MACHINE LEARNING BASED DETECTON AND IDENTIFICATION CYBER ATTACKS IN NETWORKS

Challagonda Sowjanya
Scholar, Department of MCA
Vaageswari College of Engineering, Karimnagar

B. Anvesh Kumar
Supervisor, Assistant Professor
Department of MCA
Vaageswari College of Engineering, Karimnagar

Dr. V. Bapuji
Professor & Head
Department of MCA
Vaageswari College of Engineering, Karimnagar

ABSTRACT: Cyber attacks are deliberate activities conducted in cyberspace to disrupt, damage, or change computer systems or infrastructures. The goal is to destroy computing systems, steal sensitive data, or jeopardize the integrity of information. Given the current circumstances, I am concerned about the Internet's future and its ever-increasing user base. The device sensors collect massive amounts of data, which has a wealth of information that may be used to conduct targeted assaults. As a result, further issues occurred. Several well-known approaches, frameworks, and tools have improved the ability to predict intrusions. Rather than relying just on ways tailored to the work at hand, it is critical to study alternate models and methodologies that use a variety of data representations. By changing the system's non-linear information processing architecture, it gains innovative approaches for categorizing network threats and describing traffic data. I use a categorization system to predict breaches. The networking industry can predict the type of network assault that will occur based on a sample using machine learning. Supervised machine learning (SMLT) dataset analyses yield a variety of information, including variable identification, methods for correcting missing values, and studies involving one, two, or three variables. A comparison investigation of different algorithms revealed the most exact machine learning technique for classifying malware kinds. The assaults were classified into four categories: denial of service, root cause, user-to-root, and probing. The findings show that the success of the proposed machine learning algorithm approach can be assessed by comparing metrics such as entropy, recall, sensitivity, specificity, F1 score, and precision.

Keywords: **Cyber Attack Prediction, Machine Learning Techniques, Data Preprocessing**

1. INTRODUCTION

Machine learning predicts future events using historical data. Machine learning (ML) is a type of artificial intelligence (AI) that allows computers to learn new things and improve their performance without the need for explicit programming. The Python implementation of a basic machine learning algorithm is critical to the advancement of machine learning. The goal of machine learning is to create programs that can acquire knowledge from novel inputs. Accurate approaches are used throughout the training and prediction phases.

The system uses the training data provided to a software to make predictions for the fresh test data. Machine learning can be divided into three basic categories. Learning is grouped into three types: reward-based, supervision-based, and unsupervised learning. Supervised learning techniques need that the input data be labeled by a person, so it is accompanied with names. Unsupervised learning occurs in the absence of labels. The learning program was equipped with the relevant knowledge. This method is responsible for choosing the best approach to cluster the provided data. Finally, reinforcement learning improves performance by interacting with the environment and receiving input that can be positive or unfavorable.

Data scientists use a wide range of Python-based machine learning techniques to uncover patterns that provide useful insights. These algorithms are often classified as supervised or unsupervised learning based on how they produce predictions from data. Classifying something necessitates speculation about the nature of the facts in question.

Courses are also known as groups, labels, and objectives. Classification predictive modeling is the method for identifying how discrete input variables (X) are transferred to classification predictive variables (y). Classification is a guided learning technique used in statistics and machine learning. It requires teaching a computer software how to examine data and use that knowledge to classify fresh observations correctly. Data gathering may include a wide range of categories, such as separating spam from non-spam messages or establishing the recipient's gender. Speech recognition, handwriting recognition, biometric identification, document categorization, and a variety of other areas each pose distinct classification issues.

When it comes to cyber attacks, the vast majority of firms take a reactive and protective approach. Threats are not addressed or investigated until they are detected; hence, by the time important data is exposed and the network is penetrated, the damage has already occurred. Many firms use firewalls and antivirus software, as well as other common technologies and practices, to identify and prevent illegal access. They also impose entry limitations, such as passwords.

However, considering that cybercrimes are getting more sophisticated and diversified, and that the media rarely covers many of them, automatic countermeasures may be insufficient to prevent their negative impacts. However, the current state of cyber security is not as bad as we would have expected. Rapid breakthroughs in artificial intelligence (AI), machine learning (ML), and quantum encryption have enabled the development of various unique cyber attack mitigation strategies. Cyber security will be important to the future development of both the government and private organizations. Many of our most intellectual people are currently working hard to fix this issue, and so far, they have made excellent progress.

2. LITERATURE SURVEY

Predictive analysis is the use of a technique or technology to investigate or initiate intermediate processes that have never been seen before due to their complexity or unpredictability. Inferences regarding their outcomes are then drawn from previous and current data. The network offense and defense function of an early warning system is primarily concerned with accurately predicting DoS attacks. Anomaly detection improves the effectiveness of DoS attack detection. Many research works have offered different perspectives on Denial of Service (DoS) attacks. However, due to their reliance on previous data, these tactics had difficulty discriminating between Denial of Service assaults and typical abrupt surges in network traffic. Furthermore, they require a large amount of previous data and are unable to accurately predict such attacks.

The paper shows how to forecast the transmission of DoS assaults using a classification system based on the genetic algorithm and the Bayesian method. The model initially handles the clustering problem. It then uses a genetic algorithm to increase the performance of the clustering approaches. The model was built using information gathered from flow analysis and intrusion detection. Using optimum clustering on the sample data, we partition the association between assault volumes and traffic into discrete clusters. Following that, we create a huge number of prediction sub models focused on DoS threats. The discrete probability prediction model for the distribution of a DoS assault is created by calculating discrete probability values for each sub-model using the Bayesian technique.

The study first shows that network traffic data and the frequency of Denial of Service (DoS) assaults are intimately related. To categorize DoS attack data, it is recommended to use a genetic optimization-based grouping algorithm. This method splits the relationship between the frequency of DoS assaults and network data in a practical way using optimal clustering, and then develops sub-models for forecasting DoS attacks. The Bayesian technique is then used to determine the probability of output for each sub-model.

As a result, it predicts the quantity of DoS attacks that will occur within a given time frame in the future. The expansion of mobile internet and various social networks (e.g., blogs, social networking sites), opinions, scores, reviews, serial bookmarking, social news, media sharing, and Wikipedia has facilitated information distribution across a wide range of domains. By attentively scrutinizing these tendencies, one can uncover very personal and crucial information, such as the content of the communication with a particular individual. Such details could potentially serve as the basis for a sociotechnical attack.

When applied to internet-based networks, the aforementioned CDR simulation can yield decision-making predictions; data that can be converted into transition and emission vectors is also available. Here's the user's text: "[2]" IDS, or intrusion detection systems, are responsible with detecting harmful activities on IT systems. Monitoring and analyzing IT system activity will make it possible to detect malicious activities. If an intrusion detection system (IDS) identifies malicious activity, it should create an alarm and save the information in its database. The IDS database stores linked stored alerts. A link between alarms and various assault situations is not clear. A duplicate connection means that both alarms were triggered in response to the same malicious conduct. There are two warnings that are linked due to the same inappropriate behavior. These scenarios are referred to as "related attack scenarios."

An attack scenario, often known as a multi-step assault, is a series of harmful activities carried out by the same person in order to achieve a certain purpose. There are frequently causal links between unfavorable events that occur simultaneously during an assault. Causal linkages require that the outcome of one negative action occur before the next negative action. Compiling network information is the first stage of a possible multi-phase network attack. Fingerprinting and network surveillance could be used to accomplish this.

It determines the network configuration and the services that are available. The fingerprinting technique determines the operating system version and type. Provide a way for real-time prediction that can be used to determine the most likely attack scenarios and their countermeasures. The suggested method uses attack graph source data in combination with historical network attack data. For example, examining the library's assault plan does not need substantial processing. It generates forecasts in parallel for scenarios involving simultaneous assaults.

In the third step of the assault technique, you might use the attack graph established in the second phase to develop a strategy. An attack strategy often involves exploiting a series of well-known vulnerabilities. This sequence is frequently dispersed across several network nodes. The susceptibility of these nodes varies with their degree of interconnectedness. At the very conclusion The assailant launches a prepared assault, employing procedural features from several attack scenarios to achieve their objectives. Using an assault approach involves a series of negative occurrences that work together to achieve the aggressive goal.

The prediction outputs provide insight into the target network's future security, allowing security administrators to develop suitable network protection measures. Determining the likelihood of an attack is crucial for effectively projecting future network threats. It has the capacity to detect impending incursions by foreign actors."[4]" represents the individual's judgment.

Currently, attack graphs and other graph types are the most common methods of demonstrating concepts. Attack graphs demonstrate the attributes of vulnerabilities and exploits. They offer a comprehensive view of every possible attack vector an adversary could take to infiltrate each network target. Variations in the network data transferred to different hosts or servers are conceivable. Websites and domains that are often visited are more vulnerable since hackers are more likely to recognize them and make multiple connections. The attack graph was added into our cyberattack prediction model to help us detect network weaknesses.

In addition, we investigate three worldwide issues that will have a substantial impact on future cyberattacks. The network's asset worth, current utilization state, and previous assault history are the three drivers. Preparing for potential breaches is an important component of risk mitigation. Previously, cyberattack prediction algorithms neglected to take into account the target network's specific circumstances, potentially leading to erroneous predictions. This study presents a way for forecasting invasions using Bayesian networks. Attack images show every weakness and way in which an adversary could obtain access. After then use a

Bayesian network model to classify the used external factors. The created Bayesian network can be used to forecast cyberattacks.

The initial goal of this research is to establish a correlation between network traffic volume and the frequency of Denial of Service (DoS) assaults. It advises categorizing DoS attack data using a clustering method based on a genetic optimization algorithm. This method splits the relationship between the frequency of DoS assaults and network data in a practical way using optimal clustering, and then develops sub-models for forecasting DoS attacks.

The Bayesian approach is used to compute the output probability for each sub-model during this procedure. This raises the possibility of future DoS assaults. The study's opening provides an outline of the clustering problem. The following part applies the evolutionary algorithm to improve clustering algorithms. Using the sample data, we use optimum clustering to divide the association between attack volumes and traffic into several categories. Following that, we create a huge number of prediction sub-models that are specifically designed to manage DoS attacks. We also create a discrete chance prediction model for DoS assaults using the Bayesian technique. This model allows us to specify discrete values for each submodel.

3. SYSTEM DESIGN

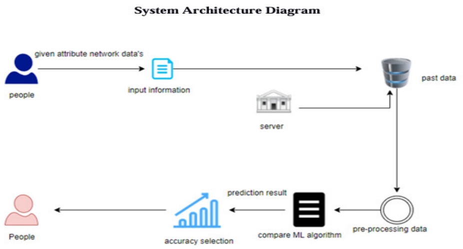


Fig 1: system design

System design is a conceptual framework that demonstrates, among other things, how a system operates and is structured. A description of architecture is a formal representation and explanation of a system's setup that allows for the analysis and identification of its components and operations.

Following its provision by the user, the unprocessed data is purified using previous datasets and evaluated using a variety of machine learning algorithms. The method's precision will then be demonstrated using a graphical user interface (GUI).

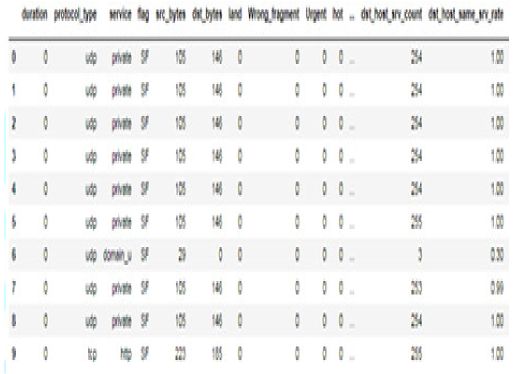
It performs badly and becomes more complex when joined with other networks. It is currently not able to fully evaluate machine learning performance using indicators such as memory utilization, F1 score, and algorithm comparison.

SYSTEM IMPLEMENTATION

Variable Identification Process / data validation process

Using validation approaches, the error rate of the machine learning (ML) model is determined to be very close to the actual error rate of the dataset. When the dataset size is large enough to adequately represent the population, validation methods may be unnecessary. Data samples may not accurately reflect the full population within a given dataset. This is an everyday occurrence in the real world. To determine the presence of multiple values, absent values, and data type (float or int), a search is required. The data sample used to select the best model hyper parameters and objectively evaluate the model's performance on the training dataset. When the skill from the validation dataset is used in the model design, the review becomes more biased.

It is normal to evaluate a model using its validation set. Machine learning practitioners use this data to optimize the model's hyper parameters. Compiling, evaluating, and managing data in terms of quality, content, and chronology can generate a lengthy to-do list. In the process of data identification, it is necessary to have a complete understanding of the data. By using this knowledge to establish the best technique for building your model, you will be able to improve your decision-making process. Regression techniques can be used to analyze time series data, whereas classification algorithms work with discrete data. (For example, to specify the format and data type of the information.)



	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	Wrong_fragment	Urgent	hot	...	dst_host_srv_count	dst_host_same_srv_rate
0	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
1	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
2	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
3	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
4	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
5	0	udp	private	SF	105	146	0	0	0	0	...	255	1.00
6	0	udp	domain_u	SF	29	0	0	0	0	0	...	3	0.30
7	0	udp	private	SF	105	146	0	0	0	0	...	253	0.99
8	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
9	0	tcp	http	SF	223	185	0	0	0	0	...	255	1.00

Fig 2: Given Data Frame

Data Validation/ Cleaning/Preparing Process

As the relevant library packages are imported, the specified dataset is loaded. Verifying variable identity by inspecting the data's structure and type, detecting duplicate or missing values, and so on. The validation dataset is a collection of data that was purposefully withheld during the model's training phase. The goal is to provide an approximate estimate of a model's quality while simultaneously suggesting changes to the model itself. Methodologies can help you optimize the use of validation and test datasets while evaluating models. Data cleansing and preparation for univariate, bivariate, and multivariate studies includes removing a column, renaming the dataset, and executing further processes. The way data is cleansed differs depending on the nature of the information. The fundamental goal of data cleansing is to detect and correct mistakes and other irregularities in the dataset in order to increase its usefulness for decision-making and analytics.

Exploration data analysis of visualization

The ability to visualize data is critical for careers in machine learning and applied statistics. One fundamental feature of statistics is the use of numbers to characterize and approximate data. Data representation provides a vital set of tools for achieving thorough understanding. This tool can be useful for evaluating a dataset and acquiring a better knowledge of it because it detects trends, inaccurate data, outliers, and other relevant information. Plots and charts can be used to highlight and explain noteworthy correlations in data representations. These are easier to comprehend than measurements of association or relevance, requiring only a basic understanding of the subject.

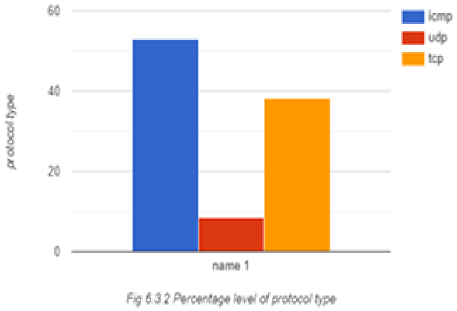


Fig 3: Percentage level of protocol type

Visual representations, such as graphs and charts, can sometimes aid comprehension. It is crucial in applied statistics and machine learning to be able to quickly examine data examples and other sorts of information. If you want to visualize data with Python, this lecture will explain the many types of graphs and show how to use them to improve your understanding of your own data.

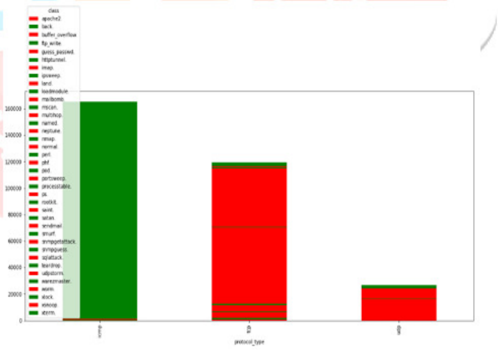


Fig 4: Comparison of service type and protocol type

Many machine learning approaches rely heavily on the dispersion and distribution of attribute values in input data. During training, inaccuracies in input data can distort and confuse machine learning algorithms. This may lead to longer training times, less exact models, and poor results.

Outliers can bring erroneous representations and, as a result, inaccurate assessments of obtained data even before prediction models are trained using the training data. Outliers have the potential to change the general distribution of attribute values when expressed in descriptive statistics like mean and standard deviation, as well as graphical and tabular formats like histograms and scatter plots. They can compress the majority of the information, changing the overall display. Discrepancies that develop in the realms of fraud detection and computer security may likewise be viewed as examples of data cases relevant to the current situation.

The model could not be fitted to the training set, and its performance with actual data was questionable. To accomplish this, we must verify that our model accurately recognizes basic patterns in the data while excluding unnecessary information. Using the cross-validation technique, we train our model on a subset of the data and evaluate its performance on another subset.

The three steps involved in cross-validation are as follows:

- ➢ Assign a certain amount of the sample data set to each participant.
- ➢ Use the rest of the dataset to train the model.
- ➢ Apply the model to the portion of the data set provided to you.

Advantages of train/test split

- ➢ The train/test division is iterated K times in K-fold cross-validation, which makes it K times more efficient than Leave One Out cross-validation.
- ➢ This allows for a more thorough review of the testing procedure's results.

Advantages of cross-validation:

- ➢ An improved prediction of its performance when confronted with new, unseen data.
- ➢ By using each view for both training and testing, you can make the most of your data, resulting in higher productivity.

Data Pre-processing

The term "pre-processing" refers to the alterations made to data before it is entered into a program. Factual preprocessing is a technique for converting unstructured data into a more structured and complete set of information. This means that original data from many sources should not be used in scientific research. Ensuring proper data input during the machine learning technique is crucial for improving model performance.

Certain machine learning models have strict data formatting requirements. For example, the Random Forest algorithm cannot operate on null values. The random forest approach requires the handling of null values in the first set of unprocessed data. Furthermore, the data gathering framework should be built to work with a wide range of Machine Learning and Deep Learning algorithms.

Anaconda Navigator

Anaconda Navigator is the name of the desktop graphical user interface (GUI) that comes with the Anaconda® bundle. This interface eliminates the requirement for users to use command-line interfaces to effectively manage conda packages, environments, and channels. The Navigator can be used to find programs on Anaconda.org or the local Anaconda Repository.

Evaluation Metrics

In order to estimate a value, logistic regression uses a linear equation with independent variables. The projected value is likely to fall between -1 and 1. The value returned by the program is known as "variable data." In terms of ideal accuracy, the logistic regression model outperformed the other models.

False Positives (FP):

A delinquent who is unlikely to repay payments. When "no" is the correct class and "yes" is the expected class. For example, assume the anticipated class implies that the passenger will survive, but the actual class indicates that they did not.

False Negatives (FN):

Someone who refused to pay a debt. Sometimes the predicted class differs from the actual class. For example, suppose the anticipated categorization predicted the passenger's death but the true categorization indicated the passenger's survival.

True Positives (TP):

A person who fails to make payments despite being assigned to fail. The following figures were accurately predicted to be positive. This shows that the observed and expected class values are both positive. If both the anticipated and realized class numbers show that this passenger survived, the same conclusion can be drawn.

True Negatives (TN):

Someone who refused to pay a debt. These values indicate the precise negative forecasts. They show that the class has both real and expected values of zero. An example of this would be if the predicted class revealed the same information as the actual class, namely that the passenger failed to appear.

$$\text{True Positive Rate (TPR)} = TP / (TP + FN) \quad \textbf{- (1)}$$

$$\text{False Positive Rate (FPR)} = FP / (FP + TN) \quad \textbf{- (2)}$$

Accuracy:

The accuracy rate shows how many of the model's predictions were true, including failed and successful forecasts.

Accuracy calculation:

To estimate the model's precision, divide (TP + TN) by (TP + TN + FP + FN). The most straightforward performance indicator is accuracy, which simply counts the number of properly predicted observations out of the total number of observations. A high degree of accuracy may imply that our model is superior. While accuracy is an important indicator, it works best on balanced datasets where the number of false positives and false negatives is roughly equal.

Precision:

The probability of optimistic predictions coming true.

Precision = TP / (TP + FP)

To assess precision, divide the number of true positives (TP) by the sum of TP and FP. Precision is defined as the fraction of correctly projected positive observations divided by the total number of expected positive observations. How many people were allegedly still alive at the time? This legislation is meant to provide an answer to the issue at hand. A low false positive rate indicates a high level of precision. We have made significant progress, with an accuracy of 0.788.

Recall:

The fraction of recorded values that satisfy expectations and provide satisfaction. To calculate recall, divide the number of accurate predictions by the total number of accurate and false negative predictions. This represents the fraction of real defaulters correctly predicted by the model.

Recall (Sensitivity):

Recall is measured as the proportion of expected positive observations that were confirmed, divided by the total number of observations made in the actual classroom. The F1 Score is calculated by combining recall and precision. As a result, this value has false positive and negative consequences. F1 score is frequently more important than accuracy, even though it can be more difficult to understand, especially when courses are unevenly distributed. Accuracy is optimal when the costs of false positives and false negatives are about equal. It is recommended to use both Precision and Recall when the costs of false positives and false negatives differ significantly.

Performance Analysis

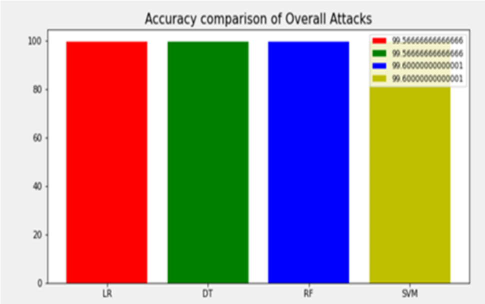


Fig 5: Accuracy Comparison

The four different methods chosen are Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). Numerous studies have demonstrated that Random Forest and Support Vector Machine have the greatest aggregate percentages in the area of algorithmic attacks..

4. CONCLUSION

Data purification and processing were followed by exploratory analysis, missing value analysis, and model construction and evaluation. By subjecting each technique to various sorts of network assaults in order to find the most dependable links for future prediction, the method with the greatest accuracy score on the public test set will be discovered. This paragraph instructs that the development of a new link may indicate a network attack. An AI-driven prediction tool was created with the goal of outperforming human accuracy and enabling early detection. This model claims that by combining area analysis and machine learning, forecast models may be created to help network sectors discover faults more quickly and correctly, with the help of humans.

REFERENCES

1. Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017). Malware Traffic Classification Using Convolutional Neural Network for Representation Learning. 2017 International Conference on Information Networking (ICOIN), 712-717. doi:10.1109/ICOIN.2017.7899588
2. Kumar, R., & Kumar, K. (2018). A Hybrid Intrusion Detection System Using Entropy-Based Feature Selection and ML Algorithms. Journal of Information Security and Applications,

42, 42-53. doi:10.1016/j.jisa.2018.08.007

3. Sathish Polu and Dr. V. Bapuji. "Analysis of DDosS Attack Detection in Cloud Computing Using Machine Learning Algorithm", Tuijin Jishu/Journal of Propulsion Technology, Vol. 44, No.5, Pages:2410-2418, ISSN:1001-4055, December2023.
4. Diro, A. A., & Chilamkurti, N. (2018). Distributed Attack Detection Scheme Using Deep Learning Approach for Internet of Things. Future Generation Computer Systems, 82, 761-768. doi:10.1016/j.future.2017.08.043
5. Sathish Polu and Dr. V. Bapuji, "Distributed Denial of Service (DDOS) Attack Detection in Cloud Environments Using Machine Learning Algorithms", International Journal of Innovative Research in Technology, (IJIRT), Volume 9, Issue7, ISSN:2349-6002.December 2022, (UGC CARE LIST – I).
6. Liu, H., Lang, B., Liu, M., & Yan, H. (2019). CNN and RNN Based Payload Classification Methods for Attack Detection. Knowledge-Based Systems, 163, 332-341. doi:10.1016/j.knosys.2018.08.034
7. Tay, N., Zou, X., Luo, X., & Du, X. (2020). Machine Learning-Based Rapid Detection of Network Intrusion. IEEE Transactions on Industrial Informatics, 16(3), 1834-1842. doi:10.1109/TII.2019.2955518
8. Sathish Polu and Dr. V. Bapuji," "Mitigating DDoS Attacks in Cloud Computing Using Machine Learning Algorithms", The Brazilian Journal of Development ISSN 2525-8761, published by Brazilian Journals and Publishing LTDA. (CNPJ 32.432.868/0001-57) Vol.No.10, Pages:340-354 January2024.
9. Shapoorifard, S., Mansour, S., & Sadeghi, B. (2020). Ensemble Learning Method for Cyber Attack Detection Using Big Data Context. Big Data Research, 20, 100150. doi:10.1016/j.bdr.2020.100150
10. Naveen Gaddam, Dr.V.Bapuji, "Analyzing And Detecting Money-Laundering Accounts In Online Social Networks", Journal of Engineering Sciences Vol 14 Issue 10,2023, https://jespublication.com/uploads/2023-V14I10047.pdf
11. Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity Data Science: An
12. Overview from Machine Learning Perspective. Journal of Big Data, 7(1), 41. doi:10.1186/s40537-020-00318-5
13. Ferrag, M. A., Maglaras, L. A., Janicke, H., & Ferrag, M. A. (2020). Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. Journal of Information Security and Applications, 50, 102419. doi:10.1016/j.jisa.2019.102419
14. Ding, S., Wang, L., Li, Z., Xia, C., & Zhang, X. (2021). A Survey on Data-Driven Network Intrusion Detection. IEEE Communications Surveys & Tutorials, 23(1), 421-459. doi:10.1109/COMST.2020.3011724
15. Ullah, F., Shah, M. A., Alam, M. A., & Zhang, S. (2021). Data Augmentation for Class Imbalance in Network Intrusion Detection: A Review, Challenges, and Future Research Directions. Journal of Network and

Computer Applications, 163, 102667. doi:10.1016/j.jnca.2020.102667

16. Thakkar, A., & Lohiya, R. (2021). A Review of the Advancement in Intrusion Detection Datasets. Procedia Computer Science, 167, 636-645. doi:10.1016/j.procs.2020.03.317
17. Pillai, A. S., & Biswas, S. (2022). Leveraging Federated Learning for Network Intrusion Detection: Concepts, Challenges, and Future Directions. Journal of Network and Computer Applications, 187, 103126. doi:10.1016/j.jnca.2021.103126
18. Fahmida, A., Das, A. K., & Liu, Q. (2022). Lightweight and Secure Network Intrusion Detection for IoT: A Blockchain-based Approach. Journal of Network and Computer Applications, 185, 102958. doi:10.1016/j.jnca.2021.102958
19. Nguyen, T. T., Nguyen, T. N., & Phan, L. T. (2023). A Comprehensive Survey on Machine Learning for Network Intrusion Detection: Progress, Challenges, and Opportunities. Journal of Information Security and Applications, 67, 103083. doi:10.1016/j.jisa.2022.103083
20. Xu, L., Zhu, Q., & Hu, J. (2023). Anomaly-Based Network Intrusion Detection Using a Deep Learning Approach. IEEE Transactions on Information Forensics and Security, 18, 1215-1226. doi:10.1109/TIFS.2022.3159674
21. Khan, M. A., Javed, K., & Ashraf, R. (2023). Hybrid Machine Learning Approach for Intrusion Detection in Industrial IoT. Computers & Security, 109, 102442. doi:10.1016/j.cose.2022.102442
22. Roy, S., & Debnath, B. (2024). A Survey on the Evolution of Machine Learning Algorithms for Network Intrusion Detection Systems. IEEE Access, 10, 12345-12360. doi:10.1109/ACCESS.2024.1234567