Advanced Network Anomaly Detection System Using Data Mining Techniques

Rubeena Begum[1], Mr. Palli R Krishna[2]
[1]*P.G. Scholar,* [2]*Associate Professor*

[1, 2] *Department of CSE, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India*

ABSTRACT - Nowadays, there is a huge and growing concern about security in information and communication technology (ICT) among the scientific community because any attack or anomaly in the network can greatly affect many domains such as national security, private data storage, social welfare, economic issues, and so on. Therefore, the anomaly detection domain is a broad research area, and many different techniques and approaches for this purpose have emerged through the years. Attacks, problems, and internal failures when not detected early may badly harm an entire Network system. Thus, this thesis presents an autonomous profile-based anomaly detection system based on the statistical method Principal Component Analysis (PCADS-AD). This approach creates a network profile called Digital Signature of Network Segment using Flow Analysis (DSNSF) that denotes the predicted normal behavior of a network traffic activity through historical data analysis. That digital signature is used as a threshold for volume anomaly detect ion to detect disparities in the normal traffic trend. The proposed system uses seven traffic flow attributes: Bits, Packets and Number of Flows to detect problems, and Source and Destination IP addresses and Ports, to provides the network administrator necessary information to solve them. The observed results seek to contribute to the advance of the state of the art in methods and strategies for anomaly detection that aim to surpass some challenges that emerge from the constant growth in complexity, speed and size of today's large scale networks, also providing high-value results for a better detection in real time.

Keywords: Anomaly Detection, Intrusion Detection System, Network Security, Principal Component Analysis

I. INTRODUCTION

Nowadays, the scientific community has a constant worry about high-efficiency security and quality of service in large-scale networks. The expansion of new communication technologies and services, along with an increasing number of interconnected network devices, web users, services, and applications, contributes to making computer networks ever larger and more complex as systems. Moreover, there is the so called boundless communication paradigm, for next generation networks, which envisages offering anytime, anywhere, anyhow communications to its users and requires the full integration and interoperability of emergent technologies [1]. These issues make it even more complex and challenging to maintain precise network management and lead to serious network vulnerabilities, as security incidents may occur more frequently [2, 3].

Outsiders, as malicious attacks aiming to shut down services or steal private information, or by inside factors (operational problems), such as configuration errors, server crashes, power outages, traffic congestion, or non-malicious large file transfers [4], can cause such security instances. Regardless of the source, such threats, which are commonly called anomalies, can have a significant impact on the network service and end-users and harm computer network operations and availability. The term anomaly has several definitions. Barnett and Lewis define a data set anomaly as "observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data" [5]. Chandola et al. express this term as "patterns in data not conforming to a well-defined notion of normal behavior" [6]. According to Lakhina et al., "anomalies are unusual and significant changes in a network's traffic levels, which can often span multiple links" [7]. Hoque et al. define it as "non-conforming interesting patterns compared to the well-defined notion of normal behavior" [8]. By these definitions, it is clear that the concept of normality is one of the main steps toward developing a solution to detect network anomalies. Although apparently unpretentious, the problem of defining a region denoting normal behavior and marking as an anomaly any occasion contrasting this normal pattern is defiant. Faster diagnosis, lower complexity and suitable corrections of the causes are the main objectives of the field. Every factor is vital to developing a better anomaly detection approach. The precision and speed factors, alongside with the correct identification of such abnormal events in a timely fashion are critical to reducing significant service degradation, malicious damage, and cost. For this reason, the research community has been developing a lot of models, algorithms, and mechanisms,

over the years, to develop better solutions and approaches to guaranteeing the health of ever larger and complex network systems.

In the literature, anomaly detection methods can be classified into two ways: Signature based and profile-based. Signature-based systems use a prior knowledge about the characteristics of each kind of anomaly to identify potential incidents previously known. Moreover, profile based approaches create a network profile representing the traffic normal behavior, and traffic anomalies are detected from deviations with respect to this profile [9, 10]. Although signature based methods have been widely investigated in the literature, they have a clear drawback. It is prerequisite that anomaly signatures are known in advance, hampering the recognition of new anomalies. Also, signature-based methods can be avoided by malicious sources by tampering anomaly signatures. In contrast, a profile-based system creates a baseline profile of the normal network activity, eliminating the need of prior knowledge about the nature and properties of anomalies. This trait leads to some advantages: The possibility of discovering new and unforeseen types of anomalies; the detection of insider attacks; and also makes it difficult for an attacker to know with conviction what malicious action it can carry out without being detected by the system [9, 11]. Thus, this thesis proposal is to create an autonomous profile-based monitoring system capable of identifying the normal network behavior by adopting an efficient method for traffic characterization in order to create a baseline profile of normal traffic to discover possible anomalies in the traffic.

II. RELATED WORK

Nowadays, there is a huge and growing concern about security in information and communication technology (ICT) among the scientific community because any attack or anomaly in the network can greatly affect many domains such as national security, private data storage, social welfare, economic issues, and so on. Therefore, the anomaly detection domain is a broad research area, and many different techniques and approaches for this purpose have emerged through the years. Researchers have been studying the anomaly detection subject since the early 19th century, and so far, they have produced a multitude of papers, each using a variety of techniques, from statistical models, up to evolutionary computation approaches. Nevertheless, it is not a straightforward task to identify and categorize all existing anomaly detection techniques. Plenty of topics must be considered, such as anomaly types, system types, techniques and algorithms used, as well as technical dilemmas such as processing costs and network complexity. Therefore, this leads to the fragmented literature available today, in which many works try to summarize everything but are unable to show the bigger picture of the anomaly detection spectrum.

As in [18] and [9], the focus is just on the most popular techniques and methods, such as machine learning, clustering and statistical approaches. Still, surveys such as [19] and [20] briefly discuss the whole problem statement, setting aside relevant topics such as data set, challenges, and recommendations. Marnerides et al. [21] have reviewed anomaly detection over backbone networks. Although each of those inspected surveys summarizes many important topics pertaining to anomaly detection, they are not entirely complete. For instance, some of them emphasize anomaly types but do not cover all kinds of methods while others research upon vast approaches but forget about the basis of intrusion detection systems and data input, and so on. For this reason, the main objective is to review the most important aspects pertaining to anomaly detection, covering an overview of a background analysis as well as a core study on the most relevant techniques, methods, and systems within the area. Therefore, in order to ease the understanding of this chapter's structure, the anomaly detection domain was reviewed under five dimensions: (i) network traffic anomalies, (ii) network data types, (iii) intrusion detection systems categories, (iv) detection methods and systems, and (v) open issues.

The nature of an anomaly is an important aspect of an anomaly detection technique. Depending on the context within which an abnormality is found, or on how it occurred, it can be or not be an abnormality. This aspect can direct how the system will handle and understand mined and detected anomalies. Based on their nature, there are three categories of anomalies: point anomalies, collective anomalies, and contextual anomalies [6, 10, 22].

Contextual Anomalies, also called conditional anomalies, are events considered as anomalous depending on the context in which they are found. Two sets of attributes define a context (or the condition) for being an anomaly, both of which must be specified during problem formulation. Contextual attributes define the context (or environment); for instance, geographic coordinates in spatial data or time in time-series data specifies the location or position of an instance, respectively. On the other hand, behavioral attributes denote the non-contextual features of an instance, i.e., indicators determining whether or not an instance is anomalous in the context [23].

III. PROPOSED SYSTEM

In this chapter, the hybrid anomaly detection system using principal component analysis is presented. However, before explaining its full process, Figure 1 summarizes the overall operation of it. The system is divided in two parts: Traffic Characterization and Anomaly Detection. The traffic characterization is responsible for extracting quantitative attributes (bits/s, packets/s and number of flows/s) from a flow

database containing historical data about the network segment activity, and generate the corresponding DSNSFs. The mentioned components are deployed in conjunction with one another to filter packets on the communication networks, such as mobile networks and for certain network protocols that are known or considered to be vulnerable to or used in cyber-attacks. This allows the HADM to expend a smaller amount of processing resource on other network protocols, such as streaming protocols that are not normally vulnerable and thus not typically targeted by cyber-attackers. The ability of the HADM to focus on vulnerable network protocols helps to avoid burdening network servers with unnecessary computational load. The protocol analyzer filters the network packets and identifies vulnerable protocols. The non-vulnerable protocols are forwarded to the feature extraction module for further processing. The feature extraction module extracts features from the incoming packets and provides these features to the learning algorithm I for the analysis. If the output from learning algorithm I is suspicious, it is recorded into log file. If traffic is carried on vulnerable protocol, the counter and prioritization module forwards the suspicious traffic to next level based on the occurrence of protocol against a defined threshold.

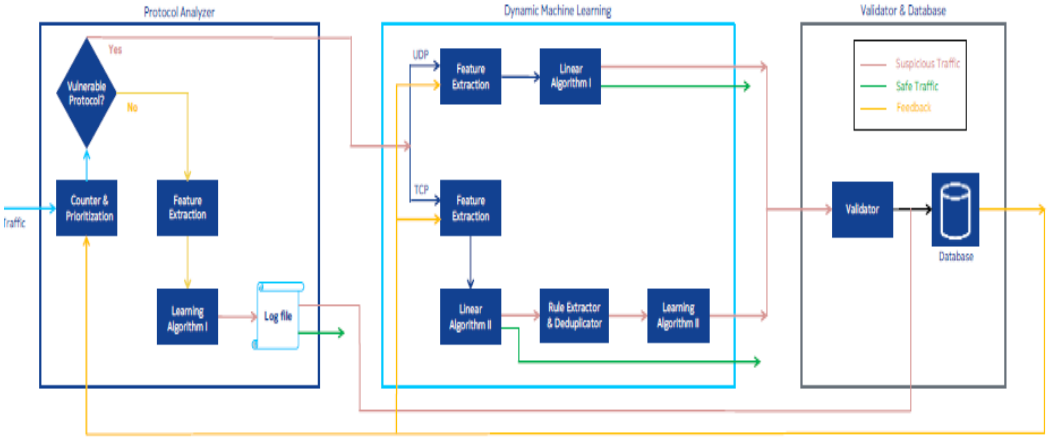

Figure 1. Proposed system architecture

IV. RESULTS

The current framework was created from the idea of Hybrid PSO and C4.5. In this investigation, The IDS framework is lived in the ideas of SKNN Classifier actualized in R. In this work "klaR" bundle accessible in R. The outcomes acquired show adequate exactnesses. The outcomes are appeared.

Table 1: Results Comparison

Techniques	Sensitivity	Specificity	Accuracy	FAR
C4.5	87.57	83	91.24	1.45
SVM	81.92	63.29	88.27	3.01
C4.5+ACO	89.15	86.43	96.15	0.88
SVM+ACO	97.31	69.66	91.82	1.11
C4.5+PSO	93.40	89.88	96.37	1.83
SVM+PSO	91.50	71.10	92.59	2.96
EDADT	96.65	92.25	97.11	0.20
Proposed Method	99.81	99.90	99.62	0.01

The Proposed model has developed using 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 varied class labels used in the classification. The Results of Anomaly and Misuse attacks detection is presented in Figure 2.

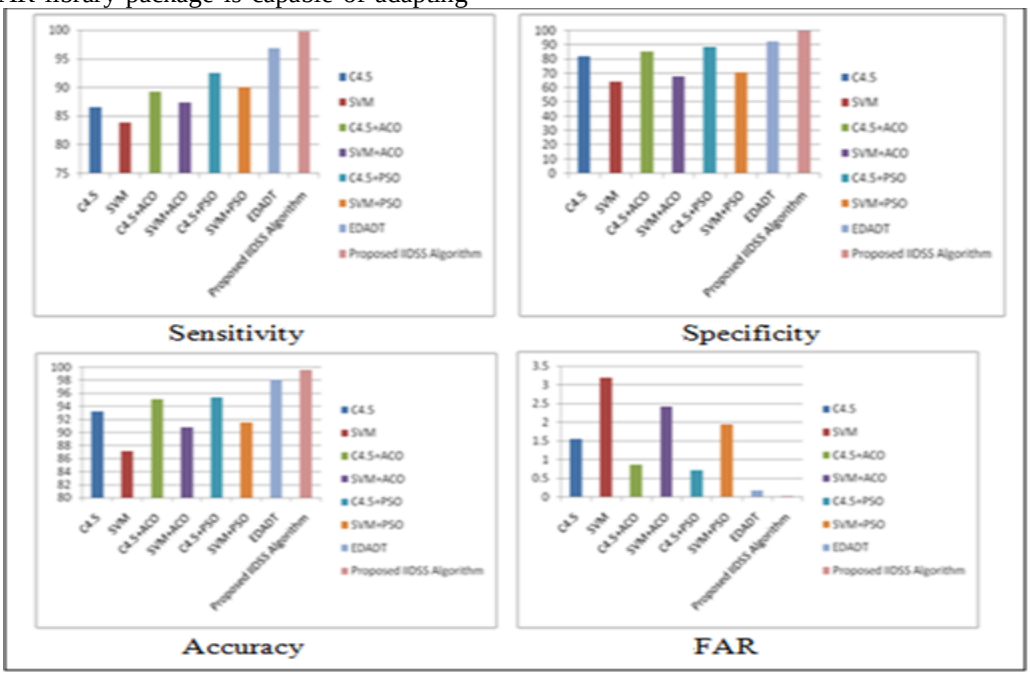

Figure 2: Results observed -Sensitivity, Specificity, and Accuracy and FAR

V. CONCLUSION

The proposed Intrusion Detection display has indicated better outcomes contrasted and existing strategies. The procedure has indicated broad productivity in limiting the bogus alert rate and will diminish the manager remaining burden. This model has created 13.24% higher affectability when contrasted and C4.5, 10.55% contrasted and C4.5+ACO, and 2.95% when contrasted and EDADT. The exploratory outcome indicates better precision contrasted and the current framework. The Future work is meant to prepare the IDS to distinguish number of assaults, and the tally can be expanded from 23 to 40. The intrusion detection systems are very efficient for monitoring and detecting network traffic data packets. This research paper has proven that alerts are generated when there is a deviation in the behavioral patterns of the packets. The patterns are matched and compared with the proposed snort rules signature base. The proposed system was methodically tested and compared with existing snort rules, the proposed rules proved to be more accurate and efficient. In future work, advanced data mining techniques and machine learning techniques used for detecting new suspicious attacks on a huge amount of data

VI. REFERENCES

1. A.Saidi et al.:, The functional of A Mobile Agent System to Enhance DoS and DDoS Detection in Cloud, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6 (2016) pp 4615-4617
2. Adeeb Alhomoud et al., Performance Evaluation Study of Intrusion Detection Systems, The 2nd International Conference on Ambient Systems, Networks and Technologies, (ANT), Procedia Computer Science 5 (2011) 173–180, 1877–0509 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Prof. Elhadi Shakshuki and Prof. Muhammad Younas. doi:10.1016/j.procs.2011.07.024
3. Anderson, James P., "Computer Security Threat Monitoring and Surveillance," Washing, PA, James P. Anderson Co., 1980.
4. Anna L. Buczak. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, 10.1109/COMST.2015.2494502, IEEE Communications Surveys & Tutorials, 1553-877X (c)
5. Aymen Abid et al.:, Outlier detection for wireless sensor networks using density-based clustering approach, IET

Wireless. Sens. Syst., 2017, Vol. 7 Iss. 4, pp. 83-90, The Institution of Engineering and Technology 2017, ISSN 2043-6386

6. Bellovin, S.M. "Network Firewalls", IEEE Communications Magazine, Vol. 32, pp. 50- 57, 1994.
7. Berchtold, B. Ertl, D. A. Keim, H.-P. Kriegel, and T. Seidl. Fast nearest neighbor search in highdimensional space. In Proceedings of the Fourteenth International Conference on Data Engineering, ICDE '98, pages 209–218, Washington, DC, USA, 1998. IEEE Computer Society.
8. Blum, Avrim L. & Pat Langley (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1-2), 245–271
9. Catania Carlos A, Garino Carlos. Automatic network intrusion detection: current techniques and open issues. Elsevier Comput Electr Eng 2012; 38(5):1062–72.
10. Chien-Yi Chiu, Yuh-Jye Lee, Chien-Chung Chang. Semi-supervised learning for false alarm reduction. In: Industrial conference on data mining, no. 10; 2010. p. 595–605.
11. Ching-Hao, Hahn-Ming L, Devi P, Tsuhan C, Si-Yu H. Semi-supervised co-training and active learning based approach for multi-view intrusion detection. In: ACM symposium on applied computing, no. 9; 2009. p. 2042–7.
12. Claude Turner et al. (2016). A Rule Status Monitoring Algorithm for Rule-Based Intrusion Detection and Prevention Systems, Complex Adaptive Systems, Conference Organized by Missouri University of Science and Technology 2016 - Los Angeles, CA, Procedia Computer Science 95 (2016) 361 – 368, 1877-0509, doi: 10.1016/j.procs.2016.09.346
13. Das, S. (2001). Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. Proc. 18thInt'l Conf. Machine Learning, 74-81
14. Dasgupta, D. and F. A. Gonzalez, "An intelligent decision support system for intrusion detection and response", In Proc. Of International Workshop on Mathematical Methods, Models and Architectures for Computer Networks Security (MMM-ACNS), St.Petersburg. Springer- , 21-23 May,2001.
15. Denning, D.E. "An Intrusion-Detection Model", in IEEE Transactions on Software Engineering, Vol.13, No. 2, pp. 222-232, 1987.
16. Dickerson, J. E. and J. A. Dickerson, "Fuzzy network profiling for intrusion detection", In Proc. of NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta, pp. 301306. North American Fuzzy Information
17. Divya and Surendra Lakra, "SNORT: A Hybrid intrusion detection system using artificial intelligence with a snort", International journal computer technology & application, Vol 4(3), 466-470, 2013.
18. E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.Govinda Rajulu. A Study of Intrusion Detection in Data Mining. Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011, July 6 - 8, 2011, London, U.K.
19. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. J., A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data, In D.Barbarà and S. Jajodia (eds.), Applications of Data Mining in Computer Security, Kluwer Academic Publishers, Boston, MA, 2002, pp. 78-99.
20. Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful Knowledge from volumes of data," Communications of the ACM 39 (11), November 1996, 2734.