2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)

2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT) | 979-8-3503-4649-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/DCOSS-IOT58021.2023.00077

The Threat of Adversarial Attacks Against Machine Learning-based Anomaly Detection Approach in a Clean Water Treatment System

Naghmeh Moradpoor*, Leandros Maglaras, Ezra Abah and Andres Robles-Durazno
School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, UK
{n.moradpoor, l.maglaras, a.roblesdurazno}@napier.ac.uk, 40482302@live.napier.ac.uk

***Abstract*—The protection of Critical National Infrastructure is extremely important due to nations being dependent on their operation and steadiness. Any disturbance to this infrastructure could have a devastating consequence on physical security, economic wellbeing, and public health and safety. To deal with the growing number of attacks, with differing degrees of impact against such systems, various machine learning-based Intrusion Detection Systems have been employed given their success in the automated detection of known and unknown cyberattacks with high degrees of accuracy. However, since machine learning models are susceptible to attacks, also known as Adversarial Machine Learning, employing such Intrusion Detection Systems has also created an additional attack vector which could potentially help hackers to evade detection. This paper explores the robustness of both traditional and non-traditional supervised machine learning algorithms by studying their classification behaviour under adversarial attacks. This includes machine learning algorithms such as Support Vector Machine, Logistic Regression, and Deep Learning models, such as Artificial Neural Network. Additionally, this contains adversarial machine learning attacks such as random & targeted label flipping, Fast Gradient Sign Method, and Jacobian Saliency Map Attack. A genuine dataset captured from a model of a clean water treatment system was used to support the experiments presented in this paper. Overall, the adversarial attacks were successful to decrease the classification performance of the machine learning algorithms but with varying degrees of influence. For example, the targeted label flipping has a stronger impact on the classification performance reduction compared with the random label flipping attacks. Additionally, Deep Learning model and Support Vector Machine both show longer fight against the adversarial attacks compared with Logistic Regression.**

***Index Terms*—adversarial attacks; machine learning; critical national infrastructure; industrial control systems; clean water treatment systems; anomaly detection**

I. Introduction

Protection of Critical National Infrastructure (CNI) such as transportation, essential utilities: like water, gas, and electricity, police systems, health services, and commerce is vital for day-to-day functioning of a society and the economy. Traditionally, these systems and their associated components were protected from cyberattacks as they were connected in air gap environments preventing them from establishing any external connections such as connecting to the Internet. However, with the appearance of Industry 4.0, which integrates traditional computer networks and CNI and includes devices with increased level of connectivity, we have witnessed an increased attack surface of such valuable assets. Criminals and state-sponsored hackers are progressively going after CNI to disturb society and the economy resulting in cyberattacks on such systems raising both in occurrence and impact.

For instance, water companies have been under cyberattacks for twenty years [4]. This includes attacks on a wastewater treatment plant in the Shire of Maroochy (Australia) in 2000 [33] by a technical contractor, attacks on a canal system in California (USA) in 2007 [34] by a former employee using unauthorised software, physical attacks on a drinking water plant in Georgia (USA) in 2013 [4] by former employees who still have keys, attacks on a public water provider in Michigan (USA) in 2016 [36] by ransomware, attacks on a distribution water company in North Carolina (USA) in 2018 [37] by ransomware, attacks on a distribution water company in Kansas (USA) in 2019 [35] by a former employee, attacks on pumping stations and water treatment facilities in Israel in 2020 [38] by suspected state-sponsored cyber criminals, attacks on agricultural water pumps in Israel in 2020 [4] by suspected state-sponsored cyber criminals, attacks on a recycled water reservoir in Israel in 2020 [39] by suspected state-sponsored cyber criminals, attack on two wastewater treatment plants in San Francisco and Florida (USA) in 2021 [40] by cybercriminals, and attacks on water treatment infrastructure in Norway in 2021 [41] by ransomware.

There has been significant increase in the Machine Learning (ML)-based cybersecurity solutions which cover a wide variety of applications and issues from traditional computer networks to wireless technology, Internet of Things (IoT), and CNI domains. This is due to their efficiency in identifying attacks and their ability to detect known and zero-day attacks with a high percentage of accuracy, particularly when they are facing enormous amounts of data. However, ML algorithms are vulnerable and susceptible to adversarial attacks. This is because hackers could exploit a ML model either during the training phase or testing phase to cause miss-classification resulting in classifying malicious events as benign, and vice versa, forcing the entire model to fail [5]. In the context of CNI, adversarial machine learning could manipulate data received from sensors/actuators to cause miss-classification, evading the detection by ML-based Intrusion Detection Systems (IDS), and

2325-2944/23/$31.00 ©2023 IEEE
DOI 10.1109/DCOSS-IoT58021.2023.00077

leading to destructive consequences. Therefore, it is crucial to evaluate a built machine learning model against adversarial examples [6-8].

The existing research on cybersecurity issues of clean water supply/clean water treatment systems are mainly focused on two testbeds and their associated datasets: SWaT [1] and WADI [2]. The former is a multi-stage water purification plant, and the latter is a consumer distribution network of testbeds. However, given our research background, we are interested in using energy consumption of the system's components to detect anomalies against clean water treatment systems and also for its adversarial attack scenarios. Therefore, the above datasets are no use for us since they did not capture any energy consumption features in their associated datasets. Therefore, in this paper we employed an authentic energy-based dataset previously captured from our implemented testbed called Virtual Napier Water Treatment System (VNWTS) [3]. The VNWTS models the water chlorination process for a clean water treatment system and was previously evaluated and published. Hence, the aim in this paper is to utilise its captured dataset for generating adversarial samples and to study the robustness of the supervised and energy-based anomaly detection algorithms against adversarial attacks.

Therefore, the main contributions of the presented research in this paper are the experimental investigation into:

- Generating adversarial samples using: 1) Label Noise Attacks (both Random and Targeted Label Flipping attacks), 2) Fast Gradient Sign Method (FGSM), and 3) Jacobian Saliency Map Attack (JSMA) from a clean water treatment dataset
- Studying the behaviour of three supervised energy-based anomaly detection algorithms (an Artificial Neural Network, a Support Vector Machine, and a Logistic Regression binary classifier) for a model of a clean water treatment system against adversarial samples

To the best of our knowledge, this is the first study which investigates the performance of a supervised machine learning model against four automatically generated adversarial attacks in the context of clean water treatment systems. The work also includes a realistic attack model and an authentic dataset directly collected from a model of a clean water treatment system.

The remainder of this paper is structured as follows: Section II reviews the related work in the field, Section III discusses the clean water treatment system and its generated dataset used for the experiments in this paper, Section IV presents the performance of a range of supervised, energy-based and binary classifiers employed to detect anomalies against the system, Section V discusses the generation of adversarial samples and their impact on the performance of the classifiers, and finally section VI concludes the paper.

II. Related Work

There are several research papers related to ML-based IDSs to detect anomalies against CNI. This includes power systems [9], wind turbines [10], water systems (mainly focused on SWaT testbed) [11] and gas pipelines [12]. However, to the best of our knowledge, the majority of the work in adversarial machine learning are related to IDSs on traditional computer networks, for example malware detection [13], and there has been less attention on adversarial attacks in the context of CNI. The following presents some of the existing work related to adversarial machine learning on CNI with a focus on power grids and water systems.

Authors in [24] analysed the impact of two adversarial machine learnings: Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and JSMA against Multilayer Perceptron (MLP) algorithm employed to detect False Data Injection Attacks (FDIAs) in power systems. The implemented attacks significantly reduced the accuracy of MLP.

Authors in [15] evaluated the strength of a ML-based prediction model for power distribution networks which is integrated with an anomaly detector to identify adversarial attacks with a particular focus on stealthy ones. The implemented attacks had a negative impact on the performance of the prediction model even when only a portion of the network was involved.

Authors in [16] assessed power grids against adversarial machine learning when the adversaries interrupt the external time series observed by the power grid's controller with the aim of increasing the energy demand and associated cost. The experiments show 8500% increase in the overall cost and 13% on the electricity demand pattern which can't be easily detected by human eyes.

Authors in [17] compared the effect of a group of adversarial attacks against their employed machine learning algorithm recruited for classification, forecasting, and control in power systems. They also implemented a set of defence mechanisms to protect their proposed ML-based solution for the above set of problems. Their results reveal the vulnerabilities of such solutions against adversarial attacks and that their implemented Generative Adversarial Network (GAN) is the most effective defence approach.

Authors in [20] proposed a Bayesian-based technique to increase the strength of their developed load forecasting algorithm for power systems using adversarial training methods. However, although they proved the effectiveness of such an approach theoretically, it doesn't illustrate the model's performance before and after adversarial training against such attacks.

Authors in [18] proposed a framework for generating adversarial attacks against their proposed data-driven invariant checkers for water treatment systems. For this they employed SWaT [1] physical testbed and have shown that these attacks increased the system's false alarm by up to 80%. They then implemented defence mechanism against such an approach showing that the invariant checkers became robust and do not raise false alarms anymore. However, the checkers are not capable of detecting attack attempts against the system.

Authors in [19] proposed two algorithms to attack a neural network-based IDS for the application of clean water treatment system by employing SWaT [1] testbed. To evaluate their proposed attack algorithms, they used artificial and real testbed

data which shows the success of both algorithms in poisoning the IDS.

Authors in [21] compared the traditional image domain of adversarial machine learning with the use of such attacks in cybersecurity domain. They also had a case study of adversarial attacks against a Long Short-Term Memory (LSTM)-based IDS applied on a dataset captured from SWaT [1]. This work is at a primary stage and the output is rather limited.

Authors in [22] developed a model of a steam condenser on Simulink with a Recurrent Neural Network (RNN) controller and proposed a gradient-based local search approach to discover adversarial samples against such a system. However, it is rather unclear if the proposed method can be applied to any models or if it is just for the systems that employ RNN with smooth activation function.

Authors in [23] proposed two types of real-time evasion attacks based on white box and black box methods. In the former approach the attacker uses an optimisation tactic with detection oracle while in the latter approach the attacker uses a convolutional neural network method to translate anomalous data into normal data. For their experiments they focused on two datasets related to water systems: WADI [2] and BATADAL [14]. Their proposed techniques significantly reduced the accuracy of the related ML-based approach, but no defence mechanism has been investigated.

The work in this paper focuses on a testbed called VNWTS [3] and its generated energy-based dataset. The testbed models the water chlorination process for a clean water treatment system. The aim in this paper is to evaluate the robustness of the previously proposed supervised and energy-based anomaly detection algorithm against a set of adversarial attacks. The work in this paper defers from the papers in the field particularly those related to water systems such as [18-19] and [21-23]. It is mainly because the work in this paper is not based on the existing dataset such as SWaT [1] WADI [2], and BATADAL [14] given that these datasets do not include energy features thus they are no use for us. Additionally, compared with the existing work, we consider a larger number of attacks, four attacks in total, while there is usually one or two in the existing work.

III. Industrial Control System Case Study: A Clean Water Treatment System

In this paper, an energy-based dataset, which was captured from a testbed called VNWTS, Fig.1 (left), is employed for the experiments. The testbed models the water chlorination process for a clean water treatment system. It was created by the authors of this manuscript and evaluated in their previous publication [3]. Although this testbed is relatively small, since it captures the core functions of such a system it is considered as being a representative example of a large clean water treatment system.

The VNWTS testbed architecture includes elements such as: an emulated water chlorination process and a UDP module in Level 0, an emulated SIEMENCE S7 – 1500 PLC, which includes all its internal components (i.e., Input, Output, working memory, and network functionalities), in Level 1, a SCADA system, a HMI, and a Python code in Level 2. The emulated water chlorination process was implemented in Simulink [26] and an emulated SIEMENCE S7 – 1500 PLC was implemented in SIMATIC S7-PLCSIM Advanced V3.0 software. The Python code provides communication between the system components (e.g., between PLC and Simulink).

The emulated water chlorination process was inspired by MPA PA Festo Rig [25] (A.K.A. RIG), Fig.1 (right), which represents a scaled-down version of a one-of-a-kind water treatment system. Hence, it has the same characteristics and dynamics of the physical elements represented in the RIG. For example, same as the RIG, the emulated process includes Pipes, one Pressure Vessel, two Pumps, one Proportional Valve, one Water Reservoir Tank, two Flow Sensors, and two Water Supplies.

TABLE I
Energy-based Features in VNWTS dataset

Cold Flow Rate It is the cold fluid/per second passing through the pipes. The cold water represents raw water that will be chlorinated.
Hot Flow Rate It is the hot fluid/per second passing through the pipes. The hot water represents chlorine that will be used to purify the raw water (i.e., cold water above).
Temperature It is the temperature of the water in the reservoir tank. This is the mixture of raw water (i.e., cold water) and chlorine (i.e., hot water).
Tank Level It shows the amount of water that should be kept in the reserve tank.
Voltage in the Warm-Water Pump It indicates the voltage which is supplied to the hot water pump.
Voltage in the Cold-Water Pump It indicates the voltage which is supplied to the cold-water pump.
Current in the Warm-Water Pump It indicates the current consumed by the hot water pump.
Current in the Cold-Water Pump It indicates the current consumed by the cold-water pump.
Class Feature It represents ‘0’ for benign and ‘1’ for attack.
Type of Attack It represents ‘0’ for benign events, ‘1’ for attacks to the level setpoint, ‘2’ for attack to the temperature setpoint, and ‘3’ for attack to multiple sensors.

A similar chlorine dosing ratio was considered to ensure a correct representation of a clean water treatment system but by replacing the chlorine element with the hot water. For the water demand model, the testbed recruited a real model of the UK energy consumption, which was fully explained in [27]. This model was implemented in a proportional valve of the VNWTS virtual process and regulated based on the water demand. For instance, high water demand symbolised by a fully open valve which consumes more energy and low water demand represented by a slightly open valve which consumes low energy.

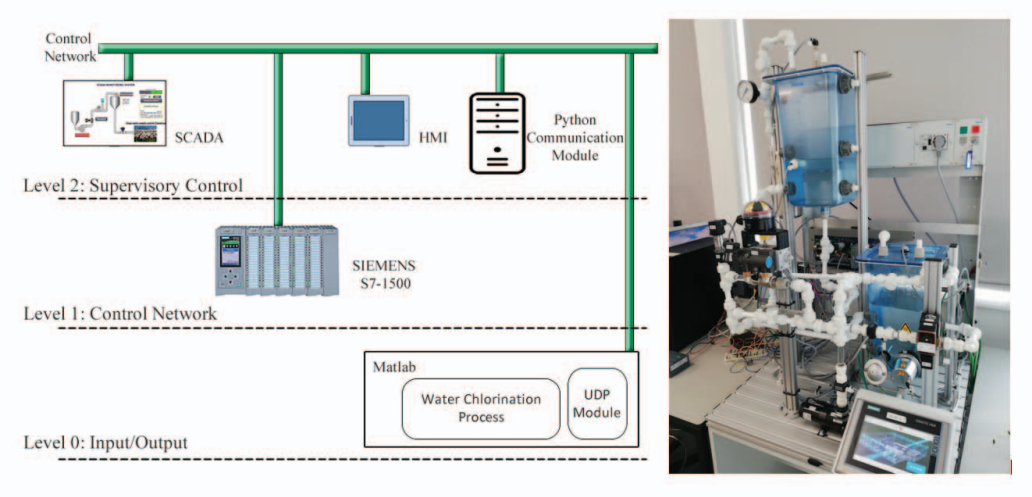

Fig. 1. VNWTS Testbed Architecture (left) [3], MPA PA Festo Rig (right) [25]

A. Dataset Overview

An energy-based dataset including both malicious and benign events was generated from VNWTS testbed. Each event consists of eight features and two labels (i.e., binary and multiclass classification features) which define the testbed based on its energy consumption, Table I. To generate the malicious events the focus was on exploiting the vulnerabilities which stem from the fixed memory space of Siemens S7-1500 PLC allocated to its input and output memory. For example, after exploiting these vulnerabilities an attacker can overwrite the memory space related to the water temperature which then leads to a disturbance in the water chlorination process. PLC's fixed memory space vulnerabilities are fully explained in our previous work [3].

The events in the dataset are categorised as either 'benign' or 'malicious' and the malicious events are further classified as 'attacks to the level setpoint', 'attack to the temperature setpoint', or 'attack to multiple sensors.' Therefore, the dataset includes two labels: 1) a Class Feature (where '0' represents benign events and '1' represents malicious events) and 2) Type of Attack (where '1' represents attacks to the level setpoint, '2' represents attack to the temperature setpoint, and '3' represents attack to multiple sensors). In this paper, we study the impact of adversarial attacks on the binary classification (A.K.A Class Feature) and thus the influence of such attacks on multiclass classification (A.K.A Type of Attack) will be considered in future work.

There are a total of 3,132,651 malicious and benign energy-based events in the captured VNWTS dataset out of which 120,000 records (4% of the dataset) were selected for the research in this paper due to the limited resources and to speed up the experiments. In the chosen 120,000 records, 60,000 events are benign, and 60,000 events are malicious (20,000 records of Attack to the level setpoint, 20,000 records of Attack to the temperature setpoint, and 20,000 records of Attack to multiple sensors). Then 70% of the selected record (48,007 benign and 47,993 malicious: total of 96,000) were used for training and 30% (11,993 benign and 12,007 malicious: total of 24,000) used for testing.

B. Data Pre-processing

For the data pre-processing phase, feature selection and normalisation have been chosen in this paper. Information Gain, Chi-Square, and Pearson's Correlation are used for the feature selection phase and MinMaxScaler is employed for the normalisation phase. The feature selection is required to identify the redundant features and remove them from the dataset which leads to a decreaseing in the computational cost of building the predictive model and the chance of overfilling, along with increasing the overall performance. The chosen feature selection techniques lowered the number of the features from eight to four, these which includeds Temperature, Tank Level, Cold Flow Rate, Voltage in the Cold-Water Pump, along with thewith the addition of Class Feature, and Type of Attack. TTherefore, the four removed features are: Hot Flow Rate, Voltage in the Warm-Water Pump, Current in the Warm-Water Pump, and Current in the Cold-Water Pump. The dataset also requires normalisation but not encoding as the data are numerical but in different formats (binary, integer, and float). For example, Class Feature is binary ('0' or '1'), Type of Attack is integer ('0', '1', '2', '3'), and Temperature, Tank Level, Cold Flow Rate, Voltage in the Cold-Water Pump are floating-point numbers (spread between 0 and 1). For normalisation, MinMaxScalar technique is used to scale the data between 0 and 1. The pre-processed dataset is called Trusted Dataset as it is not under any adversarial attacks hence it has not experienced any manipulation during training nor the testing phase.

C. Model Training & Performance Metrics

In order to reveal the impact of adversarial attacks against machine learning models, the first stage is to implement the base models also known as Trusted Models. This way a valid performance comparison will be possible. For the base models, a combination of traditional and non-traditional supervised machine learning algorithms is chosen in this paper to study and compare their classification behaviour under adversarial attacks. This includes Support Vector Machine (SVM), Logistic Regression (LR), and Deep Learning models, such as Artificial Neural Network (ANN). We follow "no free lunch" theorems [28] where the concept suggests that there is no universally best machine learning algorithm. This means the suitability of a learning model, also known as a classifier, for a problem depends on its performance for the problem and the nature of data that defines the problem. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which are confusion matrix terms, as well as accuracy, are selected to present the performance of the chosen models. These metrics will also be employed to uncover the impact of adversarial attacks on the three classifiers by comparing the adversarial examples with the trusted models. The impact of such attacks on multiclass classifications using the same SVM, LR, and ANN models will be in future studies due to the space limitation. For LR trusted model, the default LR model and its parameters in Scikit-Learn package were used except for the maximum iteration parameter where instead of the default value of 100, 1000 was used. For the SVM trusted model Keras Quasi-SVM model [29] was created to tackle the issue with the speed. This model uses a RandomFourierFeatures layer, which is set to a Gaussian kernel and can be used to Kernelize linier models by applying a non-linear transformation to the input features and then training a linear model on top of the transformed features. For the ANN trusted model, four layers were used. The first three layers use relu activation and the last one uses a softmax activation. There are 32,16,8, and 4 neurons from layer1 to layer 4 all respectively. The compiler for the ANN model uses adam optimizer, binary_crossentropy for loss calculation and accuracy metric. The models were trained using 30 epochs.

IV. Adversarial Machine Learning

In this paper, adversarial attacks are employed to study their impact on the performance of the supervised energy-based anomaly detection approach on a model of a clean water treatment system. The samples are generated to confuse the models, and this is known as a poison attack during the training phase, during the testing phase it is known as an evasion attack. Label Noise attacks (both random and targeted label flipping attacks) are used during the training phase while Fast Gradient Sign Method (FGSM) and Jacobian Saliency Map Attack (JSMA) are used during the testing phase. The three types of attacks are employed against ANN, SVM, and LR and the associated energy-based dataset.

A. Attacker Model

In this paper, it is assumed that the intruder is an insider threat attacker who has admin access privileges to the clean water treatment system (e.g., a chief network engineer). Therefore, she/he has access to the dataset and has knowledge of the associated features that the machine learning models use to detect anomalies against the local plant. However, it is assumed that the attacker does not know about the employed algorithms. It can be argued that the insider threat's lack of knowledge on the exact algorithms is due to the uniqueness, complexity, or obscure nature of the exact machine learning-based product or software. Therefore, the goal of the attacker can be: 1) exfiltrate the associated energy-based dataset to the outside of the organisation in order to give it to competitors and/or to plan further attacks against the organisation, 2) manipulate the training data to build an unreliable ML-based anomaly detection algorithm (e.g., with low performance), 3) manipulate the testing data to evade detection and bypass the anomaly detection system. Given adversary's knowledge of the system and that they know about the dataset and its features, but not the exact ML-based models, this intrusion is classified as being a grey box attack.

B. Adversarial Sample Generation Methods

There are various methods to generate adversarial samples against a ML-based algorithm. These techniques are different in terms of complexity, required generation time, performance, and mode of generation (i.e., manual or automated). In general, manual generation of perturbed samples is time consuming and less accurate, particularly for a large dataset, when compared to the automated generation of adversarial samples. From the manual category, Label Noise adversarial attacks (both Random Label Flipping & Targeted Label Flipping) and from the automated types, Fast Gradient Sign Method (FGSM) [30] and Jacobian Saliency Map Attack (JSMA) [31], are chosen for the experiments in this paper. They are also used against different models. For example, Label Noise attacks are used against SVM & LR binary classifications while FGSM & JSMA are employed against LR & ANN binary classifications. The reason for this arrangement is to study the impact of adversarial attacks against traditional ML-based approaches (i.e., SVM & LR) vs. non-traditional ones (i.e., ANN) in terms of robustness. For example, traditional ML models tend to be more vulnerable against label noise attacks (e.g., when the class label is flipped) in comparison with feature noise attacks (when perturbations are added to features). This seems to be the opposite for non-traditional ML algorithms.

In Random Label Flipping, a random portion of the labels, which is gradually increased from 0% to 70% by the scale of ten, are selected and tossed (e.g., '0' to '1' and '1' to '0') and then employed to train SVM & LR models. In Targeted Label Flipping, the absolute distance between the models and labels are calculated and the labels with the longest absolute distance from the model are selected, which is gradually increased from 0% to 70% by the scale of ten, reversed and then used to train SVM & LR models. It is expected to see that the

TABLE II
CONFUSION MATRIX TERMS FOR THE BASE MODELS WITHOUT ADVERSARIAL ATTACKS

Metrics	TP	TN	FP	FN
LR	11,981	6,306	5,701	12
SVM	11,993	8,153	3,854	1
ANN	11,993	11,393	614	1

Targeted Label Flipping attack results in a more negative impact on the model's performance particularly in the lower ranges compared with the Random Label Flipping attack.

Additionally, FGSM and JSMA are used against LR & ANN binary classifications. Given that these two attacks target the testing data to evade anomaly detection it is assumed that the intruder has access to the testing data to perform these two attacks. To generate adversarial examples on the testing data against the supervised and energy-based anomaly detection approach for the model of the clean water treatment system, an IBM-based library called Adversarial Robustness Toolbox (ART) [32] is employed in this paper. For FGSM, the epsilon, which is the degree to which the test samples are perturbed, is varied between 0.05 and 0.3. For each epsilon, a new testing set is created from the original one and further employed as a test set for the developed model. For JSMA, a fixed fraction rate of 0.1 and different perturbation rates, which varies between 0.02 and 0.2, are chosen. Hence, with the fix fraction rate of 0.1 for each perturbation rate and a new test set is created from the original one and further employed as a test set for the developed model. It is expected to see greater declines in the models' performance when epsilon in FGSM and perturbation rates in JSMA gradually increase.

V. EVALUATING SUPERVISED MODELS ON ADVERSARIAL SAMPLES (TRUSTED VS. UNTRUSTED MODELS)

In this paper, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which are confusion matrix terms, as well as accuracy, are chosen to study and compare the performance of the supervised binary classifiers with and without adversarial samples. As discussed before, the Label Noise Attacks (both Random and Targeted Label Flipping attacks) issued against LR & SVM binary classifiers, aim to manipulate the Class Feature in the training dataset, while FGSM and JSMA launched against LR & ANN binary classifiers, aim to manipulate non-class features in the testing dataset.

Table II. reveals the TP, TN, FP, and FN for the three classifiers in the trusted model when they and their associated dataset were not under influence of any adversarial machine learning attacks.

Table III. shows the TP, TN, FP, and FN for the three classifiers when they have undergone different adversarial attacks. It is interesting to see that the TP for the LR & SVM drops 100% after the Random Label Flipping (RLF) attacks. The next highest drop belongs to LR under JSMA and FGSM with almost 99% drop in TP. This is followed by LR with nearly 80% and SVM with nearly 77% reduction in TP when they both are under the Targeted Label Flipping (TLF) attack. The lowest drop in TP belongs to ANN under FGSM (nearly 66%) and JSMA (nearly 38%) attacks.

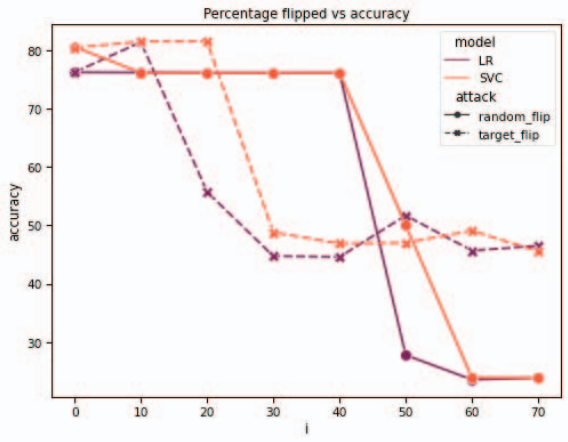

Fig. 2. Accuracy for LR vs. SVM binary classifiers after Label Noise attacks

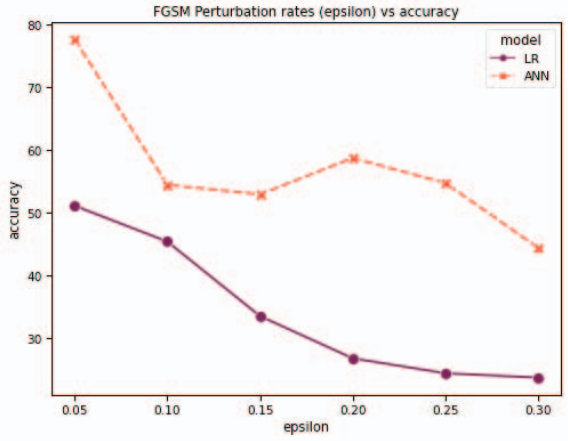

Fig. 3. Accuracy for LR vs. ANN binary classifiers after FGSM attack

Fig.2 reveals the accuracy for LR vs. SVM binary classifiers after the Random & Targeted Label Flipping attacks. For the Random Label Flipping attacks, both classifiers show almost the same performance until nearly half of the labels are flipped, after which LR shows a sharper drop in comparison with SVM.

For the Targeted Label Flipping attacks, SVM shows a longer battle in comparison with the LR. Moreover, as was expected, in the Targeted Label Flipping attacks the performance reduction starts sooner which means it starts from lower percentages (i.e., before flipping 50% of the labels) in comparison with the Random Label Flipping attacks where this happens after reversing 50% of the labels.

Fig. 3 reveals the accuracy for LR vs. ANN binary classification after FGSM attack and Fig. 4 reveals the accuracy for the same classifiers but after JSMA attack. As it is depicted, the LR classifier is weaker against both attacks in comparison

TABLE III
HERE IS A CAPTION.

Metrics	TP	TN	FP	FN
LR under RLF attacks	0 (100% down)	6,666 (5% up)	5,341 (6% down)	11,993 (99,841% up)
LR under TLF attacks	2,310 (80% down)	10,091 (60% up)	1,916 (66% down)	9,683 (80,591% up)
LR under JSMA attacks	14 (99% down)	10,987 (74% up)	1,020 (82% down)	11,979 (99,725% up)
LR under FGSM attacks	12 (99% down)	10,892 (72% up)	1,115 (80% down)	11,981 (99741% up)
SVM under RLF attacks	0 (100% down)	12,007 (47% up)	0 (100% down)	11,993 (1,199,200% up)
SVM under TLF attacks	2,719 (77% down)	8,703 (6% up)	3,304 (14% down)	9,274 (927,300% up)
ANN under JSMA attacks	7,396 (38% down)	8,428 (26% down)	3,579 (482% up)	4,597 (459,600% up)
ANN under FGSM attacks	4,015 (66% down)	9,421 (17% down)	2,586 (321% up)	7,978 (797,700 % up)

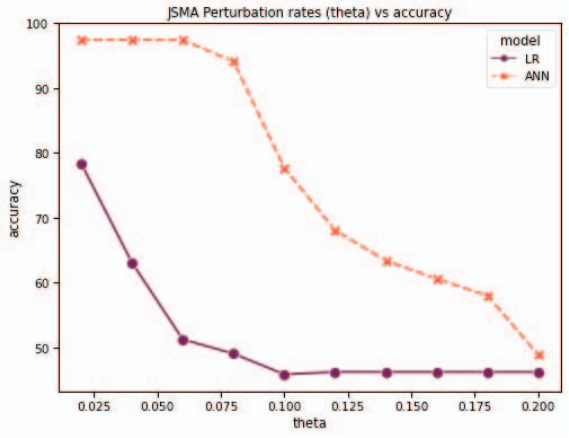

Fig. 4. Accuracy LR vs. ANN binary classifiers after JSMA attack

with the ANN since its accuracy drops quicker and stays lower than ANN's.

VI. CONCLUSIONS

This paper provided a deeper understanding of three types of adversarial machine learning attacks, called Label Noise (both Random and Targeted Label Flipping attacks), Fast Gradient Sign Method, and Jacobian Saliency Map Attack, against supervised machine learning algorithms which are used to detect anomalies in a model of a clean water treatment system. Both traditional and non-traditional learning models such as Logistic Regression, Support Vector Machine, and Deep Learning models, such as Artificial Neural Network are employed, and their robustness are evaluated against adversarial samples. Overall, SVM is tougher against Label Noise Attacks (both Random and Targeted Label Flipping attacks) in comparison with LR. The Targeted Label Flipping attacks achieve faster accuracy reduction in comparison with the Random Label Flipping attacks for both SVM and LR. Additionally, LR classifier is weaker against both FGSM & JSMA attacks in comparison with the ANN. One way to mitigate adversarial attacks is to generate adversarial samples and include them in the training set. Future work will focus on the effect of the adversarial attacks against multiclass classification, the use of physical testbed modelling a clean water treatments system instead of a simulation in addition to implementation and evaluation of possible countermeasures against adversarial attacks (for example to include adversarial samples in training set).

ACKNOWLEDGMENT

This research is supported by the School of Computing, Engineering & the Built Environment at Edinburgh Napier University.

REFERENCES

[1] A. P. Mathur and N. O. Tippenhauer, "SWaT: A water treatment testbed for research and training on ICS security," 2016 Int. Work. Cyber-physical Syst. Smart Water Networks, CySWater 2016, no. Figure 1, pp. 31–36, 2016, doi: 10.1109/CySWater.2016.7469060.

[2] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: A water distribution testbed for research in the design of secure cyber physical systems," Proc. - 2017 3rd Int. Work. Cyber-Physical Syst. Smart Water Networks, CySWATER 2017, pp. 25–28, 2017, doi: 10.1145/3055366.3055375.

[3] Durazno, A. R., Moradpoor, N., McWhinnie, J., & Porcel-Bustamante, J. (2021, December). VNWTS: A Virtual Water Chlorination Process for Cybersecurity Analysis of Industrial Control Systems. In 2021 14th International Conference on Security of Information and Networks SIN (Vol. 1, pp. 1-7). IEEE.

[4] "Twenty years of cyberattacks on the world of water", [Online], Available: https://www.stormshield.com/news/twenty-years-of-cyber-attacks-on-the-world-of-water/

[5] Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems, 30(9), 2805-2824.

[6] Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., & Roli, F. 2015. Support vector machines under adversarial label contamination. Neurocomputing, 160, 53-62.

[7] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

[8] Anthi, E., Williams, L., Rhode, M., Burnap, P., & Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in industrial control systems. Journal of Information Security and Applications, 58, 102717.

[9] Wang, D., Wang, X., Zhang, Y., & Jin, L. (2019). Detection of power grid disturbances and cyber-attacks based on machine learning. Journal of information security and applications, 46, 42-52.

[10] Hoxha, E., Vidal Seguí, Y., & Pozo Montero, F. (2019). Supervised classification with SCADA data for condition monitoring of wind turbines. In 9th ECCOMAS thematic conference on smart structures and materials (pp. 263-273).

[11] Goh, J., Adepu, S., Tan, M., & Lee, Z. S. (2017, January). Anomaly detection in cyber physical systems using recurrent neural networks. In 2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE) (pp. 140-145). IEEE.

[12] Perez, R. L., Adamsky, F., Soua, R., & Engel, T. (2018, August). 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.

[13] Aryal, K., Gupta, M., & Abdelsalam, M. (2023). Analysis of Label-Flip Poisoning Attack on Machine Learning Based Malware Detector. arXiv preprint arXiv:2301.01044.

[14] R. Taormina, S. Galelli, N. O. Tippenhauer, E. Salomons, A. Ostfeld, D. G. Eliades, M. Aghashahi, R. Sundararajan, M. Pourahmadi, M. K. Banks, B. M. Brentan, E. Campbell, G. Lima, D. Manzi, D. Ayala-Cabrera, M. Herrera, I. Montalvo, J. Izquierdo, E. Luvizotto, Jr, S. E. Chandy, A. Rasekh, Z. A. Barker, B. Campbell, M. E. Shafiee, M. Giacomoni, N. Gatsis, A. Taha, A. A. Abokifa, K. Haddad, C. S. Lo, P. Biswas, B. Pasha, M. Fayzul K.and Kc, S. L. Somasundaram, M. Housh, and Z. Ohar, "The battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks," Journal of Water Resources Planning and Management, vol. 144, no. 8, Aug. 2018.

[15] Zhou, X., Li, Y., Barreto, C. A., Li, J., Volgyesi, P., Neema, H., & Koutsoukos, X. (2019, November). Evaluating resilience of grid load predictions under stealthy adversarial attacks. In 2019 Resilience Week (RWS) (Vol. 1, pp. 206-212). IEEE.

[16] Li, P. H., Topcu, U., & Chinchali, S. P. (2022). Adversarial Examples for Model-Based Control: A Sensitivity Analysis. arXiv preprint arXiv:2207.06982.

[17] Wu, M., Roy, R., Torre, P. S., & Hidalgo-Gonzalez, P. (2022). Effectiveness of learning algorithms with attack and defense mechanisms for power systems. Electric Power Systems Research, 212, 108598.

[18] Maiti, R. R., Yoong, C. H., Palleti, V. R., Silva, A., & Poskitt, C. M. (2022). Mitigating Adversarial Attacks on Data-Driven Invariant Checkers for Cyber-Physical Systems. IEEE Transactions on Dependable and Secure Computing.

[19] Kravchik, M., Demetrio, L., Biggio, B., & Shabtai, A. (2022). Practical Evaluation of Poisoning Attacks on Online Anomaly Detectors in Industrial Control Systems. Computers & Security, 122, 102901.

[20] Zhou, Y., Ding, Z., Wen, Q., & Wang, Y. (2022). Robust Load Forecasting towards Adversarial Attacks via Bayesian Learning. IEEE Transactions on Power Systems.

[21] Zizzo, G., Hankin, C., Maffeis, S., & Jones, K. (2019, June). Adversarial machine learning beyond the image domain. In 2019 56th ACM/IEEE Design Automation Conference (DAC) (pp. 1-4). IEEE.

[22] Yaghoubi, S., & Fainekos, G. (2019, April). Gray-box adversarial testing for control systems with machine learning components. In Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control (pp. 179-184).

[23] Erba, A., Taormina, R., Galelli, S., Pogliani, M., Carminati, M., Zanero, S., & Tippenhauer, N. O. (2019). Real-time evasion attacks with physical constraints on deep learning-based anomaly detectors in industrial control systems. arXiv preprint arXiv:1907.07487.

[24] Sayghe, A., Zhao, J., & Konstantinou, C. (2020, August). Evasion attacks with adversarial deep learning against power system state estimation. In 2020 IEEE Power & Energy Society General Meeting (PESGM) (pp. 1-5). IEEE.

[25] "MPS PA Filtration Learning System", [Online], Available: https://www.festo-didactic.com/int-en/learning-systems/process-automation/mps-pa-stations-and-complete-systems/mps-pa-filtration-learning-system.htm?fbid=aW50LmVuLjU1Ny4xNy4xOC4xMDgyLjQ3ODU

[26] "Simulink", [Online], Available: https://www.mathworks.com/products/simulink.html

[27] Robles-Durazno, A., Moradpoor, N., McWhinnie, J., Russell, G., & Maneru-Marin, I. (2019). Implementation and detection of novel attacks to the PLC memory of a clean water supply system. In Technology Trends: 4th International Conference, CITT 2018, Babahoyo, Ecuador, August 29–31, 2018, Revised Selected Papers 4 (pp. 91-103). Springer International Publishing.

[28] Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. Soft computing and industry: Recent applications, 25-42.

[29] "A Quasi-SVM in Keras", [Online], Available: https://keras.io/examples/keras_recipes/quasi_svm/

[30] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[31] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In 2016 IEEE European symposium on security and privacy (EuroS&P) (pp. 372-387). IEEE.

[32] "Adversarial Robustness Toolbox (ART)" [Online]. Available: https://adversarial-robustness-toolbox.org/.

[33] "Malicious Control System Cyber Security Attack Case Study–Maroochy Water Services, Australia [Online]. Available: https://www.mitre.org/sites/default/files/pdf/08_1145.pdf.

[34] "Insider charged with hacking California canal system", [Online]. Available: https://www.computerworld.com/article/2814174/insider-charged-with-hacking-california-canal-system.html

[35] "Kansas man indicted in connection with 2019 hack at water utility", [Online]. Available: https://cyberscoop.com/kansas-ellsworth-water-district-hack-travnichek/

[36] "Ransomware Attack on Michigan Utility Provider Highlights Organizational Vulnerabilities", [Online]. Available: https://www.govtech.com/security/ransomware-attack-on-michigan-utility-provider-highlights-organizational-vulnerabilities.html

[37] "Ransomware attack hits North Carolina water utility following hurricane", [Online]. Available: https://www.csoonline.com/article/3314557/ransomware-attack-hits-north-carolina-water-utility-following-hurricane.html

[38] "Targeted attacks on Israeli water supply and wastewater treatment facilities", [Online]. Available: https://ics-cert.kaspersky.com/publications/news/2020/04/29/israel-water-cyberattacks/

[39] "What We've Learned from the Dec 1st Attack on an Israeli Water Reservoir?", [Online]. Available: https://www.otorio.com/blog/what-we-ve-learned-from-the-december-1st-attack-on-an-israeli-water-reservoir/

[40] "U.S. Water Supply System Being Targeted By Cybercriminals", [Online]. Available: https://www.forbes.com/sites/jimmagill/2021/07/25/us-water-supply-system-being-targeted-by-cybercriminals/?sh=29d4ff6e28e7

[41] "Cyber attacks and data breaches in review: May 2021", [Online]. Available: https://www.itgovernance.eu/blog/en/cyber-attacks-and-data-breaches-in-review-may-2021