An Exploratory Analysis of Effect of Adversarial Machine Learning Attack on IoT-enabled Industrial Control Systems

Conference Paper · February 2023

DOI: 10.1109/ICSCA57840.2023.10087713

4 authors:

Sandeep Trivedi

25 PUBLICATIONS **169** CITATIONS

SEE PROFILE

Tien Anh Tran

Seoul National University

75 PUBLICATIONS **418** CITATIONS

SEE PROFILE

Nuruzzaman Faruqui

Daffodil International University

28 PUBLICATIONS **282** CITATIONS

SEE PROFILE

Md Maruf Hassan

Daffodil International University

30 PUBLICATIONS **218** CITATIONS

SEE PROFILE

An Exploratory Analysis of Effect of Adversarial Machine Learning Attack on IoT-enabled Industrial Control Systems

*Note: Sub-titles are not captured in Xplore and should not be used

Sandeep Trivedi
Deloitte Consulting LLP
Houston, United States of America (USA)
Email: sandeeptrivedi@ieee.org
ORCID: 0000-0002-1709-247X

Dr. Tien Anh Tran
Department of Naval Architecture and Ocean Engineering
Seoul National University
Seoul City, South Korea
Email: anhtt.mtb@vimaru.edu.vn

Nuruzzaman Faruqui
Department of Software Engineering
Daffodil International University
Daffodil Smart City, Dhaka, Bangladesh
Email: faruqui.swe@diu.edu.bd, ORCID: 0000-0001-9306-9637

Mr. Md. Maruf Hassan
Department of Software Engineering
Daffodil International University
Daffodil Smart City, Dhaka, Bangladesh
Email: maruf.swe@diu.edu.bd
ORCID: 0000-0002-4475-2664

***Abstract*—Machine Learning (ML)-based Intrusion Detection Systems (IDS) is an effective technology to automatically detect cyber attacks in the Internet of Things (IoT) dependent Industrial Control Systems (ICS). It is faster, more efficient, and can detect attacks without human intervention. However, ML-based IDSs have introduced another security threat called Adversarial Machine Learning (AML). An AML attack may cause severe industrial infrastructural and production damage resulting in substantial financial loss. This paper presents an exploratory analysis of initiating an AML attack using adversarial samples created using a Fast Gradient Sign Method (FGSM). The research presented in this paper has been conducted from a dataset generated from a full-fledged singular module of a power distribution industry controlled by IoT-enabled ICSs. We explored the AML attack on Gradient Boosting (GB) and Iterative Dichotomiser 3 (ID3) model and discovered the average classification accuracy, precision, recall, and F1-scores are 87%, 88%, 87.5%, and 87%, respectively. The AML attack reduces the average precision, recall, and F1-score by 20.5%, 20.5%, and 22.5%, respectively, when 50% perturbations are added to 10% samples.**

***Index Terms*—Gradient Boosting, Iterative Dichotomiser 3, Adversarial Machine Learning, Intrusion Detection System, Internet of Things, Industrial Control System, Adversarial Samples.**

I. Introduction

Modern manufacturing industries are controlled by embedded systems [1]. Critical infrastructure, for example, power generation, power distribution, telecommunication Base Transceiver Station (BTS), natural resources refineries, etc., are controlled by IoT-enabled embedded systems known as Industrial Control Systems (ICS) [2]. We live in an interconnected world connected to each other through Web 2.0-enabled services [3]. And machines are interconnected through IoTs, which control and monitor the machines remotely [4]. The application of IoT also facilitates the integration of data analytic-based automation and optimization [5]. Anything connected to the internet is subject to cybersecurity vulnerabilities, including the ICSs [6]. Due to their widespread use and critical nature, these systems have become prime targets for cybercriminals [7]. As these systems regulate real-world processes, any cyber attacks on them might have far-reaching effects on the communities in which they function and the people who live in them [8]. It is, therefore, not surprising that concerns over the safety of these devices have gained international attention. As a result, it is more crucial than ever to develop foolproof, secure, and effective methods of monitoring and protecting ICS networks against cyber threats [9].

The ICSs are different from traditional IT systems [10]. These devices are specially designed to consume minimal energy. As a result, most of these have resource-constrained architecture restricting their capability to detect and defend against cyber attacks [11]. However, these embedded systems are connected to the physical infrastructure responsible for manufacturing and production. The scenario is more crucial and sensitive in a grid where power generation and distribution are controlled using network-connected ICSs [11]. Any attack on these systems can cause devastating results like nationwide blackouts causing significant damage to almost every sector [12]. That is why 24×7 monitoring to detect anomalies and defend against attacks immediately is essential for ICSs. Machine learning models are promising solutions that mimic human intelligence, recognize the signal pattern between the

Identify applicable funding agency here. If none, delete this.

ICS receives, classify them into malignant and benign classes, and defend against cyber attacks [13].

ML models trained to identify malicious and anomalous signals strengthen the security of ICSs against cyber attacks. However, a successful attack on the trained model compromises the overall security of the system [14]. Systematic attacks on trained machine-learning models employed to detect anomalies and identify cyber attacks are called Adversarial Machine Learning (AML). In AML, the weaknesses of the trained models are exploited by introducing data perturbations. Data perturbation adds noise to data carefully chosen through rigorous feature engineering so that the ML model classifies malicious or anomalous data as benign [15]. Successful misclassifications of security incidents, including but not limited to information disclosure, production loss, financial loss, infrastructural damage, and delayed detection. The purpose of using ML models is to reduce the amount of human intervention [16]. A successful AML attack can go undetected for a lengthy period of time, leaving devastating effects [17]. This is why exploring AML attacks and analyzing the potential defense mechanism is essential in modern IoT-controlled ICSs.

To our best knowledge gained from the literature review, this is the first exploratory analysis of AML attacks on Gradient Boosting (GB) and Iterative Dichotomiser 3 (ID3) using a dataset obtained from an actual full-fledged singular module of a power distribution industry controlled by IoT-enabled ICSs under testbed mode. The core contributions of this research are:

- Exploration of the intrusion detection effectiveness of Gradient Boosting (GB) and Iterative Dichotomiser 3 (ID3) models against AML attacks.
- Adversarial sample generation and effectiveness analysis.
- Studying the effect of adversarial samples on trained machine learning models.

The rest of the paper has been organized into five sections. The literature review has been presented in the second section. The third section discusses the testbed. The methodology has been studied in the fourth section. The experimental evaluation and result analysis are presented in the fifth section. Finally, the paper has been concluded in the sixth section.

II. Literature Review

Machine Learning algorithms are being adopted in different sectors, including but not limited to industrial automation [18], banking [19], education [20], healthcare [21], Human Resource Management (HRM) [22], agriculture [23], telecommunication [24], Web technology [25], mobile application [26], and many other sectors. Recently, there has been significant growth in machine learning-based intrusion detection systems for various ICS systems [27]. The literature review, presented in table I, summarizes different ICS systems and accompanying machine learning techniques to attack detection and categorization in various environments. it clearly shows the technological maturity of delegating human responsibilities of ICS management to ML models, including intrusion detection. However, it makes the system vulnerable to AML. The current published scientific literature on AML is still emerging, and there are many fields with the potentially vulnerable to AML [15].

Machine learning is all about the quality of the dataset [28]. A flaw in the dataset can impact the overall performance of the model. The characteristics of a trained model depend on the feature of the training data [29]. By altering a tiny proportion of the original training data, an adversary may exploit and effectively circumvent the machine learning algorithms used in spam filters [30]. Furthermore, R. Yumlembam et al. assess the resilience of an artificial neural network trained on the DREBIN Android malware dataset [31]. They find that simply changing a tiny number of characteristics in the training set it is easy to confuse the model. That means anyone with access to the training data can intentionally inject a set of instances that allows him to trick the model and take unauthorized access to resources. Especially malicious insiders can analyze the training data feature through feature engineering and introduce tolerance to particular inputs to the model [32].

A recent study on hostile malware as a service conducted by A. Lanz et al shows that it is possible to build thousands of adversarial applications within minutes. It is a service that constantly monitors thousands of nodes and extracts features to enrich its database. With the current enriched database, it can discover the vulnerability of the software and machine learning models and perform AML attack [33]. Another research conducted by S. Chen et al. uncovers some advanced adversarial attack strategies. Their research shows that without knowing the data features and characteristics of the machine learning models, it is possible to attack and bypass the models assigned for intrusion detection [34]. In their study, E. Alshahrani et al. use actual adversarial attacks on network intrusion detection systems. The purpose is to identify botnet traffic using machine learning classifiers. However, the adversarial samples added to the dataset during the training caused the model to fail to identify the botnet traffic. As a result, the experimenting web server crashes [35]. These observations prove that the Adversarial Machine Learning (AML) is being applied in many different sectors. It is high time we explored more about the effect of it.

The AML is a threat to any machine learning models which training dataset features or relevant information are disclosed [36]. It is easier for malicious insider to access the required information. Many research has been conducted on AML in different sector considering malicious insider the attacker. However, the effect of AML is still ignored despite few attempts [37]. I. Alarab et al. demonstrated a straightforward AML attack on a Long Short-Term Memory (LSTM) classifier using an ICS dataset [38]. It shows how a well-trained LSTM classifier with acceptable classification accuracy fails to classify anomalies. However, this research uses manually developed handcrafted adversarial samples. It is beyond the scope of this approach to attack sophisticated systems. That is why it is considered as the early stage application of AML [38]. The recent advanced approaches show that even multi-layer security layer fails when ML based

systems are under AML. That is why it is essential to analyze the effect of adversarial machine learning to understand its characteristics.

III. Testbed: A Full-fledged Singular Power Unit

The experiment has been conducted on a full-fledged singular unit of a power distribution industry controlled by IoT-enabled ICS. It has been illustrated in figure 1. There are multiple units in the industry. A fully functional unit was isolated from experimenting with the approval of the respective authority as a testbed.

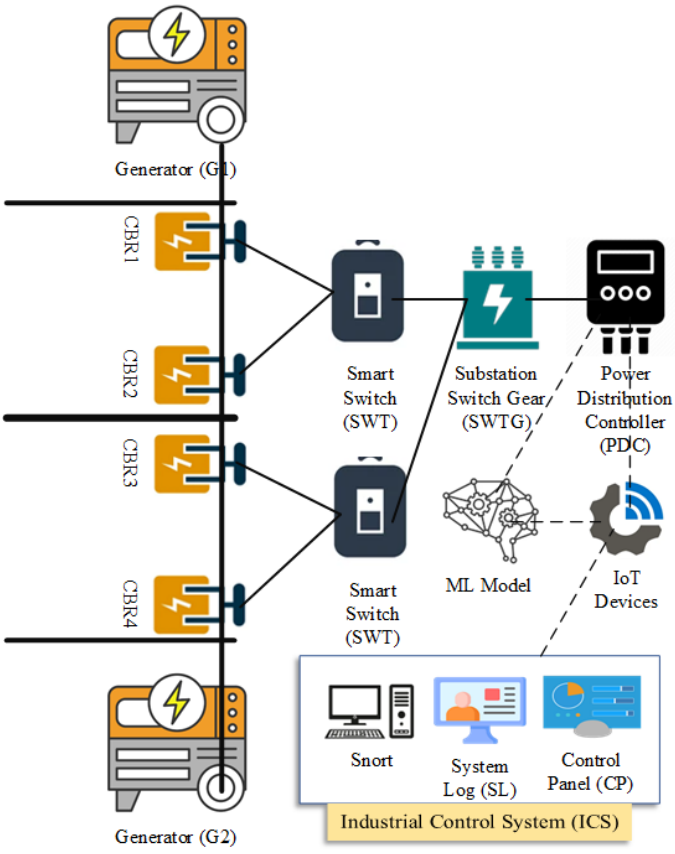

Fig. 1. Dataset and Pre-processing

The elements of the testbed, their full forms, and roles are listed in table II. There are two generators in the system. The power flow from the generator is protected by circuit breakers. There are four circuit breakers in the experimenting unit. These circuit breakers are controlled by smart switches. These switches are designed to protect the system from excessive current flow. The smart switches are connected to substation switch-gear. It can control any specific power generation node [47]. The Power Distribution Controller (PDC) uses the switch gear to control the power flow [48]. It is connected to IoT and a trained machine learning model. The ICS is connected to the PDC through an IoT device. It consists of a snort, a system log, and a control panel.

IV. Methodology

To explore how well supervised classification algorithms can learn to detect cyber attacks in an ICS environment, the performance of supervised machine learning when the corresponding data discussed in Section 4.1 was used to train the classification model and evaluated. The following Sections report the features present in the power systems dataset, as well as describe the methodology behind selecting and training the best-performing supervised classifiers.

A. Dataset

The dataset has been prepared from the testbed. It has two classes - malignant and benign. The benign class includes two types of instances. They are 'usual signal' and 'natural signal.' The usual signal refers to the regular activities with the usual amplitude and frequency of current and voltage. It also includes regular phase shifting. Sometimes natural events cause surge voltage and current, phase shift, and distorted power flow. These are also considered elements of the benign class. The malignant instances have been created by observing the behavior of the system for five different types of attack listed and described in table III

The dataset has been recorded using the System Log (SL) application. During the data collection period, the existing machine-learning model was disconnected from the system to understand the system behavior for particular attacks. The final dataset contains a total number of 60,830 instances. Among these instances, 25,126 belong to the malicious class, and 35,704 belong to the benign class. In the malicious class, the numbers of instances representing $SC_A, FM_A, RC_A, SM_A, and IJ_A$ are 4,249, 5018, 4957, 5183, and 5719, respectively. Figure 2 illustrates the different types of instances of the malignant class.

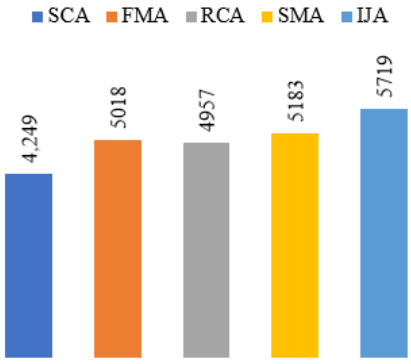

Fig. 2. Different types of instances of malignant class

B. Feature Engineering & Dataset Processing

Any machine learning model learns from the features of the dataset [49]. In feature engineering, we explored the features relevant to classifying malignant signals [50]. The analysis is entirely based on electrical signals. The pieces of equipment on the testbed are electrical components triggered

TABLE I
RECENT RESEARCH ON INTRUSION DETECTION IN INDUSTRIAL CONTROL SYSTEMS (ICSS)

Author(s)	Year	Dataset	Machine Learning Model
A. Eirini [39]	2021	Power Generation	Random Forest and J48
D. Wang et al. [40]	2019	Power System	Random Forest
E. Hoxha et al. [41]	2019	Wind Turbines	SVM
J. Gao et al. [42]	2019	SCADA Testbed	Long Short Term Memory (RNN)
SD. Anton et al. [43]	2018	Power system (synthetic)	Naive Bayes, Random Forests, SVM
RL. Perez et al. [44]	2018	Gas Pipeline	SVM, Random Forest
LA. Maglaras et al. [45]	2018	SCADA Testbed	Random Forest, J48, Logistic Regression, Naive Bayes
I. Abdallah et al. [46]	2018	Wind Turbine	Decision Trees (J48, Random Forest, CART, Ripper, etc.)

TABLE II
ELEMENTS OF TESTBED AND THEIR DESCRIPTION

Sequence	Symbol	Full Form	Role
1	G	Generator	Generating power
2	CBR	Circuit Breaker	Breaking the circuit if current overflows
3	SWT	Smart Switch	Controlling the circuit breaker
4	SWTG	Substation Switch Gear	Switch to control power distribution
5	PDC	Power Distribution Controller	Controlling the SWTG
6	Snort	Source Network Intrusion Detection System	Analyzing real-time network traffic
7	SL	System Log	Maintaining system log
8	CP	Control Panel	Displaying the controll panel dashboard
9	ML Model	Machine Learning Model	Automating intended tasks
10	IoT Devices	Internet of Things Devices	Connecting devices with internet

TABLE III
TYPES OF ATTACKS AND THEIR DESCRIPTIONS

Serial	Attack	Description
SC_A	Short-circuit	It triggers the circuit breaker to break the circuit and cause power supply interruption.
FM_A	False Maintainance	It turns off one or more power line by sending false maintenance signal.
RC_A	Remote Command Injection	It is a remote access attack allows attacker to command the PDC and control the power distribution
SM_A	Settings Manipulation	The attackers manipulate the ICS settings in this type of attacks.
IJ_A	Injection Attack	In this attack, the attackers inject manipulated values of voltage, current, and power to misguide the human operators

TABLE IV
THE FEATURES AND THEIR DESCRIPTIONS

Symbol	Feature	Description
V	Voltage	The phase angle of the voltage stored in capacitor
i_p	Phase Angle of Current	The phase angle of the AC current
i_m	Amplitude of Current	Amplitidue of AC current
v_p	Phase Angle of Voltage	The phase angle of the AC voltage
v_m	Phase Angle of Voltage	Amplitude of AC voltage
I	DC Current Polarity	Directory of the flow of DC current
I_m	DC Current Amplitude	Amplitude of DC current
F_r	Relay Frequency	Operating frequency of the relay
F_{slope}	Rate of Change of frequency	The first derivative of the relay frequency with respect to time
Z	Apparent Impedance for Relays	The apparent impedance of the relay exclusive internal resistance
Z_p	Apparent Phase Angle for Relays	The phase angle of the apparent impedance
R_s	Relay Status	Current status of the relay

and controlled by different attributes of electrical signals listed in table IV. These attributes are the features used to train the machine learning model to classify attacks on ICSs.

During studying the features, it has been observed that many instances on the dataset contain outliers. These outliers do not represent any attacks mentioned in III. That is why these instances have been removed. After cleaning the dataset, it has been observed that it is possible to create a balanced dataset from the cleaned dataset. We randomly selected 20,000 malignant and 20,000 benign instances. The malignant instances are well-balanced as well. There are 4,000 instances of each type of attack. The dataset balance instance ratio is illustrated in figure 3.

The final dataset maintains binary distribution. The instances of the malignant class are also distributed with a unity ratio. We used a training and testing ratio 65 : 45 to train and

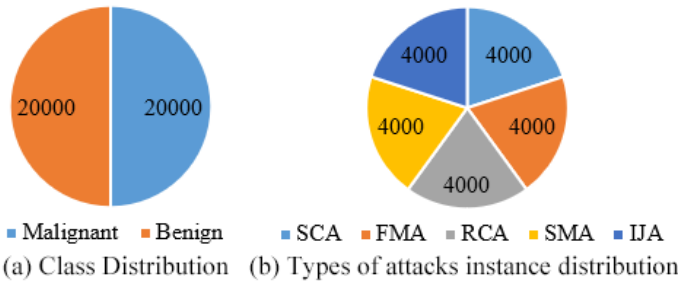

Fig. 3. The dataset distribution balance

test the models. The $k-$fold cross-validation has been used to validate the model's performance at $k = 5$.

C. Model Training

We used Gradient Boosting (GB) [51] and Iterative Dichotomiser 3 (ID3) [52] machine learning algorithms to train the model to classify the control signals of ICSs into malignant and benign classes. It has been observed that the effects of GB and ID3 in AML attacks are superficially explored. This paper aims to explore the impact of these two ML models in AML attacks. We used the dataset obtained from the full-fledged singular unit of a power distribution industry controlled by IoT-enabled ICSs illustrated in figure 1.

1) Gradient Boosting (GB): The dataset is labeled dataset. The training instances are defined by 1.

$$Traininginstance, D_t = \{(x_i, y_i)\}_{i=1}^{n} \tag{1}$$

A differentiable loss function, $L(y, F(x))$, has been used to train the model. The $F_0(x)$ is defined by 2. This equation initializes the model with a constant value.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{2}$$

Later, the pseudo-residuals are calculated for all instances using equation 3 where m represents the current instance.

$$pseudoresiduals, r_\gamma = -[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}]_{F(x)-F_{m-1}(x)} \tag{3}$$

The GB model training depends on the computational multiplier λ_m for every instance. In our approach, it has been solved using a one-dimensional optimizer defined by equation 4.

$$\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{4}$$

Finally, the model is updated using equation 5, which is repeated for M instances.

$$F_m(x) = F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{5}$$

Once the model is trained, the GB model is trained, it is expressed as $F_M(x)$, where M represents the number of instances the model was trained with.

2) Iterative Dichotomiser 3 (ID3): The initial state of the ID3 algorithm begins by considering the ground truth of any training instance $\{(x_i, y_i)\}_{i=1}^{n}$. It is considered as the root node and denoted by S. In every iteration, it calculates the entropy of unused instances defined by equation 6.

$$H(S) = \sum_{x \in X} -p(x) log_2 p(x) \tag{6}$$

Here on equation 6, X is the set of classes in S, $p(x)$ is the ratio of the number of instances in X and number of instances in S.

3) Model Performance: The models have been designed and trained using Python 3.10, Scikit-Learn 1.1.2, Matplotlib 3.5.2, and other libraries. The models were deployed in a conda environment hosted on a computer having 8192 MB of primary memory running on Windows 10 Operating System (OS) with a 3.60 GHz Intel(R) Core(TM) i3-9100 CPU. The performance of the models has been evaluated using the state-of-the-art machine learning evaluation metrics [53] accuracy, precision, recall, and F1-score, which are defined by equation 7, 8, 9, and 10, respectively, and listed in table V.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1-score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

TABLE V
THE PERFORMANCE OF THE TRAINED MODELS

Classifier	Accuracy	Precision	Recall	F1-score	Time (s)
GB	0.89	0.9	0.91	0.87	30.02
ID3	0.85	0.86	0.84	0.87	24.55

The experimental result shows that the accuracy of GB and ID3 are 89% and 85%, respectively. The precision of these two models is 90% and 86%, respectively, which indicates the quality of the positive prediction is acceptable. The performance visualization graph illustrated in figure 4 shows that the F1-score of both models is the same. However, the accuracy, precision, and recall of GB are better than ID3.

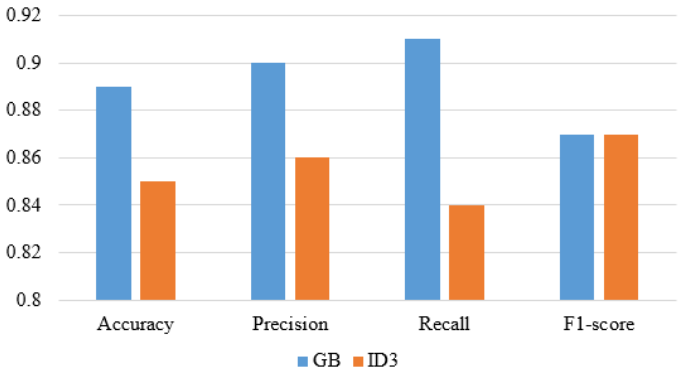

Fig. 4. The classification performance visualization

D. Attacker Model

Within the scope of our investigation, we examine an insider threat attacker with administrative access to the systems. Insider threats are among the most overlooked yet pose a considerable danger to ICSs. Identifying and managing insider risks is a complex and time-consuming process [54]. This is because insiders usually have privileged access to the network and sit beneath enterprise-level security defensive mechanisms. The adversary has access to both the dataset and its characteristics in our case. As an insider, the attacker is aware of the features used to train the ML model. However, due to the black-box nature of ML models, the attacker does not know the algorithm settings. The goals of the attackers are:

- Attack the ICS to cause service interruption,
- Misconfiguring the relays to make the system vulnerable, and
- Injecting false values of voltage, current, and power.

E. Adversarial Sample Generation Methods

Generating the adversarial samples is challenging. The AML method uses adversarial samples to perform the attack [55]. The more effective the samples are, the higher the probability of successful attacks. It is possible to produce adversarial samples using a wide variety of different approaches [56]. The level of complexity, the rate at which they generate results, and the level of performance offered by such approaches vary. The manual manipulation of the input data points is a simple method for producing such samples, but it is not an advanced one. On the other hand, manually perturbing massive datasets is a time-consuming process that may result in less accurate results [57]. Methods with a higher level of sophistication may involve the automatic analysis and identification of characteristics that provide the best discrimination between target values. The values that these features reflect are disrupted in a discrete manner such that they reflect values that are comparable to those that represent target values that are not their own.

We have used the Fast Gradient Sign Method (FGSM) to generate the adversarial samples [58]. The FGSM attack depends on adding noises to original samples X defined by equation 11.

$$X^* = X + \delta \tag{11}$$

Here on equation 11, the X^* is the adversarial samples and δ is the noise. These modified samples are classified by the trained ML models differently. These samples are further enhanced using a pre-trained Multilayer Perceptron (MLP) model. After generating the samples, the FGSM is applied, which is governed by equation 12.

$$x^* = x + \epsilon sign(\Delta_x J(\theta, x, y)) \tag{12}$$

Here is equation 12 the J is the cost function. The gradient of the cost function is used to compute the amount of noise. Here, the inputs are x, ϵ is the amount of noise, and y is the label of the ground truth. The θ in the equation is the parameters of the model.

It is more likely that the ML models will be affected if features are changed. To be more exact, an initial proportion of features, denoted by θ, is selected to be disrupted by an amount of noise denoted by λ. Thirdly, the model determines whether or not the extra noise has caused the targeted model to misclassify. This determination is made by analyzing the results of the previous two steps. If the model's performance has not been negatively impacted by the noise, a new collection of features will be chosen, and the iteration process will continue until an effective adversarial sample has been created.

V. Experimental Evaluation and Results

A. Original Samples

In the first phase of the experiment, we randomly selected instances from the dataset and evaluated the network's performance using the confusion matrix illustrated in figure 5. The instances of the dataset were unaltered. No adversarial samples were used in this phase.

The precision, recall, and F1-scores obtained from the confusion matrix of figure 5 are listed in table VI. According to these experimental results, the trained models are good enough to classify malignant and benign signals. The GB performs 4% better than ID3.

TABLE VI
Performance on random original samples

Classifier	Precision	Recall	F1-score
GB	0.88	0.88	0.88
ID3	0.85	0.84	0.84

B. Adversarial Samples

The second phase of the experiment is about exploring the response of the trained model to adversarial samples. The samples are generated at percentage of features, $\theta = 0.1$ and amount of noise, $\lambda = 0.5$. The confusion matrix analysis illustrated in figure 6 and table VII shows that the model is severely affected by the adversarial samples.

TABLE VII
Performance at $\theta = 0.1$ and $\lambda = 0.5$

Classifier	Precision	Recall	F1-score
GB	0.68	0.68	0.65
ID3	0.64	0.60	0.62

It clearly indicates that the trained models are compromised by the adversarial samples. At $\theta = 0.1$ and $\lambda = 0.5$, the average precision, recall, and F1-score have reduced by 20.5%, 20.5%, and 22.5%.

VI. Conclusion

Machine learning-based intrusion detection systems are now widely acknowledged as essential tools for the detection of cyber attacks on industrial control system (ICS) networks

(a) Gradient Boosting (GB)

(b) Iterative Dichotomiser 3 (ID3)

Fig. 5. The performance of the classifier on unaltered dataset

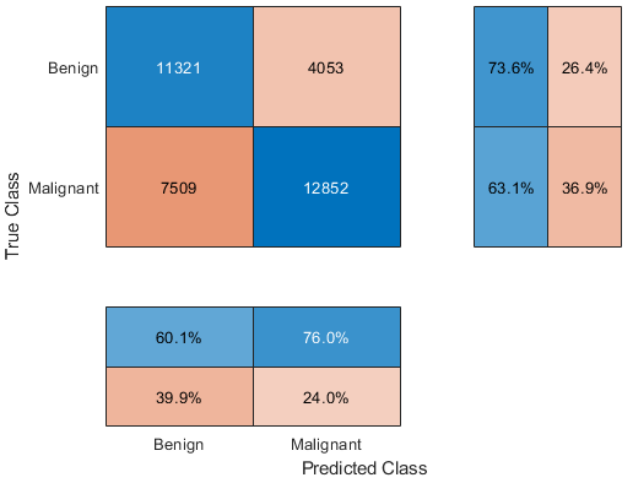

Fig. 6. The performance of the classifier on unaltered dataset

because of their efficiency and adaptability. However, such systems are susceptible to cyberattacks that, in the most prevalent form, are denoted by the acronym AML and have the potential to substantially impede or mislead their capabilities. These kinds of intrusions might have serious repercussions for ICS systems. This is because adversaries could possibly change harmful data points in order to circumvent the controllers, which would result in a delay in the attack being detected and considerable damage. In light of this, it should come as no surprise that a better knowledge of the applicability of these threats in ICS systems is required in order to design more resilient machine learning models for ICSs. However, it requires in-depth analysis to understand the weaknesses and discover a better solution. This paper lays down the foundation of such analysis. The exploratory analysis presented in this paper unearths valuable insights about the effects of AML attacks on ICS of power generation and distribution sectors. The GB and ID3 models, trained with the data obtained from real-world observation, demonstrate a remarkable performance of 89% and 85% accuracy. A classifier with this accuracy can defend most cyber attacks it is trained to defend. However, the scenario alters after the AML attack. This paper showed the adversarial sample generation using the Fast Gradient Sign Method (FGSM). After applying the adversarial samples, the performance of both GB and ID3 dramatically reduces by 21.17% on average, making the machine learning models vulnerable to five types of cyber attacks. That means a well-trained machine learning model with good performance on testing and validation datasets can be compromised by AML attacks. Cybersecurity professionals should keep it in mind before delegating the responsibilities of defending against cyber attacks to machine learning models.

REFERENCES

[1] A. Sharma and N. Singh, "Sensors, embedded systems, and iot components," in *Mathematical Modeling for Intelligent Systems*, pp. 1–15, Chapman and Hall/CRC.

[2] A. Handa and P. Semwal, "Evaluating performance of scalable fair clustering machine learning techniques in detecting cyber attacks in industrial control systems," in *Handbook of Big Data Analytics and Forensics*, pp. 105–116, Springer, 2022.

[3] C. Solsjö and S. Aronsson, "Transmedia storytelling & web 4.0—an upcoming love story: Investigating transmedia storytelling across web 2.0 & 3.0 to assess its relationship with web 4.0," 2022.

[4] S. K. Rao and R. Prasad, "Impact of 5g technologies on industry 4.0," *Wireless personal communications*, vol. 100, no. 1, pp. 145–159, 2018.

[5] P. K. Malik, R. Singh, A. Gehlot, S. V. Akram, and P. K. Das, "Village 4.0: Digitalization of village with smart internet of things technologies," *Computers & Industrial Engineering*, vol. 165, p. 107938, 2022.

[6] A. Corallo, M. Lazoi, M. Lezzi, and A. Luperto, "Cybersecurity awareness in the context of the industrial internet of things: A systematic literature review," *Computers in Industry*, vol. 137, p. 103614, 2022.

[7] A. M. Koay, R. K. Ko, H. Hettema, and K. Radke, "Machine learning in industrial control system (ics) security: current landscape, opportunities and challenges," *Journal of Intelligent Information Systems*, pp. 1–29, 2022.

[8] M. M. Hassan, A. Gumaei, S. Huda, and A. Almogren, "Increasing the trustworthiness in the industrial iot networks through a reliable cyber-attack detection model," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 6154–6162, 2020.

[9] R. Rahim, M. Ramachandran, P. Jayachandran, H. Kalyanaraman, V. Bhaskar, and A. Kumar, "Holistic feistel authenticated learning-based authorization for protecting the internet of things from cyber attacks," *Wireless Personal Communications*, pp. 1–22, 2022.

[10] K. Stouffer, J. Falco, K. Scarfone, *et al.*, "Guide to industrial control systems (ics) security," *NIST special publication*, vol. 800, no. 82, pp. 16–16, 2011.

[11] B. Craggs, A. Rashid, C. Hankin, R. Antrobus, O. Şerban, and N. Thapen, "A reference architecture for iiot and industrial control systems testbeds," in *Living in the Internet of Things (IoT 2019)*, pp. 1–8, IET, 2019.

[12] J.-P. A. Yaacoub, O. Salman, H. N. Noura, N. Kaaniche, A. Chehab, and M. Malli, "Cyber-physical systems security: Limitations, issues and future trends," *Microprocessors and microsystems*, vol. 77, p. 103201, 2020.

[13] I. H. Sarker, "Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective," *SN Computer Science*, vol. 2, no. 3, pp. 1–16, 2021.

[14] S. Mokhtari, A. Abbaspour, K. K. Yen, and A. Sargolzaei, "A machine learning approach for anomaly detection in industrial control systems based on measurement data," *Electronics*, vol. 10, no. 4, p. 407, 2021.

[15] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using rf data: A review," *IEEE Communications Surveys & Tutorials*, 2022.

[16] K. Chauhan, S. Jani, D. Thakkar, R. Dave, J. Bhatia, S. Tanwar, and M. S. Obaidat, "Automated machine learning: The new wave of machine learning," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 205–212, IEEE, 2020.

[17] S. Mitra and S. Ransbotham, "Information disclosure and the diffusion of information security attacks," *Information Systems Research*, vol. 26, no. 3, pp. 565–584, 2015.

[18] B. Maschler and M. Weyrich, "Deep transfer learning for industrial automation: a review and discussion of new techniques for data-driven machine learning," *IEEE Industrial Electronics Magazine*, vol. 15, no. 2, pp. 65–75, 2021.

[19] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[20] H. S. Alenezi and M. H. Faisal, "Utilizing crowdsourcing and machine learning in education: Literature review," *Education and Information Technologies*, vol. 25, no. 4, pp. 2971–2986, 2020.

[21] N. Faruqui, M. A. Yousuf, M. Whaiduzzaman, A. Azad, A. Barros, and M. A. Moni, "Lungnet: A hybrid deep-cnn model for lung cancer diagnosis using ct and wearable sensor-based medical iot data," *Computers in Biology and Medicine*, vol. 139, p. 104961, 2021.

[22] M. Punithavalli, "Machine learning in human resource management," in *Artificial Intelligence Theory, Models, and Applications*, pp. 281–308, Auerbach Publications, 2021.

[23] V. Meshram, K. Patil, V. Meshram, D. Hanchate, and S. Ramkteke, "Machine learning in agriculture domain: A state-of-art survey," *Artificial Intelligence in the Life Sciences*, vol. 1, p. 100010, 2021.

[24] O. G. Manzanilla-Salazar, F. Malandra, H. Mellah, C. Wette, and B. Sanso, "A machine learning framework for sleeping cell detection in a smart-city iot telecommunications infrastructure," *IEEE access*, vol. 8, pp. 61213–61225, 2020.

[25] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: A survey," *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[26] C. Vuppalapati, N. Raghu, P. Veluru, S. Khursheed, *et al.*, "A system to detect mental stress using machine learning and mobile development," in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 161–166, IEEE, 2018.

[27] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Systems*, vol. 189, p. 105124, 2020.

[28] S. Picard, C. Chapdelaine, C. Cappi, L. Gardes, E. Jenn, B. Lefevre, and T. Soumarmon, "Ensuring dataset quality for machine learning certification," in *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pp. 275–282, IEEE, 2020.

[29] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, pp. 1–26, 2020.

[30] A. E. Cinà, K. Grosse, A. Demontis, B. Biggio, F. Roli, and M. Pelillo, "Machine learning security against data poisoning: Are we there yet?," *arXiv preprint arXiv:2204.05986*, 2022.

[31] R. Yumlembam, B. Issac, S. M. Jacob, and L. Yang, "Iot-based android malware detection using graph neural network with adversarial defense," *IEEE Internet of Things Journal*, 2022.

[32] F. Aloraini, A. Javed, O. Rana, and P. Burnap, "Adversarial machine learning in iot from an insider point of view," *Journal of Information Security and Applications*, vol. 70, p. 103341, 2022.

[33] A. Lanz, D. Rogers, and T. Alford, "An epidemic model of malware virus with quarantine," *Journal of Advances in Mathematics and Computer Science*, vol. 33, no. 4, pp. 1–10, 2019.

[34] S. Chen, M. Xue, L. Fan, S. Hao, L. Xu, H. Zhu, and B. Li, "Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach," *computers & security*, vol. 73, pp. 326–344, 2018.

[35] E. Alshahrani, D. Alghazzawi, R. Alotaibi, and O. Rabie, "Adversarial attacks against supervised machine learning based network intrusion detection systems," *Plos one*, vol. 17, no. 10, p. e0275971, 2022.

[36] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, "A taxonomy and terminology of adversarial machine learning," *NIST IR*, pp. 1–29, 2019.

[37] Q. Zeng, L. Zhou, Z. Lian, H. Huang, and J. Y. Kim, "Privacy-enhanced federated generative adversarial networks for internet of things," *The Computer Journal*, 2022.

[38] I. Alarab and S. Prakoonwit, "Graph-based lstm for anti-money laundering: Experimenting temporal graph convolutional network with bitcoin data," *Neural Processing Letters*, pp. 1–19, 2022.

[39] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, p. 102717, 2021.

[40] D. Wang, X. Wang, Y. Zhang, and L. Jin, "Detection of power grid disturbances and cyber-attacks based on machine learning," *Journal of information security and applications*, vol. 46, pp. 42–52, 2019.

[41] E. Hoxha, Y. Vidal Seguí, and F. Pozo Montero, "Supervised classification with scada data for condition monitoring of wind turbines," in *9th ECCOMAS thematic conference on smart structures and materials*, pp. 263–273, 2019.

[42] J. Gao, L. Gan, F. Buschendorf, L. Zhang, H. Liu, P. Li, X. Dong, and T. Lu, "Lstm for scada intrusion detection," in *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pp. 1–5, IEEE, 2019.

[43] S. D. Anton, S. Kanoor, D. Fraunholz, and H. D. Schotten, "Evaluation of machine learning-based anomaly detection algorithms on an industrial modbus/tcp data set," in *Proceedings of the 13th international conference on availability, reliability and security*, pp. 1–9, 2018.

[44] R. L. Perez, F. Adamsky, R. Soua, and T. Engel, "Machine learning for reliable network attack detection in scada systems," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 633–638, IEEE, 2018.

[45] L. A. Maglaras and J. Jiang, "Intrusion detection in scada systems using machine learning techniques," in *2014 Science and Information Conference*, pp. 626–631, IEEE, 2014.

[46] I. Abdallah, V. Dertimanis, H. Mylonas, K. Tatsis, E. Chatzi, N. Dervili, K. Worden, and E. Maguire, "Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data," in *Safety and Reliability–Safe Societies in a Changing World*, pp. 3053–3061, CRC Press, 2018.

[47] A. Taur, S. M. Badave, S. Padmanaban, M. S. Bhaskar, V. K. Ramachandaramurthy, and J. B. Holm-Nielsen, "Testing of local control cabinet in gas insulated switchgear using design of simulation kit-revista," in *2019 IEEE 13th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, pp. 1–5, IEEE, 2019.

[48] S. S. Hossain-McKenzie, *Protecting the power grid: strategies against distributed controller compromise*. PhD thesis, University of Illinois at Urbana-Champaign, 2017.

[49] J. Ling, R. Jones, and J. Templeton, "Machine learning strategies for systems with invariance properties," *Journal of Computational Physics*, vol. 318, pp. 22–35, 2016.
[50] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in *SoutheastCon 2016*, pp. 1–6, IEEE, 2016.
[51] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The elements of statistical learning*, pp. 337–387, Springer, 2009.
[52] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.
[53] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.
[54] J. Mills, *Identify Insider Threats Using LRM*. PhD thesis, The George Washington University, 2017.
[55] Y. Deldjoo, T. D. Noia, and F. A. Merra, "A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.
[56] J. Lin, L. L. Njilla, and K. Xiong, "Secure machine learning against adversarial samples at test time," *EURASIP Journal on Information Security*, vol. 2022, no. 1, pp. 1–15, 2022.
[57] L. Kovar and M. Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 559–568, 2004.
[58] Y. Liu, S. Mao, X. Mei, T. Yang, and X. Zhao, "Sensitivity of adversarial perturbation in fast gradient sign method," in *2019 IEEE symposium series on computational intelligence (SSCI)*, pp. 433–436, IEEE, 2019.