INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

Mrs. Sumrana Siddiqui[1],Mohammed Rayees Ahmed[2], Mohammed Asjad[3], Mohammed Abdullah[4]

[1]Assistant Professor, Deccan College of Engineering and Technology Engineering, India.

[2,3,4]B. E Student, Department of CSE, Deccan College of Engineering and Technology Engineering, India.

ABSTRACT: This research presents a new method for improving the effectiveness and precision of intrusion detection systems (IDS) within the framework of dynamic wireless networks. The suggested approach seeks to maximize intrusion detection and classification by fusing Principal Component Analysis (PCA) with the Random Forest classification algorithm. By reducing the dimensionality of the dataset, PCA efficiently organizes and streamlines the data in preparation for more in-depth research. The Random Forest algorithm is then applied for classification, making use of its ensemble learning properties to raise intrusion detection accuracy.

It is shown that the suggested strategy is effective through extensive testing and analysis. PCA and Random Forest-based IDS perform better than traditional methods like Support Vector Machines (SVM), Naive Bayes, and Decision Trees, according to comparison analysis. Remarkably, the suggested approach surpasses previous approaches with an accuracy percentage of 96.78%. Furthermore, the approach demonstrates efficacy in real-time intrusion detection settings, underscoring its practical application, with a processing time of 3.24 minutes.

All things considered, this research represents a major breakthrough in network security, providing a strong defense against the growing difficulties brought on by cyberattacks. The suggested intrusion detection system (IDS) improves accuracy and guarantees fast intrusion detection and mitigation by combining PCA for data preprocessing and Random Forest for classification. This strengthens the resilience of wireless communication networks against malicious-activities.

Introduction

In today's internationally linked society, access to information, commerce, and communication are all made possible via the internet. Despite its numerous advantages, the increasing usage of internet services has resulted in significant security issues. Malicious actors conduct a range of cyberattacks, including Distributed Denial of Service (DDoS) assaults and data breaches, by taking advantage of weaknesses in programs and network infrastructure. These assaults pose a major danger to the availability, confidentiality, and integrity of online services, and robust security measures are needed to mitigate their consequences.

Among the many cyberattacks, DDoS assaults are one that is very damaging and disruptive. When malicious parties overwhelm a target server or network with traffic, they launch a denial-of-service attack (DDoS) that prevents authorized users from accessing it. This tactic interferes with vital online services and comes at a substantial financial cost to the affected companies. Therefore, it is crucial to identify and mitigate DDoS assaults in order to safeguard the stability and resilience of internet-based systems. Intrusion Detection Systems (IDS) play a major role in network protection strategies as they combat the rising danger of DDoS assaults and other undesired activities. IDS (intrusion detection systems) are specialized hardware or software designed to monitor network traffic, identify anomalous or suspicious patterns, and instantly alert management to potential security breaches. By using machine learning algorithms and powerful analytics, IDS can automatically detect and react to a wide range of security threats, enhancing the overall security posture of companies that do business online.

Given the dynamic nature of the threat landscape and the critical role that intrusion detection systems (IDS) play in preventing cyberattacks, the purpose of this project is to evaluate the effectiveness of two popular machine learning algorithms: Random Forest and Principal Component Analysis (PCA) with Support Vector Machine (SVM) in the context of DDoS attack detection. The study's particular goal is to assess how well these algorithms distinguish between malicious DDoS attack traffic and normal network traffic, with an emphasis on detection speed, computation efficiency, and accuracy. This research seeks to identify the optimal technique for enhancing DDoS detection capabilities in intrusion detection systems via a real performance comparison of the Random Forest and PCA-SVM algorithms.

Intrusion Detection Systems (IDS) play a major role in network protection strategies as they combat the rising danger of DDoS assaults and other undesired activities. IDS (intrusion detection systems) are specialized hardware or software designed to monitor network traffic, identify anomalous or suspicious patterns, and instantly alert management to potential security breaches. By using machine learning algorithms and powerful analytics, IDS can automatically detect and react to a wide range of security threats, enhancing the overall security posture of companies that do business online.

The study's findings have significant implications for DDoS attack mitigation in particular as well as cybersecurity in general. By elucidating the respective benefits and drawbacks of the Random Forest and PCA-SVM algorithms in spotting DDoS assaults, this work may aid in the creation and implementation of more dependable and effective intrusion detection systems (IDS). Additionally, businesses will be better able to choose and deploy machine learning-based DDoS detection systems, bolstering their defenses against dynamic cyberthreats, by using the insights gathered from this research. Ultimately, this research contributes to the advancement of cybersecurity practices and highlights the need of machine learning techniques for proactive threat detection and response in the digital realm.

PROBLEM STATEMENT:

The widespread availability of internet services in the modern digital age has led to an increase in cyber dangers, most notably the spread of Distributed Denial of Service (DDoS) assaults. These sneaky attacks include a planned flood of requests sent towards the intended target servers or networks, which eventually overwhelms their response capabilities and renders critical online services unavailable. Traditional security solutions are severely challenged by the sheer volume and coordinated nature of these assaults, since they often fail to discriminate between malicious payloads and normal user data.

Therefore, it is crucial to create reliable and advanced intrusion detection systems (IDS) that can quickly detect and stop these kinds of assaults. As the first line of defense against online attacks, these intrusion detection systems (IDS) constantly scan network data for unusual patterns or questionable activity. But it is intrinsically difficult and complicated to distinguish between legitimate and malicious communication in the massive volume of internet data. The inadequacy of conventional rule-based methods in identifying new and developing attack vectors emphasizes the need for more sophisticated and flexible detection systems.

In this regard, creating intrusion detection systems (IDS) that can use machine learning techniques is a viable way to improve cybersecurity defenses. Machine learning-based intrusion detection systems (IDSs) are able to automatically analyze past network traffic patterns and detect variations that may be signs of DDoS assaults by using data-driven analytics. As new dangers arise, these systems may adjust and change over time, continuously improving their detection skills. Therefore, building strong IDS is an important undertaking in protecting the continuous provision of vital online services and strengthening the resilience of internet infrastructure against growing cyber threats.

LITERATURE SURVEY

Asma Zahra , 1.Title: A Study on the 2017 Cryptowall Ransomware Utilizing Command and Control Blacklisting for IoT-Based Ransomware Growth Rate Evaluation and Detection:

This study investigates the growth rate of IoT-based ransomware and proposes a detection mechanism utilizing Command and Control (C&C) blacklisting, focusing on the 2017 Cryptowall Ransomware variant. With the proliferation of Internet of Things (IoT) devices, ransomware threats have escalated, necessitating robust detection methods. Through empirical analysis and statistical modeling, the study quantifies the growth trajectory of IoT-based ransomware infections, shedding light on the evolving threat landscape. Additionally, an innovative detection approach leveraging C&C blacklisting techniques is introduced and evaluated for its efficacy in mitigating Cryptowall Ransomware attacks.

Francesco Mercaldo, Isco Alarci, TOR Traffic Analysis and Identification: A Study on The Onion Router (TOR) :2017

This study delves into the analysis of The Onion Router (TOR) traffic, aiming to identify and understand patterns within the TOR network. By analyzing the traffic generated by TOR users, the study sheds light on the characteristics and behaviors of TOR users, as well as the types of activities conducted over the network. Through empirical analysis and data-driven insights, the research contributes to a deeper understanding of TOR's usage and its implications for privacy, security, and network monitoring.

PROPOSED APPROACH

In today's rapidly evolving digital landscape, technology-enabled enterprises face an increasing array of cyber threats, including creative and sophisticated attacks that prey on vulnerabilities before they can be identified and rectified. In response to an ever-increasing danger situation, the proposed solution uses machine learning, an innovative method that has become an essential part of modern cybersecurity efforts. Machine learning is unparalleled in its ability to analyze massive volumes of data and identify intricate patterns and connections that may go unnoticed by traditional detection methods. The technology uses supervised machine learning methods to build a robust model that can detect minute features in network traffic that may indicate zero-day and new attacks.

Machine learning's ability to adapt and evolve in real-time is a key feature of the proposed system's approach to threat recognition and response. Unlike static rule-based systems that rely on predetermined signatures or patterns, the proposed system learns from observed data and iteratively refines its algorithms to stay ahead of new threats. The system's proactive detection and mitigation of developing cyber threats shields corporate networks against possible breaches and data compromises due to its dynamic adaption. By using previously learned patterns to classify unknown network data, the system can distinguish between normal network activity and anomalous behavior associated with hostile intrusions.

Furthermore, the proposed system may enhance threat detection and response capabilities with unprecedented accuracy and efficacy via the integration of supervised machine learning algorithms. The technology analyzes historical data and applies advanced algorithms such as Principal Component Analysis (PCA) and Random Forest to acquire deeper insights into the complex interaction of components driving cyber hazards. This extensive research has allowed the system to quickly detect and neutralize new threats, lowering the risk of data breaches and ensuring that tech-enabled enterprises can remain flexible in the face of constantly evolving cyberthreats. All things considered, the proposed approach is a significant development in cybersecurity as it leverages the revolutionary potential of machine learning to fortify organizational defenses and protect digital infrastructure integrity from ongoing malware threats.

SYSTEM ARCHITECTURE

The system architecture of this projects shows the flow of the control through the system. It also shows the hardware and the software required for the execution of the program. The architecture diagram is as follows :

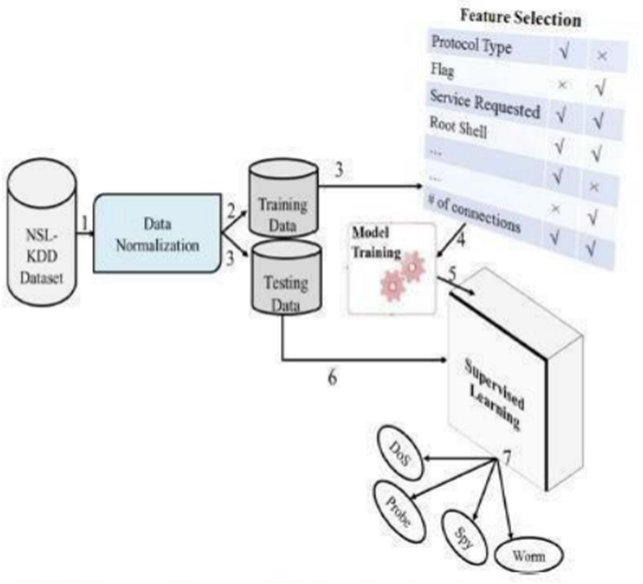


Fig 1. System architecture

Using the NSL-KDD dataset, this project offers a broad flowchart that illustrates a suggested machine learning classification approach for intrusion detection. The NSL-KDD dataset, a well-known benchmark dataset often used in network intrusion detection research, is where the study first begins. Numerous records of network traffic in this dataset have been classified as either typical or indicative of different kinds of assaults. The data required to train and assess the intrusion detection model is provided by the NSL-KDD dataset, which serves as the system's initial input. To make sure that the training process is based on a variety of genuine network traffic characteristics encompassing a broad spectrum of harmful and lawful operations, the system makes use of unique datasets.

Data rectification is a crucial next step after obtaining the NSL-KDD data collection. This stage is crucial because it transforms the unprocessed data into a format that is common and appropriate for machine learning algorithms. To make sure that the data is consistent and comparable throughout a range of values, data normalization include extending numerical values to a normal range, encoding random variables, and addressing missing values. This preprocessing stage is crucial because it mitigates size and shape distribution problems that might impair machine learning model performance. By lessening the bias brought on by the various components, data normalization increases the learning algorithm's convergence rate and the model's overall correctness.

The data is divided into training and test portions once it has been standardized. The machine learning model is taught using the training data part, and its performance is assessed using the test data section. This division makes it possible to test the model on unseen data and offers an objective evaluation of how well it generalizes to novel and unobserved network traffic. To enhance the model's performance, this procedure involves reprocessing the training data and identifying important characteristics. Finding and keeping the most crucial elements in a data collection that enable precise categorization is known as feature selection. Smaller categories are omitted from this system and characteristics like protocol type, flag, service request, root shell, and number of connections are chosen. By simplifying the model, accelerating the training process, and concentrating on more precise data, this phase may improve performance.

The classifier is constructed using a supervised learning method, and the training data comprises the characteristics used as input for the model training stage. In order to find patterns and connections between features and their labels, the model learns from the labeled training data and modifies its internal parameters in this stage. The objective is to create a model that can precisely categorize network data into common attack categories and other attack types based on these instances. Incremental optimization is used in this method to minimize training data inaccuracy while continuously enhancing the model's predictive power. Using test data, the model's performance is carefully assessed after training. Measures including precision, accuracy, recall, and F1 score are used in this assessment to gauge how well the model fits fresh data and to gauge categorization efficiency.

Ultimately, real-time intrusion detection algorithms that have been trained are used to categorize incoming network data into several assault groups. These consist of malware, spyware, denial of service (DDoS), and browser assaults. Each network marketing example is assigned by the teacher to one of these categories or marked as common using the examples that were covered throughout the training. The system can recognize and react to different types of cyberattacks thanks to this categorization process, which serves as a protective measure for network security. The whole procedure offers a methodical way to create an efficient intrusion detection system and use machine learning to defend against a variety of cyberattacks, from feature selection and data manipulation to model training and real-time classification.

TECHNOLOGIES USED

Python and Anaconda play a key part in the suggested IDS program's mix of cutting-edge technologies, which guarantee reliable and efficient cyber threat detection and mitigation. Python is the best programming language for this because of its ease of use, readability, and large ecosystem. A vast array of tools and frameworks, like TensorFlow, Keras, and scikit-Learn, make it possible to create sophisticated machine learning algorithms, which are the cornerstone of anomaly detection. These algorithms enable intrusion detection systems (IDSs) to learn from previous network traffic data, spot patterns that deviate from the norm, and find clusters that could be signs of cyberthreats.

The open-source Python and R distribution Anaconda is crucial for controlling the dependencies and environment of your project. It guarantees that libraries and tools are supported and up to date while streamlining package management and deployment. Numerous pre-installed programs helpful for data science and machine learning are included with Anaconda. These include NumPy for numerical computing, Pandas for data processing, and Matplotlib or Seaborn for data visualization. Jupyter Notebook, an integrated programming environment (IDE) from Anaconda, offers a user-friendly platform for writing and executing code, displaying data, and documenting the development process. It also makes it easier to construct and test machine learning models interactively.

The application leverages tools like Wireshark, which records real-time network traffic data, for data collecting and preprocessing. This raw data is processed by a Python script that eliminates noise, modifies the format, and extracts the required features. In this process, libraries like scikit-learn and pandas are crucial because they make it possible to handle and modify huge datasets. In order to extract significant characteristics from the raw data that are appropriate for precise machine learning analysis, feature extraction is carried out. The anomaly detection module makes use of supervised learning methods like support vector machines and decision trees, which need labeled data in order to train the model. To identify people without label data, unsupervised learning techniques like principal component analysis (PCA) and k-means clustering are also used. Combining the two methods allows machine learning to increase detection accuracy by guiding the learning process with a minimal quantity of labeled data.

Respond and Deny automates activities in response to threats discovered, resulting in a considerable reduction in response time. By integrating with your current security architecture using libraries and APIs like Pycurl for HTTP requests and Paramiko for SSH connections, the module allows rate limitation, automated blocking of dangerous IP addresses, and forwarding traffic to a honeypot for further study. Reports produced by the system and warnings facilitate manual action.

A crucial component is data visualization, which is carried out with the aid of Python packages like Matplotlib, Seaborn, and Plotly. These technologies provide dynamic, detailed dashboards that offer in-the-moment information into identified threats, network activity, and overall security posture. Managers can detect and address hazards more quickly when trends, patterns, and anomalies are easily recognized via the use of visualization.

In order to maintain the indexing modules abreast of the most recent known assault patterns, the software further incorporates threat intelligence technologies. In order to remain successful against emerging threats, this integration enables the IDS to update its attack signature database on a frequent basis. IDS projects gain from a strong, adaptable, and productive development environment when Python and Anaconda are used. This combination enables the system to maintain a high degree of network security, adapt to changing cyberthreats, and provide administrators the knowledge and resources they need to safeguard their networks. These technologies work together to provide a comprehensive, cutting-edge intrusion detection system (IDS) that can defend vital Internet services from a variety of online attacks.

DESIGN AND IMPLEMENTATION

USE A CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor.

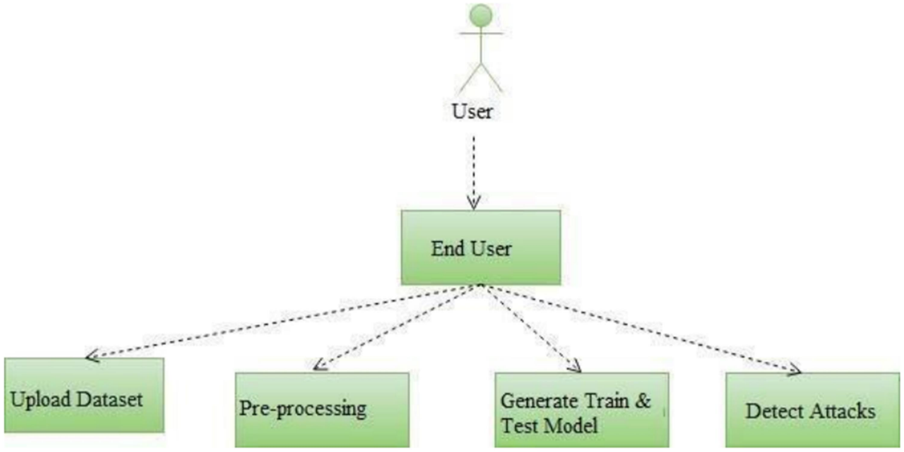


Fig 2. Use Case Diagram...

CLASS DIAGRAM

In software package engineering, a class diagram at intervals of the UnifiedModelling
Language (UML) can be the fashion of a static structure diagram that describes the
structure of a system by showing the system`s categories, attributes, operations,and additionally the relationships among objections.

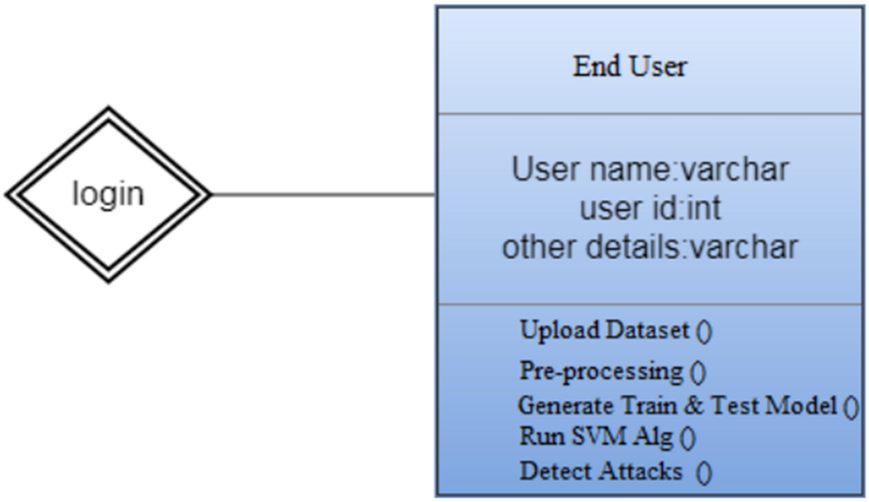


Fig 3. Class Diagram…

SEQUENCE DIAGRAM

A sequence diagram or system sequence diagram (SSD) shows object interactions arranged in time sequence within the field of software package engineering. It depicts the objects involved within the state of affairs and also the sequence of messages changed between the objects required to hold out the practicality of the state of affairs.

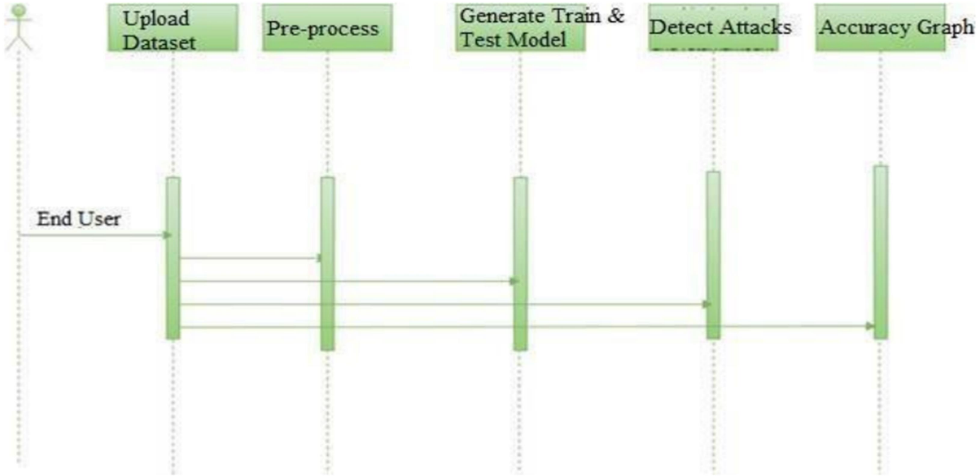


Fig 4. Sequence Diagram…

RESULTS

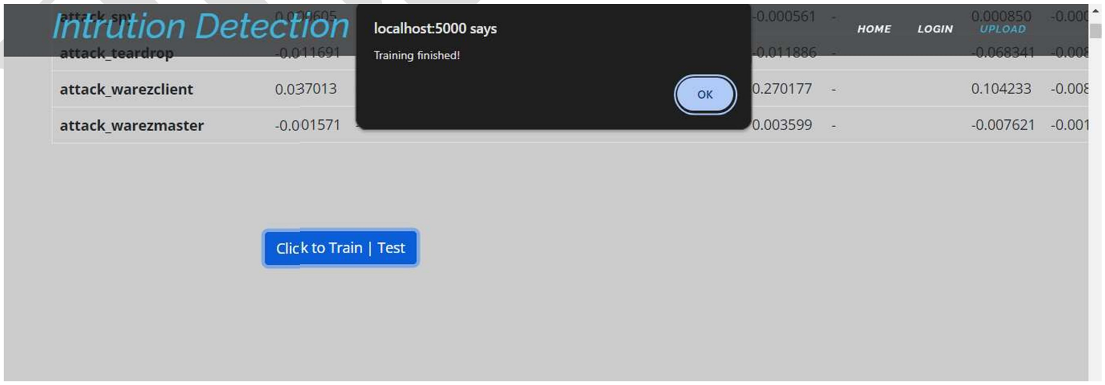
Twinter is the most ordinarily used library for developing GUI(Graphical user Interface) in Python. It`s a regular Python interface to the Tk GUI toolkit shipped with Python. As Tk and Tkinter square measure out there on most OS platforms as to well as on the Windows system, developing GUI applications with Tkinterbecomes the fastest and best.

Main Window Functionalities :

Intrusion Detection System
HOME
LOGIN
Intrusion Detection System using Machine Learning Techniques

fig 5. Main Window

Training Functionalities:



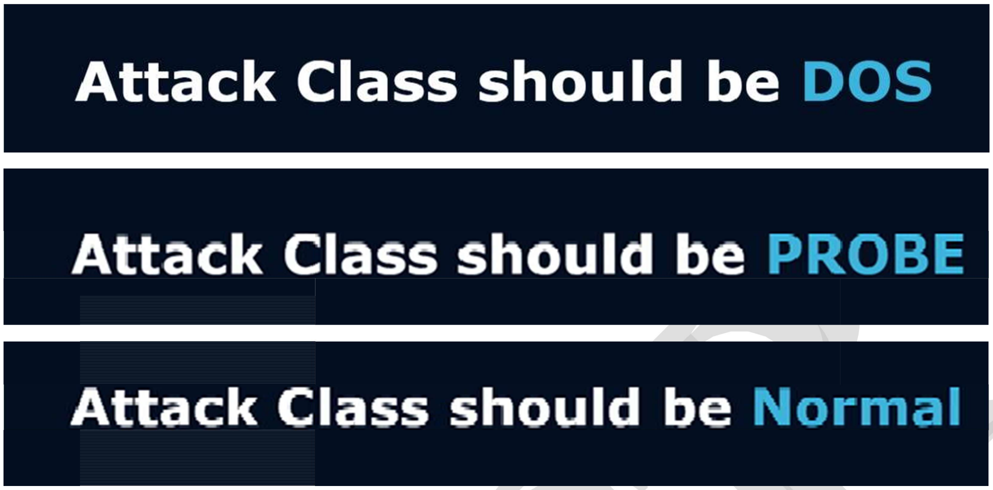
localhost:5000 says

Training finished!

attack_spy	0.009605	-0.000561	-	0.000850	-0.00
attack_teardrop	-0.011691	-0.011886	-	-0.068341	-0.008
attack_warezclient	0.037013	0.270177	-	0.104233	-0.008
attack_warezmaster	-0.001571	0.003599	-	-0.007621	-0.001

fig 6. Training and Testing

Result Functionalities:



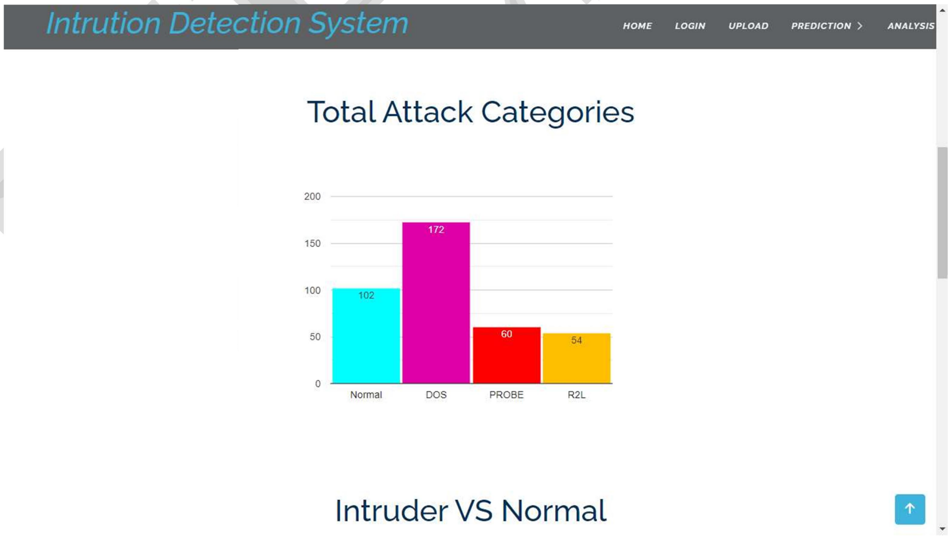
Attack Class should be DOS

Attack Class should be PROBE

Attack Class should be Normal

fig 7. Showing Results of search

Analysis of Predication:



Total Attack Categories

Intruder VS Normal

fig 8. Analysis of Prediction

CONCLUSION:

In our project, we investigated various machine learning models using various algorithms and signal selection techniques to increase the effectiveness of intrusion detection systems (IDS). Through experiments and analysis, we find the best model using Random Forest algorithm and closed segment selection methods. The model achieved an excellent detection rate of 97.02%, showing excellent performance in correctly classifying network traffic and identifying potential threats.

These results have significant implications for the development of intrusion detection systems that detect both sophisticated and novel attacks. Current IDSs struggle to detect new or outdated attacks due to high hit rates. However, our project addresses these challenges by using machine learning algorithms to improve detection accuracy and reduce false positives. By achieving high detection rates, our project demonstrates the power of machine learning approaches to improve the efficiency of intrusion detection systems.

In the future, our project will become the basis of research and development in the field of intrusion detection. The performance and adaptability of IDS can be continuously improved by improving and optimizing the machine learning models and exploring other algorithms and selection methods. Ultimately, our goal is to create an intrusion detection system capable of protecting network infrastructure from a wide range of cyber threats, including sophisticated and modern attacks.

REFERENCE:

1. H. Song, M. J. Lynch, and J. K. Cochran, “A macro-social exploratory analysis of the rate of interstate cyber-victimization,” American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2016.
2. P. Alaei and F. Noorbehbahani, “Incremental anomaly-based intrusion detection system using limited labeled data,” in Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178– 184.
3. M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, “Modeling and implementation approach to evaluate the intrusion detection system,” in International Conference on Networked Systems, 2015, pp. 513–517.
4. M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516–524, 2010.
5. A. S. Ashoor and S. Gore, “Importance of intrusion detection system (IDS),” International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.
6. M. Zamani and M. Movahedi, “Machine learning techniques for intrusion detection,” arXiv preprint arXiv:1312.2177, 2013.
7. N. Chakraborty, “Intrusion detection system and intrusion prevention system: A comparative study,” International Journal of Computing and Business Research (IJCBR) ISSN (Online), pp. 2229– 6166, 2013
8. A. S. Awad and J. M. El-Sayed, “Intrusion detection system based on machine learning techniques: A comprehensive survey,” Journal of Network and Computer Applications, vol. 103, pp. 1-25, 2018.
9. K. Revathi and K. Thangadurai, “A survey on machine learning techniques for intrusion detection

systems," International Journal of Computer Applications, vol. 181, no. 12, pp. 6-9, 2018.

10. S. Chien, J. W. Huang, and K. J. Lin, "An enhanced intrusion detection system using machine learning algorithms," Applied Sciences, vol. 8, no. 5, pp. 691, 2018.
11. G. Arora, N. Goyal, and S. Bansal, "A survey on machine learning techniques for intrusion detection system," International Journal of Computer Applications, vol. 165, no. 4, pp. 9-12, 2017.
12. Y. D. Bello and N. A. Khan, "A review on machine learning approaches for intrusion detection system," International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, pp. 168-175, 2018.