Institute of Theoretical and Applied Informatics

Polish Academy of Sciences

Extended Summary of Doctoral Dissertation

Online Self-Supervised Learning Intrusion Detection Towards Secure Internet of Things

MERT NAKIP

SUPERVISOR: PROF. DR. EROL GELENBE

Gliwice, Poland
2023

Abstract

Secure Internet of Things (IoT) systems are extremely difficult to achieve as most IoT devices are low-cost and low-maintenance devices with low computing power and human intervention to run complex security methods. Therefore, lightweight data-driven, especially Machine Learning (ML)-based, security methods have been developed particularly for IoT systems. On the other hand, these methods often learn offline from large data collected through extensive simulations, which may be time consuming and provide biased (misleading) data. This thesis investigates the open issues and ways to enable fully online and lightweight learning for ML-based intrusion detection paving the way towards secure IoT.

We first develop an Intrusion Detection System (IDS) that learns the normal traffic patterns of the IoT network and detects both malicious network traffic packets and compromised devices. This IDS is based on Deep Random Neural Network (DRNN) model combined with originally proposed traffic metrics and Statistical Whisker based Benign Classifier (SWBC). For each of malicious traffic detection and compromised device identification, we propose a set of original network traffic metrics that enable accurate recognition of Botnet traffic patterns and footprints of the attacker. In addition, we develop a new SWBC algorithm to classify traffic packets as benign and malicious by learning the classification criterion based on the DRNN outputs on the training data. We further present offline and quasi-online (incremental and sequential) learning algorithms for our IDS.

Subsequently, we evaluate the performance of our IDS with both offline and quasi-online learning algorithms for Botnet DDoS, DoS, and zero-day attacks on three public datasets. The results show the superior performance of our IDS with low computation time compared to well-known ML models. The results also reveal the potential of online learning for intrusion detection.

Finally, in order to enable fully online learning of ML-based IDS requiring no human intervention, we propose the novel Self-Supervised Intrusion Detection (SSID) framework. For the learning of utilized IDS, the SSID framework collects and labels traffic packets based only on the decisions of the IDS and their statistically measured trustworthiness. The SSID framework enables IDS to adapt time-varying characteristics of the network traffic quickly, eliminates the need for offline data collection, prevents human errors in data labeling, and avoids labor costs for model training and data collection through experiments. Therefore – as the experimental results on public datasets for malicious traffic and compromised device detection using well-known ML models also suggest – SSID is very useful and advantageous to develop an online learning ML-based IDS for IoT systems.

Publications of the Author

The publications of the author that are included in this thesis or produced from the thesis are listed below. The extended list of author's publications including those published during doctoral studies of the author but not directly related to the content of this thesis can be found in the thesis.

Journal Papers:

- M. Nakıp and E. Gelenbe, "Fully Online Self-Supervised Learning Framework for Machine Learning based Intrusion Detection," in *arXiv*, 2023, *Preprint*.
- E. Gelenbe and M. Nakıp, "Traffic Based Sequential Learning During Botnet Attacks to Identify Compromised IoT Devices," in *IEEE Access*, vol. 10, pp. 126536-126549, 2022.

Conference Papers:

- E. Gelenbe and M. Nakıp, "Real-Time Cyberattack Detection with Offline and Online Learning," *2023 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, London, United Kingdom, 2023, pp. 01-06.
- E. Gelenbe and M. Nakıp, "G-Networks Can Detect Different Types of Cyberattacks," *2022 30th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, Nice, France, 2022, pp. 9-16.
- M. Nakıp and E. Gelenbe, "Botnet Attack Detection with Incremental Online Learning," *2021 Security in Computer and Information Sciences (EuroCybersec)*, Nice, France, 2022, pp. 51-59.
- M. Nakıp and E. Gelenbe, "MIRAI Botnet Attack Detection with Auto-Associative Dense Random Neural Network," *2021 IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, 2021, pp. 01-06.

Extended Summary of Thesis

Motivation

The majority (approximately 52%) of Internet of Things (IoT) devices are low-cost and low-maintenance devices with limited resources. These devices often have either software vulnerabilities or weak login credentials (or both) making them an easier target for attackers than user-enabled devices, such that 70% of all IoT devices are considered to be vulnerable [1].

The number of such IoT devices is rapidly increasing with expanding application areas, leading to more security breaches and consequential cyberattacks. For example, it is estimated that an average smart home is targeted by $12,000$ attacks in a single week. As a result, cybersecurity of IoT networks has become one of the main concerns, such that cyberattacks on IoT devices are considered to be the primary concern by 33% of cybersecurity companies [2].

Therefore, it is crucial to enhance the cybersecurity of IoT networks in order to ensure their safe, trusted and seamless operation, and achieving secure networked systems. One can say that an Intrusion Detection System (IDS) is a very important component to enhance the cybersecurity of the networked system as it enables network management systems to take early actions and respond attacks before the damage occurs. On the other hand, IoT devices are often deployed in massive IoT networks [3] and operate with minimum (if not zero) human intervention. While systemic approaches to improving the security of cyberphysical systems have mostly been suggested [4, 5], it is difficult (if not totally impossible) to burden simple IoT devices with complex security functionalities.

One may say that it is a highly challenging task to develop advanced security methods, such as Machine Learning (ML)-based IDS, for IoT devices and networks because of the following reasons: 1) IoT devices often have insufficiently low computational resources implement complex algorithms. 2) Data-driven (e.g. ML-based) algorithms require large amounts of data that are difficult to collect as they require considerable labor, high development costs, and long deployment time. 3) The advanced security methods are mostly customized for the individual system or network to which they are applied, as their parameters are directly optimized for that system. As a result, when these algorithms need to be deployed for a new system, a significantly large amount of work has to be manually repeated.

Thesis Contributions

The main purpose of the present thesis is to research online learning for ML-based IDS towards the development of secure IoT systems. Our research aims to address the above issues and provide a lightweight, easy-to-implement algorithm for intrusion detection, which is one of the key security assurance methods.

1. We propose a novel self-supervised learning framework for ML-based IDS, called Self-Supervised Intrusion Detection (SSID) framework, that enables the fully online learning of the IDS parameters requiring no human intervention. Within the SSID framework, we statistically measure the trustworthiness of an ML-based IDS considering its generalization ability and the traffic packets that IDS learned. To this end, we present measures to estimate the generalization ability and the representativeness of the learned traffic.

   As its main advantages, the SSID framework

   - enables IDS to easily adapt time varying characteristics of the network traffic,
   - eliminates the need for offline data collection,
   - prevents human errors in data labeling, and
   - avoids labor costs for model training and data collection through experiments.

2. We develop an lightweight ML-based IDS using Deep Random Neural Network (DRNN). Through originally defined traffic measurements, IDS learns normal (benign) traffic patterns and then identifies abnormal traffic changes that may be indicative of a possible attack. To this end,

   - We determine original metrics which are easy to calculate using only the header information of the traffic packets, and highly effective to analyse the impacts of Botnet attacks on the network traffic and to capture the signatures of an attacker.
   - We develop three learning procedures for the developed IDS for offline, sequential, and incremental learning.
   - We develop a classification algorithm, called Statistical Whisker based Benign Classifier (SWBC), that identifies the malicious traffic comparing the actual traffic and the expected traffic estimated by auto-associative memory.

3. We finally develop a new system to identify compromised IoT devices (bots) based only on the network traffic without requiring access the device status or message contents. Along with the malicious traffic, it is crucial to identify compromised devices in order to pave the way to successfully prevent an attack from spreading or mitigate its impacts on the network.

Research Findings

Intrusion Detection System with Offline and Quasi-Online Learning:

In this thesis, we first develop an anomaly-based IDS with offline and quasi-online learning to detect both malicious traffic and compromised IoT devices during Botnet or zero-day attacks. This IDS, which is shown in Figure 1, is comprised of three main functionalities for extracting network traffic metrics, estimating expected metric values for normal "benign" traffic, and making a final decision on whether the analysed metrics indicate an intrusion.

In order to observe the impact of an intrusion on the network traffic and capture the signatures of an attacker, we propose original network traffic metrics specifically for each of the tasks malicious traffic detection and compromised device identification. In particular, for malicious traffic detection, three metrics

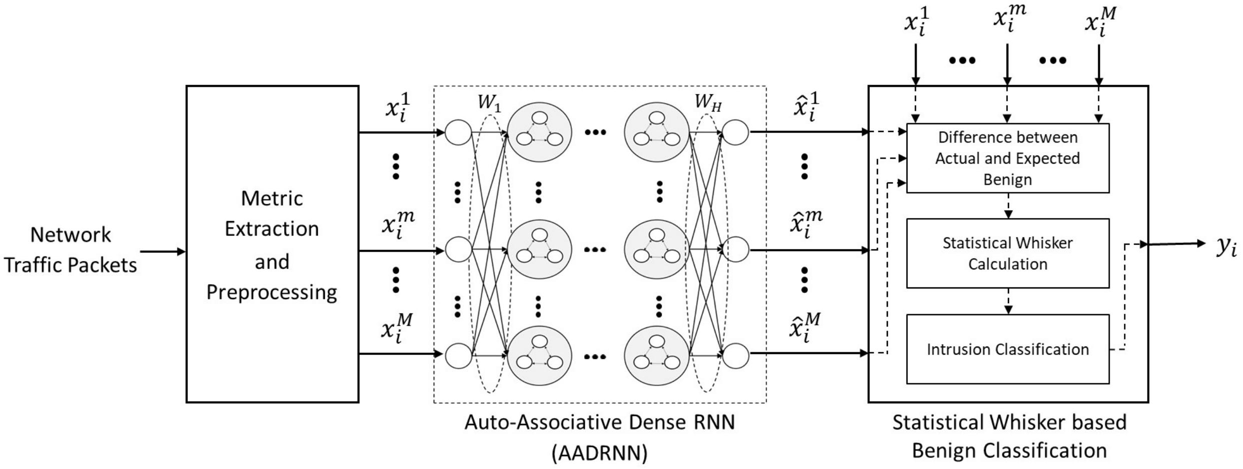
Network Traffic Packets
Metric Extraction and Preprocessing
x_i^1
x_i^m
x_i^M
W_1
W_H
$\hat{x}_i^1$
$\hat{x}_i^m$
$\hat{x}_i^M$
Auto-Associative Dense RNN (AADRNN)
Difference between Actual and Expected Benign
Statistical Whisker Calculation
Intrusion Classification
Statistical Whisker based Benign Classification
y_i

Figure 1: Architecture of the DRNN based attack detector with its three modules: Metric Extraction and Preprocessing, Auto-Associative DRNN (AADRNN), and Statistical Whisker based Benign Classification

are presented to measure the density of total network traffic while for compromised device identification, six metrics are presented to measure the density of received and transmitted traffic by an individual device.

We create an auto-associative memory using a DRNN model, called AADRNN, to estimate the metric values expected to be observed during the normal operation of the considered network. To retrieve the actual values of the metrics from their noisy versions, the DRNN model is trained using only the normal traffic packets. To this end, we develop offline and online auto-associative learning algorithms:

- **Offline Learning Algorithm** is based on the semi-supervised algorithm presented in [6]. During the learning process, the weight matrix at each layer of DRNN are determined by minimizing a cost function with L1 regularization via Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [7]. The cost function mainly measures the squared Euclidean distance between the input and the output vector of the considered layer.

- **Incremental – Quasi-Online – Learning Algorithm** enables the use of an attack detector without requiring the offline collection of benign traffic. To this end, it combines offline semi-supervised learning algorithm developed in [6] with sequential learning algorithm developed in [8]. This algorithm is comprised of initialization and incremental quasi-online learning stages. The initialization stage lasts for the transmission of I packets and is considered as the cold-start of the proposed IDS. Thus, the transmission of the first I packets are known to be benign packets, and the connection weights of AADRNN are initially learned using these packets. The incremental quasi-online learning stage is operated over time windows. At the end of each window, only the connection weights of the output layer of AADRNN are updated to learn the benign traffic collected in this window.

The final decision is made by comparing the expected metric values (i.e. the output of the AADRNN) with the actual metric values. To this end, we develop the novel Statistical Whisker-based Benign Classifier algorithm that detects an intrusion if the actual metrics differ significantly from the expected estimated metrics. The significance of the difference, as well as all parameters of the algorithm, is determined based only on the packet samples used for training, which are known to be benign traffic.

The performance of the proposed IDS is evaluated for malicious traffic detection and compromised device identification during Botnet attacks as well as for detecting zero-day (unknown) attacks. During the performance evaluation, we use publicly available datasets – namely Kitsune [9], KDD Cup'99 [10], and BotIoT [11] – and compare the performance of the proposed IDS against well-known ML models: Linear Regression (LR), Least Absolute Shrinkage and Selector Operator (Lasso), K-Nearest Neighbours Regressor (KNN), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Long-Short Term Memory (LSTM).

The results in Table 1 show that the proposed IDS (AADRNN) significantly outperforms the other methods achieving 99.82% true positive and 99.98% true negative percentages. In addition, Figure 2 reveals that the training time of the AADRNN is less than 0.1 secs while it makes online decision around 0.5 μ secs.

Table 1: Comparison of attack detection methods with respect to accuracy as well as each of the true positive, false negative, true negative and false positive percentages

Attack Detection Methods	Accuracy	True Positive	False Negative	True Negative	False Positive
AADRNN	99.84	99.82	0.18	99.98	0.02
KNN	99.79	99.79	0.21	99.75	0.25
Lasso	99.78	99.75	0.25	99.95	0.05
Simple Thresholding	93.18	93.09	6.94	93.63	6.37

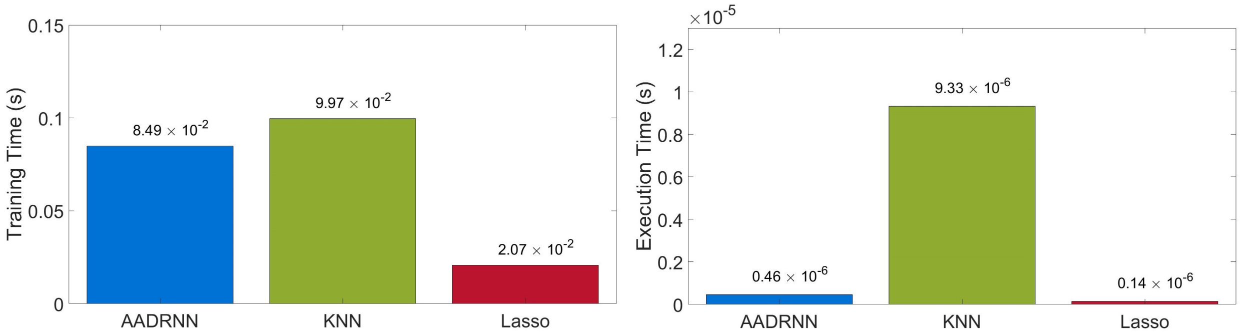

Figure 2: (left) Training times and (right) execution times of the different attack detection methods

When the proposed IDS is tested for detecting various types of cyberattacks simultaneously, the results shown in Figure 3 reveal that 1) the prediction accuracy is above 98% for 21 out of 37 attack types, and 2) the proposed IDS outperforms the state-of-the-art SVM-based One Class Classifier (SVM-OCC) by a considerable margin.

Then, the comparison between incremental quasi-online learning IDS and the offline learning IDS given in Figure 4 shows that the incremental quasi-online learning IDS achieve considerably close malicious traffic detection performance to the offline learning IDS while using a significantly lower number of packets for training. One may say that the quasi-online learning is a highly promising approach for intrusion detection in IoT networks.

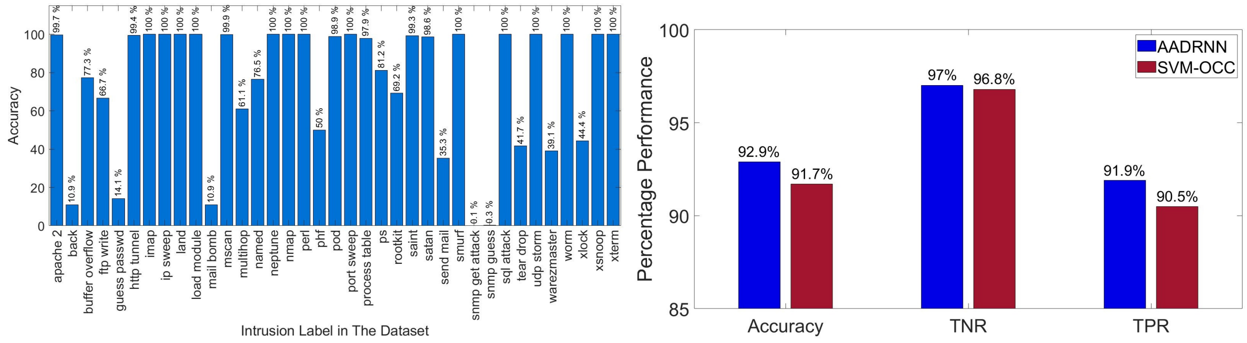

Figure 3: (left) Performance of the proposed IDS for detecting each attack type in KDD dataset with offline learning and (right) its comparison against the SVM-based One Class Classifier(SVM-OCC)

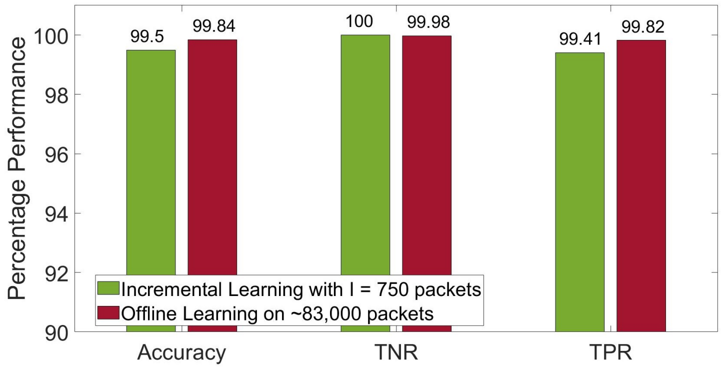

Figure 4: Performance comparison between AADRNN based IDS under Incremental Learning with 750 packets and that under Offline Learning with about 83,000 packets

Furthermore, the proposed IDS for compromised device identification, called CDIS, is evaluated for different types of DoS and DDoS attacks available online. The performance results in Figure 5 show that the CDIS can successfully detect possible malicious devices during various types of DDoS attacks where malware spreads over IoT devices.

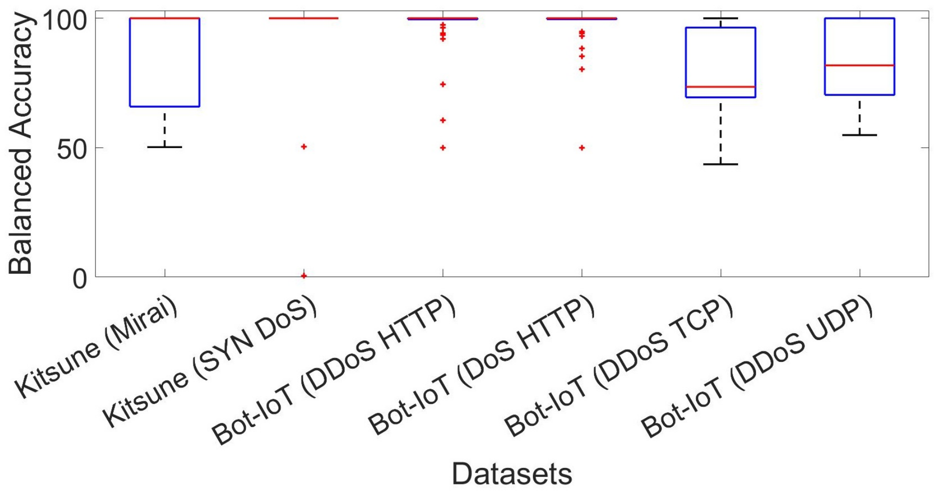

Figure 5: Performance of the proposed IDS for identifying compromised IoT devices during different types of attacks with sequential learning

In summary, the results reveal that the proposed IDS outperforms the existing methods significantly for both detecting malicious traffic and identifying compromised devices. In addition, quasi-online sequential and incremental learning algorithms have shown high potential for the development of high-performance online learning IDS, which requires low computation time and small data. On the other hand, online learning IDS still needs to be improved – as shall be provided by the Self-Supervised Intrusion Detection (SSID) framework – for more precise intrusion detection.

Fully Online Self-Supervised Intrusion Detection Framework:

As one of the main contributions of this thesis, we propose the novel SSID framework, which is designed to train any given IDS – whose parameters are calculated using the network traffic – fully online. The SSID framework automatically selects normal traffic packets for learning and decides when to update the parameters of the utilized algorithm. In this way, it completely eliminates the human intervention, the need for offline (labeled or unlabeled) data collection, and an offline training. Accordingly, the proposed framework differs sharply from existing work [12–18].

The SSID framework, as shown in Figure 6, comprises two successive learning stages initial learning and online learning. Initial learning aims to quickly adapt the IDS parameters for the network where the IDS is newly deployed, while online learning aims to update the parameters whenever an update is required to ensure the high detection accuracy of the IDS.

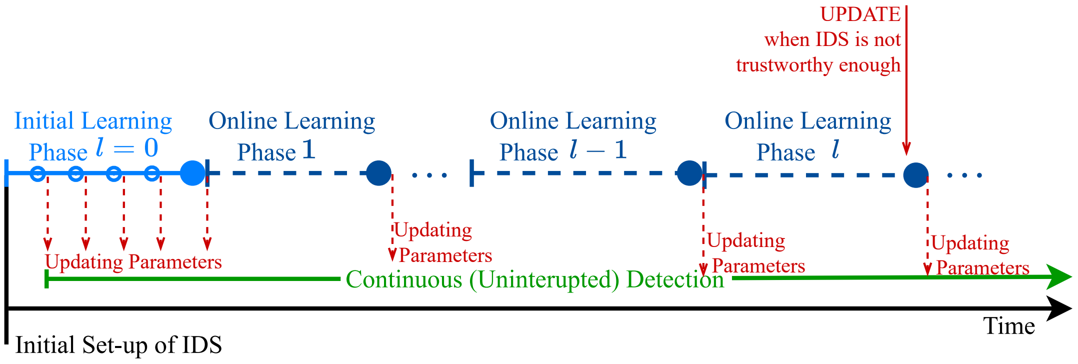

Figure 6: Detection and learning processes of IDS within the Fully Online Self-Supervised Intrusion Detection (SSID) framework

As one can see in Figure 7, during the real-time operation of the IDS, in parallel to the detection, the SSID framework performs the following main tasks:

- It continually estimates the trustworthiness of intrusion decisions to identify normal and malicious traffic, measuring the ability of the IDS to learn and generalize from data provided by SSID and the extent to which this data can represent current network traffic patterns.
- In order to provide training data for the IDS, the SSID framework selects and labels network traffic packets in a self-supervised manner based only on the decisions of IDS and the trust of SSID in those decisions.
- Considering the trustworthiness of the IDS, the selected training packets, and the latest state of network security, the SSID framework determines when to update the IDS parameters.

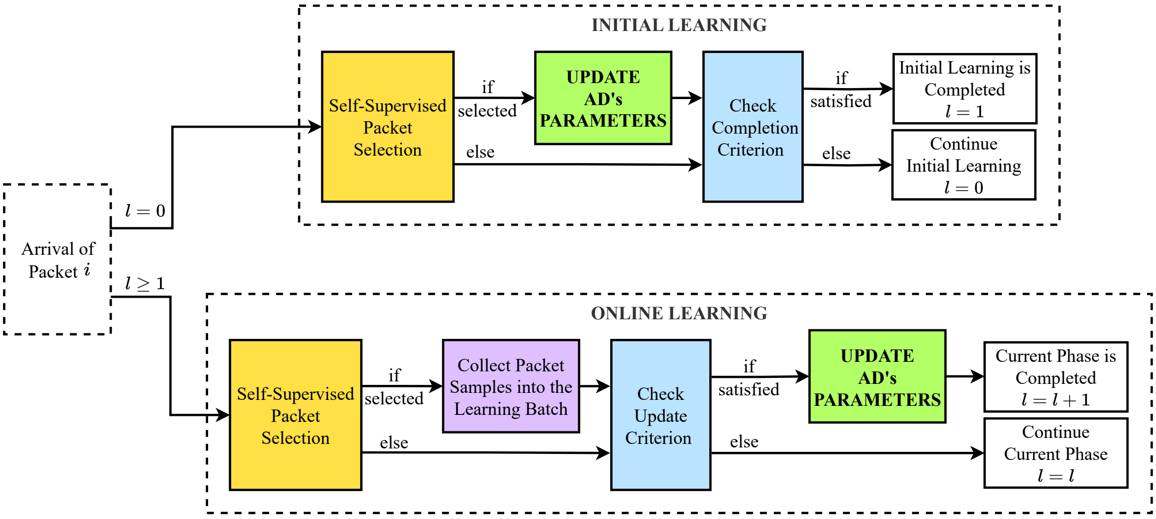
INITIAL LEARNING
Self-Supervised Packet Selection
if selected
UPDATE AD's PARAMETERS
Check Completion Criterion
if satisfied
Initial Learning is Completed $l = 1$
else
else
Continue Initial Learning $l = 0$
Arrival of Packet i
$l = 0$
$l \geq 1$
ONLINE LEARNING
Self-Supervised Packet Selection
if selected
Collect Packet Samples into the Learning Batch
Check Update Criterion
if satisfied
UPDATE AD's PARAMETERS
Current Phase is Completed $l = l + 1$
else
else
Continue Current Phase $l = l$

Figure 7: Block diagram of the learning process in SSID framework for the online self-supervised learning of the parameters of IDS

We then evaluate the performance of the SSID framework for two tasks, malicious traffic detection and compromised device identification, aiming to enhance the security of an IoT network:

For malicious traffic detection, two different ML models, DRNN and MLP, is deployed with the SSID framework and tested on the Kitsune [9] dataset. The results in Figure 8 reveal that the ML models trained under the SSID framework requiring no offline dataset achieve considerably high performance compared to the same models with offline learning.

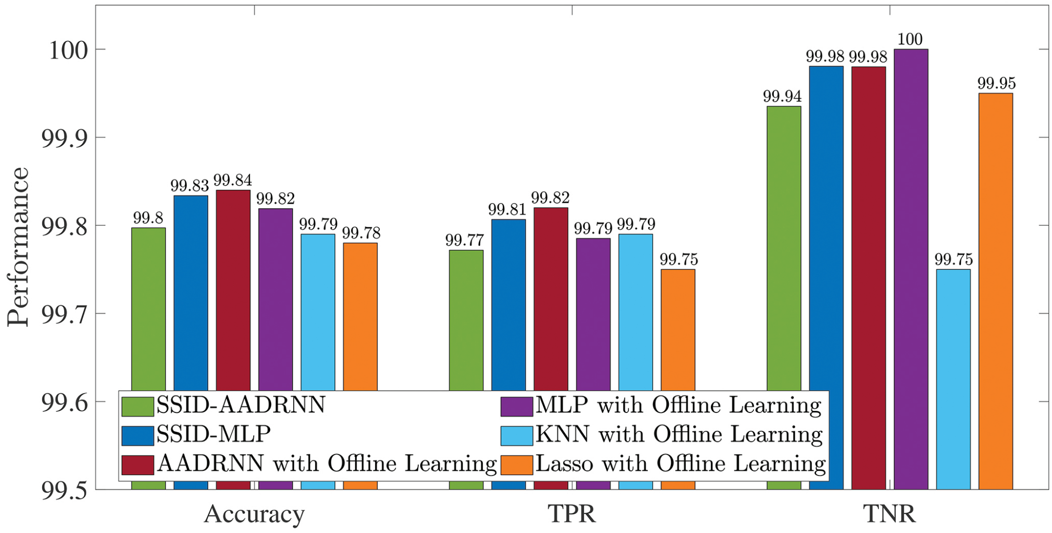

Figure 8: Performance comparison between the ML models under the SSID framework and those with offline learning

For compromised device identification, the performance of the CDIS is tested under sequential learning and the SSID framework on the data of 6 different cyberattacks provided by two public datasets Kitsune [9] and Bot-IoT [11]. The results in Figure 9 showed that the use of SSID significantly improves the performance of CDIS for the majority of cases.

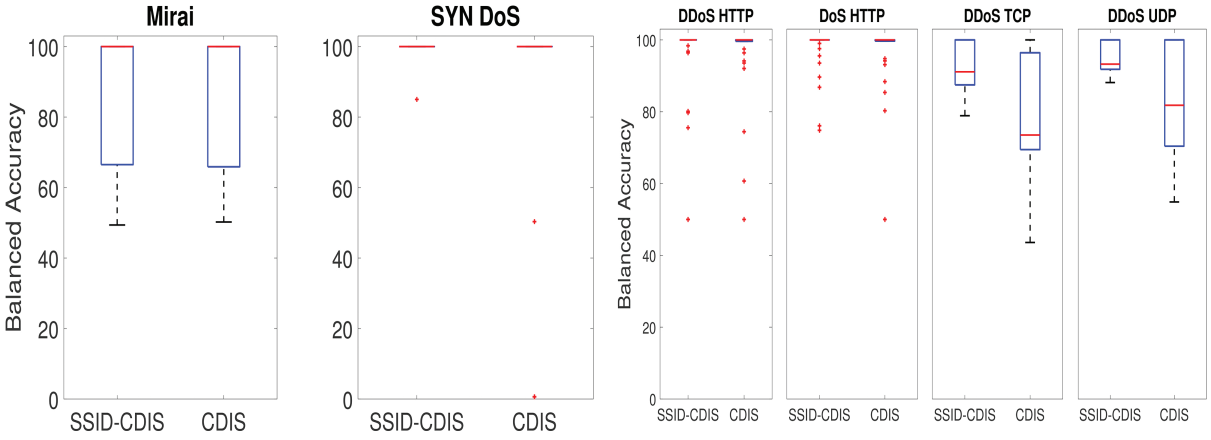

Figure 9: Performance comparison of the CDIS trained under the SSID framework with that under sequential quasi-online learning on (left) Kitsune and (right) Bot-IoT datasets

In summary, the proposed SSID framework eliminates the need for offline data collection, prevents human errors in data labeling, and avoids labor costs for model training and data collection through experiments. In addition, the SSID framework enables the IDS to easily adapt to time-varying characteristics of network traffic, significantly improving its intrusion detection performance.

Bibliography

[1] "Hp study reveals 70 percent of Internet of Things devices vulnerable to attack," 2014, accessed: 2023-03-22. [Online]. Available: https://www.hp.com/us-en/hp-news/press-release.html?id=1744676

[2] Intersog, "IoT Security Statistics: 6 Facts [Updated]," Dec 2021, accessed: 2023-03-03. [Online]. Available: https://intersog.com/blog/iot-security-statistics/

[3] Cisco, *Cisco Annual Internet Report (2018–2023)*, Mar. 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[4] G. Matta, S. Chlup, A. M. Shaaban, C. Schmittner, A. Pinzenöhler, E. Szalai, and M. Tauber, "Risk management and standard compliance for cyber-physical systems of systems," *Infocommunications Journal*, vol. 13, no. 2, pp. 32–39, June 2021.

[5] S. Maksuti, M. Zsilak, M. Tauber, and J. Delsing, "Security and autonomic management in system of systems," *Infocommunications Journal*, vol. 13, no. 3, pp. 66–75, September 2021.

[6] E. Gelenbe and Y. Yin, "Deep learning with dense random neural networks," in *International Conference on Man–Machine Interactions*. Springer, 2017, pp. 3–18.

[7] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[8] N.-y. Liang, G.-b. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1411–1423, 2006.

[9] "Kitsune Network Attack Dataset," August 2020. [Online]. Available: https://www.kaggle.com/ymirsky/network-attack-dataset-kitsune

[10] "KDD Cup 1999 Data." [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[11] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[12] H. M. Song and H. K. Kim, "Self-supervised anomaly detection for in-vehicle network using noised pseudo normal data," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1098–1108, 2021.

[13] Z. Wang, Z. Li, J. Wang, and D. Li, “Network intrusion detection model based on improved byol self-supervised learning,” *Security and Communication Networks*, vol. 2021, pp. 1–23, 2021.

[14] X. Zhang, J. Mu, X. Zhang, H. Liu, L. Zong, and Y. Li, “Deep anomaly detection with self-supervised learning and adversarial training,” *Pattern Recognition*, vol. 121, p. 108234, 2022.

[15] H. Kye, M. Kim, and M. Kwon, “Hierarchical detection of network anomalies: A self-supervised learning approach,” *IEEE Signal Processing Letters*, vol. 29, pp. 1908–1912, 2022.

[16] E. Caville, W. W. Lo, S. Layeghy, and M. Portmann, “Anomal-e: A self-supervised network intrusion detection system based on graph neural networks,” *Knowledge-Based Systems*, vol. 258, p. 110030, 2022.

[17] M. Abououf, R. Mizouni, S. Singh, H. Otrok, and E. Damiani, “Self-supervised online and lightweight anomaly and event detection for IoT devices,” *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25 285–25 299, 2022.

[18] W. Wang, S. Jian, Y. Tan, Q. Wu, and C. Huang, “Robust unsupervised network intrusion detection with self-supervised masked context reconstruction,” *Computers & Security*, vol. 128, p. 103131, 2023.