International Journal of Circuit, Computing and Networking

E-ISSN: 2707-5931
P-ISSN: 2707-5923
IJCCN 2020; 2(2): 41-46
Received: 10-05-2021
Accepted: 19-06-2021

Manas Kumar Yogi
Assistant Professor, CSE, Department of Pragati Engineering, College of Autonomous, Surampalem, East Godavari, Andhra Pradesh, India

KVV Subba Rao
Assistant Professor, CSE, Department of Pragati Engineering, College of Autonomous, Surampalem, East Godavari, Andhra Pradesh, India

Corresponding Author:
Manas Kumar Yogi
Assistant Professor, CSE, Department of Pragati Engineering, College of Autonomous, Surampalem, East Godavari, Andhra Pradesh, India

Impact analysis of using ML techniques on imbalanced datasets for leveraging security of industrial IoT

Manas Kumar Yogi and KVV Subba Rao

DOI: https://doi.org/10.33545/27075923.2021.v2.i2a.31

Abstract
Machine learning calculations have been demonstrated to be reasonable for getting stages for IT frameworks. Nonetheless, because of the basic contrasts between the industrial internet of things (IIoT) and normal IT organizations, a unique exhibition survey should be thought of. The weaknesses and security prerequisites of IIoT frameworks request various contemplations. In this paper, we study the reasons why machine learning should be coordinated into the security components of the IIoT, and where it right now misses the mark in having an agreeable exhibition. The difficulties and certifiable contemplations related with this matter are concentrated in our exploratory plan. In this paper, we advocate a novel mechanism to evaluate the various ML techniques, with the help of an IIoT testbed.

Keywords: IIoT, imbalanced, sensor, security, attack

1. Introduction
Utilizing the internet of things (IoT) innovation in the industrial control systems (ICSs), known as the industrial internet of things (IIoT), has gotten extremely famous lately. ICSs are the fundamental piece of each basic framework and have been used for quite a while to oversee industrial machines and cycles. Supervisory Control and Data Acquisition (SCADA) systems often deal with the ICSs and are considered as the biggest subset of these systems. Principle parts of these systems are to perform continuous observing and connecting with the gadgets, constant assembling and dissecting the data, and logging every one of the occasions that occur in the framework. Using IoT innovation in these systems improves the organization knowledge and security in advancement and computerization of industrial cycles. IIoTs are for the most part strategic applications with high-accessibility necessities. Their activities lead to an enormous measure of data that can be effectively overseen through huge data investigation techniques. Previously, to get ICSs from malignant external assault, these systems used to be secluded from the rest of the world. Nonetheless, ongoing advances, expanded availability with corporate organizations, and usage of internet interchanges to send the data all the more helpfully have presented the chance of digital assaults against these systems. Because of the touchy idea of the industrial application, security is the principal concern. Since intrusion is the essential security worry in IIoT, an intrusion detection system (IDS) is a basic piece of these applications to give a protected climate. Stuxnet worm, which was uncovered in 2010 and as of late returned (late December 2017), and Triton malware against the ICSs raised familiarity with the need for extraordinary thoughtfulness regarding the security of such significant system. Through the major contrasts between the ICSs and the normal IT systems, their basic weaknesses and needs are distinctive. Besides, ICSs have a specific kind of traffic and data utilizing particular IIoT communication conventions. Because of every one of these reasons, appropriate determination should be viewed as with regards to planning an IDS for ICSs.
Machine learning-based security arrangements have been generally utilized in giving security to IT systems. In any case, the reasonableness of these methods for IIoT applications is questionable. The principle security worry in IIoT devices is to detect any entrance into the system. Intrusion detection accompanies uncommon highlights, for example, significant imbalanced datasets that occasionally the prepared machine learning (ML) calculations will most likely be unable to detect the attack. The researchers have never genuinely considered the imbalanced datasets issue confronting ML calculations, where the genuine hindrances are, and how extraordinary execution metrics would respond to this issue.

In this paper, subsequent to examining how ML can be beneficial in IDS applications, we will consider the situations where current machine learning calculations miss the mark regarding giving the necessary degree of safety. All the more specifically, our fundamental spotlight is on the imbalanced dataset issue in IIoT. The metrics that can decently pass judgment on the presentation have been contrasted with measure their viability.

2. Related Work

In this segment, we audit a portion of the connected examination works. In this paper we have faced the challenge surfaced due to imbalanced IIoT dataset issue with the significantly low number of minority tests has not been contemplated yet. The intrusion detection issue in keen grids utilizing a few distinctive ML procedures has been concentrated by multiple researchers. A few countermeasures to defeat the issue of imbalanced dataset have been inspected. They have utilized ADFA-LD dataset that comprises of 12.5% attack data. Notice that this proportion isn't realistic on account of IIoT applications. Here we manage an efficient inconsistency tests in our applications, which makes the outcomes nearer to true situations.

Different inspecting methods to defeat the imbalanced dataset issue have been researched. The used datasets are separated from Github and Source forge projects with 15% lop-sidedness proportion. In recent works, many IDS have been applying amalgamation of Naive Bayes and J48 strategies. The dataset was fabricated utilizing gas pipeline system of the Distributed Analytics and Security Institute, Mississippi State University, Starkville, MS. Their dataset comprised of various kinds of attacks like reconnaissance, code injection, response injection, script injection. The J48 classifier was first utilized as a supervised trait channel. At that point, the Naive Bayes classifier was utilized to build up the abnormality based intrusion detection. The proportion of attack traffic in their examination was about 21.87%, which is far higher than a true case.

In previous works, few ML algorithms, Naive Bayes, Random Forests, One R, J48, NNge (non-settled summed up models), SVM (support vector machines) for IDS have been studied by multiple researchers. The attack traffic is created from two sorts of code injection set, script injection attacks, data injection attacks. Seven unique variations of data injection attacks were attempted to change the pipeline pressure esteems, and four distinct variations of command injection attacks to control the commands that control the gas pipeline. They utilized accuracy and review metrics to try to have a reasonable assessment notwithstanding the imbalanced dataset with about 17% attack traffic. K-means procedure, which is an unsupervised grouping calculation, for IDS has also been utilized in. An open-source virtual PLC (Open PLC stage) alongside AES-256 encryption is utilized to recreate an ICS. They have led three distinct kinds of attacks against their system, script injection, DoS, and capture attempt (listen in). In any case, they have not given any data on the level of attack data that was utilized for preparing.

One class SVM (OCSVM) as an appropriate inconsistency based IDS has also been proposed in recent years. They announce that OCSVM is a decent choice on the grounds that the dataset is imbalanced. The creators just utilized two highlights of traffic (data rate and parcel size) of an electric lattice. The prepared model did exclude any malicious attack data, and the prepared dataset was caught during ordinary activity of a SCADA system.

3. Percussions of imbalanced Datasets

The legacy IDSs can't keep up with the attackers continually evolving their methods. Additionally to face the new attacks that appear every day, or in scenarios where the attack is carried out accordingly for instance related to the man-in-the-middle attack, intelligent IDSs are required. Strong IDS that uses ML algorithms can detect any strange action if the training procedures are performed with proper care.

The intelligent IDS are based on the way that AI algorithms can detect oddity patterns that are difficult for a human to discover. Unlike rule-based IDSs, ML-based IDSs can successfully detect new types of attacks, different variations of a specific attack and obscure or zero-day attack. The zero-day exploit takes advantage of obscure vulnerabilities (i.e., no idea exists by the developers) to manipulate the processes or the system. These are altogether the reasons that ML ought to be applied in designing IDSs.

Despite the confidence in the capacity of machine learning (ML) to detect anomalies with great impact, there exist several challenges that arise when considering their applicability in IIoT. Without addressing these problems, the ML algorithms are unable to work properly. With respect to an IIoT environment, some of these challenges might be manageable and some might not, due to the nature of these systems and their associated security aspects. The foremost consideration is to choose proper features from the network traffic dataset. Sensor data in IIoT are normally obtained during an extended period from many sensors with different sampling frequencies, which results in high-dimensional datasets. Using crude data like this will add a large delay in training and detecting process. Then again, if the selected features don't fluctuate during the attacks, even the best algorithm won't detect an intrusion or a strange circumstance using that feature. Subsequently it becomes necessary to extract discriminating features to be able to use ML techniques. By employing techniques pertaining to power spectral density, Fourier analyses, the linear feature extracting method, and principal component analysis (PCA) are few instances that could be tried out to reduce the dimensionality and determine the most useful features. Then again, due to the confidentiality and user security restrictions, industrial companies scarcely release their protected network data on the intrusion attacks that might have occurred. Hence, training the ML algorithms on data collected from real networks of real industrial IoT is practically impossible. Hence, a large portion of the available research work in this area is done on commercial or public datasets that may not be specific to IIoT. Yet another boundary in utilizing these ML techniques directly in companies or industrial networks. For the same purpose, we constructed our IIoT testbed and considered all assumptions to make it as resembling as possible to a real industrial plant. It is also observed that, in any real world IIoT system, the number of intrusion attack samples is significantly low. Since the intruders don't wish to be exposed; they as a rule run their attacks randomly in brief periods of time. This leads to a very low measure of attack data to train the ML algorithm. We term this challenge as an imbalanced training dataset. In other words, imbalanced dataset means the percentage of the attack traffic compared

to the ordinary traffic in the whole dataset is very low. In the following subsection, we will discuss this challenge in more details. Machine learning techniques like other artificial intelligence classifiers generally perform best on balanced datasets. The problem of imbalanced datasets, specifically those in severe cases (i.e., considerable amount of samples from one class are compared to the other), is a critical issue in the training process. Examples of such cases include detecting rare anomalies like fraudulent bank exchanges and identification of rare diseases. There are various security concern of IIoT applications. Among them Intrusion detection is posing a challenge to security personnel. Due to the large measure of sensed data from IIoT devices (i.e., a large measure of typical traffic) from one perspective; and random, rare attack traffic (i.e., a limited quantity of attack traffic) then again, the IIoT's security suffers greatly from imbalance problem. Although various methods of mitigation are suggested for this problem, through changing the sampling method. Under sampling, over-sampling, or combinations of both are some examples. However, each of these techniques comes with several downsides. In simple terms, under-sampling means including fewer instances from the greater part class, and oversampling means including more samples of the minority class. One problem with under-sampling is the chance of losing useful information, while over-sampling might cause over fitting problems. As such, in a real IIoT system, any of these techniques might lead to an undesirable system model, the resulting models may not be accurate to solve the intrusion detection problem in practice. Besides, they result in different outcomes compared to the models trained with the full dataset. There are other challenges that should be considered when training ML techniques for intrusion detection. Here, we briefly mentioned the ones that are generally critical for IIoT applications. Due to every one of these reasons, the appropriateness of ML under different circumstances should be considered. In the next section, we study the cutoff points on the imbalanced dataset challenge on our constructed IIoT testbed to show the real restrictions, when it comes to training ML-based IDSs.

4. Proposed Mechanism

Utilization of a real testbed allows conducting real cyber-attacks and collecting a real dataset containing both normal and attack traffic. As we know the main task of an ICS is to have provisions for remote monitoring and automatically controlling the industrial processes, we have emulated a real-world IIoT control system. Fig. 1 shown below depicts the platform of our testbed. We chose a popular IIoT system that supervises the water level and turbidity quantity in the water storage tank. We have found that this type of smart model is employed in industrial reservoirs and water distribution as a part of the water treatment and distribution process. Notably the testbed has multiple elements. It has logs which show operation history, programmable logic controllers (PLCs), human-machine interference (HMI), a three-light alarm, water levels sensors and sensors to measure the turbidity, actuators (e.g., alarms, valve, pumps, and buttons), and control buttons (On, Off, Light Indicator). The primary reason for using the HMI in an ICS is to make it easy for the human controllers to monitor the current behaviour of the system, interact with the IIoT devices, and receive alarms indicating abnormal behaviours. Moreover, since the sensors and relays cannot communicate directly, PLCs are used to collect the sensed data and send commands to the actuators.

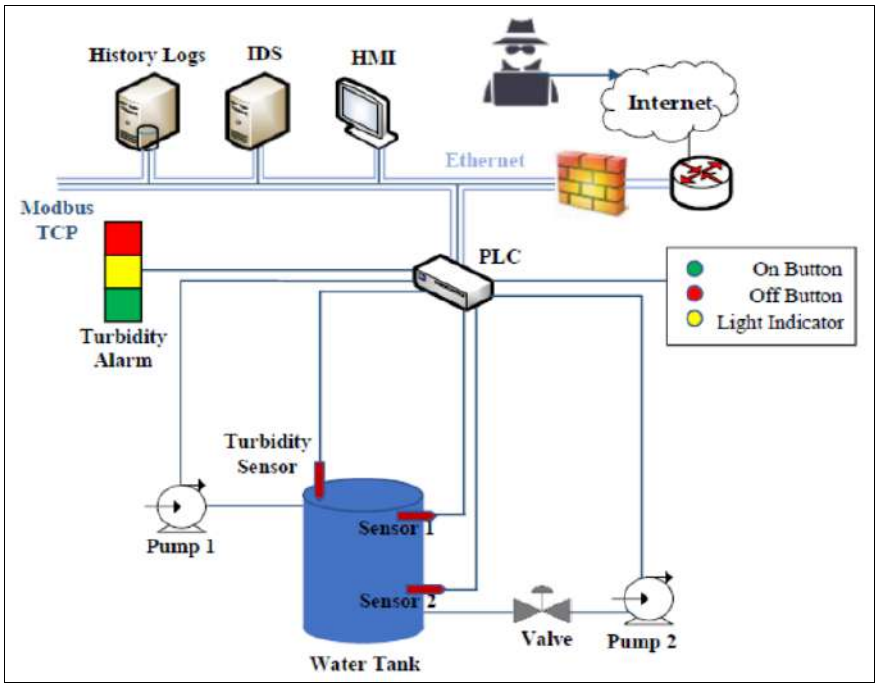

Fig 1: Proposed Mechanism of test bed

4.1 Real-World Testbed Implementation

The water storage tank has two level sensors: Sensor 1 and Sensor 2, which are used to monitor the water level in the tank. As soon as the water reaches the maximum defined level in the system, Sensor 1 sends a signal to the PLC. The PLC stops the water Pump 1 that is used to top off the tank, opens the valve, and turns on the water Pump 2 draws water from the tank. In a scenario when the water reaches the base defined level in the system, Sensor 2 sends a signal to the PLC. PLC closes the valve, kills the Pump 2, and turns on

the Pump 1 to top off the tank. Now the process repeats again when the water level reaches the maximum level. Meanwhile, there is an analog turbidity sensor that is integrated into the system to measure the turbidity of the water. On the basis of the two defined thresholds in the system, PLC illuminates one of the red, yellow or green lights of the Turbidity Alarm, green light denotes the water has an acceptable level of turbidity, red indicates that the turbidity has crossed the acceptable limits, and yellow falls between the two limits.

The under consideration IIoT testbed takes the data from sensors, and the status of the system from the PLC using the Modbus communication convention and displays them to the operator through the HMI interface. Since Modbus is one of the most popular IIoT conventions, and it's widely used by large industries, we chose this convention.

The PLC model which we have used in our testbed is Siemens S7-400 model. The Statement List (STL) is used to program the PLC. The turbidity sensor is SEN0189, and the water level sensors are Onoflot Automatic Water Level Controller. The deployed water pumps are Falcon EGP-05.

4.2 Attack Model

In this research work, we focused on the simple script injection attack in a water storage scenario to compromise the functionality of the PLC. These attacks were carried out using the Fedora Security Linux Penetration Testing Distribution using special programs for corrupt script injections in Industrial Control Systems. All data generated during the attacks as well as regular traffic (without attacks) was gathered and recorded by and Wire shark network tool.

During the script injection attack, our target is the PLC. During the initial stage of attack, the attacker tries to establish a connection to the network so as to read all the PLC register values and record them into a normal txt file. Just after the PLC registers information in the registry, the attacker rewrites some of the PLC registers that are vital to the physical process. For instance, this attack was executed while Pump 2 was supposed to draw water from the tank, it was abruptly terminated by the attacker, and Pump 1 started, and the water overflowed from the tank. Yet another scene was created, when the attacker turned on the wrong turbidity alarm light, in the way that, while the turbidity level indicated with a reading of high, and the red light was supposed to be on, the attacker reversed the operations by turning off the red light and turning on the green light instead.

4.3 Feature Selection

In this work, we used Artificial Neural Network (ANN). The model undergoes the operations like training and testing upon the imbalanced dataset gathered from the testbed, and comparison is performed on results of their performance. We must understand that in training the algorithm, important step is selecting and extracting features from the raw network traffic traces. For our work, to designing a robust system for the purpose of detecting intrusion, we chose numerous factors. These factors are common in network streams and also show a good variation during the attack phases. Some of the factors are Packet loss percentage at source and destination port, mean flow, jitter percentage etc. The influence of each attack factor depends on the type and frequency of the attack. During an attack scenario this factors behave abnormally.

5. Experimental Results

5.1 Analysis with imbalanced datasets

For us the main objective here is to examine the efficiency of ANN in detecting anomaly through different imbalance ratios. We collected a new dataset of 4.2 GB, for a total of about 80 hours. The number of attacks at each trial has been kept equal to 10000 samples, and accordingly, we added normal traffic to assemble the desired ratios. At each round of training, we divided the dataset into 80% for training and 20% for testing.

5.2 Performance Metrics

We have already observed that conventionally, the performance of the trained algorithms is measured by metrics which are derived from the confusion matrix. The confusion matrix has the below elements.

1. True Negatives (TN): Represents the number of normal packets correctly classified as normal.
2. True Positives (TP): Represents the number of abnormal packets (attacks) correctly classified as attacks.
3. False Positive (FP): Represent the number of normal packets incorrectly classified as attacks.
4. False Negative (FN): Indicates the number of abnormal packets (attacks) incorrectly classified as normal packets.

According to the confusion matrix, the metrics that are used in this work to evaluate the performance of the ML algorithms are as follows:

1. Accuracy: This metric represents the percentage of the correctly predicted samples considering the total number of predictions.

The Accuracy metric is defined as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (1)$$

2. Matthews correlation coefficient (MCC): It measures the quality of the classification. MCC is a strong metric, especially in case of imbalanced datasets, showing the correlation agreement between the observed values and the predicted values. The limitation of this technique is it is not suitable for benchmarking two hierarchical clusterings.

Matthews correlation coefficient is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2)$$

3. Fowlkes–Mallows index: So we propose to use the Fowlkes–Mallows index is an external evaluation method that is deployed to compute how much similar two clusters are, and also a metric to measure confusion matrices. This similarity index could be either among two hierarchical groups or a group and a benchmark classification. When we obtain a higher value for the Fowlkes–Mallows index we can infer a greater similarity between the clusters and the benchmark classifications.

If we represent,

TP as the number of pairs of points that are present in the same cluster in both A1 and A2.

FP as the number of pairs of points that are present in the same cluster in A1 but not in A2.

FN as the number of pairs of points that are present in the same cluster in A2 but not in A1
TN as the number of pairs of points that are in different clusters in both A1 and A2.
Then Fowlkes–Mallows index for two clusterings can be defined as

$$FM = \sqrt{\frac{TP}{TP+FP}} \times \sqrt{\frac{TP}{TP+FN}} \quad (3)$$

Where TP is the number of true positives, FP indicates the number of false positives, and FN denotes the number of false negatives. Accuracy is the most popular metric but better techniques are MCC and FM index. The FM index is more suitable for our IIoT scenario as we have to work with clusters of functional elements for industrial control system. In next section we provide the comparative analysis of FM index along with the performance of accuracy and MCC index.

5.3 Results

In this section we show the results of our proposed mechanism detecting the attacks through different ratios of imbalanced data. The graph below shows the performance of accuracy metric versus MCC and FM index. The techniques are robust to say but we can observe that the FM index technique provides a better understanding for the researchers and designers of ICS.

Fig 2: Imbalance Ratio in Dataset versus Performance Improvement

The following table summarizes the results.

Table 1: Imbalance ratio Vs. Improvement in Performance

Sl. No.	Metric	Imbalance Ratio in %	Performance Improvement in %	Data from confusion matrix included
1	Accuracy	10	6.6	Yes
2	MCC index	10	7.8	Yes
3	FM index	10	11.2	Yes

From the above results we can infer that the FM index outperforms the other two metrics when the imbalance ratio is the highest, i.e. 10%.There is nearly improvement of 5% in performance as the imbalance ratio increases from 2% to 10% when we compare the usage of accuracy metric with FM index. Hence we can fruitfully conclude that for our testbed and similar test beds using the FM index to evaluate our proposed learning model for the IDS will be relatively more beneficial.

6. Conclusion

The overall protection of the IIoT infrastructure is critical. Interruption location is the main security worry in these applications. Machine learning arrangements and enormous data analytics have been broadly used to guarantee a protected platform in these frameworks. Notwithstanding, with regards to a real-world scenario and applying these algorithms practically, they now and then fall short. The main focal point of this paper was contemplating imbalanced dataset issues and show in which broaden the machine learning algorithms are able to help.

7. References

1. Teixeira MA, Zolanvari M, Salman T, Jain R. An Intrusion Detection System Based on Deep Learning Algorithm for Industrial Control Systems, Submitted to EURASIP Journal on Information Security 2018.
2. Teixeira MA, Salman T, Zolanvari M, Samaka M, Jain R, Meskin N. SCADA System Testbed for Cyber security Research Using Machine Learning Approach, Future Internet 2018;10:76.
3. Promper C, Engel D, Green RC. Anomaly detection in smart grids with imbalanced data methods, in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI 2017.
4. Shekarforoush S, Green R, Dyer R. Classifying commit messages: A case study in resampling techniques, 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK 2017, 1273-1280.

5. Ullah I, Mahmoud QH. A Hybrid Model for Anomaly-based Intrusion Detection in SCADA Networks, 2017 IEEE International Conference on Big Data, Boston, MA 2017, 2160-2167.
6. Beaver JM, Borges-Hink RC, Buckner MA. An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications, 12th International Conference on Machine Learning and Applications, Miami, FL 2013, 54-59.
7. Alves T, Das R, Morris T. Embedding Encryption and Machine Learning Intrusion Prevention Systems on Programmable Logic Controllers, in IEEE Embedded Systems Letters 2018, 1-6.
8. Maglaras LA, Jiang J. Intrusion detection in SCADA systems using machine learning techniques, 2014 Science and Information Conference, London 2014, 626-631.
9. Siddavatam IA, Satish S, Mahesh W, Kazi F. An Ensemble Learning for Anomaly Identification in SCADA System, 2017 7th International Conference on Power Systems (ICPS), Pune 2017, 457-462.
10. Schneider PLC, M241CE40, [Online]. Available: http://www.filkab.com/files/category_files/file_3073_Bg.pdf. [Accessed August 2018].
11. Erickson KT. Programmable Logic Controllers: An Emphasis on Design and Application, Dogwood Valley Press, LLC 2011.
12. Argus, [Online]. Available: https://qosient.com/argus/. [Accessed August 2018].
13. Wire-shark, [Online]. Available: https://www.wireshark.org/. [Accessed August 2018].