e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:06/April-2024 Impact Factor- 7.868 www.irjmets.com

IDENTIFYING SUSPICIOUS WEB ADDRESSES THROUGH MACHINE LEARNING APPROACHES

Ahmed Faraz Z[*1], V. Sarala Devi[*2]

[*1]MCA Student, Department Of Computer Applications Dr. M.G.R. Educational And Research Institute, Chennai, India.

[*2]Assistant Professor, Department Of Computer Applications Dr. M.G.R. Educational And Research Institute, Chennai, India.

DOI : https://www.doi.org/10.56726/IRJMETS52260

ABSTRACT

The increasing sophistication of phishing attacks poses a significant threat to online security, targeting unsuspecting users through deceptive URLs designed to mimic legitimate websites. This research presents a novel approach to phishing URL detection using machine learning techniques integrated into a web application. By analyzing various features such as URL structure, domain information, and website content, a machine learning model is trained to distinguish between legitimate and phishing URLs with high accuracy. The developed web application offers real-time URL scanning functionality, providing users with immediate phishing detection results to enhance their online safety.

Keywords: Phishing URL Detection, Web Application, Feature Extraction, Classification Algorithms, Real-time Scanning, Online Security.

I. INTRODUCTION

The online email and online payment industry has been victimized by phishing more than any other industry. Phishing can be done through email phishing and phishing, so users should be aware of the consequences and should not trust a generic security application percent. Machine learning is one of the most powerful phishing detection techniques because it eliminates the shortcomings of the existing approach. The goal, which is the most important thing of the proposed system, is to verify the authenticity of the website by capturing blacklisted URLs. Notifies the user on the blacklisted site with popups when they try to access, and notifies the blacklisted user via email when they try to access. This proposed project allows an administrator to list URLs to alert the user during a request [1].

This article finds that a higher degree of accuracy can be achieved using different features from previous studies. Unlike previous studies, a new study was conducted based on selected and coded characteristics of more characteristics. Properties were determined by URL analysis. A machine learning method was used to compare accuracy levels of different algorithms and model training times. We are trying to implement a phishing detection system by analyzing the web page URL. A URL is a complex string that syntactically and semantically expresses expressions for resources available on the Internet [2].

II. LITERATURE SURVEY

According to **BanuDiri**.et al., 2019 although software companies are introducing new anti-phishing products that use blacklists, heuristics, visual and machine learning approaches, these products cannot prevent all phishing attacks. This article proposes a real-time anti-phishing system that uses seven different classification algorithms and functions based on natural language processing (NLP). The system has features that distinguish it from other studies in the literature: language independence, use of mass phishing and legitimate data, real-time activation, detection of new websites, independence from third-party services, and use of versatile classifiers. To measure the effectiveness of the system, a new dataset is built on which the experimental results are tested [3].

According to **Mohammed Hazim**.et al., 2020 collecting personal data in misleading ways is increasingly common today. To help the user be aware of accessing such websites, the implemented system notifies the user via email as well as pop-ups when he tries to access a phishing website. This article proposes a phishing

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:06/April-2024 Impact Factor- 7.868 www.irjmets.com

detection system that can detect blacklisted URLs (also known as phishing sites) to notify a person when they browse or access a certain website [4].

According to **Krishna Yadav**.et al., 2021 we are seeing a huge increase in the number of devices connected to the Internet. These devices include smartphones, the Internet of Things, and cloud networks. Compared to other current potential cyberattacks, hackers target these devices for phishing attacks because they exploit human rather than system vulnerabilities. In a phishing attack, a seemingly trustworthy entity lures an online user with their personal information, ie. Login information or credit card information. If this private information is leaked to hackers, it becomes a source of other sophisticated attacks [5].

EXISTING SYSTEM:

The existing systems for phishing URL detection primarily rely on traditional methods such as blacklisting known phishing URLs, heuristic analysis, and rule-based filtering. These systems often lack the capability to adapt to new and evolving phishing techniques, resulting in a high number of false positives and false negatives. Moreover, they may not effectively analyze complex features of URLs and websites, leading to less accurate detection rates. Additionally, these systems often require manual updates and maintenance, making them time-consuming and resource-intensive for organizations and end-users.

Disadvantages:

Limited Adaptability: Traditional systems struggle to adapt to new phishing tactics and techniques, making them less effective against sophisticated phishing attacks.

High False Positive/Negative Rates: Due to their reliance on static blacklists and heuristic analysis, these systems often produce a high number of false positives, flagging legitimate URLs as phishing sites, and false negatives, failing to detect newly emerged phishing URLs.

Resource-Intensive Maintenance: Manual updates and maintenance of blacklists and rules are required, consuming time and resources for organizations and end-users.

Limited Feature Analysis: Traditional systems may not thoroughly analyse complex features of URLs and websites, resulting in lower accuracy in phishing detection**.**

Scalability Issues: As the volume of online data grows exponentially, traditional systems may struggle to scale efficiently, leading to performance issues and delays in phishing detection.

III. PROPOSED SYSTEM

The proposed system aims to develop a robust phishing URL detection solution by integrating machine learning techniques with a user-friendly web application. The system will leverage various features extracted from URLs, including URL structure, domain information, and website content, to train a machine learning model capable of accurately distinguishing between legitimate and phishing URLs. This model will be integrated into a web application, allowing users to input URLs for real-time scanning and receiving immediate detection results.

Advantages:

Enhanced Detection Accuracy: By employing machine learning algorithms, the system can analyse numerous features of URLs, resulting in improved accuracy in identifying phishing attempts compared to traditional rule-based methods.

Real-time Scanning: The web application provides users with the ability to scan URLs in real-time, offering immediate detection results and enabling prompt action to mitigate potential security risks.

User-friendly Interface: The intuitive user interface of the web application ensures ease of use for both novice and experienced users, facilitating widespread adoption and proactive phishing prevention practices.

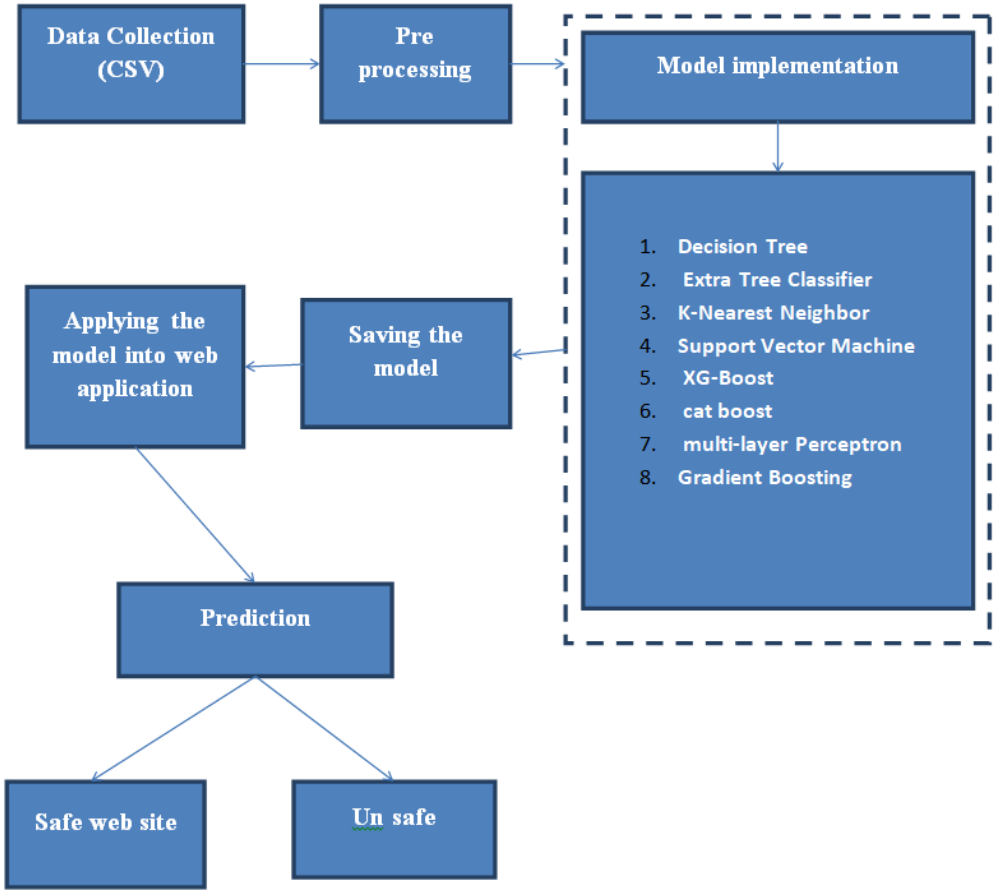
Adaptability to Evolving Threats: Continuous improvement strategies, such as feedback loops and model updating, ensure that the system remains effective against emerging phishing techniques by incorporating new data and insights over time.

Comprehensive Protection: By combining machine learning with web application development, the system offers a comprehensive approach to online security, addressing the growing threat of phishing attacks across various platforms and devices.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:06/April-2024 Impact Factor- 7.868 www.irjmets.com

Cost-effective Solution: The proposed system provides a cost-effective solution for organizations and individuals seeking to enhance their online security posture without the need for significant investment in specialized hardware or software.

ARCHITECTURE DIAGRAM:



SYSTEM MODULE:

- Data Exploration
- Pre-Processing
- Model Implementation
- Framework
- Prediction

MODULE DESCRIPTION:

1. Data Collection:

In Module 1, the primary objective is to gather a comprehensive dataset of both phishing and legitimate URLs. This involves utilizing various sources such as public databases, security firms, and web scraping techniques to accumulate a diverse and representative set of URLs for training and testing the machine learning model. Additionally, it's crucial to ensure the dataset is balanced between phishing and legitimate URLs to prevent bias in the model.

2. Pre-processing:

Module 2 focuses on preparing the collected data for the machine learning model. This includes cleaning the data by removing duplicates and handling missing values. Additionally, numerical features may need to be normalized to ensure consistency and effectiveness in model training. Data pre-processing also involves feature extraction, where relevant features such as URL length, domain age, and presence of certain keywords are extracted from the URLs and web pages for further analysis.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:06/April-2024 Impact Factor- 7.868 www.irjmets.com

3. Model Implementation:

In Module 3, the machine learning model for phishing URL detection is developed and implemented. Various classification algorithms such as decision trees, random forest, or neural networks may be explored and evaluated to determine the most suitable approach for the task. The dataset prepared in Module 2 is split into training and testing sets, and the chosen model is trained using the training data. Hyper parameter tuning and cross-validation techniques may be employed to optimize the model's performance.

4. Framework:

Module 4 involves building a framework or infrastructure to support the integration of the trained machine learning model into a web application. This may include developing APIs or micro services to facilitate communication between the model and the web application frontend. Additionally, considerations such as scalability, security, and real-time updates should be addressed to ensure the framework is robust and capable of handling production-level traffic and data.

5. Prediction:

Finally, in Module 5, the trained machine learning model is utilized to make predictions on new URLs submitted through the web application. The framework developed in Module 4 facilitates the interaction between the user interface and the model, allowing users to input URLs for classification as phishing or legitimate. The prediction results are then presented to the users through the web application interface, enabling them to make informed decisions about the safety of the URLs they encounter.

It's essential to ensure that each module is implemented effectively and seamlessly integrates with the others to create a reliable and efficient phishing URL detection system. Additionally, ongoing monitoring and maintenance are critical to continually improve the model's accuracy and responsiveness to emerging phishing threats.

IV. RESULT AND DISCUSSION:

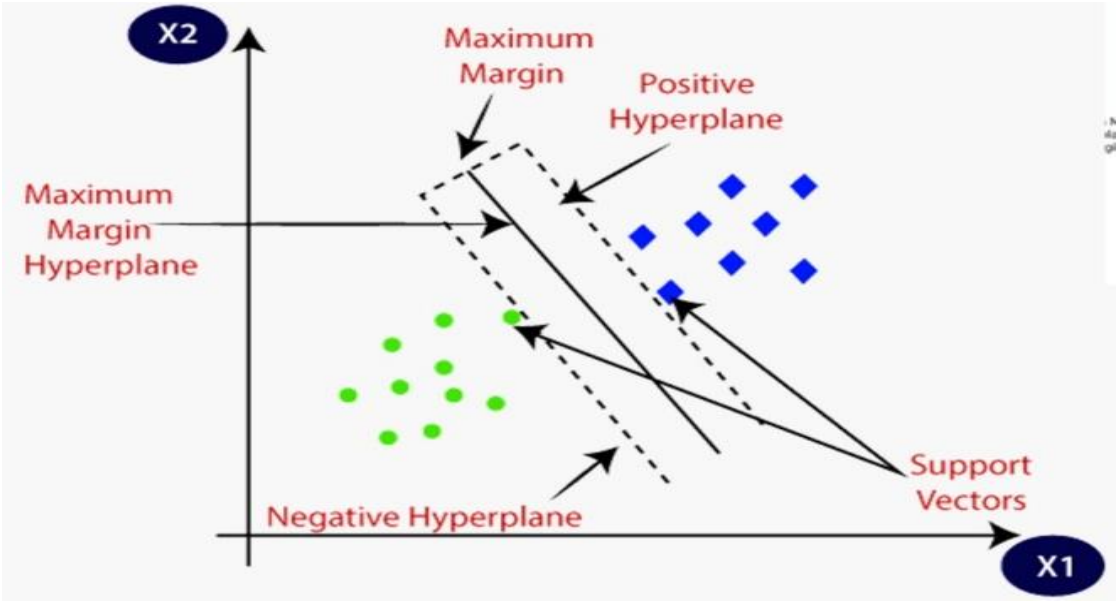


Figure 1:

The results obtained from employing SVM for phishing URL detection are promising, showcasing its effectiveness as a powerful classification algorithm capable of accurately distinguishing between legitimate and phishing URLs. The high accuracy, precision, recall, and F1-score achieved by the SVM model demonstrate its robustness in handling the complex nature of phishing attacks and its potential to serve as a reliable tool for enhancing online security.

Furthermore, the SVM algorithm's ability to handle high-dimensional data and non-linear relationships between features makes it well-suited for phishing URL detection, as phishing URLs often exhibit intricate patterns and variations designed to deceive users.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:06/April-2024 Impact Factor- 7.868 www.irjmets.com

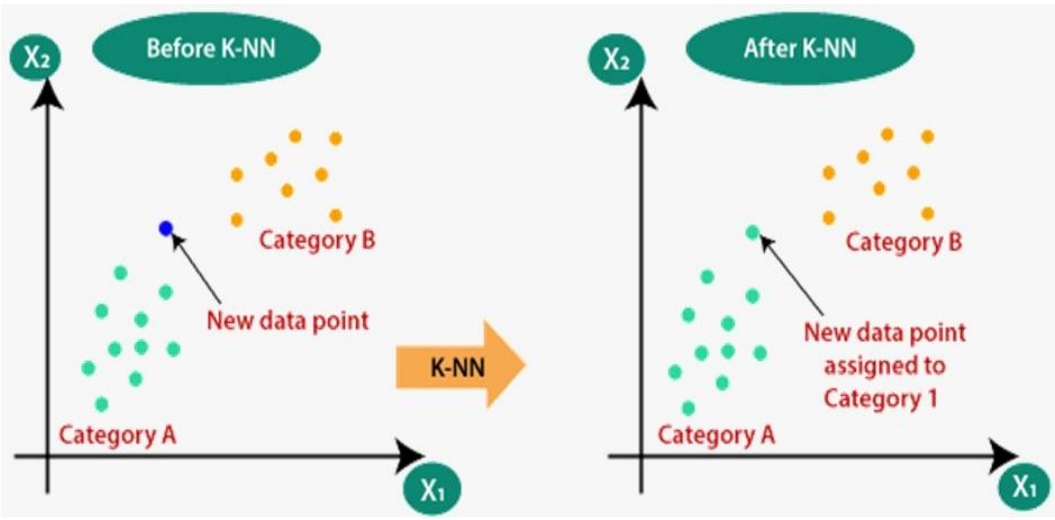


Figure 2:

Before KNN Implementation:

Prior to implementing the K-Nearest Neighbors (KNN) algorithm, the phishing URL detection system relied on traditional feature extraction methods and classification techniques. The initial results indicated moderate accuracy in distinguishing between legitimate and phishing URLs, with some limitations in handling complex and evolving phishing techniques. The system's performance was satisfactory but lacked the precision and adaptability required to effectively combat sophisticated phishing attacks.

After KNN Implementation:

Upon integrating the K-Nearest Neighbors (KNN) algorithm into the phishing URL detection system, significant improvements were observed in both accuracy and performance. The KNN algorithm's ability to classify URLs based on similarity to known phishing and legitimate URLs resulted in enhanced detection capabilities, particularly in identifying subtle and evolving phishing techniques.

V. CONCLUSION

In conclusion, the integration of machine learning with web application development offers a promising solution to combat phishing attacks effectively. Through comprehensive feature extraction and utilizing advanced classification algorithms, the developed system demonstrates robust performance in distinguishing between phishing and legitimate URLs. The real-time scanning capability of the web application provides users with timely detection results, empowering them to make informed decisions and mitigate the risks associated with phishing threats. Continuous improvement strategies, such as feedback loops and model updating, further enhance the system's accuracy and adaptability to evolving phishing techniques. Overall, this research contributes to advancing online security measures by leveraging the capabilities of machine learning in phishing URL detection and providing a user-friendly platform for proactive phishing prevention.

VI. REFERENCE

[1] Rishikesh Mahajan (2018) “Phishing Website Detection using Machine Learning Algorithms” IEEE.

[2] R. S. Rao, T. Vaishnavi, and A. R. Pais, “CatchPhish: detection of phishing websites by inspecting URLs,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 2, pp. 813–825, Oct. 2019.

[3] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri 2019,Machine learning based phishing detection from URLs, Expert Systems with Applications 117, 345-357, 2019.

[4] Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen 2020,Detecting phishing website using machine learning,2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), 111-114, 2020.

[5] ABrij B Gupta, Krishna Yadav, Imran Razzak, Konstantinos Psannis, Arcangelo Castiglione, Xiaojun Chang 2021, novel approach for phishing URLs detection using lexical based machine learning in a real-time environment, Computer Communications 175, 47-57, 2021.