2024 5th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI) | 979-8-3315-3399-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCBD-AI65562.2024.00050

A Multi-stage based Approach for Zero-day Attack Detection

1st Simeng Li
Cyberspace Security Academy
Chengdu University of Information Technology
Chengdu,China
18200518645@163.com

2nd Guogen Wan*
Cyberspace Security Academy
Chengdu University of Information Technology
Chengdu,China
wanguogen@cuit.edu.cn

***Abstract*—Intrusion detection systems (IDS) have traditionally been effective for security monitoring. However, with the continuous advancement of digitalization, the number and variety of connected devices are increasing, leading to the emergence of new and unknown threats. Current IDS struggle to effectively detect zero-day attacks, allowing these threats to bypass detection and execute their malicious tasks.To address this issue, this paper proposes a new multi-stage intrusion detection method. In the first stage, residual networks are employed to distinguish between benign and abnormal samples. The second stage uses a random forest algorithm to identify known attack types, while the third stage differentiates zero-day attacks from other threats.Performance was evaluated using the publicly available benchmark datasets CIC-IDS-2017 and CIC-IDS-2018.The findings demonstrate that our proposed approach is not only capable of effectively identifying zero-day attacks but also achieves higher classification performance compared to existing methods. Additionally, the multi-stage approach reduces the consumption of computing resources and bandwidth. The optimal performance model, balancing the threshold set, correctly classified 92.7% of zero-day attacks (38 out of 41), while reducing bandwidth requirements by 71%.**

Keywords-residual network; random forest; anomaly detection; multi-stage detection; multi-classification;

I. INTRODUCTION

Increased cybersecurity risks have made it difficult for existing security monitoring systems to detect threats in real time, and to mitigate these risks, the research community needs improved defenses[1].Traditional intrusion detection systems (IDS) rely heavily on single machine learning models,which have limitations in identifying unknown or zero-day attacks. Current methods, methods like signature-based detection work well for known threats but struggle with new, novel attacks. Additionally, deploying a single model increases computational costs and latency, especially in distributed networks.To address these shortcomings, new approaches like ensemble learning and distributed IDS are being explored, but they still struggle with the detection of unknown threats. This paper proposes a multi-stage hierarchical intrusion detection method that effectively identifies zero-day attacks and enhances classification capabilities, aiming to overcome the limitations of existing methods.

II. RELATED WORK

With the increase in network security threats, detecting malicious traffic is particularly important in time-sensitive environments. To ensure timely detection and response to malicious traffic, it is critical to develop efficient and accurate detection methods.

A. Single-stage multi-level malicious traffic detection methods

Networks are deployed in environments that can use unique methods for malicious traffic detection based on different layers of devices, synthesizing the output from each layer to obtain a final prediction. For example, [3] employ a multi-layered approach to detecting cyber-attacks by monitoring different features and using various authentication and detection techniques to improve detection accuracy. However, it is important to note that in time-sensitive environments, the efficiency of traffic detection must also be considered to ensure timely detection and response to malicious traffic.

B. Multi-stage malicious traffic detection methods

The studies [4, 5] employ a two-stage classification method that begins with feature selection., followed by classification using the plain Bayesian algorithm and CF-KNN. In contrast, Al-Yaseen et al. [6] connect multiple stages together to form a waterfall pattern, with each stage having a classifier for detecting specific attack types.

In addition to using multiple detection methods to improve classification performance, some research depends on a blend of anomaly detection, typically utilizing unsupervised machine learning methods, along with multi-class classifiers. This approach filters out suspicious samples in a lightweight manner by combining anomaly detection with classification and sends them to a more complex stage for classifying the type of attack [9]. The study[10] uses a multi-tiered model where OC-SVM and self-organizing mapping are employed for classification. Bovenzi et al. [11] proposed a two-stage hierarchical approach that uses multimodal DAEs and soft-output classifiers for classification, training multi-class classifiers to detect unknown attacks through an open-set approach. Verkerken et al. [8]'s three-stage hierarchical approach employs OC-SVM and RF classifiers, using anomaly scores to distinguish between normal samples and zero-day attacks. These methods not only improve classification accuracy and efficiency but also detect unknown and zero-day attacks.

C. Comparison between the proposed method and other methods

The method proposed by Zhang et al. [3] detects both malicious and benign traffic without pre-processing, which increases the bandwidth and computing requirements of the system while reducing the efficiency of traffic detection. This approach is not suitable for time-sensitive scenarios. In the method proposed in this paper, malicious traffic is distinguished from benign traffic in the first stage

979-8-3315-3399-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICCBD-AI65562.2024.00050

of detection, and only the malicious traffic data is sent to the following stage for further detection.This not only enhances the system's detection efficiency but also decreases bandwidth consumption and computational requirements.

The two-stage malicious traffic detection method proposed in [4, 5, 6] has greatly improved efficiency compared to single-stage methods. However, a significant issue is that normal traffic entering the second stage for multi-classification is often misclassified as unknown attacks, despite originally being normal traffic. This misclassification leads to a high false positive rate in the system. In contrast, this paper adds an extended phase beyond the two existing phases to better differentiate between normal traffic and unknown attacks, thereby reducing the false positive rate.When the multi-stage detection method proposed in [8, 11] selects the dividing line between benign and malicious traffic, it is often challenging to find a suitable threshold for differentiation, resulting in a high false positive rate at the end of the process. In this paper, we use a residual network to obtain anomaly scores. The key difference is that the residual network enables end-to-end learning of these scores.It directly learns the anomaly score using a limited number of labeled anomalies, instead of following the traditional two-step method that involves first learning a new representation and then using an anomaly metric to determine the score.This method can effectively separate benign samples from abnormal samples.

III. MULTI-STAGE ANOMALY DETECTION MODEL

This section presents a new multi-stage layered intrusion detection approach. First, the design decisions made during the development process are thoroughly outlined, and the overall architecture is presented. Next, the separate stages that make up the overall architecture are examined individually. Lastly, the benefits of a layered deployment are emphasized..

A. *Overall structure*

The multi-stage anomaly detection model proposed in this paper consists of three stages: the first stage involves data preprocessing and traffic pre-classification; the second stage focuses on classifying malicious traffic and screening out unknown traffic attacks to send to the expansion stage; the third stage entails re-verifying unknown attacks and ultimately categorizing them into zero-day attacks or benign traffic. What follows is a detailed description of the process at each stage and the algorithms used.

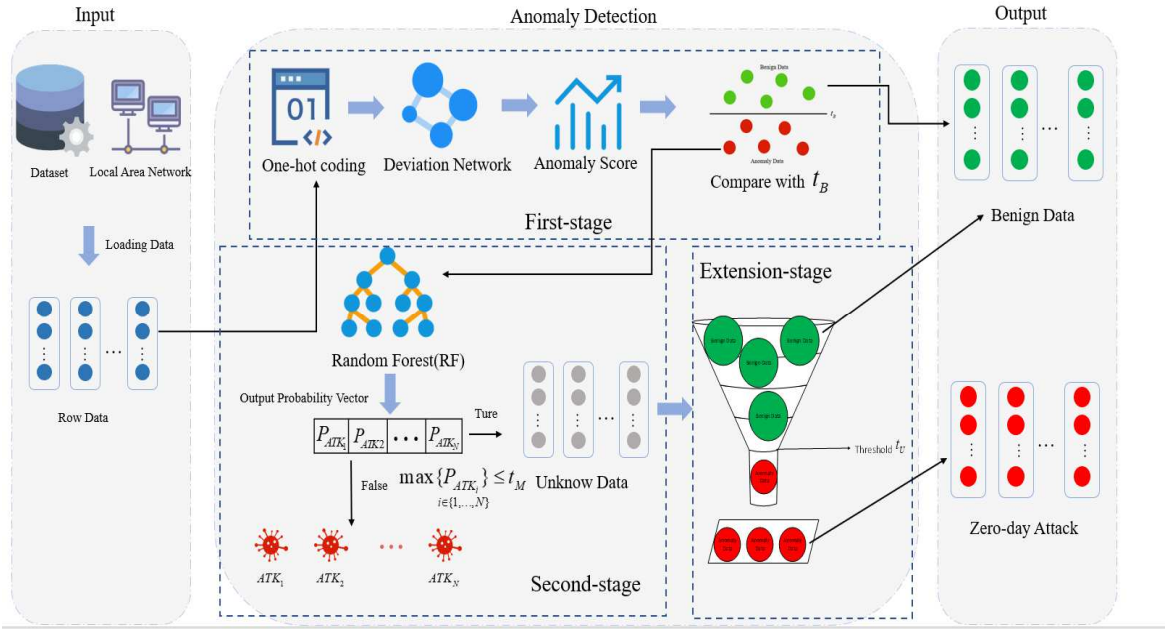


Figure 1. Overall frame diagram of the system

B. *Phase I: anomaly detection*

The main objective of the first stage is to classify the initial set of samples into benign and abnormal categories, and then pass the abnormal samples to the second stage for further detection. To achieve this goal, a residual network is employed to generate an anomaly score for the samples, which serves as the basis for determining whether they are benign or anomalous. A detailed description of this stage of anomaly detection follows.

1) *Residual networks*

The residual network (DevNet) introduced here for anomaly score generation utilizes a Gaussian prior along with a Z-Score-based residual loss. This approach allows for the direct optimization of anomaly scores via an end-to-end neural network [7]. After optimization, the initial stage can distinguish between benign and anomalous samples based on their anomaly scores, facilitating initial filtering.

2) *Anomaly detectors*

Here, we define a threshold value t_B used to distinguish abnormal samples from normal samples. When the anomaly score of a sample X is greater than the threshold t_B, it is classified as abnormal; otherwise, it is classified as normal. The choice of the threshold value determines the number of samples passed into the second stage, as well as the detection accuracy of the system. Therefore, adjusting the threshold can greatly impact the overall performance and accuracy of the entire system.

C. Phase II: Multi-classification

1) Random Forest

A random forest (RF) is composed of several decision trees. [14], each built using either the entire dataset or a subset of it. The final prediction is determined by averaging the outputs of all the trees or, in classification tasks, by selecting the class with the most votes. The optimized hyperparameters include the number of trees, the proportion of subsamples, and the number of features taken into account for each split in a tree. This paper employs the scikit-learn implementation for these processes.

2) Multiple classifiers

The main objective of the second phase is to classify the anomalous samples incoming from the first phase as known attack types or as unknown attack types that proceed to the third phase. The classifier in this phase uses the Random Forest (RF) algorithm. Its training set consists solely of individual attack type samples, which are trained to learn the features of each attack type. After training, when an anomaly sample X is input, the classifier outputs a vector $[P_{ATK_1}, P_{ATK_2}, \cdots, P_{ATK_N}]$ indicating the probability of the anomaly sample X belongs to each attack type. Here a threshold t_M is introduced; if all values of the anomaly sample X with the highest probabilities are below the threshold t_M, it is classified as an unknown attack sample and forwarded to the third stage. Otherwise, the sample is assigned to the attack type with the highest probability.

D. Phase III: Expansion

The third phase focuses on reclassifying the unknown attack samples from the second phase, without using additional classifiers, and reusing the anomaly scores obtained in the first phase. The primary goal of this phase is to minimize false positives, specifically the number of normal samples misclassified as anomalous, while also enabling the detection of zero-day attacks.

E. Phase IV: Detect zero-day attacks

The multi-stage anomaly detection model presented in this paper is structured into three distinct stages, each employing different algorithms to achieve specific objectives. The overall architecture is illustrated in Figure 1. Initially, the system receives an eigenvector X, which is processed in the first stage. This stage features an anomaly detector AS_B that takes the input eigenvector X and produces an anomaly score as the detection outcome. If the anomaly score is below the threshold t_B, it signifies that X is a benign sample and does not require further analysis in the second stage. Conversely, if the anomaly score exceeds the threshold t_B, vector X proceeds to the second stage for additional evaluation.The second stage utilizes a multi-class classifier to compare the incoming sample X from the first stage against known attack types (ATK_i) to determine if it corresponds to any recognized threats. If it does not match any known attack type, the sample is forwarded to the third stage.The third stage, referred to as the expansion stage, does not introduce a new classifier; instead, it utilizes the anomaly score AS_B output from the first stage. If this score is below the threshold, it is classified as benign to correct any errors made in the first stage. However, if the score is above the threshold t_U, the sample is classified as either an unknown attack sample or a zero-day attack.

IV. Experimental

In this section, the datasets used, the methods of data preprocessing, and the evaluation metrics used to assess the models' performance are outlined.

A. Data sets

The CIC-IDS-2017[2] dataset is a contemporary network intrusion detection dataset based on flow data. It includes both legitimate and malicious network flows. The dataset employs statistical methods and machine learning to simulate benign traffic and generate malicious flows using existing attack tools. The resulting data is provided in both CSV and PCAP formats, making it suitable for machine learning applications.

After data cleansing, the original attack categories were merged into six higher-level categories. For example, the original SSH and FTP cracking patterns were unified into the brute force cracking attack category. Since the data samples are highly unbalanced, sampling techniques were employed to mitigate the impact of data imbalance on the experimental results. Stratified sampling was used to ensure that the number of samples for each attack type was 1,948, allowing for uniform representation of each attack sample in the final dataset. The total number of samples in the dataset for the attack infiltration and cardiac hemorrhage categories are 11 and 36, respectively, making them particularly suitable as unknown or zero-day attacks. So, these two categories are combined and collectively referred to as the unknown attack category, which is used to assess the model's ability to detect zero-day attacks.

The dataset acquired through stratified sampling is split into three segments: training, validation, and testing.The malicious attack sample dataset is split into a 70% training set and a 30% test set, while the data from unknown attack categories are directly included in the test set. The first phase uses benign samples to train the anomaly detector, and the validation phase contains 5% malicious traffic. The second stage uses a dataset consisting solely of malicious traffic to train a multi-classifier, with the validation stage comprising 50% benign traffic and 50% malicious traffic. The final test set comprises 95% benign traffic and 5% malicious traffic, simulating a realistic scenario with predominantly benign traffic. All datasets are derived from the balanced processed dataset through stratified sampling.

B. Experimental environment and evaluation indicators

The following hardware and software platforms were used for the experiments: Intel(R) Core(TM) i7-10700 CPU, Windows 11 Professional (64-bit), NVIDIA GeForce RTX 2080 Ti, NVIDIA CUDA 11.1, Python 3.8.18, and the Scikit-learn library (version 1.1.1).

Recall, precision, and f1-score are used to evaluate the model.

C. Experimental results

1) Performance of the multi-stage malicious traffic detection model

The effectiveness of the multi-stage malicious traffic detection model can be assessed through three methods. The first aspect is the overall performance of the three classification stages; the second aspect is the reduction of the computational and bandwidth requirements of the system in the case of layered deployment; and the third aspect is the capability to identify zero-day attacks. However, the experimental results indicate that there is no single threshold arrangement that can achieve good results in all three aspects simultaneously, making it necessary to choose reasonable thresholds based on the actual environment of the model deployment or the desired goals.

The multi-stage malicious traffic detection model must not only separate malicious traffic from data streams that are predominantly benign but also match the separated malicious traffic with known attack types and ultimately detect zero-day attacks. Therefore, it is important to consider the model's overall performance comprehensively. Average accuracy, average recall, and average F1-score are used as reference metrics to evaluate the system's performance.

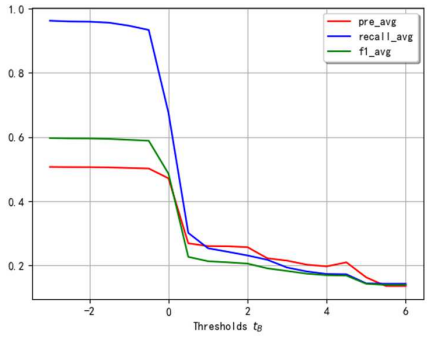
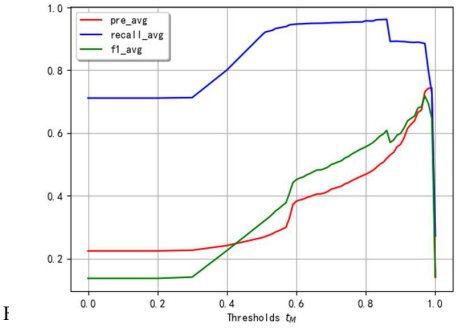


Figure 2. Mean accuracy, mean recall and mean f1-score with threshold t_B

From Fig. 2, we can see that pre_avg, recall_avg, and f1_avg decrease as the threshold t_B increases. Notably, in the small interval where the threshold t_B shifts from negative to positive, there is a steep decrease in the values. The three metrics reach their optimal values when the threshold t_B is around -2.



threshold t_M

From Fig. 3, we can see that pre_avg and f1_avg gradually increase with the increase of the threshold t_M. However, when the threshold t_M is very close to 1, both indicators experience a steep drop. For the indicator recall_avg, it gradually increases with the threshold value t_M in the range of [0, 0.85], while in the range of [0.85, 1.0], it gradually decreases as the threshold value increases, with a steep drop occurring as it approaches 1. It can be concluded that pre_avg and f1_avg reach their maximum values when the threshold is close to 1, while recall_avg reaches its maximum value when the threshold is close to 0.85.

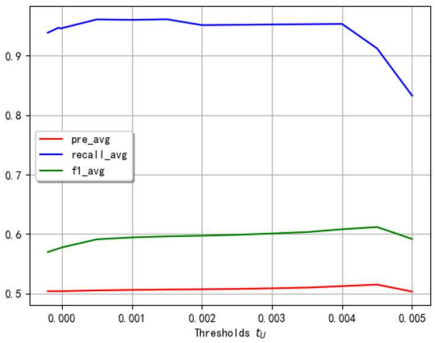


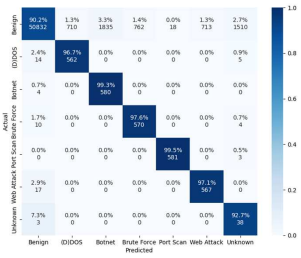
Figure 4. Average accuracy, average recall and average f1-score with threshold t_U

From Fig. 4, we can see that pre_avg, f1_avg, and recall_avg do not change significantly overall with the increase of the threshold t_U, and only recall_avg increases in magnitude when it is in the range [0.004, 0.005]. This is because, as noted earlier, the threshold t_U is used in the expansion phase to distinguish between benign traffic and unknown attack traffic, primarily to reduce false positives. Therefore, its value will mainly impact the recall rate, resulting in the observations shown in Fig. 4.

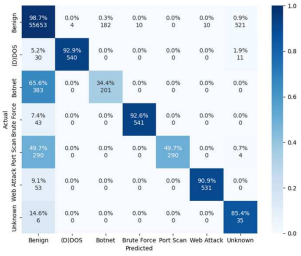
2) Comparison with other methods

The confusion matrix is employed to assess the effectiveness of the multi-stage malicious traffic detection

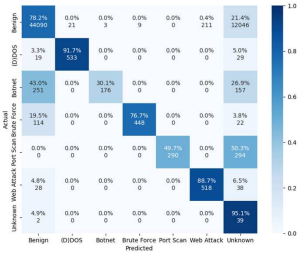
model. In this matrix, each column represents a predicted category, with the total number of columns indicating the total predictions for that category. Each row corresponds to the actual category of the data, with the total number of instances in each row reflecting the true count for that category.



(a) Multi-stage anomaly detection approach



(b) Verkerken et al.[8]



(c) Bovenzi et al . [11]

Figure 5. Confusion matrix

As shown in Fig. 5, the three confusion matrices represent the final results of the multi-stage anomaly detection method, the method proposed in paper[8], and the method proposed in paper[11]. on the CIC-IDS-2017 dataset, respectively. From this, we can conclude that our proposed method outperforms the other two methods in recognizing known attack types and also demonstrates the capability to identify zero-day attacks.

3) Robustness of the model

The model's capability to identify zero-day attacks is assessed using the CIC-IDS-2018 dataset. The method accurately classified 110,872 samples as unknown attack types, resulting in a recall rate of 86.73%. The method proposed in paper[8] correctly classified 100,199 samples, with a recall of 78.38%. The method proposed in paper[11] correctly classified 111,125 samples, achieving a recall of 86.99%.

V. DISCUSSION

A. *Advantages of the expansion phase*

The advantages of the expansion phase are divided into two main aspects. The first aspect is that it allows the first phase to make mistakes; that is, benign samples marked as anomalous by the first phase can be corrected in the third phase, enabling more samples to be identified as anomalous in the first phase. The second aspect is that it improves the performance of the system by allowing anomalous samples that are not detected in the second stage to be labeled and classified as unknown attack types or benign samples, thus enhancing the detection performance of the system and reducing the number of false alarms.

B. *Thresholds* t_B , t_M , t_U

Figure 6 depicts the distribution of benign and abnormal samples based on anomaly scores when the residual network is utilized as a binary classifier. The thresholds t_B and t_M are indicated by vertical dashed lines. The left dotted line represents the threshold t_B ;. samples with anomaly scores below the threshold t_B are categorized as benign, while those on the right are marked as suspicious samples that need additional testing. The selection of the threshold t_B is very important for the detection results. Setting it too low consumes resources and bandwidth and may misclassify benign samples as malicious ones. Conversely, setting it too high will reduce detection accuracy. The second dotted line represents the threshold t_U which defines the cutoff value for zero-day attacks. Samples above t_U are classified as unknown attacks, and samples between t_B and t_U are categorized as benign samples. during the expansion phase. Setting the threshold t_U too low may result in benign samples being misclassified as zero-day attacks, whereas setting it too high could hinder the detection of zero-day attacks. Consequently, selecting the appropriate threshold is essential for effective zero-day attack detection.

The threshold t_M is the cutoff value used to control the prediction confidence of the second-stage attack classifier. When the threshold $t_M = 0$, no samples will be categorized as unknown attack types, and none will proceed to the expansion phase. Conversely, if the threshold $t_M = 1$, only samples that are absolutely determined by the attack classifier will be classified as known attack types. Therefore, if the t_M setting the threshold too low may result in benign samples being incorrectly classified as known attack types, while setting it too high will prevent the identification of known attacks.

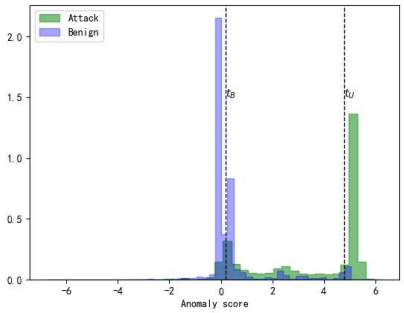


Figure 6. Histogram of anomaly scores for the output of the first stage for benign over anomalous samples and possible thresholds t_B , t_U

VI. LIMITATIONS

Our multi-stage anomaly detection approach performs well in identifying known attacks and zero-day attacks, but it suffers from two limitations. First, it requires a substantial amount of labeled samples for pre-training, which is not suitable for scenarios with fewer samples or unlabeled datasets. Second, our method does not consider the impact of adversarial samples[12,13], and deep learning algorithms are very sensitive to adversarial samples, which may lead to a degradation in classification performance. Therefore, our model is not applicable in scenarios where adversarial samples are present.

VII. FUTURE WORK

Future improvement work can focus on two aspects. Firstly, to address the multi-stage anomaly detection model's reliance on a large number of labeled samples, small-sample learning methods can be explored to learn the features of target samples using a small number of labeled or zero samples, thereby decreasing the reliance on labeled data. Secondly, to tackle the model's sensitivity to adversarial samples, pre-training can be conducted using datasets that include adversarial samples to enhance the model's robustness and improve the system's adaptability and transferability. These improvements will enhance the performance and application scope of multi-stage anomaly detection models.

VIII. CONCLUSION

This study introduces a novel multi-stage layered intrusion detection approach. The overall architecture and design choices are initially outlined to support the proposed method. Following the introduction of the new technique, it is evaluated using two contemporary network intrusion datasets: CIC-IDS-2017 and CSE-CIC-IDS-2018. The experimental findings demonstrate that our method not only improves performance in terms of classification capability but also effectively detects zero-day attacks. In particular, 38 out of 41 zero-day samples, or 92.7%, are accurately classified from the CIC-IDS-2017 dataset. Furthermore, the robustness of zero-day detection is validated by correctly identifying 110,872 zero-day samples, representing 86.73% of the 127,844 zero-day samples in the CSE-CIC-IDS-2018 dataset.

REFERENCES

[1] Pour M S, Nader C, Friday K, et al. A comprehensive survey of recent internet measurement techniques for cyber security[J]. Computers & Security, 2023, 128: 103123.

[2] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterisation," in Proc. 4th Int. Conf. Inf. Syst. Security Privacy Vol. 1 (ICISSP), 2018, pp. 108-116.

[3] A. Ali and M. M. Yousaf, "Novel three-tier intrusion detection and prevention system in software defined network," IEEE Access, vol. 8, pp. 109662-109676, 2020.

[4] H. H. Pajouh, G. Dastghaibyfard, and S. Hashemi, "Two-tier network anomaly detection model: a machine learning approach," J . Intell. Inf. Syst. vol. 48, no. 1, pp. 61-74, Feb. 2017.

[5] SH. H. Pajouh, R. Javidan, R. Khayami, A. Dehghantanha, and K.-K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks," IEEE Trans. Emerg. Topics Comput. vol. 7, no. 2,pp. 314 -323, Apr.-Jun. 2019.

[6] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Syst. Appl. vol. 67, pp. 296-303, Jan. 2017.

[7] Guansong Pang, Chunhua Shen, and Anton van den Hengel. 2019. Deep Anomaly Detection with Deviation Networks. in The 25th ACM SIGKDDConference on Knowledge Discovery and Data Mining (KDD '19), August 4-8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 10 pages.

[8] Verkerken M, D'hooge L, Sudyana D, et al. A novel multi-stage approach for hierarchical intrusion detection[J]. IEEE Transactions on Network and Service Management, 2023.

[9] L. Yang, A. Moubayed, and A. Shami, "MTH-IDS: A multitiered hybrid intrusion detection system for Internet of Vehicles," IEEE Internet Things J., vol. 9, no. 1, pp. 616-632, Jan. 2022.

[10] M. F. Umer, M. Sher, and Y. Bi, "A two-stage flow-based intrusion detection model for next-generation networks," PLoS One,. vol. 13, no. 1, Jan. 2018, Art. no. e0180945.

[11] G. Bovenzi, G. Aceto, D. Ciuonzo, V. Persico, and A. PescapØ, "A hierarchical hybrid intrusion detection approach in IoT scenarios. " in Proc. IEEE Global Commun. Conf., Dec. 2020, pp. 1-7.

[12] Sheatsley R, Papernot N, Weisman M J, et al. Adversarial examples for network intrusion detection systems[J]. Journal of Computer Security, 2022, 30(5): 727-752.

[13] Jmila H, Khedher M I. Adversarial machine learning for network intrusion detection: a comparative study[J]. Computer Networks, 2022, 214: 109073.

[14] P. A. A. Resende and A. C. Drummond, "A survey of random forest based methods for intrusion detection systems," ACM Comput. Surv. vol. 51, no. 3, pp. 1-36, May 2018.