IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IDENTIFICATION OF SECURITY THREATS AND LEGITIMATE STATUS

[1]Mrs. K. S. Sawant, [2]Mahak Chawla, [3]Sejal Bora, [4]Prajakta Khairnar, [5]Shravani Thakare
[1]Professor, [2]Student, [3]Student, [4]Student, [5]Student
1Computer Department, 1BVCOEW, Pune, India

Abstract: Identifying security dangers and confirming the legitimacy of URLs are the main topics of this study paper, which focuses on phishing and malware. By utilizing the potent SVM algorithm, we put forth a sturdy approach for identifying and categorizing harmful URLs linked to malware infections and phishing scams. Our methodology consists of analysing general features included in malware and phishing URLs that are taken from large datasets that are made available by reliable sources. Designed for end-host deployment, the SVM algorithm shows its effectiveness in identifying malware URLs and phishing URLs, both known and unknown. Our solution provides precise and timely detection of URLs by examining their subtle properties, which enables proactive mitigation of cyber threats. Our methodology's efficacy is demonstrated through experimental validation, exhibiting a low rate of false positives and a high detection accuracy. By offering a sophisticated and effective method for recognizing and differentiating between security risks and authentic URLs, this research advances cybersecurity. In this case, the use of SVM is a major step in strengthening the resistance of digital ecosystems against malware and phishing attempts.

Keywords: Legitimacy, URLs, Phishing, Malware, SVM algorithm, Cyber threats

1.INTRODUCTION

When it comes to cybersecurity, identifying security threats and verifying the authenticity of URLs are critical goals for securing digital environments. This study explores these important areas, concentrating on the widespread dangers posed by malware and phishing. Our technology, which is based on the powerful Support Vector Machine (SVM) algorithm, offers a reliable way to identify and classify malicious URLs linked to malware infections and phishing scams. The study's methodology entails a thorough examination of the general characteristics present in both malware and phishing URLs. By utilizing vast datasets obtained from dependable sources, our study tackles threats that are both recognized and unidentified. The SVM algorithm, which was created especially for end-host deployment, has proven to be successful in identifying the minute characteristics of malware and phishing URLs, enabling the proactive mitigation of online threats. This research contributes to the field of cybersecurity and strengthens digital ecosystems against the ubiquitous threats of malware and phishing by offering a sophisticated and efficient method for identifying and distinguishing between security risks and legitimate URLs. Specifically, applying the SVM algorithm turns out to be a crucial first step in improving digital environments' ability to fend off advanced cyberattacks.

2. LITERATURE REVIEW

The system proposed by Mohammad Nazmul Alam, Ishita Saha, Dhiman Sarma, Rubaiath-E-Ulfath, Farzana Firoz Lima, Sohrab Hossain [1] utilizes ML algorithms, including Random Forest (RF) and Decision Tree (DT), on a legitimate dataset obtained from Kaggle. Feature selection algorithms like Principal Component Analysis (PCA) are employed for dataset analysis. The experimental results demonstrate the efficacy of the SVM algorithm, achieving a maximum accuracy of 97% through the RF algorithm.
Abdul Basit et al. [2] proposed a novel ensemble machine learning method for detecting phishing attacks on websites, emphasizing the increasing threat of such attacks during the COVID-19 pandemic. The authors employ three machine learning classifiers, namely Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Decision Tree (C4.5), in an ensemble method with the Random Forest Classifier (RFC). The ensemble model demonstrates superior accuracy in detecting phishing attacks compared to existing studies, achieving an accuracy of 97.33%.

Lakshmanarao et al. [3] collected a dataset from the UCI repository and applied various machine learning algorithms, including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest. They also introduced two priority-based algorithms, PA1 and PA2, to assign priorities to the base classifiers based on True Positive Rate (TPR) and True Negative Rate (TNR). The results showed that Random Forest achieved the highest priority in both algorithms.

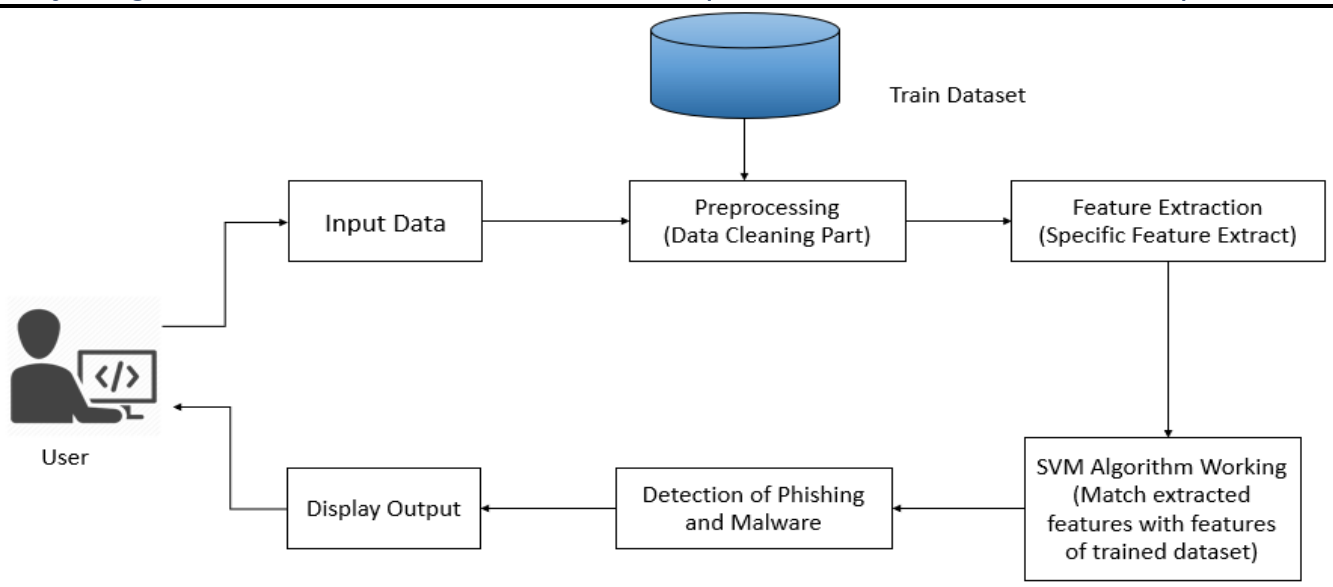
Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen, And Rusyaizila Ramli [4] addressed a system using various machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine, Neural Network, and Linear Model are applied for phishing detection. Results indicate high accuracy, with Random Forest achieving 97.14% accuracy. The paper concludes that machine learning is a highly effective approach, surpassing 99% accuracy, to analyze and classify phishing websites based on the latest datasets.

Md Jobair Hossain Faruk, Hossain Shahriar et al. [5] introduced a novel GCN-based malware detection system, leveraging API call sequences and graph convolutional networks for feature extraction and classification. The system achieves a high accuracy of 98.32%. Additionally, a lightweight system for identifying unknown malware on Android devices is proposed, using machine learning techniques with feature extraction based on static and dynamic analyses.

The authors Colin Galen and Robert Steele [6] investigated the performance maintenance over time of machine learning-based malware detection models, with a specific focus on Random Forest-based models. The paper evaluates model performance in terms of accuracy, area under the receiver operator characteristic curve (AUC), precision, and recall. The Random Forest model stands out as the best-performing model, with an AUC exceeding 0.9, considered excellent discriminative performance, even up to the end of the dataset period in 2018.

Sanket Agarkar and Soma Ghosh [7] explored behaviour-based identification of ransomware, considering not only the file's identity but also its intended operations over time. The practical work and results compare the performance of Decision Tree, Random Forest, and Light GBM classifiers. Light GBM demonstrates the highest accuracy (99.50%), outperforming the other models.

The authors Saiful Islam Rimon and Md. Mokammel Haque [8] proposed a hybrid learning approach combining Random Forest and K-Nearest Neighbor Classifier to effectively detect and classify malware. The study employs a dataset comprising 10,000 samples of both malware and benign files, featuring 78 distinctive feature values and spanning six diverse malware classes. The authors conduct training for both binary and multi-class classification, comparing their results with existing methods to determine the superior approach. The proposed hybrid model exhibits promising results, achieving 98% accuracy in binary classification and 93% accuracy in multi-class classification.



3. SYSTEM ARCHITECTURE

The system architecture for the identification of the legitimate status of phishing and malicious URLs using the Support Vector Machine (SVM) algorithm typically involves several key components. Here's a high-level overview of the system architecture:

1.Data Collection: Collect a comprehensive dataset containing both phishing and malicious URLs. This dataset will be used for training and testing the SVM model.

2.Data Preprocessing: Extract relevant features from the URLs, such as the domain, path, length, presence of special characters, etc. Transform the raw URL data into a format suitable for SVM training. This may involve techniques like one-hot encoding, tokenization, or other methods based on the chosen features.

3.Training Phase: Utilize the pre-processed dataset to train the SVM model. The SVM algorithm works by finding the hyperplane that best separates the phishing and legitimate classes in the feature space.

4.Model Evaluation: Assess the SVM model's performance using a validation set that the model has not seen during training. Adjust the model parameters if necessary to improve its performance.

5.Testing Phase: Evaluate the SVM model on a separate testing dataset to assess its generalization to unseen data.

6.Real-Time URL Classification: Deploy the trained SVM model to predict whether a given URL is phishing, malicious, or legitimate in real-time based on its extracted features.

7.Monitoring and Maintenance: Periodically update the SVM model with new data to ensure its effectiveness against evolving threats. Implement a monitoring system to track the model's performance and trigger alerts for retraining if necessary.

This architecture provides a structured approach to building a system for identifying the legitimacy of URLs using the SVM algorithm.

4. CONCLUSION

The goal of this research paper is to address cybersecurity issues associated with malware and phishing by identifying security threats and verifying the authenticity of URLs. The study suggests a reliable method for identifying and categorizing malicious URLs connected to malware infections and phishing scams that makes use of the Support Vector Machine (SVM) algorithm. The methodology uses large datasets from reputable sources to analyse general features found in malware and phishing URLs. Designed for end-host deployment, the SVM algorithm shows promise in detecting known and unknown malware as well as phishing URLs. The study highlights the importance of promptly and accurately identifying URLs based on their subtle characteristics in order to prevent cyberattacks. The suggested methodology's effectiveness is demonstrated through experimental validation, which highlights the high detection accuracy and low false positive rates. The research makes a substantial contribution to cybersecurity by offering a sophisticated and effective technique for identifying and separating reputable URLs from security threats. In order to defend advanced cyberattacks against digital ecosystems, the SVM algorithm is essential.

5. FUTURE SCOPE

1.Expansion to Other Threats: Expand the approach to tackle novel cybersecurity risks beyond malware and phishing, taking into account the dynamic nature of cyberattacks.
2.Dynamic Analysis: Use dynamic analysis approaches to improve the system's comprehension of malware and phishing behaviors over time, allowing it to adjust to novel attack patterns.
3.Real-Time Threat Intelligence: Incorporate real-time threat intelligence feeds to enhance the system's responsiveness to new and emerging threats by keeping it updated with the most recent information on known threats.
4.User Education and Awareness: In order to enable users to identify and report possible threats, think about incorporating user education and awareness components into the cybersecurity framework.
5.Cross-Platform Adaptability: To offer thorough protection across a range of digital environments, improve the system's adaptability to different platforms, including mobile devices and different operating systems.

6. REFERENCES

[1] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. -E. -. Ulfath and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1173-1179, doi: 10.1109/ICSSIT48917.2020.9214225.

[2] A. Basit, M. Zafar, A. R. Javed and Z. Jalil, "A Novel Ensemble Machine Learning Method to Detect Phishing Attack," 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-5, doi: 10.1109/INMIC50486.2020.9318210.

[3] A. Lakshmanarao, P. S. P. Rao and M. M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1164-1169, doi: 10.1109/ICAIS50930.2021.9395810.

[4] M. H. Alkawaz, S. J. Steven, A. I. Hajamydeen and R. Ramli, "A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods," 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, 2021, pp. 82-87, doi: 10.1109/ISCAIE51753.2021.9431794.

[5] M. J. Hossain Faruk et al., "Malware Detection and Prevention using Artificial Intelligence Techniques," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 5369-5377, doi: 10.1109/BigData52589.2021.9671434.

[6] C. Galen and R. Steele, "Performance Maintenance Over Time of Random Forest-based Malware Detection Models," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2020, pp. 0536-0541, doi: 10.1109/UEMCON51285.2020.9298068.

[7] S. Choudhary and A. Sharma, "Malware Detection & Classification using Machine Learning," 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3), Lakshmangarh, India, 2020, pp. 1-4, doi: 10.1109/ICONC345789.2020.9117547

[8] Rimon, S.I., Haque, M.M. (2023). Malware Detection and Classification Using Hybrid Machine Learning Algorithm. In: Vasant, P., Weber, GW., Marmolejo-Saucedo, J.A., Munapo, E., Thomas, J.J. (eds) Intelligent Computing & Optimization. ICO 2022. Lecture Notes in Networks and Systems, vol 569. Springer, Cham. https://doi.org/10.1007/978-3-031-19958-5_39

[9] S. Mahdavifar and A. A. Ghorbani, "DeNNeS: deep embedded neural network expert system for detecting cyber-attacks," (in English), Neural Computing & Applications, Article; Early Access p. 28.

[10] N. A. Azeez, B. B. Salaudeen, S. Misra, R. Damasevicius, and R. Maskeliunas, "Identifying phishing attacks in communication networks using URL consistency features," (in English), International Journal of Electronic Security and Digital Forensics, Article vol. 12, no. 2, pp. 200-213, 2020.

[11] "Phishing activity trends report," https://docs.apwg.org//reports/apwg trends report q4 2016.pdf, 2020, accessed: 28-Aug-2020.

[12] N. Abdelhamid, F. Thabtah, and H. Abdel-jaber, "Phishing detection: A recent intelligent machine learning comparison based on models content and features," in 2017 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2017, pp. 72–77.

[13] A. .K.S., "Impact of malware in modern society," Journal of Scientific Research and Development, vol. 2, pp. 593– 600, 06 2019.