A Two-Stage approach with CVAE and extreme value theory for an intrusion detection system

Vaka Chiru Anand[1], Mohammad Tameem Ansari[2], Srikanth Yadav.M[3]
[1, 2, 3] *Department of IT, VFSTR Deemed to be University, Guntur, A.P., India*

ABSTRACT - This research aims to provide the framework for developing an intelligent intrusion detection system capable of classifying known and unknown attacks to protect organizations and their related information systems from catastrophic loss. Specifically, we reduce the identification risk of inferring unknown attacks by first formulating the problem of fine-grained known/novel intrusion detection as a two-stage minimization problem, where the first stage seeks a score measure for minimizing the empirical risk of misclassifying the known attacks. We developed a hierarchical intrusion detection system based on class-conditioned auto-encoders due to the complex nature of the problem. In the second phase, extreme value theory describes the distribution of reconstruction mistakes to make distinguishing between unknown and known attacks easier since the former tend to have more significant reconstruction errors. We constructed a benign clustering module to study the multimodal distribution of benign traffic to reduce the number of false positives. The proposed method is evaluated using two widely used intrusion detection datasets, with positive results showing improved detection rates for previously undiscovered attacks while maintaining a low false positive rate.

Keywords: Auto-encoder, Intrusion detection, Machine learning, NSLKDD

I. INTRODUCTION

Cybercrime evolves at a rate proportional to the rate at which network infrastructure and related services grow. Internet Security Threat Report estimates that in 2018, almost 250,000,000 unique malware versions and 155 unique targeted attack groups were detected, with 23% exploiting zero-day vulnerabilities. The rise of innovative network technologies like cryptocurrency and the Internet of Things will only solidify this trend (IoT). Therefore, it is vital to establish comprehensive network security monitoring in light of the challenges of the ever-increasing volume of malware, network traffic, and malware sophistication, especially for as-yet-undisclosed novel attacks. In this study, we provide an AI-enhanced network attack detection system that can both recognize known incursions and infer previously unseen attacks.

To create effective anomaly-based detections that can infer fresh assaults, machine-learning approaches have been frequently employed in the literature. However, because it was trained with generic labels, the anomaly-based system contains a lot of spuriously positive results and insufficient diagnostic data. In contrast, a fine-grained attack categorization system can effectively train security experts to handle risks by giving appropriate diagnostic information. The use of machine learning and deep learning for precise attack detection has been extensively debated in academic circles. Most of this research, however, assumes that all assaults seen at detection time are also present in the training set and therefore assesses the detection accuracy using a closed, fixed set of attacks.

Misclassifying an unknown assault, which has a unique pattern that cannot be mistaken for a known attack, may lead to false conclusions about the nature of the problem. For this reason, intrusion detection systems might have a hard time keeping up with the constantly evolving threat landscape that is cyber security. As a result, creating intelligent intrusion detection systems capable of solving open-set issues is essential. The goal is to decrease the possibility of falsely labeling an unknown assault as recognized. Several approaches have been suggested for the open-set problem of network intrusion detection, including the Weibull-calibrated Support Vector Machine (W-SVM) and the Extreme Value Machine (EVM). Maintaining classification accuracy when using EVT and SVM to approximate a single score metric for unknown attack detection is difficult. In this research, we suggest a hierarchical architecture for tackling this problem by splitting intrusion detection into two distinct but interrelated activities: preserving the precision of classification for known assaults and discovering new threats. The two-step process of employing class conditional auto-encoders and EVT accomplishes this. By recasting the known/unknown intrusion detection problem as a two-stage minimization problem, using Conditioned Variational Auto-Encoder (CVAE) and EVT to build a hierarchical detection framework, and clustering and relabeling benign flows to lower the false positive rate, the proposed method outperforms benchmark methods relying on a single score measure. Using two popular datasets, we show that our suggested technique performs much better in experiments.

II. RELATED WORK

For Network Intrusion Detection Systems (NIDS), researchers have previously looked at machine learning and deep learning, emphasizing choosing suitable learning models and traffic characteristics. Studies in recent years have investigated the efficacy of using several machine learning models, such as Support Vector Machines (SVM), Random Forest (RF), and Extreme Learning Machines (ELM), to enhance intrusion detection capabilities. Comparisons between deep learning and more conventional approaches have shown that the former excels at evaluating big data sets and doing more thorough system analysis. Novel multilevel semi-supervised learning frameworks have been presented to solve issues like class imbalance and non-identical distribution in IDS. Model selection is crucial, but feature selection is just as important when perfecting intrusion detection systems. Effective representations for intrusion detection have been studied via packet- and flow-level characteristics and direct analysis of network traffic payload. Self-Taught Learning (STL) IDS and Sparse AutoEncoder (SAE) are two more approaches that have been worked on for autonomous feature learning. An adversarial statistical learning mechanism for anomaly detection has been presented as a solution to the problem of evasion and poisoning attacks against ML or DL-based intrusion detection systems; this mechanism may self-adapt in the face of data poisoning assaults.

One of the most promising aspects of machine learning-based intrusion detection is its ability to identify previously undiscovered threats, sometimes known as zero-day assaults. An unknown assault may be detected by learning the typical communication pattern and looking for any variations that signal an attack. This is known as a one-class classification issue. In this context, auto-encoders are the most popular model choice, with several studies investigating their potential value in detecting previously unseen attacks. Online bidirectional principal component analysis, tensor factorization, Bayesian network analysis, and ensemble classification techniques are only some of the other machine learning methods presented. Furthermore, known assaults may fine-tune unsupervised models with known threat data or retrain the detection model in real time on manually marked misclassified flows. Both can enhance anomaly detection performance.

Zero-day assaults are a massive worry for businesses and academics, so machine learning-based intrusion detection is essential. Unknown attack detection is often framed as a one-class classification issue, in which a model is trained on typical traffic and then alerted to any variations as potential threats. Alternatively, the detection model may be prepared using information about known assaults. The most popular model for this task is the auto-encoder. Research has used diverse approaches, including sparse auto-encoders, stacked non-symmetric deep auto-encoders, and Bayesian networks. Tensor factorization and symbolic sequence analysis are only two of the other systems that have been investigated. Multiple methods have been suggested, such as semi-supervised learning and ensemble classification.

III. PROPOSED ARCHITECTURE

The two-stage architecture shown in Fig. 1 tackles the hierarchy-based problem by training hierarchical models and doing hierarchical intrusion detection. The two-stage model training process is shown in Fig. 1(a), which involves the simultaneous training of both a general attack categorization and an unknown assault identification. Figure 1(b) illustrates the two-stage procedures for determining whether an assault is known or unknown. Training data is biased toward labeling all occurrences of benign traffic as belonging to the same category. However, various app categories show off relatively low-key traffic patterns. Due to the possibility of a minor separation between benign activity and anomalous behavior, false alarms and misclassifications may occur more often for benign traffic. We use K-means and other unsupervised clustering techniques to address this issue to classify benign traffic flows that share behavioral characteristics. We buy several flow labels for a more secure transportation system, one for each category. Later, the detection model is trained with the relabeled regular traffic.

We adopt the two-stage architecture for hierarchical model training and intrusion detection shown in Fig. 1 to solve the hierarchy problem. First, we attempt to decrease the empirical risk of misclassifying known attacks by training a discriminative model to approximate the classification function g1. To achieve this, we maximize the probability that the actual value y is given the input instance x ($P(y|x)$) by using a Conditional Variable Approximation Engine (CVAE) (a). The information about the latent z is crucial for the next stage of training, which is termed identifying attacks that have not yet been seen.

As mentioned, we employ the first-stage classification function g1 to make inferences about previously undetected attacks while limiting the identifying risk associated with doing so in the second phase (2). The generative model may learn an approximation of the identification function g2 by training on the outputs of the discriminative model, namely the intrinsic distribution of the instance x from each class, $P(x|y)$. The reconstruction error between the input and generated instances is determined by the trained generative model using the label predicted by the discriminative model. Since a significant deviation usually indicates an unanticipated attack, it might help separate it from a typical one.

Training the generative model is aided by the latent feature space's much lower dimensional representation than the input feature space. Learning the underlying distribution of previously observed attacks in the latent feature space over z

rather than the input feature space over x improves the generative model's performance.

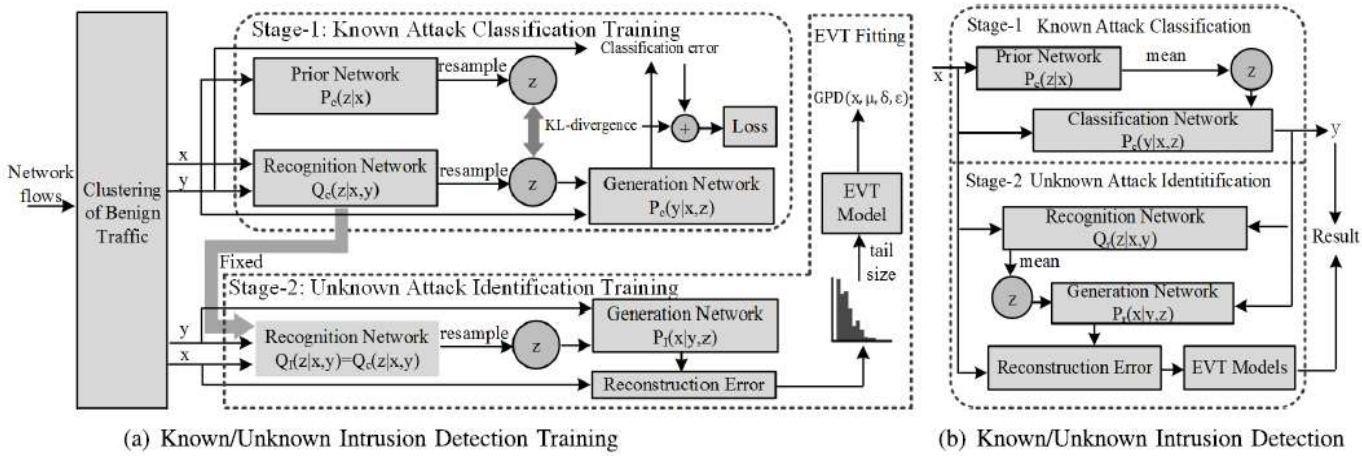

Figure 1: Proposed IDS Architecture

Algorithm 1: Intrusion Detection System

Input: Labeled known data {xi, yi} n i=1.

Output: Trained neural networks QC (x, y), PC(x), PC(x, z), and PI(y, z), and K EVT models with parameters {ui, ξi, σi}Ki=0.

1. Cluster benign traffic using K-means.
2. Stage-1: Train the known attack classification model
3. For each epoch from 1 to E do
4. For each batch with size M do
5. Compute the mean and variance using Eq. (10) and Eq. (11).
6. Sample z using Eq. (13).
7. Predict results and compute loss using Eq. (16).
8. Update the weights of the prior network, recognition network, and classification network using Stochastic Gradient Descent (SGD).
9. End for
10. End for
11. Stage-2: Train the unknown attack recognition model
12. For each epoch from 1 to E do
13. For each batch with size N do
14. Compute the mean and variance using Eq. (10).
15. Sample z using Eq. (13).
16. Get reconstructed results and compute loss using Eq. (21).
17. Freeze the parameter of QC(x, y).
18. Update the weights of PI(y, z) using SGD.
19. End for
20. End for
21. Compute the reconstructed error R and obtain {Ri}Ki=0.
22. For i from 0 to K do
23. For each sample rk in Ri do
24. Compute the mean excess function using Eq. (24).
25. End for
26. Find the linear region of (rk, Ek), where ui is set to the minimum of r in the linear region.
27. Get reconstruction errors larger than ui: Ti = {r | r > ui, r ∈ Ri}.
28. Estimate the location and shape parameters: ξi, σi = GPDFi(t) (Ti).
29. End for.

IV. RESULTS AND OBSERVATION

To test our approach, we conducted evasion assaults on the CICIDS2017 Setting1 dataset using the fast gradient sign method (FGSM) [38], a well-established technique for producing adversarial samples. We used FGSM to fine-tune our attacks on the test set not to be detected while providing a meaningful test (known and unknown assaults). Even though we built adversarial examples using FGSM without thinking about whether they could be used successfully, the confusion matrices of our two-stage detection strategy are provided in Fig. 2. Research shows that most known assaults are mistakenly assessed as innocuous during the known attack categorization stage when evasive attacks are included, as illustrated in Fig. 2. Figure 2 shows that the vast majority of malicious traffic is incorrectly labeled as unknown. As a result, with this degree of assault detection, the situation is better. Although we successfully dampened the effects of a standard FGSM attack, this does not render our system impervious to adaptive ones. We aim to learn more about the method employed to address this issue, which involves retraining a neural network to agree with the detection model's classification results.

Model	Accuracy	Precision	Recall	F1-Score
CVAE	0.918	0.924	0.922	0.921
EVT	0.908	0.909	0.91	0.91
CVAE-EVT*	0.933	0.937	0.937	0.937
SVM	0.897	0.898	0.896	0.897
Random Forest	0.911	0.913	0.913	0.913

Fig 2. The performance comparison of the proposed work

V. CONCLUSION

This article aims to detect new assaults while retaining high classification accuracy for known traffic, a challenge known as open-set intrusion detection. We approached this issue by categorizing known attacks and identifying undiscovered ones into its parts. As a result of developing a hierarchical issue formulation, we could imagine a two-stage intelligent framework for identifying known and undiscovered attacks. We used a two-stage framework that included training and testing methods, and we did so using CVAE and EVT. By examining reconstruction mistakes, our technique may detect previously unseen assaults. The outcomes of our experiments demonstrate the viability and efficiency of our CVAE-EVT* assisted strategy. We also compared our method to the best available benchmarks and found it significantly outperformed them. However, obstacles still exist due to extremely asymmetrical network traffic and a lack of insight into potential assaults.

VI. REFERENCES

[1]. C. Ting, R. Field, A. Fisher, and T. Bauer, "Compression analytics for classification and anomaly detection within network communication," IEEE Trans. Inf. Forensics Security, vol. 14, no. 5, pp. 1366–1376, May 2019.

[2]. H. Alipour, Y. B. Al-Nashif, P. Satam, and S. Hariri, "Wireless anomaly detection based on IEEE 802.11 behavior analysis," IEEE Trans. Inf. Forensics Security, vol. 10, no. 10, pp. 2158–2170, Oct. 2015.

[3]. S. Cruz, C. Coleman, E. M. Rudd, and T. E. Boult, "Open set intrusion recognition for fine-grained attack categorization," in Proc. IEEE Int. Symp. Technol. Homeland Security. (HST), Apr. 2017, pp. 1–6.

[4]. W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 11, pp. 2317–2324, Nov. 2014.

[5]. J. Henrydoss, S. Cruz, E. M. Rudd, M. Gunther, and T. E. Boult, "Incremental open set intrusion recognition using extreme value machine," in Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2017, pp. 1089–1093.

[6]. E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 3, pp. 762–768, Mar. 2018.

[7]. I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," IEEE Access, vol. 6, pp. 33789–33795, 2018.

[8]. B. Dong and X. Wang, "Comparison deep learning method to traditional methods used for network intrusion detection," in Proc. 8th IEEE Int. Conf. Commun. Softw. Netw. (ICCSN), Jun. 2016, pp. 581–585.

[9]. H. Yao, D. Fu, P. Zhang, M. Li, and Y. Liu, "MSML: A novel multilevel semi-supervised machine learning framework for an intrusion detection system," IEEE Internet Things J., vol. 6, no. 2, pp. 1949–1959, Apr. 2019.

[10]. N. Moustafa, K.-K.-R. Choo, I. Radwan, and S. Camtepe, "Outlier Dirichlet mixture mechanism: Adversarial statistical learning for anomaly detection in

the fog," IEEE Trans. Inf. Forensics Security, vol. 14, no. 8, pp. 1975–1987, Aug. 2019.
[11]. M. E. Ahmed, S. Ullah, and H. Kim, "Statistical application fingerprinting for DDoS attack mitigation," IEEE Trans. Inf. Forensics Security, vol. 14, no. 6, pp. 1471–1484, Jun. 2019.
[12]. M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, "Deep learning approach combining sparse autoencoder with SVM for network intrusion detection," IEEE Access, vol. 6, pp. 52843–52856, 2018.
[13]. S. Naseer et al., "Enhanced network anomaly detection based on deep neural networks," IEEE Access, vol. 6, pp. 48231–48246, 2018.
[14]. W. Wang et al., "HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection," IEEE Access, vol. 6, pp. 1792–1806, 2018.
[15]. Z. Zhang, Q. Liu, S. Qiu, S. Zhou, and C. Zhang, "Unknown attack detection based on zero-shot learning," IEEE Access, vol. 8, pp. 193981–193991, 2020.
[16]. Topics Comput. N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," IEEE Trans. Emerg. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.
[17]. V. L. Cao, M. Nicolau, and J. McDermott, "Learning neural representations for network anomaly detection," IEEE Trans. Cybern., vol. 49, no. 8, pp. 3074–3087, Aug. 2019.
[18]. Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," in Proc. Netw. Distrib. Syst. Secur. Symp., Feb. 2018, p. 2.
[19]. K. Xie et al., "Online anomaly detection with high accuracy," IEEE/ACM Trans. Netw., vol. 26, no. 3, pp. 1222–1235, Jun. 2018.
[20]. [20] K. Xie et al., "Fast tensor factorization for accurate Internet anomaly detection," IEEE/ACM Trans. Netw., vol. 25, no. 6, pp. 3794–3807, Dec. 2017.