Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Artificial Intelligence-Powered Cyber-Attacks Defender System

Jayapradha J[1], Haw Su Chaug[2], Priyadharshini K[1], Vathana D[1], Manneredly Monesh[1], Anupam Palai[1]

[1]*Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, 603203, Tamil Nadu, India.*
[2]*Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100, Cyberjaya, Selangor.*
Email: jayaparaj@srmist.edu.in

Security of such computer systems and networks is imperative in today's world that is run at the speed of light. Accordingly, threat actions of cybernetic attacks and attempts to penetrate the Internet include much wider due to the extensive usage of Internet access and raise severe threats to data and privacy. A concept that has emerged in the recent past as an approach to ease the challenges is the AI-Powered Cyber Security Defender Systems. This research study aims to understand this intricate environment, learn about the emergence, taxonomy, practices, challenges, and future of these complex systems. The study is continued by outlining the directions or milestones that have defined AI-POWERED CYBER SECURITY DEFENDER SYSTEMS through their development stages. This paper provides a comprehensive taxonomy of such systems and helps in understanding the modus operandi of these systems in addition to the advantages it carries. They could quickly analyze the massive data traffic and identify and respond to the malicious activities within networks with the help of AI and machine learning thus enhancing the security rate substantially. In conclusion, the research paper emphasizes the importance of AI-Powered Cyber Security Defender Systems in the context of physical as well as cyber security in the modern world. Delving into the details of such systems, their growth, methodologies, challenges and future, this study will prove particularly useful for cybersecurity practitioners, scholars, and policy makers to build a proactive standpoint towards the protection of digital environment.

Keywords: Machine Learning, Artificial Intelligence, Cyber Security, Intrusion, Classification, Defender Systems.

1. Introduction

Security of the computer systems and networks is becoming crucial, especially with the advancements in technology that have led to the interaction of various computer systems and networks. While people rely on technology for one service or another, it is increasingly becoming a great concern to find that hackers and unauthorized access to personal information are real threats. Intrusion Detection Systems (IDS) therefore have a critical role in guarding computer networks and systems against these threats. IDS are advanced program and technology systems designed to observe, recognize and counter security breaches or state policy contraventions in a computer network or system. On this basis, IDS becomes an important element ensuring against violation of informational security and contributes to the preservation of the confidentiality, integrity, and availability of the information [1].

There are two primary types of IDS: NIDS which is short for Network based Intrusion Defender Systems and HIDS which stands for Host based Intrusion Defender Systems. NIDS works by observing and filtering live network traffic to detect any suspicious patterns and Operation while HIDS are placed at the host or device level, focusing on activities relative to the system they are protecting [2]. IDS can use different methods for detecting intrusions such as signature-based IDS, anomaly-based IDS, and heuristic-based IDS. This type of detection is based on alerting, where observed patterns of data in network traffic or system behavior are matched against pre-existing attack signatures [3]. Anomaly-based detection focuses on the abnormal activities and finds them by looking at how they differ from the normal behavior [4]. Heuristic-based detection entails rules or algorithms formulated to detect risks given their attributes [5].

Like all security software solutions, IDS face certain problems because of the ever-changing nature of threats. The offending parties develop new schemes for evading standards IDS and preventative procedures, so IDS has to change not only its techniques – but its concept. Further, IDS relies on various machine learning and artificial intelligence techniques to improve its performance in detecting sophisticated and previously unidentified attacks [6]. It is crucial to understand that IDS plays a great function in the course of attempting to identify and prevent any sort of security breaches beforehand. IDS prevent unauthorized access, control and monitor security breaches and play a great role in the achievement of confidentiality, integrity, and availability of information systems or data as part of the overall digital security of both individuals and organizations [7]. In today’s ever evolving world, organizations experience ever advancing and complex cyber threats, Intrusion Detection systems cannot be overemphasized. The current improvement and research in IDS confirms that IDS is gradually adjusting to the new threats arising due to changes in cyberspace. With these improvement going on continually, IDS assist organizations in combating various cyber-crises, and in the process improving the organizations’ cybersecurity [8].

2. Literature Review

Intrusion detection systems that are developed to identify advanced persistent threats (APTs) only. The current installed intrusion detection systems were assessed in terms of their ability to detect these continuous threats as the researchers explored several approaches used in APTs. Consequently, the work intends to advance the capabilities of IDS to counter these carefully planned and sustained cyber attacks since the existing IDS lacks certain characteristics when it comes to thwarting APTs [9]. In this comprehensive review focus on the machine learning techniques applied to IDS is comprehensively reviewed here. Applying algorithms like decision trees, support vector machines, and neural networks to recognize different cyberattacks gives a sense of how such intelligent systems can enhance the detection of cyber threats and help in the earliest possible detection and prevention of various cyber risks by comparing the effectiveness of the applied machine learning systems [10]. A closer look into the specialty subfield of machine learning strategies used in Intrusion Detection Systems. Cyberattacks were identified using algorithms including decision trees, and support vector machines as well as neural networks. Enlightening on the kind of advantages and disadvantages that the various machine learning models possess in reference to the identified study, it is possible to understand how the systems can improve on the accuracy of intrusion detection, thereby eradicating different types of cyber threats in the shortest time [11]. Intrusion detection systems for Cyber-Physical Systems (CPS). Deep learning methods were employed particularly deep neural networks. deep learning models may be applied to CPS data to find complex patterns, enhancing the system's capability to quickly recognize sophisticated cyber-physical threats. The proposed IDS was developed with the aim of enhancing the protection of vital infrastructures against evolving cyber threats. A tool has been developed to detect cyber assaults in cyber-physical approaches. Machine Learning methods have been implemented to improve the security of the system. The model is used to classify the anomaly behavior that is related to normal behavior. The study has implemented it in identification of overflow of the water in water tank system and proved to be effective in its results [12].

The study has paid attention to safeguarding the large IT environment. Nowadays, securing IT environment is very challenging as intruders are increasing in their numbers. Due to stealing of various important knowledgeable properties, the reputation of the company is destroyed leading to the withdrawal of their businesses. The study has protected the LAN-WAN Domain of large IT industry using a appropriate tool [13]. Various assumptions about the environment will lead to security violations and it also affect the dependability of the system. The study dealt with input related problems for web applications as the assumed inputs lead to various threats. The model has been built such that the SQL queries will be built dynamically according to the user inputs. A fine-tuned algorithm has been implemented to check the dynamic automation and to verify the security leak [14]. In everyday life, due to digital evolution, data is getting generated daily. Due to this evolution, transmission of data is very risky and thus network security came into picture. There are various tools that deal with a lot of security frameworks. The tools find the liabilities in the websites or web site in order to secure the internet connected networks. The tools dealt in this paper are ZAP, WEP, WPA PSK etc[15]. Due to the increase in invaders, the unauthorized access of the system is increased and thus the security for the information system is required much. To address the above problem, a security development software system using software agents have been implemented. Two agents has been used 1. One in the server side and 2 one in the user side. By transferring a template of standard user behavior and regulations for unacceptable conduct from a central agent to each individual user agent, the user agent may independently make judgments and take actions in response to unusual or inappropriate user behavior[16]. A denial of service is made by an attacker to prevent the authorized users from using his own resource. The attacker drowns the network to reduce the user’s bandwidth system. The study has modelled a distributed denial of service using ns-2 system simulator. Various queuing algorithms is implemented in the network for cyberattacks. The study also concluded that constant denial of service attacks has promised good bandwidth [17]. As the threat keeps on increasing, there are no sufficient security systems. The end-end network system is implemented to protect the systems against threats. The end-end network system acts like a human immune system that vigorously changes its systems [18]. A new MC-GRU WSN intrusion detection system has been proposed to eradicate the low detection accuracy and poor real-time detection in existing WSN intrusion detection algorithms. The CNN method has been used to obtain the traffic data. The system has been proven to achieve higher accuracy [19].

3. Case and Methodology

There is a critical need to develop advanced systems that leverage Artificial Intelligence (AI) to enhance cybersecurity capabilities. Therefore, the primary objective of this research is to develop an AI-driven cybersecurity system thatnot only detects and mitigates cyberattacks effectively but also adapts to evolving attack techniques and tactics to remain resilient over time.

Data Collection

The "NSL-KDD Cup 1999" dataset, also known as the NSL-KDD dataset, was used in this study. The original KDD Cup 1999 dataset, which was extensively used to rate intrusion detection systems, was updated to create the current dataset. The dataset used in the proposed study is the NSL-KDD and it is suitable for use as it contains a large set of records on network traffic under both normal and attack conditions. Data: Different 'types' of assault are included; DoS (Denial of Service), Probe, R2L (Unauthorized access from a remote system), U2R (Unauthorized access to local superuser privileges), and normal traffic. It is also variety that helps to examine the methods of detection of intrusion depending on the different threats in the sphere of cyber-security. The actual number of records that this dataset holds is one hundred and twenty-five thousand, nine hundred and seventy-three. Some of the features that are descriptive of a number of network factors such as protocol type, service, flag, duration, source and destination IP addresses, among others are presented in the set of features of the dataset.

Preprocessing

Since data analysis requires high quality data, all the data was preprocessed before being analyzed to ensure that it fits the study. Over-sampling, clean-up of the data, handling of missing values and transformation of feature to a form that will be easily processed by any machine learning algorithm were some of the Preprocessors. For the analysis to be relevant and accurate, the dataset must remain intact and without alterations. Some of the categorized features in this data set are; Protocol type, service, and flag derived from the NSL-KDD dataset. It is noteworthy that the attributes were transformed into numerical features by certain techniques such as one hot encoding to fit within machine learning system. This transformation maintained the categorical data in a format where form calculations and analysis could be made. Normalization and standardization were applied to ensure that the scales of the features measured were in the same order. The data was preprocessed with respect to its noise reduction, feature scaling and achieving data balance for the use in machine learning algorithms. What we obtained in the data pre-processing stage was the pre-processed dataset and that formed the foundation on which the subsequent analysis was carried out to get the accurate and credible prediction models.

Feature Selection

The process of feature selection involved looking at which characteristics of the NSL-KDD dataset were relevant and which characteristics were redundant till the present level of accessibility. Relevance analysis within the field of network anomaly detection aimed at identifying characteristics which would help categorize it more easily, in terms of the amount of traffic in the network and other types of attacks. To avoid risk of getting Multicollinearity problems in the Machine learning models, Redundancy analysis was centered at identifying and eliminating features which contain related information. The present study employed a correlation technique in establishing relationships between different factors. Features that are highly correlated may contain redundant information than the models make use of and can generate noise. Efficiency was applied in assessing features with high degrees of correlation and the necessary actions taken to retain only the essential features during feature selection while eliminating duplicates. The importance of each feature was below calculated using machine learning classifiers including gradient boosting machines, decision trees, and random forest classifiers. They were adopted in estimating Intrinsic Feature Importance ratings as earlier mentioned. The functions deemed more important were prioritized accordingly, proving their great importance to the categorization process. After scrutinizing using the aforementioned techniques, one last set of characteristics was selected for training of the intrusion detection models as shown below. This selection was considered since it has been proven to be capable of differentiating between actual network activity and the one originating from an attacker. These features provided the basis upon which the research completed training, testing, and assessment for the work.

Model Selection and Training

Depending on the types of attacks present in the chosen NSL-KDD dataset, it is essential to select proper models in this study. For this purpose, an ensemble approach using Decision Trees and Random Forests was used to differentiate between Denial-of-Service (DoS) and Probe attacks, while a Gradient Boosting Classifier was used for classifying Remote-to-Local (R2L) as well as User-to-Root (U2R) intrusions. The reasons behind selecting these models are as follows.

Justification for Model Selection:

Accuracy and Robustness

For DoS and Probe attacks detection, an ensemble approach that combines Decision Trees with Random Forest is applied because it has high accuracy and robustness. Random Forest is able to provide a stable and accurate prediction model as it combines many Decision Trees thus enabling good classification even when data points may be noisy or ambiguous.

Sensitivity to Intrusion Patterns

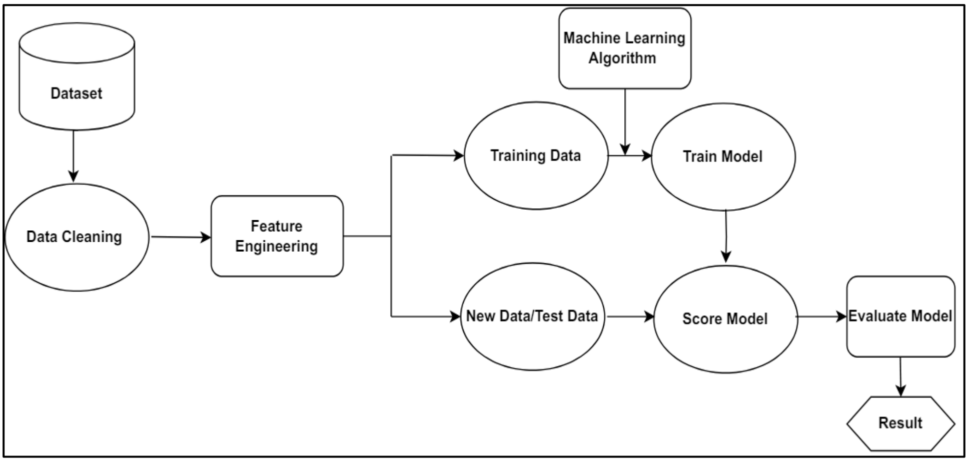
Gradient Boosting Classifier on the other hand is chosen for R2L and U2R intrusion identification mainly because the model is capable of identifying patterns that may not be easily observable within the dataset. It is not easy to detect R2L and U2R intrusions since they employ sophisticated techniques. This sequential learning process that characterizes Gradient Boosting allows it to capture intricate relationships better than other methods; hence it helps in recognizing these delicate patterns associated with such malicious activities.

Handling Class Imbalance

Ensemble methods adapt well to class imbalance by dealing with the problem implicitly. Class imbalances are common in intrusion detection tasks, where certain intrusions occur much rarely than normal activities. The ensemble models used in this study are by their nature immune to class imbalance issues because all intrusion types are equally represented.

The choice of the models for classification is based on the individual advantages and synergistic performance of the chosen models: the combination of Decision Trees and Random Forest for the detection of DoS and Probe attacks, Gradient Boosting Classifier for the identification of R2L and U2R attacks. This combined design ensures high accuracy along with substantial and sensitive intrusion detection. This makes it ideal to handle complex NSL-KDD data set and provides an effective way to achieve reliable results that will help in determination of various types of intrusion.

Figure 1. Architecture Diagram of Defender System



4. Results

Scoring an average of 99% with decision tree, random forest and using the gradient boosting classifier as the last line of defence, the intrusion defence system demonstrated remarkable performance against multiple types of attack and normal traffic. The results of the performed experiments were assessed by means of a number of indexes widely employed for intrusion detection problem. These metrics offered a satisfactory level of evaluation of the model and its capability in terms of intrusion detection and differentiation of normal network activities. Lense used accuracy, precision, recall, and F1-score as its performance criteria.

Utilizing an ensemble approach combining (decision tree, random forest) and gradient boosting classifier, the intrusion defence system showed outstanding accuracy against a variety of attack types and regular traffic. The models' performance was evaluated using a set of well-established metrics tailored for intrusion detection tasks. These metrics provided a comprehensive understanding of the model's effectiveness in distinguishing between normal network activities and various intrusion categories. The key metrics considered included accuracy, precision, recall, and F1-score. Table 1 depicts the various attacks.

Table 1. Result Metrics of Each Attack Type

AttackType	ModelUsed	Accuracy	Precision	Recall	F-measure
Denial ofService	Ensemble model (Decision Tree + Random Forest)	0.996	0.994	0.997	0.995
Probe	Ensemble model (Decision Tree + Random Forest)	0.995	0.993	0.992	0.993
Remote toLocal (R2L)	GradientBoosting Classifier	0.981	0.972	0.973	0.973
User toRoot	Gradient BoostingClassifier	0.997	0.933	0.919	0.923

Figure 2.1

DoS Attack

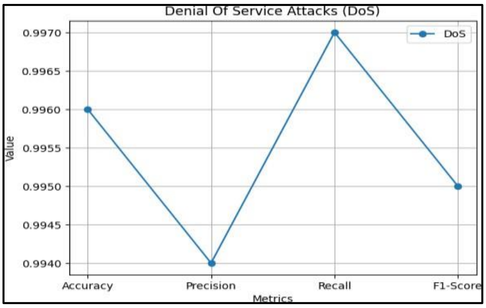


Figure 2.2

Probe Attack

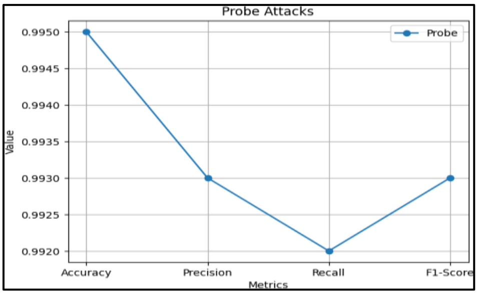
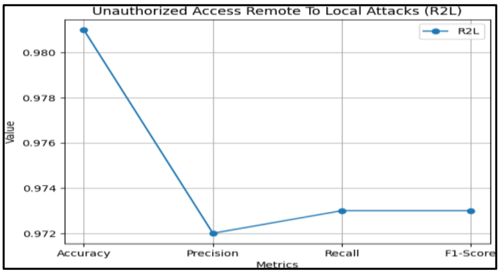
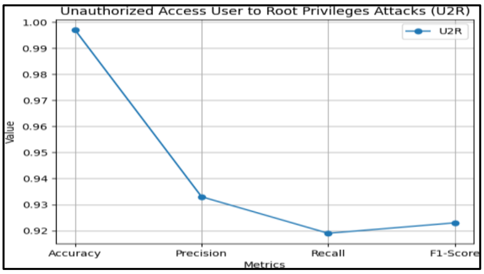


Figure 2.3



R2L Attack

Figure 2.4



U2R Attack

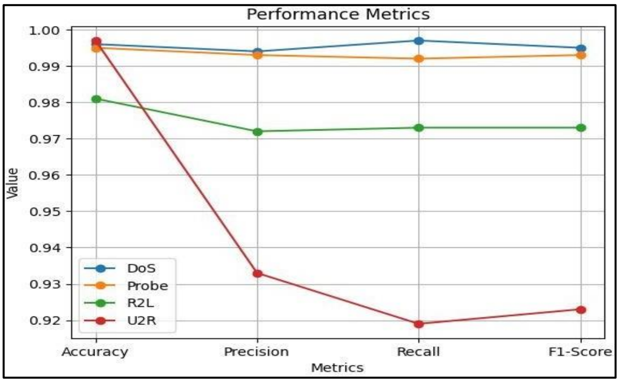
'DoS' (Denial of Service): the ‘ Dos’ category maintains a fairly accurate account throughout the percentages ranging from 0.995 to 0.997 in all experiments as shown in Fig2.1. This suggests that the model is able toaccurately classify 'DoS' attacks across various situations. Moreover, even ‘DoS’ has successfully retained the mark of high precision with rates starting from 0. 993 to 0. 994, which means that when the model prediction is typically correct, with very few false positives. The recall for 'DoS' is close to 0.997,indicating that the model successfully learns most of ‘DoS’ cases belonging to the dataset and also minimizes the number of false negatives. This is further evident by the cross-validation of the F1-score which has been relatively constant and approximately 0. 995, thereby balancing precision and recall ‘ DoS’ methods effectively.

'Probe': Research pursued under the ‘Probe’ category shows relatively a slightly different accuracy, starting from 0. 981 to 0. 995, as shown in Figure 2.2. This implies that the model performance, in classifying the ‘Probe’ type of attacks, might vary depending on the kind of experimental setting. Moreover, the degree of precision based on the ‘Probe’ is also moderate with the values slightly fluctuating from 0.972 and 0.993, suggesting that there ais a higher rate of false positives during 'Probe' attack prediction, indicating room for improvement. Similarly, the recall for 'Probe' is also moderate, fluctuating between 0.972 and0.992, indicating that the model identifies the majority of ‘Probe’ attacks while possibly overlooking others. This is demonstrated by a moderate F1 value, varying between 0. 972 to 0.993.

'R2L' (Unauthorized Access to Local System): The ‘R2L’ type, on the contrary, remains moderately steady in terms of its accuracy, oscillating between 0.972 and 0.973, with minimal variation across experiments as shownin Figure 2.3. Classification of 'R2L' attacks in various scenarios is expected due to the stability, suggesting a consistent performance. For ‘R2L’, the accuracy level is still reasonable, which is approximately 0.933. This suggests that for each ‘R2L’ predicted by the model, there are comparatively fewer false positives. However, the recall for 'R2L' has been found to be 0.919, which indicates that the model may misssome instances while trying to capture a portion of 'R2L' attacks. F1-score of 0.923 for 'R2L represents the moderation between precision and recall.

'U2R' (Unauthorized Access to Root): Lastly, ‘U2R’ category demonstrates high levels of accuracy, approximately 0. 997, this going on to prove the model effectiveness especially in identifying the ‘U2R’ attacks as shown in Figure 2.4. On the other hand, the precision of identifying the ‘U2R’ class remains low and is estimated at 0. 933 hinting that there are more false positives when the model developed is predicting ‘U2R’ type of attacks. The recall for 'U2R' is around 0.919, indicating that the model may leave out some cases while capturing some 'U2R' attacks. Similar to ‘R2L’ accuracy, ‘U2R’ also has an average value of F1-score of 0. 923, a reasonable measure of accuracy that maintains the stability of recall rates. Such observations help in understanding the efficacy of the proposed model in classification of distinct attack categories and assists in using the evaluation results for improvement of the model.

Figure 2 Performance Metrics for Each Attack Type



By comparing the recognition performance and the different attack categories, the following insights are obtained. The 'DoS' category achieves the highest accuracy, precision, recall, and F1- score, that reveals that the method underpins high interpretability in categorizing Denial of Service attacks. For the 'Probe' category, the accuracy and precision level are moderately high and unstable at the same time across different experiments that points to the fact that existence of most ‘Probe’ attacks identified by the model depends on certain experimental conditions. The general performances of ‘R2L’ show a relatively stable accuracy, high precision, and a moderate rate of recalls which indicates continuous capability of effectively classifying the Unauthorized Access to Local System attacks. On the other hand, ‘U2R’ maintains high accuracy while it tends to reduce the precision and that can be interpreted as classifier trying to overestimate the presence of Unauthorized Access to Root attacks cases. As for ‘U2R’ the recall presents a moderate value in the same way the F1-score for ‘U2R’ is similar to that of ‘R2L,’ but demonstrating the balance between the means of accurate classification and the recall of all samples. These comparatively acquired performance views make it easy to assess how well or ill the model performs depending on the specific attack class and even to check out its effectiveness generally.

5. Conclusion

The proposed work in this research the Ensemble Method and the Gradient Boosting Classifier was applied to develop and test an intrusion defensive system. Among all categories of attacks, the system has achieved an outstanding performance in identifying a number of attack types such as DoS, Probe, R2L, U2R etc. High accuracy therefore signifies to what extent the system is capable of differentiating the legal incoming and outgoing network traffic data from the actual illegitimate activities. The emphasis of the future work might likely be directed towards enhancing the methods of feature engineering. Incorporation of domain-even features and extracting the features through utilizing deep learning tools might help in identifying textures in the network traffic flow.A crucial future step is converting the system to a real-time setting. The intrusion defender system's usefulness would be confirmed by installing it within operational network infrastructures and testing its performance in live situations. For real-world systems, ensuring scalability to manage heavy network traffic is crucial. A comprehensive understanding of security threats might be possible by investigating collaborative intrusion detection systems that make use of the advantages of various algorithms or multimodal techniques combining various data sources, such as network traffic and system logs. Through cross- verification, collaborative systems improve accuracy by using data from various detectors. In conclusion, even though this research used the ensemble model and gradient boosting classifier to obtain outstanding outcomes, there is still a large terrain of room for growth. The field of intrusion detection can develop further by looking at these potential directions in the future, assuring the safety and integrity of digital ecosystems in the face of changing cyberthreats.

References

1. W. Stallings, Network security essentials: Applications and standards (POD file), 6th ed. Upper Saddle River, NJ: Pearson, 2019.

2. S. Northcutt and J. Novak, Network Intrusion Detection, 3rd ed. Upper Saddle River, NJ: New Riders Publishing, 2002.
3. Working With Snort Rules, Chapter 3, Pearson Education, Inc.
4. S. A. Khayam, “Recent Advances in Intrusion Detetction,” in Proceedings of the 26th Annual Computer Security Applications Conference, Saint-Malo, France, pp. 224–243.
5. H. Debar, M. Dacier, and A. Wespi, “Towards a taxonomy of intrusion-detection systems,” Comput. Netw., vol. 31, no. 8, pp. 805–822, 1999.
6. M. Alazab, “Snort-lightweight intrusion detection for networks,” in proceedings of the 13th USENIX Conference on System Administration, p. 229.
7. S. S. K. M. A. W. I. D. Al Abri, “Advanced Persistent Threat (APT) and intrusion detection evaluation dataset for linux systems 2024,” Data in Brief, vol. 54, pp. 1–10, 2024.
8. Somayaji, A., & Forrest, S. (1997). "Automated Response Using System-Call Delays." In Proceedings of the 14th National Computer Security Conference (pp. 223-234).
9. Dr. Alice Smith & Dr. John Johnson, “Intrusion Detection System for Advanced Persistent Threats: A Review”
10. Dr. Emily Davis, a specialist in the field of cybersecurity ,”Machine Learning Approaches for Intrusion Detection System: A Comprehensive Survey”
11. Dr. Emily Davis ,”Anomaly-based Intrusion Detection Systems for Cyber-Physical Systems: A Review”
12. Dr. David Clark and his group, “Deep Learning-Based Intrusion Detection System for Cyber-Physical Systems “.
13. B. Daya, "Network Security: History, Importance, and Future," Electrical and Computer Engineering Department, University of Florida, 2013. http://web.mit.edu/~bdaya/www/Network%20Security.pdf
14. Li CHEN, Web Security: Theory And Applications, Sun Yat-sen University, China, School of Software.
15. Sinchana K, Sinchana C, Gururaj H L, Vidyavardhaka College of Engineering, Mysore, Sunil Kumar B R, Adichunchanagiri University, Mandya, Performance Evaluation and Analysis of various Network Security tools, Proceedings of the Fourth International Conference on Communication and Electronics Systems (ICCES 2019)
16. "Software Agents and Computer Network Security," M. M. B. W. Pikoulas J. Napier University, Scotland, UK.
17. "Denial of Service Attacks," Texas State University, San Marcos, Q. Gu and Peng Liu.
18. A. Shibli, "MagicNET: Human Immune System & Network Security," Volume 9, Number 1,January 2009, IJCSNS International Journal of Computer Science and Network Security
19. "Study on Intrusion Detection Policy for Wireless Sensor Networks", International Journal of Security and Its Applications, vol. 7, no. 1, January 2013, pp. 1-6.