

The development of a new system for generating training data of AI-based anomaly detection

Thi My Truong *, Won Seok Choi *, Jeong Jang Hyeon **, Seong Gon Choi *

* College of Information and Communication Engineering, Chungbuk National University, Chungdae-ro 1, Seowon-gu

** JJ SOLUTION INC, Chungbuk National University, Cheongju-si, Chungcheongbuk-do, South Korea

mytruong@cbnu.ac.kr, wschoi@cbnu.ac.kr, jjsol210120@gmail.com, sgchoi@cbnu.ac.kr

Abstract— This paper proposes a method and system for generating training data to support AI based anomaly detection. The use of AI in abnormal behavior detection systems is becoming increasingly popular, with active research on AI-based anomaly detection methods using machine learning. In general, existing research relies on open datasets provided by various laboratories like Swat, WaDI, SMAP and MSL for testing and validation purposes. Since the types of normal and malicious packets depend on the specific network to which they are applied, verifying AI-based anomaly detection methods using an open dataset may yield different results than when applied in real-world scenarios. In other words, open datasets captured from specific networks may not be suitable for applying AI-based abnormal detection methods to other networks. In addition, AI-based datasets may be insufficient for learning, leading to the use of simulated attacks. Open datasets are difficult to provide sufficient data for training and often contain malicious packets using simulated attack packets. Since malicious attacks are always transformed into new forms and developed in types, it is necessary to prepare a database for new malicious attacks and to learn about them. Therefore, one of the major challenges in developing effective anomaly detection systems is acquiring an appropriate dataset. To address this issue, we propose a system for extracting training data by collecting packets from the actual network to apply AI-based abnormal detection. Our proposed system offers the advantage of accurately reflecting the network's packet characteristics by gathering data from live networks for AI-based abnormality detection and dataset creation. Furthermore, as it incorporates a dataset for the latest malicious attacks within the network, it enables more practical anomaly detection compared to the use of existing datasets. We simulated and tested the proposed system at the laboratory level to confirm its behavior.

Keywords— anomaly detection, artificial intelligence (AI), dataset, training data, cybersecurity

I. INTRODUCTION

Anomaly detection, a fundamental component of network security, plays a pivotal role in identifying abnormal patterns and potential threats within a given environment. In this context, anomaly detection refers to the process of discerning deviations from expected or typical behavior in a system or dataset [1]. These deviations, often referred to as anomalies, can indicate a wide range of irregularities, including network intrusions, cybersecurity threats, equipment malfunctions, or fraudulent activities.

The integration of artificial intelligence (AI) and machine learning techniques has revolutionized the field of anomaly detection, allowing for more highly advanced and adaptable solutions. AI-based anomaly detection methods leverage the power of algorithms and statistical models to learn and recognize patterns in data, enabling them to automatically detect anomalies that might be difficult to be identified through traditional approaches.

While AI-based anomaly detection methods have made significant progress, they are only as effective as the data they are trained on. Existing research in this domain has often relied on open datasets offered by various laboratories and research institutions, such as Swat, WaDI, SMAP and MSL. These datasets have been used for testing and validating anomaly detection algorithms. However, a critical limitation arises when applying AI-based anomaly detection methods in the real-world network environments. The challenge lies in the diversity among networks, each with its distinct attributes and traffic patterns. Therefore, a method that performs well on one network may yield different results when applied to another, making the transition from research to practical implementation complex.

Moreover, to train AI models efficiently, the datasets must be diverse, comprising of both normal and malicious packets. However, the aforementioned open datasets do not produce a sufficiently diverse and comprehensive set of data for training. The landscape of malicious attacks is constantly evolving, giving rise to new forms and types of threats. To build effective anomaly detection systems, it is essential to have access to data that accurately reflects the latest network conditions and includes the most recent malicious attacks.

This paper addresses these challenges and proposes a novel method and system for generating training data specifically to support AI-based anomaly detection. Our approach is grounded in the collection of real-world network traffic data, offering a distinct advantage in accurately reflecting the unique characteristics of the network under consideration. Furthermore, our system is designed to incorporate data related to the latest malicious attacks within the network, ensuring that AI-based anomaly detection methods are well-equipped to handle the dynamic nature of cybersecurity threats.

In the subsequent sections of this paper, we will present in detail our method for extracting training data from live

networks, the mechanisms for dataset creation, and the results of simulating and testing our proposed system at the laboratory level.

II. RELATED WORK

A. Existing Anomaly detection research

Machine learning based approach in anomaly detection is favored by researchers for flexibility and applicability to any network structure in comparison to the traditional detection [2]. Machine learning has been applied to detect abnormal behavior in various types of networks, using a variety of model types, including supervised learning, unsupervised learning, reinforcement learning, and deep learning models.

Chen et al. [3] analyzed network traffic and employed machine learning methods to identify abnormal behavior and detect malicious apps. They used imbalanced classification methods, including the Synthetic Minority Oversampling Technique (SMOTE) combined with Support Vector Machine (SVM), SVM Cost-Sensitive (SVMCS), and C4.5 Cost-Sensitive (C4.5CS) methods. While this approach performed well with highly imbalanced training data, its performance became unstable when the dataset's imbalance ratio was under 1000.

Hamamoto et al. [4] utilized Genetic Algorithms to analyze the network and subsequently employed a Fuzzy Logic scheme to determine whether an instance represents an anomaly. This method exhibits high performance in Denial of Service (DoS) and Distributed Denial of Service (DDoS) attack detection, but it is associated with a high false-negative rate.

Alauthman et al. [5] employ the output of a supervised learning model as the state in the reinforcement learning model, both the SL and RL model are improving through this interaction. This method has a good accuracy rate when the input data is reduced in the model, leading to reduced training time. However, it has been validated in MATLAB using three datasets and has not been implemented in a real network.

Wei et al. [6] employ Convolutional Neural Networks (CNN) to learn spatial features in the data and use Recurrent Neural Networks (RNN) with long-short term memory to learn temporal features. Subsequently, the original datasets DARPA1998 and ISCX2012 undergo preprocessing. The advantage is the improved performance achieved, this approach has only been validated using a fixed dataset, which can be considered a limitation.

All these studies have encountered data limitations, such as unbalanced data, insufficient data, and a lack of validation in real network environments.

B. Available open datasets

SWAT (Secure Water Treatment) [7] simulates the operations of a real-world industrial water treatment plant. SWaT was run and data were collected over an 11-day period. The first 7 days were dedicated to normal data collection without any attacks or errors, while the remaining 4 days involved 36 attacks created by the research team. This dataset includes physical properties relevant to the plant and the water

treatment process, as well as network traffic within the testbed. It comprises 51 channels, encompassing sensors such as flow meters, level transmitters, conductivity analyzers, and actuators like motorized valves and pumps.

The WADI dataset [8] was collected from the WADI testbed, which is an extension of the SWAT testbed. This dataset comprises data from 1233 sensors and actuators and was collected over a 16-day period. Of these 16 days, 14 days were dedicated to normal data collection, while the remaining 2 days involved 15 attacks.

The SMAP and MSL datasets [9] consist of expert-labeled telemetry anomaly data from NASA's Soil Moisture Active Passive (SMAP) satellite and Mars Science Laboratory (MSL) rover. The number of SMAP variables is 1375, while the that of MSL features is 1485, making them significantly more extensive compared to single-entity datasets. To effectively handle this increased complexity, the data has been systematically organized into 55 distinct entities for SMAP and 27 entities for MSL. This categorization facilitates a structured methodology for the analysis and detection of anomalies.

These open datasets only reflect the characteristics of the specific networks to which they are applied. While SWAT and WADI datasets pertain to networks in a simulated water plant, SMAP and MSL datasets deal with telemetry data.

III. PROPOSED METHOD

The core of our system revolves around a three-step process described in Figure 1. First, we collected log information from security devices in public institutions. Then, we analyzed the collected IP addresses, comparing them to Threat Intelligence (TI) information to check for potential threats. When a threat is detected, the system is automatically applied to the relevant policy to agencies with API connection, while the rest without API connection receive email notifications about the threat.

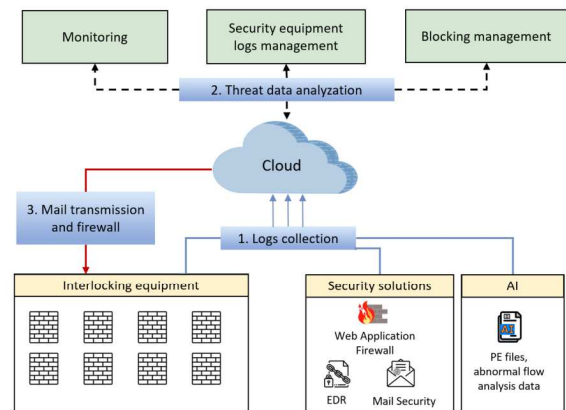


Figure 1. Overall System Architecture

The data collecting process is described in Figure 2. In each target network, packets are transmitted from the switch to the data collecting servers using packet control techniques such as mirroring or inline. The output of the data collecting servers consists of log files or extracted files with unique network

characteristics. These files are subsequently transmitted to a virtual machine via the Internet. Before being sent to the virtual machine, a firewall is configured to receive only the traffic directed to a specific port number and the IP address of the collection sensor associated with the relevant agency. Finally, the extracted files have been successfully transmitted by sending an acknowledgment message.

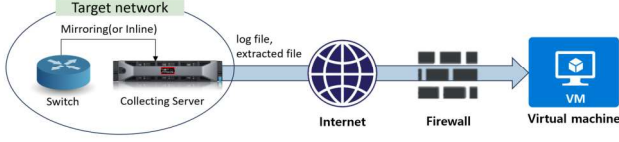


Figure 2. Data collecting process

The core of our system is the data collecting process that takes place at the collecting server as Figure 3. Packets are transmitted from the Packet Collecting module to the Metadata extracting module, where the network traffic is analyzed. This process extracts application protocols and information of flow, including IP addresses, port numbers, packet counters, and additional details corresponding to the protocol type. The output of Metadata extracting module, in the form of a string (raw data), is transmitted through a pipeline and stored in a temporary file in minutes. During the data processing, this temporary file is read, and the timestamp is converted to UNIX time format. After removing any abnormal lines, such as those lacking proper newline delimiters, the information is saved to a final file, and the temporary file is deleted. Statistical information regarding the extracted protocols is updated on a minute-by-minute basis.

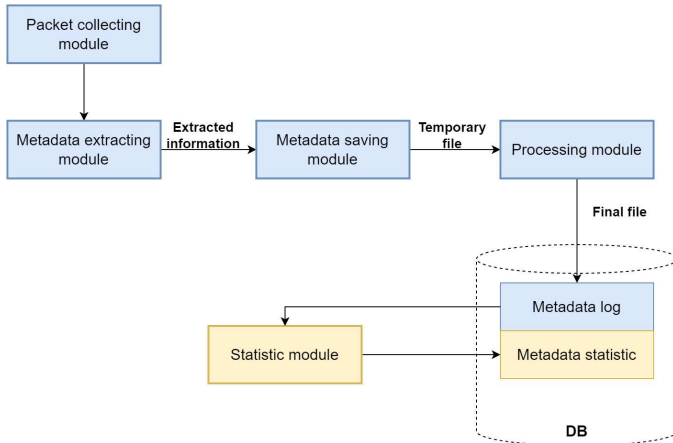


Figure 3. Collecting server

Each metadata record is comprised of 44 features, described in Table 1. Features 1-3 explain the primary protocol, sub-protocol, and Layer 4 protocol. Features 4-7 pertain to flow start and end times in seconds and sub-seconds. Information about the IP addresses and port numbers of both the source and destination is provided in features 8-11.

Features 12-19 provide information regarding the number of packets, packet size, and actual valid data throughput in both directions (source to destination and vice versa, destination to source). Features 20-41 represent the count of packets with specific flags, such as TCP Congestion Window Reduced, ECN-Echo, Urgent, Acknowledge, Push, Reset, Synchronized Sequence Numbers, and the Finish flag, enabled within the flow, in each direction from source to destination and from destination to source. Additional flow-related information is provided in the final feature.

TABLE 1. METADATA OF THE DATASET

Number	Field name	Number	Field name
1	proto_app	23	sum_ece_cnt
2	proto_master	24	s2d_ece_cnt
3	proto_l4	25	d2s_ece_cnt
4	epoch_t_first	26	sum_urg_cnt
5	micro_s_first	27	s2d_urg_cnt
6	epoch_t_last	28	d2s_urg_cnt
7	micro_s_last	29	sum_ack_cnt
8	src_ip	30	s2d_ack_cnt
9	dst_ip	31	d2s_ack_cnt
10	src_port	32	sum_psh_cnt
11	dst_port	33	s2d_psh_cnt
12	sum_packets	34	d2s_psh_cnt
13	sum_bytes	35	sum_rst_cnt
14	s2d_packets	36	s2d_rst_cnt
15	s2d_bytes	37	d2s_rst_cnt
16	s2d_goodput	38	sum_syn_cnt
17	d2s_packets	39	s2d_syn_cnt
18	d2s_bytes	40	d2s_syn_cnt
19	d2s_goodput	41	sum_fin_cnt
20	sum_cwr_cnt	42	s2d_fin_cnt
21	s2d_cwr_cnt	43	d2s_fin_cnt
22	d2s_cwr_cnt	44	info

IV. EXPERIMENT RESULT

Real data collection is illustrated in Figure 4. The system detects a variety of application protocols, including TLS, OpenDNS, Radius, Azure, Microsoft 365, BitTorrent, HTTP, Google, DNS, etc. IP address and port number information has been masked for security purposes. Actual traffic information is retained for the training of AI-based anomaly detection models.

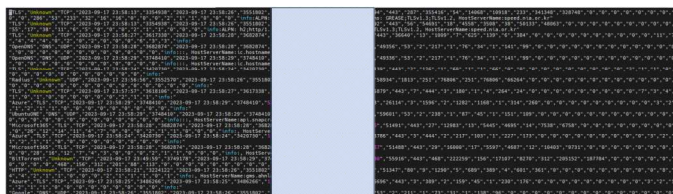


Figure 4. Real Data Collection (Private IP Addresses)

V. CONCLUSIONS

While AI-based anomaly detection is gaining widespread attention and application, the reliance on open datasets for research and testing has certain limitations. Existing open datasets, collected from specific networks, may not be directly applicable to other network environments due to variations in normal and malicious packet behaviors. As a solution, we have proposed a system that collects packets directly from live networks, produced a more accurate representation of the network's unique characteristics. This approach to data collection not only enhances the performance of AI-based anomaly detection but also contributes to the ongoing development of more adaptable systems.

ACKNOWLEDGMENT

This work was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462, 33%) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945, 34%). This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1006167, 33%).

*corresponding author is S.G. Choi(choisg@cbnu.ac.kr).

REFERENCES

- [1] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 1-58.
- [2] Wang, S., Balarezo, J. F., Kandeepan, S., Al-Hourani, A., Chavez, K. G., & Rubinstein, B. (2021). Machine learning in network anomaly detection: A survey. *IEEE Access*, 9, 152379-152396.
- [3] Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., & Yang, B. (2018). Machine learning based mobile malware detection using highly imbalanced network traffic. *Information Sciences*, 433, 346-364.
- [4] Hamamoto, A. H., Carvalho, L. F., Sampaio, L. D. H., Abrão, T., & Proença Jr, M. L. (2018). Network anomaly detection system using genetic algorithm and fuzzy logic. *Expert Systems with Applications*, 92, 390-402.
- [5] Alauthman, M., Aslam, N., Al-Kasassbeh, M., Khan, S., Al-Qerem, A., & Choo, K. K. R. (2020). An efficient reinforcement learning-based Botnet detection approach. *Journal of Network and Computer Applications*, 150, 102479.
- [6] Wei, G., & Wang, Z. (2021). Adoption and realization of deep learning in network traffic anomaly detection device design. *Soft Computing*, 25(2), 1147-1158.

- [7] Mathur, A. P., & Tippenhauer, N. O. (2016, April). SWaT: A water treatment testbed for research and training on ICS security. In *2016 international workshop on cyber-physical systems for smart water networks (CySWater)* (pp. 31-36). IEEE.
- [8] Ahmed, C. M., Palleti, V. R., & Mathur, A. P. (2017, April). WADI: a water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd international workshop on cyber-physical systems for smart water networks* (pp. 25-28).
- [9] Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (2018, July). Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 387-395).



Thi My Truong received B.S. degree in College of Information & Communication Engineering from Chungbuk National University in 2023. She is currently pursuing the Master degree in Radio Communication Engineering, Chungbuk National University. Her research interests include cybersecurity, Blockchain, AI.



Won Seok Choi received B.S. and Ph.D. degree in the College of Electrical and Computer Engineering, Chungbuk National University, Korea in 2008 and 2014 respectively. He is currently researcher in Research institute of Computer and Information Communication, Chungbuk National University. His research interests include Vehicle network, Energy saving network, SDN, NFV and NGN.



Jang Hyeon Jeong received B.S. and M.S. degree in the College of Electrical & Computer Engineering, Chungbuk National University, Korea in 2019 and 2021. His research interests include Network Security, Smart Grid. He is currently researcher in Xabyss Inc and CEO in JJsolution Inc. His research interest is network security.



Seong Gon Choi received B.S. degree in Electronics Engineering from Kyungpook National University in 1990, and M.S. and Ph.D. degree from KAIST in Korea in 1999 and 2004, respectively. He is currently a professor in College of Electrical & Computer Engineering, Chungbuk National University. His research interests include V2X, AI, smart grid, IoT, mobile communication, high-speed network architecture and protocol.