INTRUSION DETECTION OF IMBALANCED NETWORK TRAFFIC BASED ON MACHINE LEARNING AND DEEP LEARNING

Ms. M. Kowsalya[1], M. E., Ms. A. Meena[2], Ms. D. Mohanapriya[3], Ms. C. Sowmiya[4]
Assistant Professor 1, Student 2,3,4, Department of Computer Science and Engineering,
Erode Sengunthar Engineering College (Autonomous)
Thudupathi, Erode, Tamil Nadu, India.

ABSTRACT

The dynamic issues in cyber security are examined via the lens of intrusion detection, using the ADT-SVM (Adaptive Decision Tree-Support Vector Machine) method. In the context of a fast evolving cyber threat scenario assisted by the Internet, the research investigates the use of Machine Learning (ML) approaches, highlighting the importance of data. The researchers index, study, and analyse publications presenting various ML algorithms, with an emphasis on temporal or thermal correlations, while also highlighting widely used network datasets and the issues connected with ML in cybersecurity. Using the KDD dataset as a benchmark, the project uses the ADT-SVM method to divide data properties into four categories: Basic, Content, Traffic, and Host. Evaluation measures, such as Detection Rate (DR) and False Alarm Rate (FAR), are then used to evaluate the performance of an Intrusion Detection System.

Keywords: Network Security, Cyber Threats, Anomalies , Machine Learning

1. INTRODUCTION

The mitigation and identification of cyber threats have become critical in the ever changing field of network security. Innovative strategies are often required since traditional tactics are unable to keep up with the sophistication of contemporary attackers. Because it can identify patterns and abnormalities in large datasets, machine learning (ML) has become a powerful tool for strengthening network defenses. In order to improve the proactive detection of possible attacks, this introduction investigates the integration of machine learning techniques in the context of network security. Through the utilization of sophisticated algorithms, machine learning (ML) presents the prospect of more flexible, effective, and expandable solutions, ushering in a new phase of increased cyber-attack resistance.

1.1 NETWORK SECURITY

Network security is an essential component of modern digital environments, protecting the availability, confidentiality, and integrity of data transferred between linked systems. Given the widespread reliance on interconnected networks in today's world, it is more important than ever to protect these infrastructures from a wide range of potential dangers. Network security is a complex field that includes the use of intrusion detection systems, encryption protocols, firewalls, and other strong defenses. Malicious actors' techniques also evolve with technology, therefore network security strategies must stay innovative and constantly changing. This introduction explores the core significance of network security and clarifies its function as the first line of defense against a wide range of cyber threats that aim to take advantage of weaknesses in the complex web of interconnected digital ecosystems.

1.2. CYBER THREATS

Cyber dangers are a real concern to the integrity and security of information systems in our digitally connected and interconnected society. Cyber threats comprise a broad range of malevolent actions planned by individuals, collectives, or states with the aim of jeopardizing privacy, causing disruptions, and taking advantage of weaknesses in computer networks. Cyber dangers are constantly changing in terms of sophistication and scope, ranging from advanced hacking methods to social engineering strategies. Our digital infrastructure is interconnected, which increases the potential effect of these risks and makes them a widespread worry for governments, businesses, and individuals. This introduction examines the complexity of cyber threats and highlights the need for all-encompassing cybersecurity solutions to reduce risks and protect the integrity of our data-driven and more interconnected society.

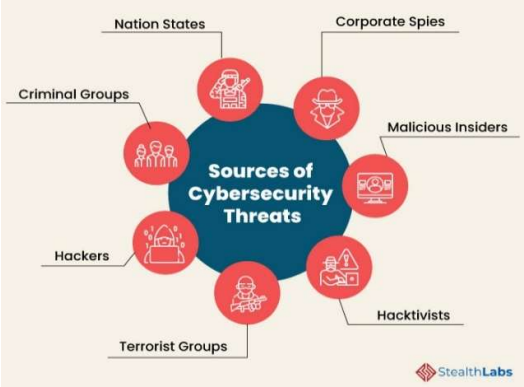


Figure 1.Cyber security threats

1.3 ANOMALIES

Anomalies are variations or abnormalities from the expected or usual patterns in a variety of disciplines, including data analysis, system monitoring, and network security. These variations may point to underlying problems, possible

hazards, or areas in which more research may be needed. Anomalies can indicate anything from mistakes in data collection to new and previously unknown patterns, making them important signals that require attention. Finding and interpreting abnormalities is critical in a variety of domains, from spotting possible security breaches in network data to detecting irregularities in financial transactions. The relevance of anomalies as departures from the norm is examined in this introduction, with a focus on how they can reveal hidden patterns, possible dangers, and areas that need more research in a variety of analytical and surveillance domains.

1.4 MACHINE LEARNING

In this era of digitalization, we are constantly surrounded by an overwhelming amount of data. Whether it's the online activities we engage in, the sensors embedded in our smartphones, or the extensive databases that support modern businesses, data has become the life force of the information age. However, amidst this data-driven revolution, the true power lies in our ability to convert raw data into actionable knowledge, paving the way for extraordinary technological advancements. At the core of this transformative process lies the field of "Machine Learning." The driving force behind machine learning is our inherent need to comprehend the intricate and expansive data ecosystems that shape our contemporary world.

OBJECTIVE

- Detect network intrusions with high accuracy. This means that the IDS system should be able to identify both known and unknown attacks with a low false positive rate.
- Reduce over fitting and also improve adaptability and flexibility.

2. LITERATURE REVIEW

2.1 The Evolution Of Ethernet Passive Optical Network (EPON) and Future Trends

Felix Obiteet.al. has presented in this research paper that the significant growth in Internet traffic confirms the shift of the telecommunications backbone from time division multiplexing (TDM) to a focus on Ethernet solutions. Ethernet PON, which combines low-cost Ethernet and fiber infrastructures, has emerged as the dominant technology in a market previously dominated by DSL and cable modems. This new technology is characterized by its simplicity, affordability, and scalability, enabling the delivery of massive data services to end-users over a single network. The paper provides an overview of the evolution of Ethernet Passive Optical Network (EPON), with a particular emphasis on the ongoing

development of future high-data-rate access networks such as Next-Generation Passive Optical Network Stage 2 (NG-PON2), Wavelength Division Multiplexing (WDM) PON, and Orthogonal Frequency Division Multiplexing (OFDM) PON. Additionally, the recently concluded 100 Gb Ethernet Passive Optical Network (100G-EPON) is reviewed to highlight the latest advancements in the field. This comprehensive and up-to-date review aims to equip network operators and interested practitioners with a clear understanding of common priorities and timelines. Furthermore, the study aims to identify technical solutions for future investigation. The exponential increase in data traffic and the growing number of online users, who spend more time online and engage in bandwidth-intensive applications, necessitate broadband services that can support high-speed internet transmission. Therefore, future access networks must possess large bandwidth capacity and mobility to accommodate new and real-time broadband applications. DSL and cable modems are inadequate to meet such demands.

2.2 Revisiting Wireless Internet Connectivity: 5G VS Wi-fi 6

In recent times, there has been a notable focus on the fifth generation of wireless broadband connectivity, commonly referred to as '5G', which is currently being implemented by Mobile Network Operators. However, the attention given to 'Wi-Fi 6', the new IEEE 802.1ax standard within the Wireless Local Area Network technology family, specifically designed for private, edge-networks, has been surprisingly limited. This article reevaluates the effectiveness of cellular and Wi-Fi technologies in providing high-speed wireless internet connectivity. Both technologies aim to offer significantly improved performance, enabling faster wireless broadband connectivity and supporting the Internet of Things and Machine-to-Machine communications. Consequently, these two technologies can be seen as technical alternatives in various usage scenarios. Our conclusion is that both will play crucial roles in the future, simultaneously acting as competitors and complements.

2.3. Subjective Of 360-Degree Virtual Reality Videos and Machine Learning Predictions intrusion Detection Systems In The Internet of Things: A Comprehensive Investigation

Recently, Somayye Hajiheidariet.al. proposed a system that introduces a new aspect of intelligent objects by reducing the power consumption of electrical appliances. This advancement allows everyday physical objects to be

enhanced with electronic devices, enabling them to connect to the internet and possess local intelligence. This concept is referred to as the Internet of Things (IoT), which encompasses these intelligent objects. However, due to their direct connection to the internet, these objects are susceptible to attacks from malicious individuals. The accessibility of resource-constrained devices through public internet access exposes them to potential intrusions. These intrusions, known as internal attacks, do not explicitly damage the network but infect internal nodes to carry out attacks on the network. Therefore, the implementation of Intrusion Detection Systems (IDSs) in the IoT is crucial. Despite the significance of this topic, there is currently no comprehensive and systematic review that discusses and analyzes the mechanisms of IDSs in the IoT environment. Hence, this paper presents a Systematic Literature Review (SLR) of IDSs in the IoT environment.

2.4. Ensemble Learning For Intrusion Detection Systems: A Systematic mapping study and Cross Benchmark Evaluation

Bayu Adhi Tamaet.al. has proposed a system that emphasizes the importance of Intrusion Detection Systems (IDSs) in preventing cyberattacks. In order to enhance the detection rate, there is a need to develop an improved detection framework, especially when utilizing ensemble learners. The process of designing an ensemble faces two main challenges: selecting appropriate base classifiers and combiner methods. This research paper provides an overview of how ensemble learners are utilized in IDSs through a systematic mapping study. We have gathered and analyzed 124 significant publications from the existing literature. These publications have been categorized based on the year of publication, publication venues, datasets used, ensemble methods, and IDS techniques. Additionally, this study presents an empirical investigation of a novel classifier ensemble approach called "stack of ensemble" (SoE) for anomaly-based IDS. The SoE is an ensemble classifier that employs a parallel architecture to combine three individual ensemble learners, namely random forest, gradient boosting machine, and extreme gradient boosting machine, in a homogeneous manner. The performance of different classification algorithms is statistically evaluated using metrics such as Matthews correlation coefficients, accuracies, false positive rates, and area under the ROC curve.

2.5. Deep Abstraction and Weighted Feature Selection For Wi-fi Impersonation Detection

Muhamad Erza Amina et al. have presented a system that addresses the security challenges posed by the widespread use of IoT-enabled devices in our daily lives, thanks to recent advancements in mobile technologies. The main concern lies in the vulnerability of wireless mediums like Wi-Fi networks, which are open in nature. An impersonation attack occurs when an adversary disguises themselves as a legitimate party within a system or communication protocol. The abundance of connected devices generates a vast amount of high-dimensional data, making simultaneous detections complex. However, feature learning can mitigate potential issues arising from the large volume of network data. In this study, a novel approach called Deep-Feature Extraction and Selection (D-FES) is proposed, which combines stacked feature extraction and weighted feature selection. By reconstructing relevant information from raw inputs, stacked autoencoding enhances the meaningfulness of representations.

3. EXISTING SYSTEM

Advancements in multimedia technologies have raised concerns regarding the security of digital data, leading researchers to explore modifications to existing security protocols. However, numerous encryption algorithms proposed in the past few decades have been found to be insecure, posing significant risks to critical data. The selection of an appropriate encryption algorithm is crucial for protecting data, but individually evaluating each algorithm can be time-consuming. To tackle this issue, we propose an approach for detecting the security level of image encryption algorithms using a support vector machine (SVM). Furthermore, we have developed a dataset that incorporates standard encryption security parameters, such as entropy, contrast, homogeneity, peak signal-to-noise ratio, mean square error, energy, and correlation, extracted from various cipher images. These parameters serve as features, and the dataset is categorized into three security levels: strong, acceptable, and weak.

3.1 DISADVANTAGES

- The models are generally computationally expensive to train and deploy. This may be a limitation for organizations with limited resources.
- This model require large amounts pf training data to achieve good

performance. This may be a limitation for organizations that do not have access to large datasets of network traffic data.

- Deep learning models are often considered to be black boxes, meaning that it can be difficult to understand how they make predictions. This can make it difficult to debug the model and identify false positives.

4. PROPOSED SYSTEM

The proposed system for Content-Based Image Retrieval (CBIR) incorporates Reversible Data Hiding (RDH) using the Triple DES algorithm to encode and categorize visual image attributes such as color, shape, texture, and spatial arrangement. Ongoing CBIR research is focused on enhancing methodologies for analyzing, interpreting, organizing, and indexing image databases, while also evaluating the performance of retrieval systems. This project introduces an innovative steganography approach through reversible texture synthesis, where small texture images, whether artistically designed or photographically captured, are resampled to generate new textures of different sizes. This process utilizes a patch-based technique to embed secret messages, ensuring that the original texture can be recovered during message extraction.

4.2 ADVANTAGES

- This system is used to detect network intrusions with high accuracy. This means that the IDS system should be able to identify both known and unknown attacks with a low false positive rate.
- Reduce over fitting. Over fitting occurs when the IDS system learns the training data to well, which can lead to poor performance on new data. The MRF algorithms uses bagging and random feature selection to reduce over fitting.
- Improve adaptability and flexibility. The proposed system automatically selects the studied parameter values according to the used training dataset, which makes the system more adaptable to different types of network traffic and intrusions patterns.

5.MODULES

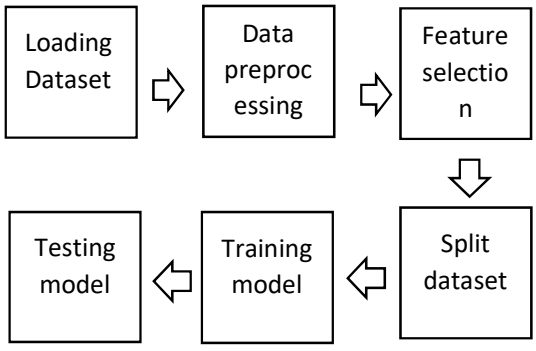
- Probability Model
- Calculating the link-Anomaly score
- Change point Analysis and DTO
- ADT-SVM Detection Method

5.1 PROBABILITY MODEL

This lesson focuses on the creation and use of a probability model for interpreting network data. The probability model most likely evaluates the possibility of specific events or patterns in the data, offering a basic knowledge of the baseline behavior. By creating a probability distribution, anomalies may be found by deviating from predicted patterns, allowing the system to detect possibly malicious activity.

5.2 CALCULATING THE LINK-ANOMOLY SCORE

In this module, the system computes link anomaly scores to measure the irregularity of network links or connections. The calculation entails examining several properties related with network connections, such as traffic patterns, communication frequencies, and data transfer volumes. A higher link-anomaly score may suggest suspicious or anomalous behavior , alerting the intrusion detection system to possible security concerns within the network.



5.3 CHANGE POINT ANALYSIS AND DTO

This subject covers change point analysis and Dynamic Time Warping (DTO) methodologies. Change point analysis seeks to uncover changes or variations in the statistical features of data, which may indicate possible security events. DTO, on the other hand, includes assessing sequence similarity across time to help in the discovery of temporal patterns. Integrating these strategies improves the system's capacity to adapt to changing cyber threats and detect abnormal activity.

5.4 ADT-SVM DETECTION METHOD

The ADT-SVM Detection Method module applies the Adaptive Decision Tree-Support Vector Machine (ADT-SVM) algorithm to intrusion detection. This technique combines the flexibility of decision trees with the classification capability of support vector machines. The ADT-SVM model is trained using labeled data to discriminate between normal and abnormal network behavior. Once trained, it is used to classify incoming data properties into preset categories such as Basic, Content, Traffic, and Host, making it easier to identify possible security concerns on the network. The module will most likely include fine-tuning and improving

the ADT-SVM settings to achieve optimal detection performance.

6. RESULT ANALYSIS

Strong results are obtained from the suggested intrusion detection system's result analysis. The application of three different machine learning models—Random Forest, ADT-SVM, and Linear Regression—shows strong predictive powers. Linear regression, Random Forest, and ADT-SVM have accuracy values of 84%, 80%, and 85%, respectively, highlighting how well the system detects and categorizes network intrusions. The ADT-SVM component's exceptionally high accuracy score attests to its remarkable ability to identify and mitigate harmful actions within computer networks. The study of the results highlights the strategic importance and dependability of the suggested system, confirming its potential as an effective instrument for bolstering cybersecurity in the face of dynamic threats to network integrity.

Algorithm	Accuracy	Precision	Recall	F1 score
Rf	0.8	0.7	0.97	0.81
LR	0.84	0.76	0.92	0.83
ADT-SVM	0.85	0.75	0.96	0.84

Table 1.Table comparison

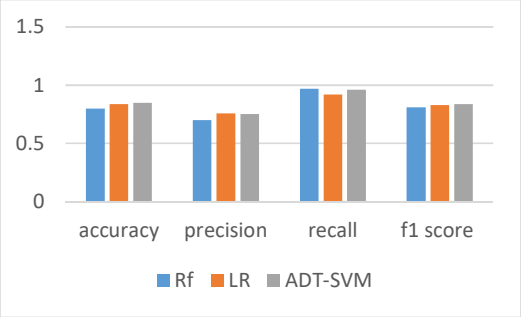


Figure 2. Comparison Graph

7. CONCLUSION

In conclusion, A network intrusion detection system (IDS) using a modified random forest (MRF) algorithm is a promising approach for detecting network intrusions with high accuracy, reduced over fitting, improved adaptability and flexibility, and robustness against new and emerging threats. MRF-based IDS systems are relatively easy to implement and train, and they are scalable, meaning that they can be used to monitor large networks. MRF-based IDS systems can be used to detect a wide range of network attacks, including denial-of-service attacks, port scanning attacks, and malware attacks. However, it is important to note that no IDS system is perfect. MRF-based IDS systems are just as vulnerable to evasion techniques as other IDS systems. Additionally, MRF-based IDS systems can be computationally expensive to train and operate.

8. FUTURE WORK

MRF classifiers are known for their high accuracy in classification tasks, but there is still room for improvement. Future work could focus on developing new MRF algorithms that are even more accurate and efficient. Attackers are constantly developing new techniques to evade IDS systems. Future work could focus on developing MRF classifiers that are more robust against these evasion techniques. It can be computationally expensive to train and operate. Future work could focus on developing new training algorithms and optimization techniques that can reduce the computational cost of MRF classifiers.

9. REFERENCES

1. S. B. Atitallah, M. Driss, W. Boulila, and H. B. Ghézala conducted a survey on utilizing deep learning and IoT big data analysis to support the development of smart cities. Their research was published in the journal "Computer Sci. Fire up." in November 2020, under the article number 100303.

2. M. Al-Sarem, W. Boulila , M. Al-Harby, J. Qadir, and A. Alsaeedi conducted a systematic survey on deep learning-based rumor detection on microblogging platforms. Their findings were published in the journal "IEEE Access" in 2019, with the volume number 7 and pages 152788-152812.

3. M. S. Anwar, J. Wang, W. Khan, A. Ullah, S. Ahmad, and Z. Fei investigated the subjective quality of experience of 360-degree augmented reality videos and AI predictions. Their research was published in the journal "IEEE Access" in 2020, with the volume number 8 and pages 148084-148099.

4. T. B. Dijkhuis, F. J. Blaauw, M. W. van Ittersum, H. Velthuijsen, and M. Aiello proposed an AI approach for personalized physical work coaching. Their work was published in the journal "Sensors" in February 2021, with the volume number 18, issue number 2, and page number 623.

5. A. Roy, A. P. Misra, and S. Banerjee developed a chaos-based image encryption technique using vertical-cavity surface-emitting lasers. Their research was published in the journal "Optik" in January 2022, with the volume number 176 and pages 119-131.

6.M.Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in 2009 IEEE symposium on computational intelligence for security and defense applications. IEEE, 2021, pp. 1–6.

7. I. Shara faldin , A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new

intrusion detection dataset and intrusion traffic characterization." In ICISSP, 2020, pp. 108–116.

8.L.v.d. Maaten and G. Hinton, "Visualizing data using t-sne," Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2021.

[41] X. Ma and W. Shi, "Aesmote: Adversarial reinforcement learning with smote for anomaly detection," IEEE Transactions on Network Science and Engineering, 2020.

9.P. Bedi, N. Gupta, and V. Jindal, "I-siamids: an improved siam-ids for handling class imbalance in network-based intrusion detection systems," Applied Intelligence, pp. 1–19, 2020.

10. G. Caminero, M. Lopez-Martin, and B. Carro, "Adversarial environment reinforcement learning algorithm for intrusion detection," Computer Net-works, vol. 159, pp. 96–109, 2022.

11. A. K. Verma, P. Kaushik, and G. Shrivastava, "A network intrusion detection approach using variant of convolution neural network," in 2022 International Conference on Communication and Electronics Systems(ICCES), 2022, pp. 409–416.

12. J.-T. Wang and C.-H. Wang, "High performance wgan-gp based multiple-category network anomaly classification system," in 2019 InternationalConference on Cyber Security for Emerging Technologies (CSET). IEEE,2020, pp. 1–7.

13. M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Con-ditional variational autoencoder for prediction and feature recovery ap-plied to intrusion detection in iot," Sensors, vol. 17, no. 9, p. 1967, 2021

14. A. Liu, J. Ghosh, and C.E. Martin, "Generative oversampling for mining imbalanced datasets." In DMIN,2021,pp. 66-72.

15. Dina AS, Siddique AB, Manivannan D (2022) Effect of balancing data using synthetic data on the performance of machine learning classifiers for intrusion detection in computer networks. CoRR arXiv: abs/2204.00144https://doi.org/10.48550/arXiv.2204.00144