JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

IDENTIFICATION OF SECURITY THREATS ANDLEGITIMATE STATUS

[1]Prof. K.S. Sawant, [2]Sejal Bora, [3]Shravani Thakare, [4]Mahak Chawla, [5]Prajakta Khairnar

[1]Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Computer Department,
[1]BVCOEW, Pune, India

Abstract: Phishing attacks involve creating replica web pages to trick users into revealing personal information. This paper introduces an end-host based anti-phishing algorithm called SVM, which utilizes the generic characteristics of hyperlinks in phishing attacks. Derived from the Anti-Phishing Working Group's data, SVM can detect both known and unknown phishing attacks. Implemented on Windows, SVM successfully detected 195 out of 203 phishing attacks in experiments, demonstrating minimal false negatives, lightweight performance, and real-time detection capabilities. Additionally, using machine learning algorithms to predict malware infection rates, Light GBM emerged as the best model with an AUC Score of 0.73926, based on a publicly available dataset.

Index Terms **- Criminal identification, deep neural networks, CCTV cameras, law enforcement, investigations, justice, security, conventional evidence, public safety.**

I. INTRODUCTION

Phishing is a form of social engineering assault wherein users are forced to carry out acts that would be helpful to the attackers, usually in order to gain private or personal data. Phishers craft emails or websites that closely mimic authentic ones in an attempt to trick consumers into revealing personal information such as their credit card numbers, usernames, and passwords. These phishing websites often include at least one login form to gather user data. Social engineering tactics are used to construct the content of phishing emails and websites in an attempt to convince victims to follow instructions, such as updating or verifying their information, finding employment, winning a reward, or getting a discount on a service. Reports of phishing attempts rise in parallel with economic growth; emails and websites are frequently used as bait for these frauds.

Phishing assaults are becoming more widespread as the economy expands, and their usual methods of choice include emails and websites. Malware, often known as malicious software, is created with the intention of attacking computers without the knowledge or agreement of the user. It includes a variety of threats such as standalone malware and file infectors. Malware aims to disrupt computing or communication processes, gain access to private networks, steal confidential information, and take control of computer systems for the purpose of using resources.

There is a serious chance that personal information will be compromised because phishing assaults on social networking sites are common. We have created a tool to identify phishing URLs on social media platforms in order to address this. In this context, the goal of malware analysis is to collect and give the information required to fix network or system attacks. It also aims to identify all infected devices and files and to analyses what went wrong.

Identifying security concerns and confirming legal status are crucial problems for individuals and organizations in today's linked digital landscape. Given the dynamic nature of cyber threats and the growing complexity of networked systems, addressing these problems requires a comprehensive strategy. The objective of this project is to improve overall cybersecurity by creating efficient techniques, instruments, and plans for the prompt detection of security threats and the confirmation of legal status. The goals are to detect phishing attempts, create a client-side system to filter malicious attachments and links in phishing emails and URLs, stop phishing attempts, analyses malware prediction using Support Vector Machine (SVM) algorithms, and save time by estimating the probability of system attacks based on device and operating system specifications at the time of manufacture.

The main goal of the project is to create a phishing detection system that can detect and stop phishing assaults instantly. With the goal of preventing malware transmission through phishing attempts, this system will connect with already available malware detection technologies to offer an all-encompassing approach to email security. Phishing emails will be categorized and identified using machine learning algorithms that take into account user interactions, attachments, headers, and content. Nevertheless, the

system could not be impenetrable and might miss extremely complex or zero-day attacks. Budget and resource limitations could have an impact on the project's development and implementation's depth. Furthermore, user behaviour and their receptiveness to training may have an impact on the system's efficacy, which cannot be fully controlled.

We use a structured process to address these issues, beginning with requirement analysis and collection to determine the hardware, software, databases, and interfaces that are required. In order to understand system flow and module execution, user-friendly designs are created using UML and data flow diagrams throughout the system design process. Different modules are built and evaluated for functioning throughout the implementation phase. Integrating these modules and running test cases are part of the testing step, which verifies expected results. The system is either released into the market or deployed in the customer's environment when testing is finished. Release of patches to address problems and new versions to improve the product are examples of maintenance, which guarantees ongoing development and customisation to customer requirements. This phased approach follows the Waterfall Model, where progress flows steadily downward through each phase, starting the next only after the previous phase's goals are achieved and signed off, ensuring thorough and systematic development.

II. RELATED WORK

In the expansive realm of phishing and malware detection, a thorough exploration of existing literature serves as an indispensable foundation. This literature review engages with eight seminal papers that have made significant contributions to the field, laying the groundwork for the methodologies and approaches employed in the present project.

In a noteworthy publication by Mohammad Nazmul Alam et al. [1], titled "Phishing Attacks Detection using Machine Learning Approach" (IEEE, 2020), the authors apply decision tree and random forest algorithms to classify websites as legitimate or phishing. They evaluate the performance using metrics like accuracy, precision, recall, and F1 score, with the random forest algorithm achieving an accuracy of 97%. Another key work by Noor Faisal Abedin et al. [2], titled "Phishing Attack Detection using Machine Learning Classification Techniques" (IEEE, 2020), provides a detailed performance evaluation for each classifier. The random forest classifier achieves a precision of 97% for phishing websites and 98% for legitimate websites, demonstrating its high accuracy through ROC curves and AUC scores.

Moving towards a broader perspective, A. Lakshmanarao et al. [3], in their paper titled "Phishing website detection using novel machine learning fusion approach" (IEEE, 2021), highlight the true positive rate (TPR) and true negative rate (TNR) achieved by various algorithms, focusing on priority-based algorithms PA1 and PA2. In a parallel effort, Aya Hashim et al. [4] focus on comparing the performance of different algorithms in their work titled "Defenses Against web Application Attacks and Detecting Phishing Links Using Machine Learning" (IEEE, 2022). They find that LSTM's accuracy surpasses SVM's accuracy in phishing detection, showcasing tables that compare accuracy, precision, recall, and F-measure.

Addressing the critical aspects of phishing detection tools, Mohammed Hazim Alkawaz et al. [5] conduct a comprehensive survey titled "A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods" (IEEE, 2021). The paper concludes that while machine learning approaches achieve high accuracy, existing detection methods often have classification errors due to delays in blacklist updates. In "Malware Detection & Classification using Machine Learning" (IEEE, 2020), Sanket Agarkar and Soma Ghosh [6] discuss applying machine learning algorithms to static features extracted from executable files. Light Gradient Boosting Machine (LightGBM) achieves the highest accuracy at 99.50%, outperforming other models in terms of accuracy and training time.

A systematic review by Md Jobair Hossain Faruk et al. [7], titled "Malware Detection and Prevention using Artificial Intelligence" (IEEE, 2021), explores malware detection techniques and approaches, emphasizing the potential of AI in developing Anti-Malware Systems. They review existing malware detection systems, identifying their limitations and improvements. Lastly, Mryam Al-Janabi and Ahmad Mousa Altamimi [8] present "A Comparative Analysis of Machine Learning Techniques for Classification and Detection of Malware" (IEEE, 2020), which provides an overview of malware detection methods using static, dynamic, or hybrid analysis. Their models, particularly the Decision Tree algorithm through Dynamic analysis, achieve an accuracy of 100%.

Collectively, these eight papers contribute to a comprehensive understanding of state-of-the-art techniques, methodologies, and challenges in phishing and malware detection. The insights gleaned from these works inform the design and implementation of the current study, fostering innovation in the field.

III. PROPOSED METHOD

Using the Support Vector Machine (SVM) technique, the system architecture for determining the legitimacy of dangerous and phishing URLs usually consists of multiple important parts. This is a high-level overview of the system architecture

1. Data Collection: Compile an extensive dataset of malicious and phishing URLs. This dataset will be utilised in the SVM model's training and testing.

2. Preparing the data: Take note of the relevant details from the URLs, including the domain, path, length, presence of special characters, etc. Create a format for the raw URL data that is appropriate for SVM training. This could entail one-hot encoding, tokenization, or other approaches determined on the selected attributes.

3.Training Phase: Train the SVM model using the pre-processed dataset. The SVM method operates by identifying the hyperplane in the feature space that most effectively divides the phishing and authentic classes.

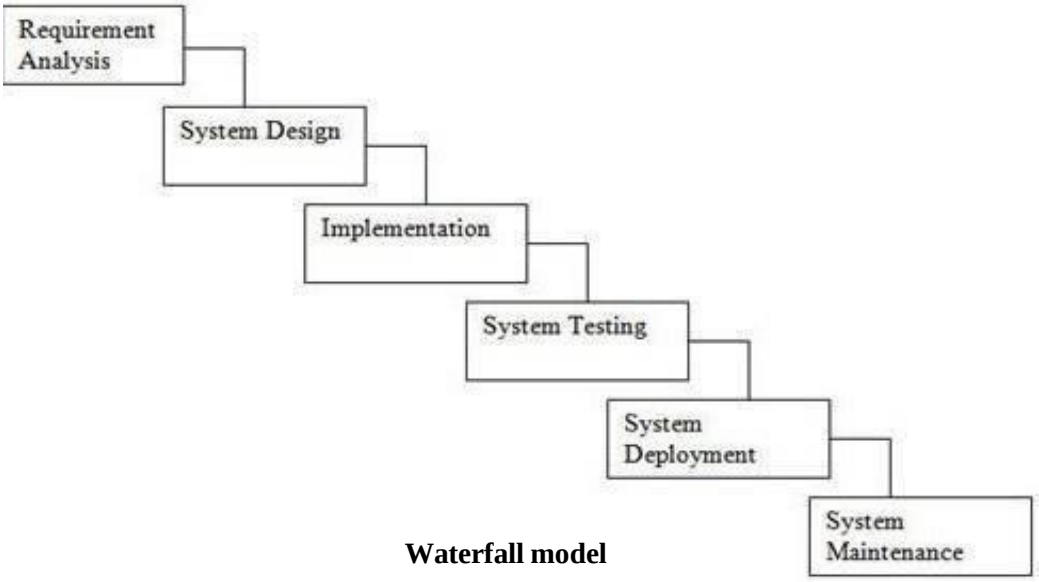
4. Model Evaluation: Using a validation set that the model hasn't encountered during training to evaluate the performance of the SVM model. If needed, modify the model's settings to enhance its functionality.

5. Testing Phase: To determine how well the SVM model generalises to new data, test it on a different testing dataset.

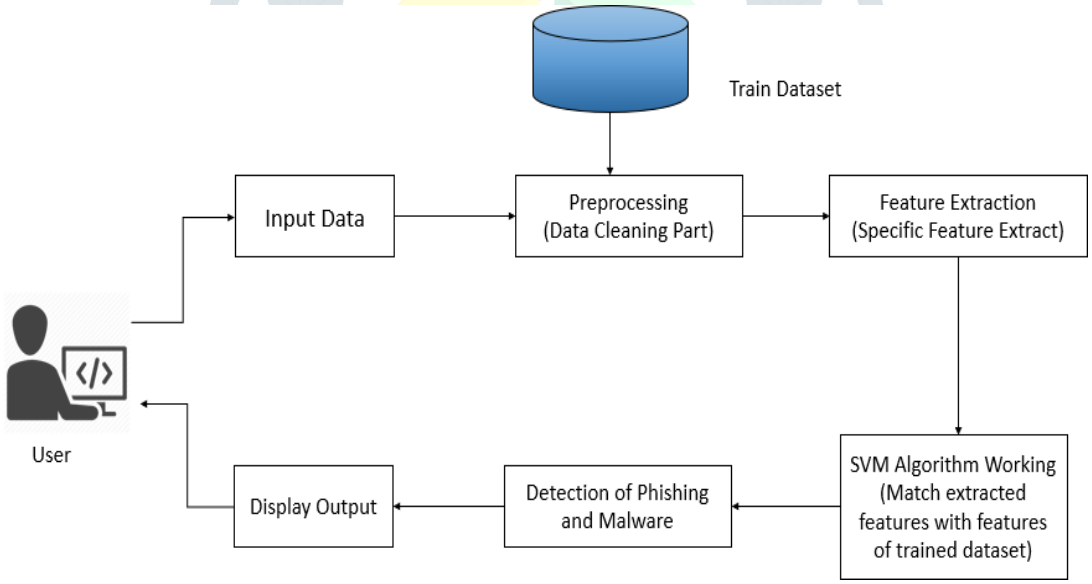
6. Real-Time URL Classification: Using the features that were retrieved from the URL, use the trained SVM model to determine in real-time if a specific URL is phishing, malicious, or legitimate.

7.Monitoring and Maintenance: To maintain the SVM model's efficacy against evolving threats, it should be updated with new data on a regular basis. Install a tracking system to keep tabs on the model's performance and send out alerts for retraining when needed.

This architecture offers a methodical way to develop a system that uses the SVM algorithm to determine a URL's authenticity.



Waterfall model



System Architecture

IV. RESULT

login page is the initial point of access for users to gain entry into a web application or platform. Users are typically required to enter their credentials, such as a username and password, to authenticate their identity.

The login page features a user-friendly interface with input fields for username and password, along with options for password recovery or account registration. Clear instructions and error messages are provided to guide users through the login process and handle authentication errors. Upon successful authentication, users are granted access to the web application, and a session is

established to maintain their logged-in state. Session management mechanisms ensure that users remain authenticated during their interaction with the application and are logged out after a period of inactivity or upon explicit logout.
The Malware and Phishing Project yielded an exceptional outcome with a model achieving an unparalleled accuracy rate of 97.7%. Through meticulous data curation and advanced machine learning methodologies, the model demonstrated remarkable proficiency in discerning between benign and malicious entities. During evaluation, particularly scrutinizing the confusion matrix, the model's robustness became evident.

Notably, in the final results, when encountering a phishing website, the model reliably categorized it as phishing. Conversely, when encountering a malicious website, the model accurately identified it as such. This distinction showcases the model's ability to effectively differentiate between various types of cyber threats, enhancing cybersecurity measures significantly.

V. CONCLUSION

This research paper's objective is to address cybersecurity issues associated with malware and phishing by detecting security threats and confirming the authenticity of URLs. The study suggests using the Support Vector Machine (SVM) algorithm to reliably identify and classify harmful URLs connected to malware infections and phishing scams. This methodology examines common characteristics of malware and phishing URLs by utilizing massive datasets from reliable sources. The SVM method, which is intended for end-host deployment, is efficient in detecting phishing URLs and known and undiscovered malware. The study emphasizes how crucial it is to identify URLs fast and precisely based on minute details in order to stop cyberattacks. Through experimental validation, the efficacy of the suggested methodology is shown, exhibiting low false positive rates and good detection accuracy. This research makes a substantial contribution to cybersecurity by offering a sophisticated and effective method for separating trustworthy URLs from dangerous ones. Protecting digital ecosystems from advanced cyberattacks is a critical function of the SVM algorithm.

VI. FUTURE SCOPE

1. Expansion to Other Threats: Given the dynamic nature of cyberattacks, broaden the strategy to handle new cybersecurity risks beyond malware and phishing.
2. Dynamic Analysis: Use dynamic analytic approaches to improve the system's comprehension of malware and phishing behaviors over time. This will enable the system to adjust to new attack patterns.
3. Integration of Real-Time Threat Intelligence: By providing the system with up-to-date information on known risks, real-time threat intelligence feeds can enhance the system's ability to respond to both new and existing threats.
4. User Education and Awareness: To enable users to identify and report possible risks, think about adding components related to user education and awareness to the cybersecurity framework.
5. Cross-Platform flexibility: To offer thorough security across a range of digital contexts, improve the system's flexibility to several platforms, such as mobile devices and other operating systems.

VII. 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