

Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles

J. Yamini Devi¹, Sanjana Bandaru², Duggirala Preethi³, Nagandla Mounika⁴

¹ Assistant Professor, Department of IT, GRIET, Hyderabad, Telangana, India.

^{2,3,4} UG Student, Department of IT, GRIET, Hyderabad, Telangana, India.

Abstract- A significant challenge in cybersecurity is automating cyber threat detection efficiently. Our proposed method converts multiple security events into individual profiles and utilizes deep learning for enhanced identification within an AI-SIEM system. Integrating event profiling with various neural network methods like FCNN, CNN, and LSTM distinguishes true positives from false positives, enabling rapid threat response. Benchmark experiments (NSLKDD, CICIDS 2017) and real-world datasets compared our approach with five traditional machine learning techniques. Results indicate our method's effectiveness for network intrusion detection, even outperforming conventional methods in real-world scenarios.

KEYWORDS

Cybersecurity, Cyber-threat detection, Artificial neural networks, Deep learning, AI-SIEM system, Network intrusion detection

1. INTRODUCTION

The advancement of artificial intelligence (AI) has significantly enhanced learning-based methods for cyberattack detection, showing notable progress in various studies. However, the ever-changing landscape of cyber threats poses challenges to safeguarding IT systems effectively. To combat diverse intrusions and malicious activities, robust defenses and security measures are imperative. Traditionally, intrusion prevention systems (IPS) utilize signature-based methods within enterprise networks to generate intrusion alerts, which are managed by Security Information and Event Management (SIEM) systems. SIEM plays a pivotal role in collecting and analyzing security events and logs, offering a reliable solution for security operations. Learning-based approaches offer valuable insights into analyzing large datasets to determine potential attacks. These approaches, categorized into analyst-driven and machine learning-driven solutions, rely on expert-defined rules and detect anomalous patterns. However, existing methods encounter limitations. Firstly, they depend on labeled data for model training and evaluation, posing challenges in obtaining sufficient labeled datasets. Secondly, features used in theoretical studies often lack generalization for practical deployment in real-world network security systems. Recent advancements in intrusion detection leverage deep learning technologies for automation, evaluated using benchmark datasets like NSLKDD and CICIDS2017. Yet, these datasets may not fully represent real-world scenarios due to feature limitations. Overcoming these challenges requires evaluating learning models using datasets collected from actual environments. By addressing these limitations, learning-based models can be refined for more effective cyber threat detection in real-world applications.

2. REQUIREMENT ENGINEERING

2.1 Software Requirements: To ensure optimal performance, our cybersecurity solution requires a modern OS, a multi-core processor, storage, and a stable internet connection. It supports Python 3.8+ with libraries like TensorFlow, databases like MySQL, and integration with security tools (firewalls, IDS/IPS, SIEM). A web-based interface using React or similar frameworks and compatibility with Docker and Kubernetes are also necessary, to ensure robustness, efficiency, and scalability.

- Operating System: Windows 10
- Technology: Python
- Debugger and Emulator: Any Browser

2.2 Hardware Requirements: For our cybersecurity system to function effectively, it requires specific hardware components. These include a processor with ample processing power, such as an Intel i5 or AMD Ryzen 5 and above, to ensure smooth operation and efficient handling of tasks. Having 16 GB or more is preferable for optimal performance, especially when dealing with large datasets or multiple processes simultaneously. These hardware requirements collectively ensure that the cybersecurity system operates smoothly, and efficiently, and is capable of meeting the demands of modern cybersecurity needs.

- Processor: Pentium IV or higher
- RAM: 256 MB
- Space on Hard Disk: minimum 512MB

3. RELATED WORKS

The paper "Enhanced Network Anomaly Detection Based on Deep Neural Networks" delves into the heightened significance of information network security, prompted by the proliferation of Internet applications. It examines the effectiveness of deep learning methodologies in fortifying anomaly-based intrusion detection systems. Evaluating diverse deep neural network structures like CNNs, autoencoders, and RNNs on NSLKDD datasets, the study showcases promising real-world applicability. Furthermore, both deep and conventional machine learning IDS models undergo rigorous assessment employing established classification methods and evaluation metrics.

In "Network Intrusion Detection Based on Directed Acyclic Graph and Belief Rule Base," the focus shifts to the pivotal role of intrusion detection in network situation awareness. While existing approaches address intrusion detection, they often struggle to leverage expert knowledge and quantitative data optimally. To bridge this gap, the paper introduces the DAG-BRB model, amalgamating a directed acyclic graph (DAG) with a belief rule base (BRB). Enhanced optimization via CMA-ES fine-tunes DAG-BRB model parameters, mitigating constraint issues within the BRB. A comprehensive case study validates the model's efficiency, showcasing superior detection rates over alternative methods and affirming its real-world applicability.

Lastly, "HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks" tackles the ongoing challenge of crafting effective anomaly-based intrusion detection systems (IDS) in cybersecurity. Introducing HAST-IDS, the solution leverages deep neural networks to extract spatial and temporal features from network traffic autonomously. By deploying deep CNNs for low-level spatial feature learning and LSTMs for high-level temporal feature extraction, HAST-IDS achieves remarkable accuracy in intrusion detection. Its automated feature learning mechanism significantly curtails false alarms, a common hindrance in IDS deployment. Rigorous

experimental evaluations corroborate HAST-IDS's superior performance vis-à-vis existing methods, marking a promising stride in IDS technology with heightened accuracy and reliability in identifying network anomalies while minimizing false alarms.

4. PROBLEM STATEMENT

The project addresses the urgent need for threat detection and response in cloud infrastructure-dependent IT organizations. It emphasizes swift and accurate identification of network, application, or asset threats to ensure robust cybersecurity. Key aspects include regular testing, training, and updates to adapt to evolving threat landscapes. Leveraging learning-based methods, particularly machine learning-driven solutions, enhances threat detection capabilities within extensive datasets. This approach offers scalability and efficiency, complementing analyst efforts and bolstering security in cloud-based IT environments.

5. TECHNOLOGY

The project integrates machine learning algorithms, big data analytics, cloud security solutions, continuous monitoring tools, and threat intelligence feeds to enhance threat detection and response in cloud-dependent IT organizations. Decision trees, support vector machines, and deep learning models analyze extensive datasets, while Apache Hadoop and Apache Spark uncover malicious activity patterns. Cloud-native security services and third-party solutions strengthen security, with SIEM systems and IDS providing real-time visibility into security events. Through this integration, the project aims to swiftly identify and mitigate threats in cloud environments, establishing a robust cybersecurity infrastructure.

6. IMPLEMENTATION

Modules

Data Parsing: In this module, the input dataset undergoes parsing, a process of analyzing the dataset's structure and content. This parsing action aims to extract relevant information from the raw dataset and organize it into a structured format, forming a raw data event model. By parsing the dataset, the system prepares the data for further processing and analysis, ensuring that relevant information is accurately captured and utilized in subsequent stages of the cybersecurity system.

TF-IDF: TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical technique used to evaluate the importance of a word within a document relative to a collection of documents. In the context of the cybersecurity project, the TF-IDF module converts the parsed raw data into an event vector. This event vector represents each event in the dataset and includes information about the occurrence of specific terms (words or phrases) within each event. By calculating TF-IDF scores for each term, the module identifies and emphasizes terms that are distinctive to individual events, distinguishing between normal and attack signatures within the dataset.

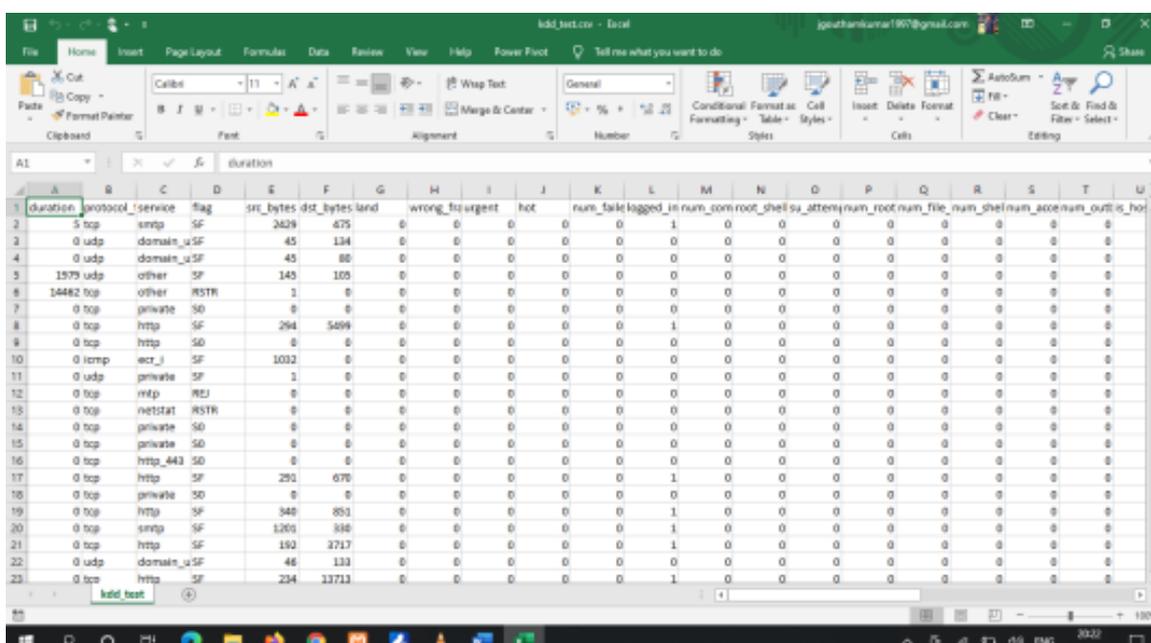
Event Profiling Stage: Following TF-IDF processing, the dataset is further refined through the event profiling stage. During this stage, the processed data is partitioned into separate subsets known as training and testing models. These subsets are created based on distinct profiling events, allowing the system to capture and learn from various patterns and characteristics present within the dataset. The training model serves as a reference dataset used to train machine learning algorithms, while the testing model is utilized to evaluate the performance and effectiveness of the trained models in detecting and classifying attacks.

Deep Learning Neural Network Model: This module applies CNN and LSTM algorithms to both training and testing data, generating a training model. The trained model is then tested to calculate prediction scores, Recall,

Precision, and FMEA scores. Through iterative learning, the algorithm refines its accuracy, ultimately selecting the most effective model for deployment in real systems for attack detection.

7. RESULTS

The dataset was collected from two enterprise systems, ESX-1 and ESX-2, over varying periods. ESX-1 data spans 5 months, while ESX-2 spans 30 days. Security analysts recorded threat information, including occurrence time, attack details, response, IP addresses, and victim network data. ESX-1 recorded 798 cyber threats, comprising 240 scans, 547 system hacks, and 11 worms. ESX-2 recorded 941 scans, 3,077 system hacks, and 51 worms, categorized manually. System hacking includes cross-site scripting, DDoS, brute force, and injection attacks; scanning includes trojan and backdoor attacks. Experimental results indicate the efficacy of the proposed learning-based models for intrusion detection, showing comparable performance to conventional methods on benchmark datasets. However, in real-world scenarios, conventional methods' accuracy degraded, whereas the EP-ANN model maintained performance despite data volume and feature differences. This highlights the robustness of EP-ANN models in real-world intrusion detection scenarios.



duration	protocol	service	flag	src_bytes	dst_bytes	land	wrong_flags	urgent	hot	num_talk	logged_in	num_com	root_shell	su_attempts	num_root	num_file	num_shell	num_acc	num_out	is_hot
5	top	snmp	SF	2429	475	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	udp	domain_uSF		45	134	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	udp	domain_uSF		45	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1579	udp	other	SF	145	109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14482	top	other	RSTR	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	private	SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	http	SF	294	5499	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	top	http	SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	icmp	ecr_i	SF	1032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	udp	private	SF	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	rnp	RZJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	netstat	RSTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	private	SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	private	SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	http_443	SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	http	SF	291	670	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	top	private	SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	http	SF	340	851	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	top	snmp	SF	1201	816	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	top	http	SF	192	3717	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	udp	domain_uSF		46	133	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	top	http	SF	234	13713	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Figure 1. Datasets

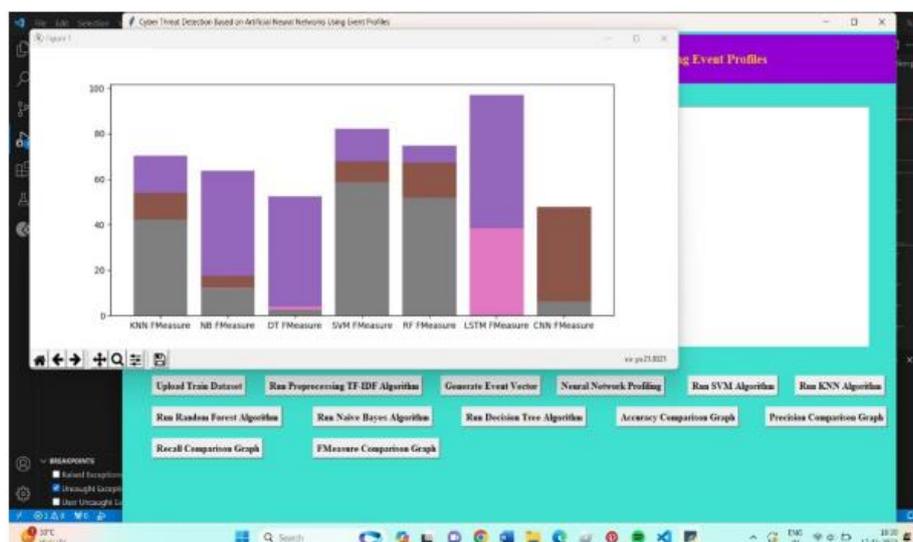


Figure 2. F-measure graph

8. CONCLUSION AND FUTURE ENHANCEMENTS

In this paper, we introduced the AI-SIEM system, leveraging event profiles and artificial neural networks to enhance cyber-threat detection capabilities. Our approach innovatively condenses extensive data into event profiles, employing deep learning methods for superior threat detection. The AI-SIEM system empowers security analysts to efficiently handle significant security alerts by comparing long-term security data, aiding in rapid response to dispersed cyber threats and reducing false positive alerts. Performance evaluation using benchmark datasets (NSLKDD, CICIDS2017) and real-world datasets showcased the efficacy of our approach, surpassing conventional machine learning methods in accurate classification. Future enhancements will focus on refining threat predictions through multiple deep-learning approaches to discern long-term patterns in historical data. Additionally, efforts to improve labeled datasets for supervised learning and the establishment of purpose-built test beds for performance evaluations will be prioritized. Real-world IPS data collection over several months will further augment our system's capabilities in addressing evolving cyber-attack challenges.

9. REFERENCES

- [1] Sheraz Naseer; Yasir Saleem; Shehzad Khalid; Muhammad Khawar Bashir; Jihun Han; Muhammad Munwar Iqbal, "Enhanced Network Anomaly Detection Based on Deep Neural Networks"2018.
- [2] B. Zhang, G. Hu, Z. Zhou, Y. Zhang, P. Qian, L. Chang, "Network Intrusion Detection Based on Directed Acyclic Graph and Belief Rule Base", ETRI Journal, vol. 39, no. 4, pp. 592-604, Aug. 2017
- [3] W. Wang, Y. Sheng, and J. Wang, "HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection," IEEE Access, vol. 6, no. 99, pp. 1792-1806, 2018.
- [4] S. Sandeep Sekaran, K. Kandasamy, "Profiling SIEM tools and correlation engines for security analytics," In Proc. Int. Conf. Wireless Com., Signal Prove. and Net. (Wisp NET), 2017, pp. 717-721.
- [5] Hubbell and V. Surya Narayana False alarm minimization techniques in signature-based intrusion detection systems: A survey," Compute. Common., vol. 49, pp. 1- 17, Aug. 2014.

[6] A. Naser, M. A. Majid, M. F. Zolile and S. Anwar, "Trusting cloud computing for personal files," 2014 International Conference on Information and Communication Technology Convergence (ICTC), Busan, 2014, pp. 488-489.

[7] Y. Shen, E. Marconi, P. Verviers, and Gianluca Stringham, "Tiresias: Predicting Security Events Through Deep Learning," In Proc. ACM CCS 18, Toronto, Canada, 2018, pp. 592-605. 49

[8]Mahmood Lavallee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghobadi, "A detailed analysis of the kid cup 99 data set," In Proc. of the Second IEEE Int. Conf. Comp. Int. for Sec. and Def. App., pp. 53-58,2009.

[9]Kyle Soaks and Nicolas Christin, "Automatically detecting vulnerable websites before they turn malicious," In Proc. USENIX Security Symposium., San Diego, CA, USA, 2014, pp.625-640.

[10]K. Veerama Channid, I. Arnaldo, V. Koraput, C. Basis, K. Li, "AI2: training a big data machine to defend," In Proc. IEEE Bigdata Security HPSC IDS, New York, NY, USA, 2016, pp. 49-54