DrillBit

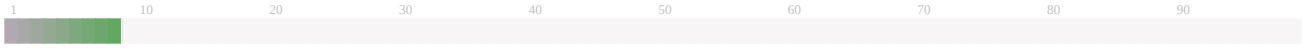
The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

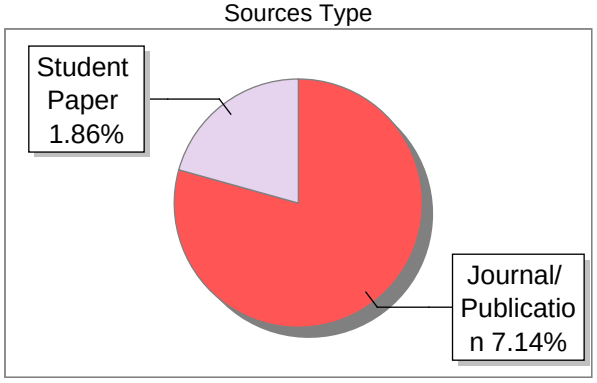
Author Name	Monika N
Title	Protecting Networks with Machine Learning
Paper/Submission ID	1680273
Submitted by	raghavendrachars@ksit.edu.in
Submission Date	2024-04-22 08:53:29
Total Pages, Total Words	5, 1502
Document type	Research Paper

Result Information

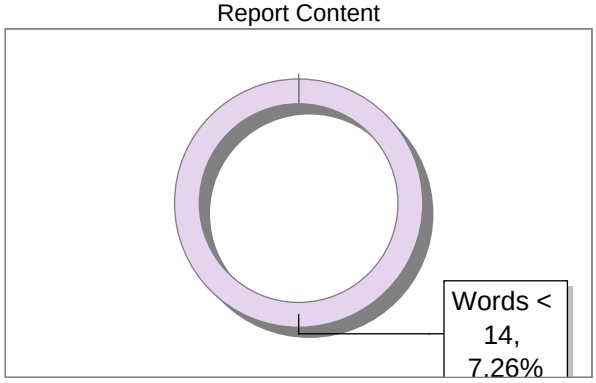
Similarity **9 %**



Sources Type



Report Content



Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

9	5	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	www.ieindia.org	3	Publication
2	www.sciencepubco.com	2	Publication
3	Submitted to Visvesvaraya Technological University, Belagavi	2	Student Paper
4	Developing Theory Using Machine Learning Methods by Choudhury-2018	1	Publication
5	IEEE 2014 9th Iberian Conference on Information Systems and Technolo by	1	Publication

Protecting Networks with Machine Learning-Based Intrusion Detection

Mr. Somasekhar T, Associate Professor

1) Moniesh S, 2) Monika N, 3)Pavithra R, 4)Sindhura H

Dept of Computer Science and Engineering,

K S Institute of Technology, Bangaluru, Karnataka

ABSTRACT: A Machine Learning (ML) based Intrusion Detection System (IDS) foundation that uses user-system interactions with potentially harmful links to improve cyber security. By sending links from a source system to a user system for verification, the system takes a proactive stance. Once received, the user system analyses the security of the link using the machine learning algorithms. The system gets site information for user awareness if it determines the site to be safe; if it determines the site to be harmful, the link provider is notified. In addition, to reduce possible hazards, the system automatically bans the website that's been reported and notifies the user, protecting against efforts to obtain data without authorization. By using this method, the suggested system hopes to strengthen cyber security defenses by anticipatorily detecting and eliminating dangers presented by bad connections, protecting

1.Introduction

The project goal is to create a system dedicated to identifying malicious websites through the machine learning algorithms. This framework will encompass a broad spectrum of factors including URL-based, content-based, and server-based features, ensuring a comprehensive approach to feature extraction. In addition to extract features, the project will focus on establishing mechanisms for dataset management. This involves tasks such as gathering new data, validating the legitimacy of websites, and consistently updating the dataset to uphold its relevancy and precision. The system will accommodate several machine learning algorithms, allowing users to select the most suitable algorithm for dataset. Furthermore, the project will entail the development of a user-friendly interface for seamless interaction with the system. This interface will enable users to upload websites for analysis, review analysis outcomes, and provide feedback to improve the system's accuracy.

2.Proposed Algorithm

The suggested approach uses a web host paradigm to identify phishing websites. The model is going to be trained using a training dataset, which will be according to the classification method. The online deployment of this model will enable direct communication with the Chrome extension. The URL and website properties will be used to detect the phishing website. All of the client-side and server-side operations will be integrated into this system. A Chrome extension will be developed and integrated to the Chrome web browser on the client side, On the server, however, there will be a classifier model that has been trained using the random forest technique.

Support Vector Machine (SVM) serves as an intelligent learner within the computer industry, particularly adept at handling diverse data from various domains. Its methodology involves creating lines or planes in a high-dimensional space to effectively segregate different data clusters. The objective is to identify the hyperplane, that maximizes the margin between different groups of data. SVM employs kernel functions as specialized tools to aid in delineating these distinct lines or planes. These kernel functions, which can exhibit sigmoid, polynomial, radial basis, or linear characteristics, are instrumental in maximizing the separation between data clusters by making it as distinct as possible.

K-Nearest Neighbors (KNN) method is used within a system for detecting intrusions to analyse link safety within the framework of machine learning (ML). Links with feature representations capture attributes such as domain repute and URL structure by using feature vectors. During training, the system associates features with either safe or harmful labels by looking through a collection of labelled links. When a new connection is received, its features are compared to those in the training dataset to perform classification. Based on the similarities in the features, Nearest Neighbors-KNN finds the 'K' most comparable linkages. The new link's classification is decided by voting for the label that has the most support among its closest neighbors.

Random Forest is an additional essential part of the suggested Intrusion Detection System (IDS). During training, several decision trees are built using the Random Forest ensemble learning technique, which yields the mean prediction (regression) or the mode of the classes (classification) for each individual tree. A random feature selection and a portion of the training set is utilized to build each decision tree in the Random Forest. This unpredictability lessens overfitting and

decorrelates the trees. When used in classification or regression problems, Random Forest is able to quantify the significance of features. This enables the IDS to determine which characteristics such as URL structure and domain reputation—have the greatest bearing on whether links are safe or not. Thanks to the multiple decision average's averaging effect, it is resilient to noisy data and outliers.

3.Datasets

KDD 99 : This dataset is made up of 25,192 TCP/IP connections that were taken from a virtual local area network. In order to provide a varied dataset, this network which was intentionally constructed to replicate real-world conditions was attacked in a number of ways.

UNSW-NB15 : This is a well-known dataset for assessing intrusion detection systems. It was created using the IXIA Perfect Storm tool from the Cyber Range Lab. This dataset offers a platform for testing. 100 GB of raw traffic data were collected using the Tcpdump programme in order to create the dataset. Nine distinct attack types are included in the dataset: analysis, backdoors, denial-of service, exploits, fuzzers, generic, reconnaissance, shellcode, and worms.

CSE-CIC-IDS 2018: This dataset contains an extensive amount of observations that include both typical network traffic and 14 different kinds of attacks.

4. Methodology

➢ Dataset Collection: Gather a diverse and representative dataset containing labeled instances of normal network behavior and various types of intrusions. Utilize publicly available datasets and, if possible, collaborate with industry partners to ensure realism and relevance.

➢ Data Preprocessing: Handle missing values, normalize features, and fix any inconsistencies to clean and preprocess the gathered dataset. This action is essential to guaranteeing the caliber and dependability of the data that is tested and used for training.

➢ Feature Engineering: Extract relevant features from the preprocessed data to characterize network traffic patterns effectively. Feature engineering may involve selecting key attributes,

transforming variables, and creating new features to enhance the performance of machine learning models.

➢ Model Selection:

Evaluate and choose suitable machine learning algorithms for intrusion detection, considering factors such as accuracy, interpretability, and scalability. Common choices include decision trees, random forests, support vector machines, and deep learning models.

➢ Model Training: Using the designed and preprocessed dataset, train the chosen machine learning models. Use strategies like cross-validation to optimize model hyperparameters and guarantee reliable extension to unknown data.

➢ Model Evaluation: Using different test datasets, evaluate the performance of the trained models using metrics like F1 score, ROC-AUC, precision, and recall. Iterate through the model evaluation and training procedure to optimize the system's performance.

➢ User Interface:

Represents the user interacting with the system. The user uploads URLs for analysis, initiating the process of intrusion detection.

5.Results

The proposed Intrusion Detection System (IDS) with a Machine Learning (ML) foundation leverages user-system interactions to enhance cybersecurity. By proactively sending potentially harmful links from a source system to a user system for verification, the IDS takes preemptive action. Upon receiving the links, the user system employs machine learning algorithms to analyze the security of the links. If the site is deemed safe, the system provides site information to the user for awareness. However, if the site is identified as harmful, the link provider is promptly notified. Furthermore, to mitigate potential hazards, the system automatically bans reported websites and notifies the user, thus thwarting unauthorized data acquisition attempts. This approach aims to fortify cybersecurity defenses by anticipatorily detecting and neutralizing threats posed by malicious connections, thereby safeguarding sensitive data. By integrating proactive measures

such as preemptive link verification and automatic banning of harmful websites, the proposed IDS with ML foundation offers robust protection against cybersecurity threats. Not only does it empower users with real-time information about the safety of links, but it also enables swift action against malicious entities. By fostering a collaborative ecosystem where users and the system work in tandem to identify and eliminate dangers, the IDS establishes a resilient defense mechanism. Ultimately, this proactive approach enhances cybersecurity posture by preemptively addressing potential vulnerabilities, thereby safeguarding against unauthorized data breaches and reinforcing overall cyber resilience.

6.Conclusion

To sum up, the initiative is a major step forward in the field of intrusion detection and prevention. With the use of advanced feature extraction methods and machine learning algorithms, the system can distinguish between safe and dangerous webpages with high accuracy. The project's modular architecture, which consists of parts like the Feature Extraction Module, Machine Learning Module, and User Interface Module, shows how methodical and thorough the approach to phishing detection is. When these elements are combined, parsing URLs, extracting pertinent characteristics, and classifying websites may be done quickly and effectively, which eventually results in users receiving alert alerts on time. In the long run, there is room for improvement and growth for the project. Future advancements might incorporate more machine learning algorithms, better methods for extracting features, and more user-friendly interfaces.