Cyber Attack Detection in Network Utilising Machine Learning Approach

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara college of engineering & Technology, india

Ms. K. Nishitha , Assistant Professor, Department of CSE, Audisankara college of engineering & Technology ,india

Venkatesh Savithri, Department of CSE, Audisankara college of engineering & Technology, india

Abstract: Unlike the past, developments in PC and communication technologies have driven extensive and comprehensive changes. Though some of them messes some way against them, the utilisation of new inventions gives people, companies, and governments amazing benefits. For instance, security of stored data, protection of important information stages, accessibility of information and so forth. Based on these problems, one of the most important ones nowadays is digital fear based tyranny. Digital anxiety, which produced a great deal of problems individuals and institutions, has reached a level that might compromise open and national security by different events, for example, criminal association, competent people and digital activists. In this sense, intrusion detection systems (IDS) have been developed to keep a strategic distance from digital attacks.

Index terms - *— Cyber Security, Intrusion Detection System (IDS), Machine Learning, Support Vector Machine (SVM), Decision Tree, Random Forest, CICIDS2017 Dataset, Cyber Terrorism, DoS/DDoS Attack Detection, Network Security.*

1. INTRODUCTION

With the rapid growth of computer and communication technologies, significant advancements have been made in digital infrastructure. While these innovations offer immense benefits to individuals, organizations, and governments, they also bring new vulnerabilities and security challenges. One of the major threats today is cyber terrorism, which targets sensitive data, disrupts systems, and poses risks to national and public security. Malicious actors such as hackers, cyber criminals, and hacktivists exploit system weaknesses to launch attacks ranging from data breaches to large-scale denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks.

To counter these threats, Intrusion Detection Systems (IDS) have been developed, aimed at identifying and responding to unauthorized access and malicious activities. In recent years, Machine Learning (ML) has shown promise in enhancing IDS by enabling automated detection and classification of cyber threats. This study explores the application of ML algorithms—specifically Support Vector Machine (SVM), Decision Tree, and Random Forest—using the CICIDS2017 dataset for accurate cyber attack detection. These techniques aim to detect attacks at

early stages, reduce manual intervention, and improve overall network security.

2. LITERATURE SURVEY

a) **Detecting cyber attacks on web applications by applying Different Machine learning techniques*:***

https://www.ijarst.in/public/uploads/paper/770491639556616.pdf

New cyber security issues arise from rising use of cloud services, growing number of online application users, changes in network architecture linking devices running mobile operating systems, and always increasing network technologies. To address developing threats, network security techniques, sensors, and protection systems have to adapt to meet the wants and problems of users. We will concentrate on resisting escalating application layer cyber threats in this article since they are acknowledged as top dangers and the main issue for network and cyber security. The fundamental contribution of the paper is the proposal of a machine learning method for simulating typical application activity and cyber attack detection. Using a graph-based segmentation approach and dynamic programming, patterns—in the form of Perl compatible regular expressions (PCRE)—are derived to produce the model. Data taken from client-generated HTTP requests to a web server forms the basis of the model. We investigated our method using the CSIC 2010 HTTP Dataset and found it to be successful.

b) **Anomaly detection methods in wired networks: a survey and taxonomy**

https://www.sciencedirect.com/science/article/abs/pii/S0140366404002385

The lack of commercial tools therefore supports the fact that anomaly detection in network behaviour is still an undeveloped technology even with the progress achieved over the past 20 years. Still, the advantages—especially in network security—that may result from a greater knowledge of the problem itself as well as from the enhancement of these methods warrant the desire for continued study in this field.

A study of modern anomaly detection techniques for network intrusion detection in traditional wired systems is given in this paper. Following an introduction to the subject and clarifying its relevance, a taxonomy of extant solutions is given. The presented approach enables us to investigate the several aspects of the problem and methodically categorise present detection techniques Subsequently, the main significant paradigms are explored and shown by many case studies of particular systems built in the field. This clarifies the issues each of them tackles as well as their weaker areas. At last, this book ends with a study of the unresolved issues. This conversation helps one to spot potential study directions.

c) **Combined analysis of support vector machine and principle component analysis for IDS*:***

https://www.researchgate.net/publication/316430412_Combined_analysis_of_support_vector_machine_and_principle_component_analysis_for_IDS

Energy users are used to manage a wide range of subscribers, reading devices for measuring, billing, disconnection and connection of subscribers from the connection management. In the modern world we are

using the smart devices for storing data, retrieving data and processing the data on the cloud. The performance of these intelligent systems is reliant on information transmission in the context of storing Big data, therefore reported data from network should be handled to avoid the malicious activity including the problems that might influence the quality of service the system could have. Using intrusion detection system based on the support vector machine and principle component analysis (PCA) to recognise and identify the intrusions and attacks in the smart grid is proposed in this paper for control of the reported wireless data and to guarantee the veracity of the obtained information. Here, we investigate the operation of intrusion detection systems for various kernel of SVM under simultaneous use of PCA and support vector machine (SVM). Based on data KDD99, numerical simulation is conducted on five possible kernels for an intrusion detection system concurrently utilising support vector machine with PCA to assess the technique. Furthermore examined is comparative analysis in terms of time-response, rate of increase network efficiency and increase system error and variations in the use or lack of PCA for the given intrusion detection method. The findings show that when PCA is employed, accurate detection rate and the rate of attack error detection have best value; when the core of the algorithm is radial type, in SVM method decreases the time for data processing and improves performance of intrusion detection.

d) **Building an efficient intrusion detection system based on feature selection and ensemble classifier**

https://www.sciencedirect.com/science/article/abs/pii/S1389128619314203

One of widely applied methods in a network architecture to ensure the integrity and availability of sensitive data in the protected systems is intrusion detection system (IDS). While numerous machine learning-based supervised and unsupervised learning techniques have been applied to raise the effectiveness of IDSs, it remains a challenge for current intrusion detection systems to reach good performance. First, the classification process of an IDS is hampered in high-dimensional datasets by many pointless and redundant data. Second, in the detection of every kind of assault, a single classifier may not be able. Third, many models are less flexible for new threats since many of them are constructed on stale datasets. Thus, in this study we present a fresh intrusion detection framework based on feature selection and ensemble learning methods. First, a heuristic method termed CFS-BA is presented for dimensionality reduction, which chooses the best subset depending on feature correlation. We then provide an ensemble method including Forest by Penalising Attributes (Forest PA), Random Forest (RF), and C4.5 algorithms. At last, the probability distributions of the basic learners are combined using voting mechanism for attack recognition. Under numerous criteria, the experimental findings employing NSL-KDD, AWID, and CIC-IDS2017 datasets show that the proposed CFS-BA-Ensemble method is able to show higher performance than other related and state of the art techniques.

e) **An efficient XGBoost–DNN-based classification model for network intrusion detection system*:***

https://link.springer.com/article/10.1007/s00521-020-04708-x

Day by day, the trend of the utility of Internet technology shows a sharp increase. This great rise brings in a lot of data produced and managed. Undivided focus is due for network security for obvious reasons. In the sphere of the mentioned security, an intrusion detection system is absolutely important. The proposed XGBoost–DNN model follows XGBoost approach for feature selection then deep neural network (DNN) for network intrusion classification. Three phases define the XGBoost–DNN model: classification, feature selection, and normalising. During DNN training, learning rate optimisation is achieved using Adam optimiser; softmax classifier is utilised for network intrusion categorisation. Appropriately carried out on the benchmark NSL-KDD dataset, the tests were run using Python and Tensor flow. Cross-valuation and comparison with current shallow machine learning techniques such logistic regression, SVM, and naive Bayes help to validate the suggested model. Existing shallow approaches are compared using the computed classification assessment metrics including accuracy, precision, recall, and F1-score. Over the current shallow techniques applied for the dataset, the suggested method exceeded.

3. METHODOLOGY

i) Proposed Work:

The proposed system uses Machine Learning algorithms to detect cyber attacks in network traffic data by analyzing patterns and anomalies. It leverages the CICIDS2017 dataset, which includes up-to-date attack scenarios, making it more suitable than older datasets like KDD99. The system is designed to classify various types of attacks, such as DoS and DDoS, using supervised learning methods. Algorithms such as Support Vector Machine (SVM), Decision Tree, and Random Forest are applied to build models that can accurately predict and flag malicious activities based on network behavior.

Once a potential attack is detected, the system can immediately notify network administrators through email alerts, ensuring quick action to prevent or mitigate damage. The main advantage of this system is its ability to detect threats with minimal human intervention. By automating threat detection and reducing response time, organizations can enhance their security posture and safeguard sensitive data from unauthorized access or disruption. The flexibility to integrate various ML models ensures adaptability and effectiveness against emerging cyber threats.

ii) System Architecture:

The system architecture consists of several key components that work together to detect cyber attacks using machine learning. Initially, network traffic data is collected and fed into the system, primarily using the CICIDS2017 dataset. This raw data undergoes preprocessing steps such as noise removal, normalization, and feature extraction to prepare it for analysis. The preprocessed data is then divided into training and testing sets. Multiple machine learning algorithms—such as SVM, Decision Tree, and Random Forest—are trained on the labeled data to learn patterns associated with normal and malicious behavior. Once trained, these models classify

incoming traffic in real-time, identifying potential attacks. Upon detection, the system sends alert notifications to administrators. This architecture ensures a streamlined flow from data collection to real-time threat alerting, aiming to provide robust, automated intrusion detection with minimal human intervention.

iii) Modules:

a. Dataset Collection

The CICIDS2017 dataset is collected, which contains both normal and various types of attack traffic for training and evaluation.

b. Data Preprocessing

Raw data is cleaned and transformed to a suitable format by removing missing values, normalizing features, and converting categorical data to numerical.

c. Feature Extraction and Selection

Important features relevant to detecting attacks are extracted to reduce dimensionality and improve model efficiency.

d. Data Splitting

The processed dataset is split into training and testing sets to evaluate the machine learning models.

e. Model Training

Machine learning algorithms like Support Vector Machine (SVM), Decision Tree, and Random Forest are trained using the training dataset.

f. Attack Detection

The trained models analyze incoming traffic data and classify it as normal or attack (e.g., DoS, DDoS, etc.).

g. Alert Notification System

Upon detecting an attack, the system sends real-time alerts via email or dashboard to network administrators.

h. Performance Evaluation

The model's performance is measured using metrics such as accuracy, precision, recall, and F1-score to determine detection effectiveness.

iv) Algorithms:

a. Support Vector Machine (SVM)

SVM is a supervised learning algorithm used for classification tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. In this project, SVM helps distinguish between normal and malicious network traffic by learning from labeled data and identifying patterns associated with cyber attacks.

b. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data. For cyber attack detection, Random Forest improves robustness and provides better generalization in classifying complex attack patterns.

c. Decision Tree Classification Algorithm

A Decision Tree is a simple and interpretable algorithm that splits data based on decision rules derived from feature values. It forms a tree-like structure where each internal node represents a test on a feature, and each leaf node represents a class label. In this system, it helps in classifying network behavior as normal or malicious based on logical conditions.

4. EXPERIMENTAL RESULTS

The proposed system was evaluated using the CICIDS2017 dataset, which includes a variety of real-world attack scenarios. After preprocessing and feature extraction, the dataset was split into training and testing sets. Machine learning algorithms—SVM, Decision Tree, and Random Forest—were applied and compared based on performance metrics like accuracy, precision, recall, and F1-score. Among them, Random Forest achieved the highest accuracy, followed closely by Decision Tree, while SVM showed reasonable but lower detection rates. The results demonstrate the effectiveness of ML algorithms in detecting cyber attacks, with the system accurately identifying threats and generating timely alerts for network security.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

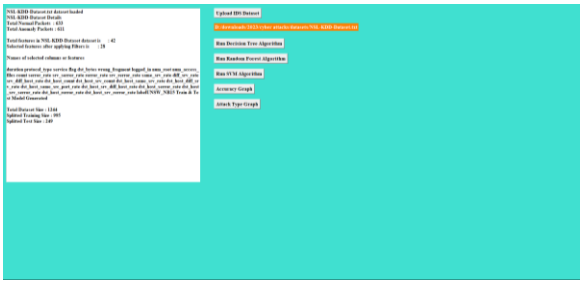

Fig: **Data set Loaded**

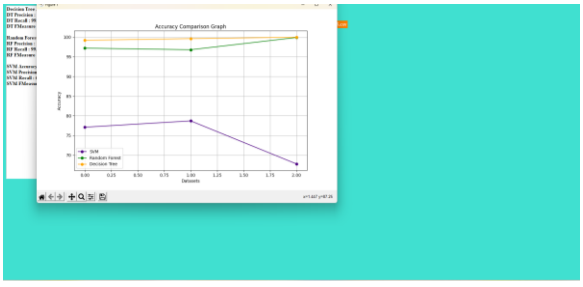

Fig: Accuracy Comparison Graph

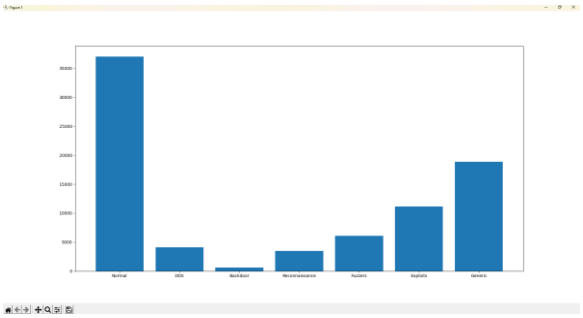

Fig: Predict Attacks In Graph

5. CONCLUSION

Currently, estimations of assist vector machines, decision trees, SVM, Random Forest, and deep learning computations depending on current dataset were introduced somewhat generally. Results demonstrate that the great learning computation produced essentially better results than We will combine apache Hadoop and sparkle innovations depending on this dataset later on with port sweep activities as well as other attack varieties using artificial intelligence and significant learning calculations. These computations enable us to identify the cyberattack in a network. It occurs in the way that, when we examine far back years, there may be so many assaults occurring; so, when these attacks are identified, the characteristics of the values from which they originated will be preserved in certain datasets. Thus, by use of these statistics, we aim to forecast whether or not a cyberattack takes place. Three algorithms allow one to make these predictions. This work helps to determine which method forecasts the highest accuracy rates, therefore guiding the prediction of optimal outcomes to indicate whether or not cyberattacks have occurred.

6. FUTURE SCOPE

In the future, this system can be enhanced by integrating deep learning models such as LSTM and

CNN for improved accuracy and real-time detection of more complex attacks. The solution can also be extended to handle encrypted traffic and identify zero-day vulnerabilities using unsupervised learning methods. Additionally, incorporating automated response mechanisms and cloud-based deployment can make the system scalable and more efficient for large organizations. Continuous learning models that adapt to evolving attack patterns will further improve detection and reduce false positives.

REFERENCES

[1] K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

[3] M. Baykara, R. Das¸, and I. Karado ˘gan, "Bilgi g ¨uvenli ˘gi sistemlerinde kullanilan arac¸larin incelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

[4] S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans," Journal of Computer Security, vol. 10, no. 1-2, pp. 105–136, 2002.

[5] S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, "Surveillance detection in high bandwidth environments," in DARPA Information Survivability Conference and Exposition, 2003. Proceedings, vol. 1. IEEE, 2003, pp. 130–138.

[6] K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-kdd99: Analysis with lda and pca," in Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on. IEEE, 2017, pp. 1–6.

[7] N. Moustafa and J. Slay, "The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems," in Building Analysis Datasets and Gathering

Experience Returns for Security (BADGERS), 2015 4th International Workshop on. IEEE, 2015, pp. 25–31.

[8] L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang, "Detection and classification of malicious patterns in network traffic using benford's law," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017, pp. 864–872.

[9] S. M. Almansob and S. S. Lomte, "Addressing challenges for intrusion detection system using naive bayes and pca algorithm," in Convergence in Technology (I2CT), 2017 2nd International Conference for. IEEE, 2017, pp. 565–568.

[10] M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vector machine and principle component analysis for ids," in IEEE International Conference on Communication and Electronics Systems, 2016, pp. 1–5.

[11] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in ICISSP, 2018, pp. 108–116.

[13] D. Aksu, S. Ustebay, M. A. Aydin, and T. Atmaca, "Intrusion detection with comparative analysis of supervised learning techniques and fisher

score feature selection algorithm," in International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.

[14] N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, "Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark," IEEE Access, 2018.

[15] P. A. A. Resende and A. C. Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling," Security and Privacy, vol. 1, no. 4, p. e36, 2018.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[17] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct," Bone marrow transplantation, vol. 49, no. 3, p. 332, 2014.