Abstract

The project focuses on creating IDS (short for intrusion detection systems) by the use of machine learning and deep learning techniques to enhance network security . IDS provides detection and mitigation of attacks in real time by detection of abnormal patterns in the network traffic or by recognizing the signature of the given attack . The system developed here uses KDD 1999 dataset , for network intrusion detection , containing a wide range of network traffic features . This model uses various machine learning models like SVM(Support Vector Machine) and XGBoost for advanced classification and regression tasks .Various data preprocessing techniques such as PCA and RobustScaler for scaling , Decision trees ,KNearest neighbors ,etc, are used for classification of data. The models are evaluated using accuracy , precision ,F1 and recall score to assess their ability to detect different intrusions . Through the use of ML and DL this project helps demonstrate the usefulness of ML and DL in the areas of cyber threats and detection of previously unseen attack patterns by doing scalable , accurate and real-time detection proving to be more useful than the traditional signature based detection methods.

Introduction

A) Problem Statement:

As we all know, our world is shifting further into the digital realm, a shift that brings with it an ever-increasing danger of attack, be it hacking, malware or denial-of-service (DoS) type attacks. Most of these attacks are prone to causing data loss and tarnishing images which translates to extreme amounts of cash being lost. This has created a scenario where both businesses and individuals have to find methods of protecting their computer networks, devices and files from illegal access and other destructive attacks. It is very important, in the context of computer networks and systems, to have good solutions that can be used in preventing or detecting any system-intrusion. This has resulted in the growing importance of Intrusion detection Systems (IDS). IDS stands for intrusion detection system, which is broadly a security system implemented to observe a network or system's actions for illicit behaviours or breaches of policies. In this way, IDS is a very important security measure for prevention planning, where attacks are anticipated and measures taken to deal with them so as to reduce their destructive nature.

Key features of the IDS include:

• Real-time Monitoring: They are the real-time active threats that are detected and countermeasures are made, with the help of continuous network activity.
• Advanced Analytics: Preprocessing in this case is done through unsupervised machine learning where the end helps in modeling network traffic with a view of escalating the likelihood of identifying security breach attempts.
• Comprehensive Alert System: Instant security breaches once they are spotted and users are informed on time in order to prevent or take appropriate corrective action.
• Integration with Existing Security Infrastructure: Additional safety devices and programs can be added without problems to complement the total security plan.

The importance of IDS for instance is very high. A report set up recently suggested that with an approximate per capita GDP and proved a data breach could lead to damages of about $3.9 million A study also suggested that the losses from cybercrime could be around $6 trillion by 202. Also, owing to the heinous rise in the cases of different cyber related criminals, it has become pertinent that several measures should be adopted by institutions and individuals that will safeguard computer networks and systems against intrusive hazards.

B) Objectives:

The objective of this project is therefore to develop a prototype Intrusion Detection System (IDS) based on machine learning techniques to be able to detect anomalous network traffic.

• Develop & Deploy an ID: Intrusion Detection System.
• Inclusion of KDD Cup 1999 Dataset in the model training phase.
• Data preparations to be done in preparation of modeling to try and have the model perform better.
• Assess the performance of both classifiers: SVM and XGBoost in regard to their efficiency.
• Review models based on conventional models evaluation metrics.
• Proving the potential of the proposed system in terms of Scalability and Adaptability to perform real time Intrusion Detection.
• Practical IDS Models as papers presented to the Cybersecurity Research proceeding within the field. With the help of constant network activity monitoring.
• Advanced Analytics: Unsupervised machine learning techniques are applied to model network traffic, improving the identification of security breaches attempts.
• Comprehensive Alert System: Security breaches as soon as they are discovered and users are notified promptly in order to ensure prompt repair and prevention.
• Integration with Existing Security Infrastructure: Other security tools and systems can be integrated easily to complete the overall security concept.

C) Significance:

This project is important as it shows how the efficiency and effectiveness of intrusion detection systems can be enhanced through the application of machine learning based algorithms. Particularly, instead of relying on pre-defined signatures like other types of IDS, the proposed system will be able to learn new attack patterns from the network traffic. The XGBoost and SVM model are used in the IDS which enhances the cross-sectional model hence it is able to execute a huge volume of information by detecting intrusion instantaneously which makes it applicable in addressing current challenges in network security. The lessons learnt in this project will help in the provision of ideas on how better and efficient informed and robust cyber security systems can be developed in the future frameworks.

Literature Review

Recent research suggests that network intrusion detection is one of the important mechanisms that can help to reduce the levels of security threats that have emerged. The task of type of intrusion classification is presented in a new light, after the deep learning technology was added, detection rate increased tremendously. The results of the experiments using KDD demonstrate that the aforementioned methods execution indeed is more effective than standard approaches of machine learning (Kumar et al., 2020, Li et a1., 2020). There is also a paper analyzing the intrusion detection task using RNNs including a comparison of various classification tasks, learning rate, neuron count (Zhang et al., 2020). This speaks volumes on the growing role of deep learning in designing solutions for network security problems.

The literature reviewed emphasised the role of anomalies detection, features selection and hybrid approaches in IDS research. Studies have been devoted for example to the effective methods of anomaly detection including statistical analysis, clustering and even neural networks (Garcia-Teodoro et al., 2019). Studies were also carried out in searching feature selection and engineering techniques to achieve improvements in deployed IDS detection performance (Zhang et al., 2020). There have also been suggested Hybrid IDSs, which combine both signature and anomaly detection methods in order to take the advantages of the two (Kumar et al., 2020). These developments have improved the accuracy and the scalability of IDS detection but issues like false alarms, missed detections or evasion techniques are yet to be resolved.

Dataset Description: KDD Cup 1999 (or similar data set on Intrusion Detection)

Apart from the collection of network traffic data, the KDD Cup 1999 dataset is also part of attack intrusion detection based data which acts as an evaluation criterion. It was obtained from the darpa intrusion detection system evaluation data set, and it includes data that has been classified as normal from an attack. This dataset has good coverage of different network features as well as different attack types thus it is very good for machine learning model training and intrusion detection model testing.

Usage of the Dataset in the Project:-

Feature Selection: There are various main purposes for structural analysis of the database, among which we can mention taking the network traffic features from the duration, protocol and content for modeling and training machine learning models in the project.

Some of such key features are:

1. Basic features: protocol type, service, flag, and duration
2. Content features: src_bytes, dst_bytes, and num_failed_logins
3. Traffic features: count, srv_count, and dst_host_count
4. Host features: dst_host_srv_count and dst_host_same_srv_rate

The dataset is labeled with one of five classes:

1. Normal: normal network traffic
2. Probe: probing attacks
3. DoS: denial-of-service attacks
4. U2R: user-to-root attacks
5. R2L: remote-to-local attacks

Data Preprocessing:

- **Data cleaning:** The mean substitution technique was used to handle missing values for numerical features, while the mode substitution technique was used for categorical features.
- **Feature scaling:** The functionalities were scaled using the Min-Max Scaler so that the range of values in all of the properties is similar and one property does not dominate others.
- **Feature selection:** The Recursive Feature Elimination (RFE) algorithm was applied to rank the predictors and only the top 20 were retained to reduce the dimensionality of the data so that the model performs better.
- **Data normalization:** The data was standardized using the Standard scaler that had zero mean and unit variance, enhancing the performance of machine learning algorithms.
- **Data splitting:** The entire dataset was divided into two groups; training data (80%) for model training and testing data (20%) for evaluation of proposed IDS model.
- **Attack Detection:** The model objectives are training them to classify the traffic as either normal or attack while focusing on detecting DOS, Probe, R2L, and U2R types of attacks.