https://doi.org/10.46335/IJIES.2024.9.7.17 **e-ISSN: 2456-3463**
Vol. 9, No. 7, 2024, PP. 84 -87
International Journal of Innovations in Engineering and Science, www.ijies.net

Enhancing Cybersecurity through Machine Learning: A Review of Malicious Website Detection Methods

Nityay Kherde, Anushka Kale, Govind Kurup , Asmit Meshram ,Prof Leena Mandurakar

Dept of Computer Science & Engineering (Data Science), St. Vincent Pallotti College of Engineering and Technology, Nagpur

asmitmeshram123@gmail.com

***Received on**: 5 May,2024* ***Revised on:** 30 June,2024* ***Published on**: 03 July ,2024*

Abstract- *This paper presents a review of machine learning-based approaches for identifying and classifying malicious websites. Various studies have tackled this challenge by extracting features from URLs and website content to train machine learning models. Methods include lexical, host-based, and content-based feature extraction, as well as domain and Alexa-based analyses. Different algorithms, including Gradient Boosting, Random Forests, and Neural Networks, have been utilized, achieving high accuracy in classifying malicious URLs. Real-time concept drift detection and retraining models have also been proposed to adapt to evolving cyber threats. The paper concludes by highlighting the importance of feature extraction in bolstering cybersecurity.*

Keywords: Malicious websites, Machine learning, Feature extraction, Cybersecurity, Classification algorithms.

INTRODUCTION

In today's digital world, keeping our online activities safe from cyber threats is more important than ever. One major threat comes from malicious websites, which can do harmful things like spreading viruses, stealing personal information, or tricking people into giving away money. With so many websites out there, it's hard to tell the good ones from the bad ones. That's where machine learning comes in. It's a way for computers to learn from data and make decisions without being explicitly programmed. Researchers have been using machine learning to help identify and classify malicious websites. By looking at things like the web address (URL) and other features, they can train computer algorithms to recognize signs of danger.

In this paper, we'll explore different studies that have tackled the challenge of spotting malicious websites using machine learning. We'll look at what methods they used, what they found, and what challenges they faced. By understanding these efforts, we can learn more about how to keep ourselves safe online in an ever-changing digital landscape

As technology continues to evolve, so do the threats we face online. By studying how researchers are using machine learning to combat malicious websites, we can better understand how to protect ourselves and our data from cyber attacks.

II- LITETURE REVIEW

Feature Extraction

This is the overall review about the feature exatraction.Various methodologies have been proposed in the research domain to extract features aimed at detecting malicious websites. These methodologies are crucial in discerning between benign and malicious websites by identifying

https://doi.org/10.46335/IJIES.2024.9.7.17 e-ISSN: 2456-3463
Vol. 9, No. 7, 2024, PP. 84 -87
International Journal of Innovations in Engineering and Science, www.ijies.net

pertinent characteristics within URL strings and website content.

A recurring approach among the studies involves the categorization of features into distinct groups. These categorizations span a range of dimensions including lexical, host-based, and content-based factors, as well as domain-based, Alexa-based, and obfuscation technique-based criteria. Additionally, some researches employ broader classifications such as alpha-numeric data analysis, keyword analysis, security analysis, domain identity analysis, and rank-based analysis. Such categorizations facilitate a systematic exploration of the diverse aspects of website URLs and content.

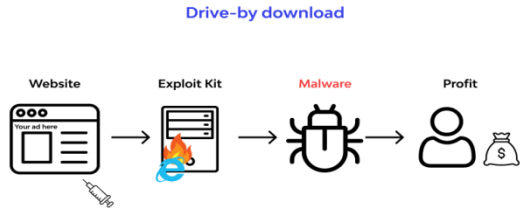

Fig 1: Drive by download

The above fig shows the Drive by download [6]. The techniques utilized for feature extraction exhibit variability across the researches. For instance, certain studies emphasize the extraction of lexical, host-based, and content-based attributes such as URL length, host length, HTTPS presence, and JavaScript code analysis. Others delve into extracting features from domain-based sources through WHOIS queries, Alexa rankings, and JavaScript obfuscation techniques. In contrast, one research employs sophisticated methods like Linear Discriminant Analysis and Principal Component Analysis to reduce dimensionality, thereby expediting the processing of features for detecting malicious websites. Another study adopts a comprehensive approach, amalgamating software codes, online processing, and manual efforts to amass a diverse set of 133 features related to website URLs.[1][2][3][4]

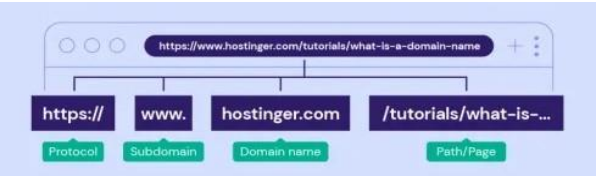

Fig 2: URL Components

The Fig 2 shows the Basic components of the URL [5]. Each research contribution offers distinctive insights into feature extraction methodologies for malicious website detection. Certain papers provide detailed explanations elucidating the significance of specific feature types in identifying malicious behaviour. Others focus on technical intricacies and broader feature categorizations. Integrating insights from these diverse researches holds promise for enhancing the accuracy and efficiency of malicious website detection systems.

In conclusion, feature extraction stands as a pivotal component in the realm of detecting malicious websites. Through systematic extraction and analysis of various features from URL strings and website content, these methodologies significantly contribute to the development of robust machine learning models for bolstering cybersecurity applications.

Firm size, firm age, firm risk, profitability, leverage, current ratio, and dividend pay-out to gauge the effect.

In the following section, we provide an overview of the selected papers chosen for review. These papers delve into the realm of machine learning classification techniques, encompassing a diverse array of algorithms and methodologies for feature extraction and data categorization. Through a concise examination of each paper's contributions and findings, we aim to elucidate the advancements and insights offered within the field of classification research.

[1] The paper presents a detailed review of a machine learning-based approach for classifying malicious and benign websites using URL features, including the training process on curated datasets, real-time concept drift detection, and the effectiveness of the Gradient Boosting Algorithm with 96.4% accuracy.

The techniques used for the model in this research include feature extraction using Lexical, Host-Based, and Content-Based features, supervised learning with algorithms like Random Forests, Gradient Boosted Trees, and Feed Forward Neural Networks, and real-time concept drift detection with retraining of the model. Neural Networks are also utilized for detecting malicious websites.

[2] It discusses the common cybersecurity threat of malicious URLs, the need for reliable solutions to classify and identify them, the use of machine learning classifiers for classification, and the importance of balanced data for accuracy.
And it primarily addresses the identification of malicious URLs through binary classification employing a range of machine learning classifiers, including Logistic Regression, Stochastic Gradient Descent, Random Forest, Support Vector Machine, Naïve Bayes, K-Nearest Neighbours, and Decision Tree. It highlights the significance of distinguishing between benign and malicious URLs, discussing the methodologies involved in classification, such as encoding categorical features and dividing data for training and testing. The study conducts a comparative analysis of the classifiers' performance and underscores the potential for enhancing accuracy through training on more balanced datasets.

[3] The techniques used for the model in this research include ANOVA test and the XGBoost algorithm for feature selection and model training, with a focus on improving accuracy and stability in classifying malicious URLs, which achieved an accuracy percentage of 99.98%. The dataset used in the experiments consists of both benign and malicious URLs, ensuring balance.
The original 41 features are narrowed down to the top 17 most significant ones. Various features are extracted, including domain-based, Alexa-based, and obfuscation technique-based features, totaling 41 dimensions. The paper highlights the importance of feature selection and reduction to enhance model efficiency and reduce training time.

It Offers valuable insights into the realm of cybersecurity, showcasing how machine learning techniques can be harnessed to tackle the growing threat of malicious URLs. The meticulous approach to dataset curation, feature extraction, and model optimization demonstrates a comprehensive understanding of the subject matter.

[4] The research in Mustafa Aydin, Kemal Bicakci, Ismail Butun, Nazife Baykal (2020) extensively uses data mining algorithms, classifier algorithms, attribute-based feature selection methods, and machine learning algorithms, specifically evaluating the performance of Naïve Bayes, J48, and SMO algorithms in detecting fraudulent website URLs based on URL features. The study concludes that J48 and SMO algorithms are more effective in detecting fraudulent websites compared to Naïve Bayes.

TABLE 1: Analysis of Techniques used in the Research paper

Ref No.	Feature Extraction Techniques	Classification Techniques Used
[1]	Lexical, Host-Based, Content-Based	Various Machine Learning Classifiers
[3]	Domain-Based, Alexa-Based, Obfuscation	XGBoost Algorithm
[4]	Attribute-Based Feature Selection (Gain Ratio, Relief)	N/A

CONCLUSION

When it comes to identifying fraudulent websites, feature extraction is essential. These approaches aid in the creation of strong machine learning models for cybersecurity applications by carefully studying features from URL strings and webpage content. Additional incorporation of knowledge from many research projects could improve harmful website detection systems' precision and effectiveness. Continued innovation and research in this area are essential to protecting online activity as cyber threats change

REFERENCES

[1] *S. Singhal, U. Chawla and R. Shorey, "Machine Learning & Concept Drift based Approach for Malicious Website Detection," 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, 2020, pp. 582-585, doi: 10.1109/COMSNETS48256.2020.9027485. keywords: {Feature extraction;Uniform resource locators;Forestry;Supervised learning;Malware;Training data;Machine learning;URL Feature Extraction;Malicious Website Detection;Concept Drifts;Feature Vectors;Gradient Boosted Trees;Random Forest;Feedforward Neural Networks},*

[2] *Shantanu, B. Janet and R. Joshua Arul Kumar, "Malicious URL Detection: A Comparative Study,"*

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1147-1151, doi: 10.1109/ICAIS50930.2021.9396014. keywords: {Uniform resource locators;Training;Phishing;Predictive models;Malware;Random forests;Web search;Malicious URL;Machine learning;Phishing;Spamming;Malware;Spoofing},

[3] *Y. -C. Chen, Y. -W. Ma and J. -L. Chen, "Intelligent Malicious URL Detection with Feature Analysis," 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 2020, pp. 1-5, doi: 10.1109/ISCC50000.2020.9219637. keywords: {Uniform resource locators;Training;Machine learning algorithms;Virtual assistants;Computer architecture;Feature extraction;Classification algorithms;malicious URL;JavaScript detection;artificial intelligence;feature analysis},*

[4] *M. Aydin, I. Butun, K. Bicakci and N. Baykal, "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0774-0779, doi: 10.1109/CCWC47524.2020.9031125. keywords: {Phishing;Feature extraction;Uniform resource locators;Classification algorithms;Machine learning algorithms;Machine learning;Attribute-based feature selection;Cyber theft;Data analysis;Fraudulent website detection;Machine learning algorithms},*