is applied randomly:

- It is likely that some benign samples are located close to attack samples (and likewise, some attack samples may be surrounded by benign samples). Therefore, based on the proposed methodology, these genuine benign or attack data points will be relabeled.
- There may be cases where all the neighboring data points are poisoned. For example, a genuine benign sample is surrounded by poisoning points that are benign in nature, but their label is flipped as an attack (or the same on the other side, a genuine attack sample is surrounded by poisoning benign samples so that it is relabeled). Moreover, in some cases, a poisoning benign (attack) sample may be surrounded by other poisoned benign (attack) samples and as a result, it is not detected and relabeled.

Based on the paper, the first problem is likely to happen in the regions where benign and attack samples overlap (close to hyperplane for SVM). It is also discussed that (for the first problem) the fraction of benign and attack samples that are relabeled is expected to be the same. However, this problem is still discussed as a shortcoming in [41]. The experiments in this section are conducted on a dataset same as what was discussed in section IV-A2. For each data point in the data set, 10 neighbors (using Sklearn NearestNeighbors) are extracted.

The second problem (that happens when all the neighboring data points are poisoning) seems to be more important. In this case, when there is a genuine benign or attack data instance surrounded by poisoning points, the label of data instance is flipped wrongly. Moreover, there may be regions in the data space where label of all instances is flipped. In this case, the algorithm is unable to detect data instances with wrong labels. As an example, table IX shows the case in which a genuine attack data instance is surrounded by poisoning benign samples so that its label is flipped. The noise level here is 30%.

TABLE IX. A GENUINE ATTACK DATA INSTANCE IS SURROUNDED BY POISONING BENIGN SAMPLES SO THAT ITS LABEL IS FLIPPED

Neighbors	Neighbor Category
1^{st} Neighbor	DDoS (Attack in real)
2^{nd} Neighbor	DDoS (Benign is real)
3^{rd} Neighbor	DDoS (Attack in real)
4^{th} Neighbor	DDoS (Benign in real)
5^{th} Neighbor	DDoS (Attack in real)
6^{th} Neighbor	DDoS (Benign in real)
7^{th} Neighbor	DDoS (Benign in real)
8^{th} Neighbor	DDoS (Benign in real)
9^{th} Neighbor	DDoS (Benign in real)
10^{th} Neighbor	DDoS (Benign in real)

An experiment is conducted to show the weaknesses of the proposed defense for the random label flipping attack. This experiment includes looking for the genuine attack samples that are changed to benign wrongly, the genuine benign samples that are changed to attack wrongly, poisoned attack samples that are not detected, and poisoned benign samples that are not detected. The applied random label noise in this experiment is 30%. The results are demonstrated in table X. According to the results, the K-NN method is not a powerful defense mechanism against the random label flipping attack.

VI. CONCLUSION

Adversarial attacks have become concerns for ML models not only used in IoT but also in other domains. In this paper, a comprehensive review of the recent research works regarding Adversarial Machine Learning (AML) in the IoT domain is presented. Moreover, a classification of adversarial attacks in IoT is proposed to assist the researchers for their future works in this domain. As mentioned earlier, IoT environments are vulnerable to poisoning attacks such as label flipping. As another contribution of this paper, we have investigated the effect of label flipping on an IoT dataset in three different scenarios to determine how choosing samples to flip their labels contributes to more negative effects. These experiments can lead to designing more efficient attack and defense strategies. Based on the results, choosing different samples to flip their labels can cause distinct impacts. Moreover, it is observed that

TABLE X. RESULTS FOR THE EXPERIMENT AIMING AT DEMONSTRATING THE DATA INSTANCES SURROUNDED BY SAMPLES WITH POISONING LABELS.
A = GENUINE ATTACK DATA INSTANCES ARE CHANGED TO BENIGN.
B = ATTACK DATA INSTANCES THAT ARE FLIPPED TO BENIGN ARE NOT RELABELED.
C = BENIGN DATA INSTANCE THAT ARE FLIPPED TO ATTACK ARE NOT RELABELED.
D = GENUINE BENIGN DATA INSTANCE ARE CHANGED TO ATTACK.

Case	Round 1	Round 2	Round 3	Round 4	Round 5	Average
A	368	424	366	371	384	384
B	95	134	128	143	115	123
C	12	5	14	9	6	9
D	40	34	39	22	17	30

as the number of samples from one particular attack category whose labels are flipped increases, the number of misclassified samples from the same category increases as well. In terms of the defense against the random label flipping attack, the K-NN method did not show promising results. However, this defense mechanism is likely to be a good choice for cases where specific samples are chosen to flip their labels. Our future work would focus on other types of poisoning attacks that affect ML models.

ACKNOWLEDGMENT

The authors graciously acknowledge the support from the Canadian Institute for Cybersecurity (CIC), the funding support from the National Research Council of Canada (NRC) through the AI for Logistics collaborative program, the NSERC Discovery Grant (no. RGPIN 231074), and Tier 1 Canada Research Chair to Dr. Ghorbani. The authors also would like to thank Dr. Philippe Lamontagne for all his help.

REFERENCES

[1] W. Z. Khan, M. Rehman, H. M. Zangoti, M. K. Afzal, N. Armi, and K. Salah, "Industrial internet of things: Recent advances, enabling technologies and open challenges," *Computers & Electrical Engineering*, vol. 81, p. 106522, 2020.

[2] S. Sapre, P. Ahmadi, and K. Islam, "A robust comparison of the kddcup99 and nsl-kdd iot network intrusion detection datasets through various machine learning algorithms," *arXiv preprint arXiv:1912.13204*, 2019.

[3] McKinsey, "What's new with the inter-net of things," 2017. [Online]. Available: https://www.mckinsey.com/industries/semiconductors/our-insights/whats-new-with-the-internet-of-things

[4] Z. B. Celik, E. Fernandes, E. Pauley, G. Tan, and P. McDaniel, "Program analysis of commodity iot applications for security and privacy: Challenges and opportunities," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–30, 2019.

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.

[6] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[7] "The bot-iot dataset," https://research.unsw.edu.au/projects/bot-iot-dataset, accessed: 2021-11-4.

[8] S. Dadkhah, H. Mahdikhani, P. K. Danso, A. Zohourian, K. A. Truong, and A. A. Ghorbani, "Towards the development of a realistic multidimensional iot profiling dataset," in *2022 19th Annual International Conference on Privacy, Security Trust (PST)*, 2022, pp. 1–11.

[9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[10] P. Kantartopoulos, N. Pitropakis, A. Mylonas, and N. Kylilis, "Exploring adversarial attacks and defences for fake twitter account detection," *Technologies*, vol. 8, no. 4, p. 64, 2020.

[11] E. Tabassi, K. Burns, M. Hadjimichael, A. Molina-Markham, and J. Sexton, "A taxonomy and terminology of adversarial machine learning," *NIST IR*, 2019.

[12] X. Wang, J. Li, X. Kuang, T. yu an, and J. Li, "The security of machine learning in an adversarial setting: A survey," *Journal of Parallel and Distributed Computing*, vol. 130, 04 2019.

[13] A. Paudice, L. Muñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 5–15.

[14] Y. Wu, "Robust learning-enabled intelligence for the internet of things: A survey from the perspectives of noisy data and adversarial examples," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9568–9579, 2021.

[15] S. S. Swarna Sugi and S. R. Ratna, "Investigation of machine learning techniques in intrusion detection system for iot network," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 1164–1167.

[16] A. Ferdowsi and W. Saad, "Deep learning for signal authentication and security in massive internet-of-things systems," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1371–1387, 2019.

[17] W. Li, Q. Li, and R. Liu, "Iot devices identification based on machine learning," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 770–777.

[18] P. M. Mutescu, A. Lavric, A. I. Petrariu, and V. Popa, "Evaluation of a new spectrum sensing technique for internet of things: An ai approach," in *2022 International Conference on Development and Application Systems (DAS)*, 2022, pp. 91–94.

[19] H. Albataineh, M. Nijim, and D. Bollampall, "The design of a novel smart home control system using smart grid based on edge and cloud computing," in *2020 IEEE 8th International Conference on Smart Energy Grid Engineering (SEGE)*, 2020, pp. 88–91.

[20] A. Schwarzschild, M. Goldblum, A. Gupta, J. P. Dickerson, and T. Goldstein, "Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9389–9398.

[21] O. Suciu, R. Marginean, Y. Kaya, H. Daume III, and T. Dumitras, "When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1299–1316.

[22] O. Taran, S. Rezaeifar, and S. Voloshynovskiy, "Bridging machine learning and cryptography in defence against adversarial attacks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[23] J. Xu, Z. Cai, and W. Shen, "Using fgsm targeted attack to improve the transferability of adversarial example," in *2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE)*, 2019, pp. 20–25.

[24] A.-U.-H. Qureshi, H. Larijani, N. Mtetwa, M. Yousefi, and A. Javed, "An adversarial attack detection paradigm with swarm optimization," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

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

[26] M. Guarino, P. Rivas, and C. DeCusatis, "Towards adversarially robust ddos-attack classification," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2020, pp. 0285–0291.

[27] N. Ghosh, K. Maity, R. Paul, and S. Maity, "Outlier detection in sensor data using machine learning techniques for iot framework and wireless sensor networks: A brief study," in *2019 International Conference on Applied Machine Learning (ICAML)*, 2019, pp. 187–190.

[28] K. R. Dalal, "Analysing the role of supervised and unsupervised machine learning in iot," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 75–79.

[29] P.-Y. Chen, S. Yang, and J. A. McCann, "Distributed real-time anomaly detection in networked industrial sensing systems," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3832–3842, 2015.

[30] J. Branch, B. Szymanski, C. Giannella, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," in *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, 2006, pp. 51–51.

[31] T. Luo and S. G. Nagarajan, "Distributed anomaly detection using autoencoder neural networks in wsn for iot," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[32] Z. Bao, Y. Lin, S. Zhang, Z. Li, and S. Mao, "Threat of adversarial attacks on dl-based iot device identification," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[33] H. Sun, X. Wang, R. Buyya, and J. Su, "Cloudeyes: Cloud-based malware detection with reversible sketch for resource-constrained internet of things(iot) devices," *Software Practice and Experience*, vol. 47, 05 2016.

[34] A. A. Cardenas, T. Roosta, and S. Sastry, "Rethinking security properties, threat models, and the design space in sensor networks: A case study in scada systems," *Ad Hoc Networks*, vol. 7, no. 8, pp. 1434–1447, 2009.

[35] N. Baracaldo, B. Chen, H. Ludwig, A. Safavi, and R. Zhang, "Detecting poisoning attacks on machine learning in iot environments," in *2018 IEEE international congress on internet of things (ICIOT)*. IEEE, 2018, pp. 57–64.

[36] W. Ding, X. Jing, Z. Yan, and L. T. Yang, "A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion," *Information Fusion*, vol. 51, pp. 129–144, 2019.

[37] Z. Luo, S. Zhao, Z. Lu, Y. E. Sagduyu, and J. Xu, "Adversarial machine learning based partial-model attack in iot," in *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, 2020, pp. 13–18.

[38] Y. E. Sagduyu, Y. Shi, and T. Erpek, "Iot network security from the perspective of adversarial deep learning," in *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2019, pp. 1–9.

[39] M. Wang, C. Perera, P. P. Jayaraman, M. Zhang, P. Strazdins, R. Shyamsundar, and R. Ranjan, "City data fusion: Sensor data fusion in the internet of things," in *The Internet of Things: Breakthroughs in Research and Practice*. IGI Global, 2017, pp. 398–422.

[40] W. Yan, Z. Wang, H. Wang, W. Wang, J. Li, and X. Gui, "Survey on recent smart gateways for smart home: Systems, technologies, and challenges," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 6, p. e4067, 2022.

[41] R. Taheri, R. Javidan, M. Shojafar, Z. Pooranian, A. Miri, and M. Conti, "On defending against label flipping attacks on malware detection systems," *Neural Computing and Applications*, vol. 32, no. 18, pp. 14 781–14 800, 2020.

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

[43] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and B. Xi, "Adversarial support vector machine learning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1059–1067.

[44] Y. Shi, T. Erpek, Y. E. Sagduyu, and J. H. Li, "Spectrum data poisoning with adversarial deep learning," in *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*. IEEE, 2018, pp. 407–412.

[45] Y. Shi, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu, and J. H. Li, "Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2018, pp. 1–6.

[46] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 1, pp. 2–14, 2018.

[47] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, 2020, pp. 1–6.

[48] A. Singh and B. Sikdar, "Adversarial attack and defence strategies for deep-learning-based iot device classification techniques," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2602–2613, 2022.

[49] M. Sadeghi and E. G. Larsson, "Physical adversarial attacks against end-to-end autoencoder communication systems," *IEEE Communications Letters*, vol. 23, no. 5, pp. 847–850, 2019.

[50] Y. Sharaf-Dabbagh and W. Saad, "On the authentication of devices in the internet of things," in *2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2016, pp. 1–3.

[51] B. Kaur, P. Kumar, P. P. Roy, and D. Singh, "Impact of ageing on eeg based biometric systems," in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2017, pp. 459–464.

[52] D. Jing and H.-B. Chen, "Svm based network intrusion detection for the unsw-nb15 dataset," in *2019 IEEE 13th International Conference on ASIC (ASICON)*. IEEE, 2019, pp. 1–4.

[53] K. Tbarki, S. Ben Said, R. Ksantini, and Z. Lachiri, "Rbf kernel based svm classification for landmine detection and discrimination," in *2016 International Image Processing, Applications and Systems (IPAS)*, 2016, pp. 1–6.

[54] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

[55] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, "A bio-signal based framework to secure mobile devices," *Journal of Network and Computer Applications*, vol. 89, pp. 62–71, 2017.

[56] 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.

[57] P. Papadopoulos, O. T. v. Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, and W. J. Buchanan, "Launching adversarial attacks against network intrusion detection systems for iot," *Journal of Cybersecurity and Privacy*, vol. 1, no. 2, pp. 252–273, 2021.

[58] E. Bisong, "Introduction to scikit-learn," in *Building machine learning and deep learning models on Google cloud platform*. Springer, 2019, pp. 215–229.

[59] M. M. Ahsan, M. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021.

[60] scikit-learn developers, "sklearn.svm.linearsvc," https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC.decision_function, accessed: 2021-11-12.

[61] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," in *ECAI 2012*. IOS Press, 2012, pp. 870–875.

[62] P. Vidnerová and R. Neruda, "Evolutionary generation of adversarial examples for deep and shallow machine learning models," in *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, 2016, pp. 1–7.

[63] T. Lin, "Deep learning for iot," in *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*, 2020, pp. 1–4.

[64] Z. Yang, I. A. Abbasi, F. Algarni, S. Ali, and M. Zhang, "An iot time series data security model for adversarial attack based on thermometer encoding," *Security and Communication Networks*, vol. 2021, 2021.

[65] E. Anthi, L. Williams, A. Javed, and P. Burnap, "Hardening machine learning denial of service (dos) defences against adversarial attacks in iot smart home networks," *computers & security*, p. 102352, 2021.

[66] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing adversarial attacks against deep learning for intrusion detection in iot networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[67] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial attacks against network intrusion detection in iot systems," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 327–10 335, 2020.

TABLE XI. SUMMARY OF THE RESEARCH WORKS REVIEWED IN THE LITERATURE REVIEW SECTION

Author	Investigated Attack(s)	Related Categories	Used Models	Dataset	Description
Lin (2020) [63]	Attack against data filter operation retrieval in cyber security operation centers (SOCs)	Evasion	RNN	NA	• An RNN-based data retrieval method for IoT data analysis is proposed. • Data retrieval solutions are investigated to avoid adversarial hacking.
Yang et al. (2021) [64]	Attack against IoT time series data (FGSM attack is investigated)	Evasion	Encode-decode model - ResNet	Time series datasets in UCR archive including Coffee dataset	Encode-decode joint training model is proposed to construct a robust IoT classification model for time-series data.
Taheri et al. (2020) [41]	Label manipulation against android malware detection	Poisoning	K-means (as the clustering technique used in the proposed attack) - CNN	Drebin, Contagio, and Genome	• A label flipping attack is proposed using silhouette clustering. • Two defense methods are proposed.
Papadopoulos et al. (2021) [57]	Label flipping and FGSM attacks against IoT network intrusion detection system	Poisoning and Evasion	SVM - ANN	BoT-IoT	SVM and ANN-based network intrusion detection systems are evaluated against on label flipping and FGSM attacks.
Anthi et al. (2021) [65]	JSMA and FGSM attacks against the classifiers used for detecting Denial of Service (DoS) attacks in smart home IoT environment	Evasion	J48 Decision Tree, Random Forest, Naïve Bayes, Bayesian Network, SVM, Zero R, and One R	Data collected from an IoT testbed	A rule-based approach is proposed for generating AML attack samples.
ibitoye et al. (2019) [66]	FGSM, BIM, and PGD attacks against IoT network intrusion detection systems	Evasion	Self-normalizing Neural Network and FNN	BoT-IoT	• The adversarial robustness of Self-normalizing Neural Networks (SSN) and Feedforward Neural Networks (FNN) are evaluated and compared. • The effect of feature normalization on adversarial robustness of models is investigated.
Qui et al. (2020) [67]	Attack against Iot network intrusion detection systems (using iterative FGSM)	Evasion	Kitsune network intrusion detection system (the autoencoder part is replicated)	Mirai and VideoInjection	Proposed a new adversarial attack based on model extraction and saliency maps.