Performance Evaluation of a Combined Anomaly Detection Platform Using Machine Learning Techniques

M. Srikanth[1], D. Prasanna[2]

[1]*Assoc. Prof, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India*

[2]*PG Scholar, Dept of CSE, Tirumala Engineering College, Narasaraopet, Guntur, A.P., India*

ABSTRACT - Hybrid Anomaly Detection Model (HADM) is a platform that filters network traffic and identifies malicious activities on the network. The platform applies data mining techniques to tackle the security issues effectively in high-load communication networks. The platform uses a combination of linear and learning algorithms combined with a protocol analyzer. The linear algorithms filter and extract distinctive attributes and features of the cyber-attacks, while the learning algorithms use these attributes and features to identify new types of cyber-attacks. The protocol analyzer in this platform classifies and filters vulnerable protocols to avoid unnecessary computation load. Using linear algorithms in conjunction with learning algorithms and protocol analyzer allows the HADM to achieve improved efficiency in terms of accuracy and computation time to detect cyber-attacks over existing solutions. While the authors' previous paper evaluated HADM efficiency (accuracy and computation time) against related studies, this paper concentrates on HADM robustness and scalability. For this purpose, five datasets, including ISCX-2012, UNSW-NB15 Jan, UNSW-NB15 Feb, ISCX-2017, and MAWILab-2018 with various sizes and diverse attacks, have been used. Different feature selection methods are applied to find the best features. The feature selection methods are selected based on the algorithms' computation time and detection rate. The best algorithms are then selected through a benchmark on applied datasets and based on the metrics such as cross-entropy loss, precision, recall, and computation time. The result of the HADM platform shows robustness and scalability against datasets with different sizes and diverse attacks.

Keywords: Covid-19, Machine learning, Forecasting

I. INTRODUCTION

There is no doubt that the Internet plays an essential role in different aspects of life these days. For example, it has been found that social networking such as Facebook, Twitter, and Linked-in have a remarkable impact in bringing people from different parts of the world together (Muila, 2010). Although it has changed the world, it has raised the possibility that malicious users gain illegal access to organizations to steal confidential information they are interested in or destroy it by injecting malware applications. Those applications are created to give malicious users the ability to control organizations' computers remotely. Malicious users get illegal access to those organizations by exploiting weaknesses and vulnerabilities in organizations' networks or web applications. The impact of attacks can lead to delaying delivery services in some organizations causing financial damages. A survey made by Statistica (2015) provides information on the distribution of costs for external consequences of targeted cyber-attacks on companies in global markets in 2014.

Williams (2014) reported that cyber-attacks were estimated to cost the global economy around $445 billion annually. She also reported that those attacks affected more than 800 million people in 2013. An annual study was conducted by the Ponemon Institute (2014) in seven countries, including the United States, United Kingdom, Germany, Australia, Japan, France, and the Russian Federation. The study involves a total benchmark sample of 257 organizations. Figure 1.2 presents the estimated average cost of cyber-attacks for each country; it has been found that the US sample achieved the highest total average cost at $12.7 million while the Russian Federation sample got the lowest total average cost at $3.3 million. The figure also that the cost of cyber-attacks went up in six countries during the past year compared to 2013 (apart from the Russian Federation), the highest increase was found in the United Kingdom (22.7%), while the lowest increase was found in Japan (2.7%). The study also reported that cyber-attacks target all industries but at different levels. The study pointed out that organizations providing energy and financial services experience higher cyber-attack costs than organizations providing services in media, life sciences, and healthcare.

Regardless of the wide advancement of information development, detection mechanisms and frameworks for detecting intrusions are not escalated. The amounts of hacking and interference scenes are extending year on year as development takes off. The Security peril originates from external interlopers just as from inward customers as maltreatment. Thus, on the off chance that there is a necessity to allow an opening to a framework, at that point a firewall

which is a static guideline based, unfit to shield from interruption endeavours. The firewall will almost certainly break the framework, and it can open the structure into the framework and is unfit to separate between positive or negative movement.

On the other hand, Intrusion Detection Systems can analyse and detect security breaches. Interestingly, Intrusion Detection Systems (IDS) can stare at the hostile activity on these systems. The conventional representation of IDS is portrayed in figure 1.1.

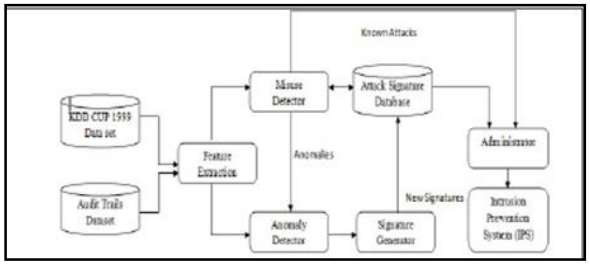

Figure 1: Representation of Generic IDS

II. RELATED WORK

Di Pietro et al. [3] apply machine learning algorithms such as k-nearest Neighbor (k-NN) and Support Vector Machine (SVM) to detect anomalies. Furthermore, a Deep Packet Inspection (DPI) mechanism is utilized to define rules for capturing packets. However, the rules, protocols, and details of the process are not explained. In addition, the authors have not discussed their model scalability and robustness, neither any experimental result is presented in this study.

Vasseur et al. [4] propose a supervised learning classifier to detect DDoS attacks. This study applies Deep Neural Networks (DNN) classifier, which mainly concentrates on optimizing the training process to provide labeled data. However, this study introduces a combined method, and In addition, the authors have not discussed their model scalability and robustness neither any experimental result is presented.

Pietro et al. [5] apply a machine learning-based model comprising ANN to compare received traffic with expected traffic. The presented model is trained with expected traffic, and upon receiving the input data, the data signature is compared with the expected traffic. If they are different, a signature for the attack class will be generated, and the model will be trained with new information. This study doesn't discuss any types of attack, neither the implementation result is presented.

Yadav et al. [6] propose a Virtual Machine (VM) based analytic model to detect anomalies within the network traffic based on the dynamic modelling of network behavior. They have applied honeypot to collect malicious traffic. Though the model comprises unsupervised and supervised machine learning algorithms, the honeypot relies only on received attacks and not the other attacks. In this study, applied algorithms are not disclosed, and the experimental result is not presented. In addition, the authors have not discussed the scalability and robustness of their model.

III. PROPOSED ARCHITECTURE

With simple Alternatives of previous work, polynomial and Prophet have been applied and evaluated. Polynomial is similar to the linear model. However, it works well on exponential growth, and Facebook developed prophet with new evolution on the forecasting models. It is a forecast time series addictive model which automatically fits the trends like weekly, yearly, daily. Etc. By giving value to object(periods). Some features are Accurate and fast, likelihood implementation. The proposed architecture is shown in Figure 1.

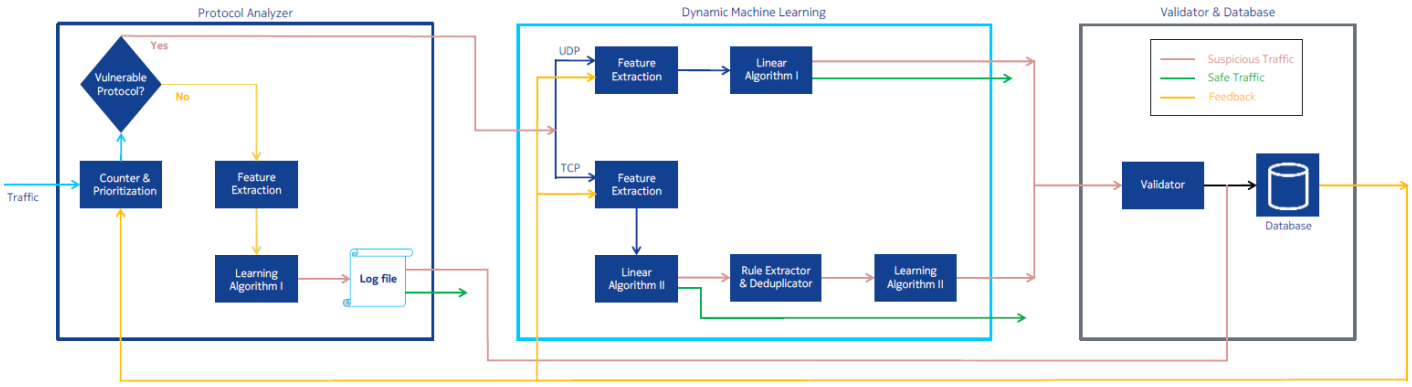

Figure 1. Proposed architecture

IV. RESULTS AND OBSERVATION

The Proposed model has been developed using the SKNN Classification model and Statistical analysis tool; R programming language is used for analytical and classification activities. The KLAR library package is capable of adapting various class labels used in the classification. The Results of Anomaly and Misuse attacks detection are presented in Figure 2.

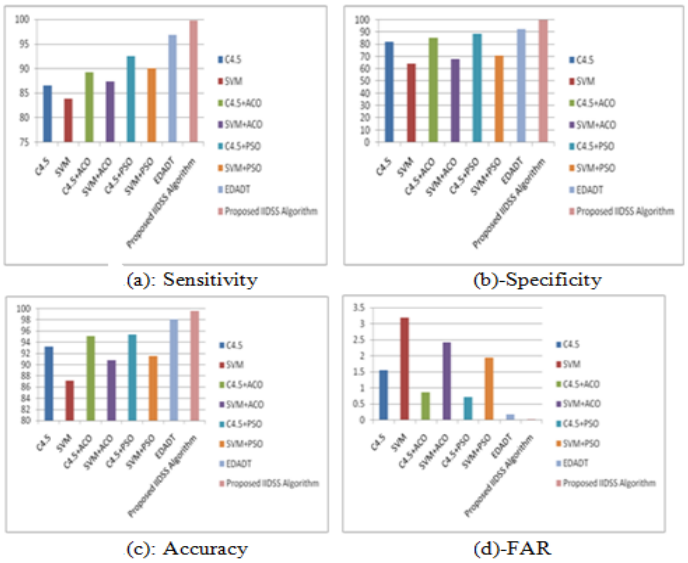

Figure 2. Results obtained

V. CONCLUSION

Even though it has been challenging to find reliable and publicly available datasets to measure the model robustness and scalability over the previous study, the model has been tested with various datasets. In this paper, various feature selection methods have been applied with several algorithms to achieve the highest efficiency. The experimental results show that the SKNN algorithm and SVM online feature selection improve User Datagram Protocol (UDP) Denial of Service (DoS) detection accuracy and reduced computation time. Similarly, the Decision Tree (DT) algorithm with SVM online feature selection method gives higher efficiency for other attacks. The results show that HADM did not have a tremendous increase in computation time nor a considerable decrease in detection factors. In contrast, various datasets with different sizes and diverse attacks have been used. This shows that the proposed model is scalable and robust.

VI. REFERENCES

[1] K. S. Desale and R. Ade, “Genetic algorithm-based feature selection approach for an effective intrusion detection system,” in 2015 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, pp. 1-6, 2015.

[2] M. Monshizadeh and Z. Yan, “Security Related Data Mining,” in IEEE International Conference on Computer and Information Technology, Xi’an, pp. 775-782, 2014.

[3] A. D. Pietro et al., “Dynamic deep packet inspection for anomaly detection,” US Patent 2017099310 (A1), 6 Apr. 2017.

[4] J. Vasseur et al., “Anomaly detection in a network coupling state information with machine learning outputs,” US Patent 20170104774 (A1), 13 Apr. 2017.

[5] A. D. Pietro et al., “Signature creation for unknown attacks,” US Patent 20160028750 (A1), 28 Jan. 2016.

[6] N. Yadav et al., “Network behavior data collection and analytics for anomaly detection,” US Patent 20160359695 (A1), 8 Dec. 2016.

[7] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, “From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods,” in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 3369-3388, 2018.

[8] B. G. Atli, “Anomaly-based intrusion detection by modeling probability distributions of flow characteristics,” 2017.

[9] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436-444, 2015.

[10] G. Huang, Q. Zhu, and C. Siew, “Extreme learning machine: theory and applications,” in Neurocomputing, vol. 70, no. 1-3, pp. 489-501, 2006.

[11] N. Moustafa and J. Slay, “The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set”, Information Security Journal: A Global Perspective, vol. 25, no. 1-3, pp. 18-31, 2016.

[12] C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to Information Retrieval,” in Natural Language Engineering, 16(1), pp. 100-103.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” in Machine Learning, vol. 46, no. 1-3, pp. 389–422, 2002.

[14] P. Laskov, C. Gehl, S. Krüger and K. Müller, "Incremental support vector learning: Analysis, implementation and applications," in The Journal of Machine Learning Research, vol. 7, pp. 1909-1936, 2006.

[15] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, “Toward developing a systematic approach to generate benchmark datasets for intrusion detection,” in Computers & Security, Volume 31, Issue 3, May 2012, pp. 357-374.

[16] N. Moustafa and J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” in 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, 2015, pp. 1-6.

[17] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, 2018, pp. 108-116.
[18] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking," in ACM CoNEXT 2010, Philadelphia, PA, 2010, pp. 8:1-8:12.
[19] M. M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," in International Journal of Machine Learning and Computing vol. 3, no. 2, pp. 224-228, 2013.
[20] scikit-learn Machine Learning in Python, https://scikit-learn.org/stable/