Prediction of Cyber-attacks (IDS/IPS) using Machine Learning

Abstract:

The goal of this project is to predict network attacks using a variety of classifiers. The project involves pre-processing the dataset, selecting relevant features, and splitting the data into training and testing sets. Different classifiers such as Decision Trees, Random Forests, Support Vector Machines (SVMs), Neural Networks, and others are trained on the training set and their performance is evaluated on the testing set. The performance metrics considered include accuracy, precision, recall, F1 score, and ROC-AUC curve. The results obtained from different classifiers are compared to select the best-performing one for network attack prediction. The outcome of this project can be useful for cybersecurity professionals to develop better defense strategies against network attacks.

Proposed Methodology:

Dataset Selection: We will make use of The CICIDS 2017 dataset, a suitable dataset for network attack prediction. It is a dataset that is openly accessible and contains information about several kinds of cyberattacks on network traffic.

Data Pre-processing: Pre-processing the dataset includes cleaning it, getting rid of any duplicated or missing information, and putting it into a format that can be used by various classifiers.

Feature Selection: Once the data is pre-processed, relevant features from the dataset will be selected and will be used to train the different classifiers.

Data Splitting: The dataset will then be split into training and testing sets. The different classifiers are trained using the training set, and their performance is evaluated using the testing set.

Model Selection: After splitting the data, different AI/ML models such as Decision Trees, Random Forests, Support Vector Machines (SVMs), Neural Networks, and others are selected for classification.

Model Training: Once the models are selected, then they are trained using the training set. This involves fitting the model to the training data and tuning the hyperparameters to achieve the best possible performance.

Model Evaluation: After training the models, their performance is evaluated using the testing set. Performance metrics such as accuracy, precision, recall, F1 score, and ROC-AUC curve are calculated to determine how well the model performs.

Model Comparison: Finally, the performance of each classifier is compared, and the best-performing classifier is selected for Network Attack Prediction.

Dataset:

The University of New Brunswick's CICIDS 2017 dataset has a traffic flow summary that is frequently used in data science research on information security. Specialists ordinarily use the summed-up information to assemble administered AI models and assess their classification performance. A network of computers that simulate both victim and attacker nodes is used to create the CICIDS dataset. In addition to regular network traffic, the network is subjected to various attack scenarios over the course of five consecutive days. This gives researchers a wide variety of data to work with and examine in order to create and test intrusion detection systems utilizing machine learning techniques. The CSV files contain various traffic flows such as:

Num	Traffic Flow	Description
1.	Benign	Normal Traffic
2.	FTP Patator	Brute force FTP password cracking using the tool Patator
3.	SSH Patator	Brute force SSH password cracking using the tool Patator
4.	DoS Slowloris	DoS attack using the tool Slowloris
5.	DoS Slowhttptest	DoS attack using the tool Slowhttptest
6.	DoS Hulk	DoS attack using the tool Hulk
7.	DoS Goldeneye	DoS attack using the tool GoldenEye
8.	Heartbleed	Exploit Heartbleed bug using the tool Heartleech
9.	Web Attack Brute Force	Exploit vulnerable application DVWA using Brute Force
10.	Web Attack XSS	Exploit vulnerable application DVWA using XSS
11.	Web Attack SQL Injection	Exploit vulnerable application DVWA using SQL Injection
12.	Infiltration	Emulate infiltration using Metasploit
13.	Bot	Botnet emulation using the tool Ares
14.	PortScan	Portscan using the tool LOIC
15.	DDoS	DDoS attack using the tool LOIC

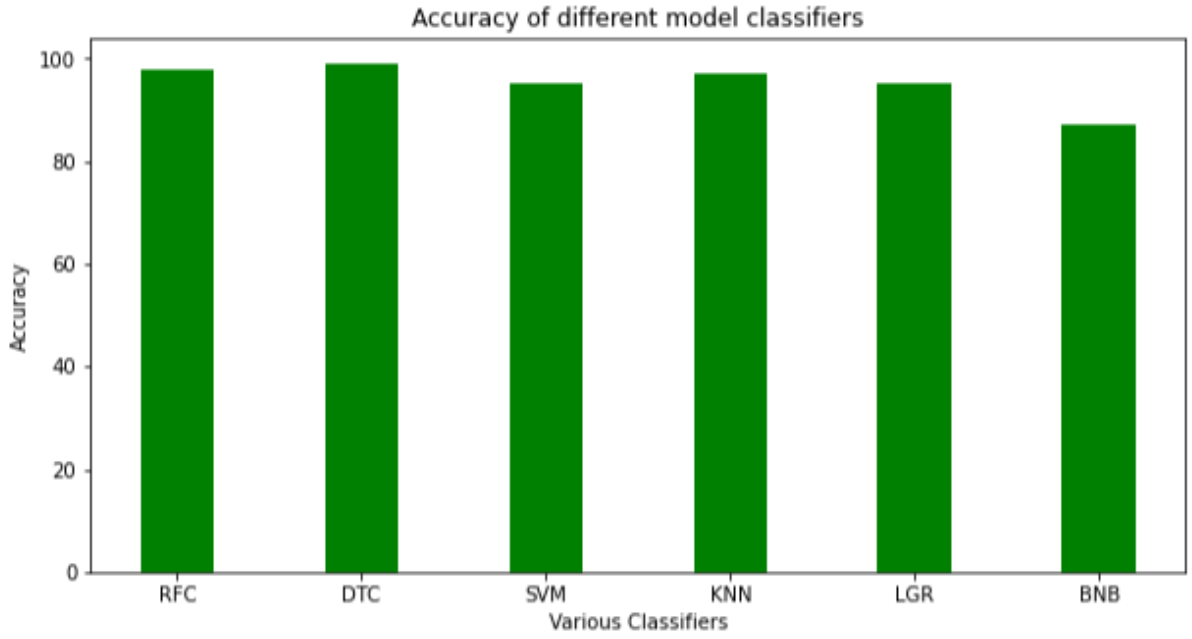
IP addresses (or other comparable identifiers) for the source or destination are absent from the CSV files. These addresses can be used to identify malicious and malfunctioning nodes.

Additionally, neither the originating TCP/UDP port number nor the IP protocol identity are present in the CSV files. When combined with the originating IP address, this data can be helpful in identifying various types of assaults.

Related Work/Literature Survey:

The frequency and seriousness of cyber-attacks and network security breaches have rapidly increased as businesses rely more on networked systems and online data storage. Massive volumes of private data are often compromised as a result of data breaches that take place without the affected individuals being aware of them. It is essential to provide a strong software solution that can detect and identify network intrusions in order to protect computer networks from unwanted access considering this difficult environment. In this article, we suggest using the CICIDS dataset to identify whether a connection is targeted or not using machine learning approaches. Our goal is to research ensemble learning voting classifier algorithms to improve packet connection transfer forecasting and accurately predict large-scale attacks, DOS and probe attacks with high accuracy, recall, and F1 score. Our results demonstrate that the proposed AI-based approach outperforms other machine learning algorithms, highlighting its effectiveness in accurately predicting network intrusion events.

Results:



Conclusion:

The results of this project show that creating a prediction model with machine learning methods can improve accuracy and offer early network attack detection. According to the results, applying machine learning techniques can decrease human error in the diagnostic procedure, increasing

the model's efficacy. The research successfully predicted network attacks and significant assaults using AI-based methodologies, yielding acceptable accuracy, recall, and F1 Score. These findings highlight the potential advantages of applying the suggested AI calculation technique to the field of network security.

References:

1.Arif Yulianto, Parman Sukarno and Novian Anggis Suwastika. Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. 2019 https://iopscience.iop.org/article/10.1088/1742-6596/1192/ 1/012018/pdf

2. VMware NSX Distributed IDS/IPS. 2019. VMware Inc. Retrieved from https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/ docs/vmware-nsx-distributed-ids-ips-tech-white-paper.pdf

3. Ranjit Panigrahi and Samarjeet Borah. A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems. 2018. International Journal of Engineering and Technology.

https://www.researchgate.net/publication/329045441 A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems

4. Network Traffic Analysis (NTA): A Cybersecurity Quick Win. 2020. https://www.cisco.com/c/dam/en/us/products/ collateral/security/stealthwatch/stealthwatch-esg-wp.pdf

5. https://www.unb.ca/cic/datasets/index.html

6. Kim K D., and Kumar P. (2012). Cyber–physical systems: A perspective at the centennial, in Proc.IEEE.

7. Zhang H., Shu Y., Cheng P., and Chen J.(2016). Privacy and performance trade-off in cyber-physical Systems, IEEE Network.

8. Manandhar K., Cao X., Hu F., and Liu Y.(2014). Detection of faults and attacks including false data injection attack in smart grid using Kalman filter, IEEE Transactions on Control of Network Systems.

9. Pasqualetti F,. D¨orfler F., and Bullo F.,(2013). Attack Detection and Identification in Cyber-Physical Systems, IEEE Transactions on Automatic Control.

10. Iman Sharafaldin, Arash Habibi Lashkari and Ali A. Ghorbani. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. 2018. 4th International Conference on Information Systems Security and Privacy (ICISSP).

https://www. scitepress.org/Papers/2018/66398/66398.pdf