International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2511316

An Efficient Hybrid Feature Select Technique towards Prediction of Suspicious URLs in IoT Environment

Battula Manideep, Gandham Pavan Kumar Reddy, Golla Ajay Kumar, Kalle Shiva Shankar, Dr. Dhanaraj Cheelu

Department of Artificial Intelligence and Machine Learning, Dr K V Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted : 05 May 2025

Published: 08 May 2025

Publication Issue

Volume 11, Issue 3

May-June-2025

Page Number

255-260

ABSTRACT

the evolving landscape of the Internet of Things (IoT), securing network-connected devices from cyber threats has become a critical concern. Among these threats, malicious URLs pose a significant risk by facilitating phishing, data theft, and malware attacks. This paper proposes an efficient hybrid feature selection technique aimed at enhancing the prediction of suspicious URLs within an IoT environment. The hybrid approach combines filter and wrapper-based methods to extract the most relevant features from large and complex URL datasets. By applying machine learning classifiers such as Random Forest and Support Vector Machines (SVM), the system demonstrates improved accuracy, reduced false positive rates, and faster detection times. The proposed technique optimizes feature space, reduces computational cost, and increases prediction reliability, making it highly suitable for real-time threat detection in resource-constrained IoT devices.

Keywords: Suspicious URLs, IoT Security, Hybrid Feature Selection, Machine Learning, Cyber Threat Detection

Introduction

Phishing has become the most serious problem, harming individuals, corporations, and even entire countries. The availability of multiple services such as online banking, entertainment,education,softwaredownloading,andso cialnetworkinghasaccelerated the Web's evolution in recent years. As a result, a massive amount of data is constantly downloaded and transferred to the Internet. Spoofed emails pretending to be from reputable businesses and agencies are used in social engineering techniques to direct consumers to fake websites thatdeceive users into giving financial information such as usernamesandpasswords.Technicaltricksinvolvetheins tallationofmalicioussoftware oncomputerstostealcredentialsdirectly,withsystemsfreq

uentlyusedtointerceptusers' online account usernames and passwords.

Therapid proliferation ofInternet of Things (IoT) devices has significantly increased the attack surface for cyber threats, making IoT environments more vulnerable to malicious activities, including phishing and malware distribution through suspicious URLs. Traditional URL detection methods often suffer from inefficiency due to high- dimensional feature spaces, leading to increased computational complexity and reduced accuracy in identifying threats. Existing feature selection techniques either lack adaptability or fail to achieve an optimal balance between computational efficiency and prediction accuracy.

Toaddressthesechallenges,thisprojectproposesanefficienthybridfeatureselection technique that integrates multiple selection strategies to refine relevant features, thereby improving the detection and prediction of suspicious URLs in IoT environments. The objective is to develop a robust, scalable approach that enhances security mechanisms by accurately identifying malicious URLs while maintaining computational efficiency, ultimately contributing to safer IoTecosystems..

LITERATURESURVEY

Many scholars have done some sort of analysis on the statistics of phishing URLs. Our technique incorporates key concepts from past research. We review past work in the detection of phishing sites using URL features, which inspired our current approach.

Happydescribephishingas"oneofthemostdangerousways forhackerstoobtainusers' accountssuchasusernames,accountnumbersandpasswords,withouttheirawareness." Users are ignorant of this type of trap and will ultimately, they fall into Phishing scam. Thiscouldbeduetoalackofacombinationoffinancialaidandpersonalexperience,as wellasalackofmarketawarenessorbrandtrust.Inthisarticle,Mehmetetal.suggested methodforphishingdetectionbasedonURLs.Tocompare theresults,theresearchers utilizedeightdifferent algorithmsto evaluate theURLsof threeseparate datasetsusing various sorts of machine learning methods and hierarchical architectures. The first method evaluates various features of the URL; the second method investigates the website'sauthenticitybydeterminingwhereitishostedandwhooperatesit;andthethird method investigates the website's graphic presence. We employ Machine Learning techniques and algorithms to analyze these many properties of URLs and websites. Garera et al. classify phishing URLs using logistic regression over hand-selected variables. The inclusion of red flag keywords in the URL, as well as features based on Google's Web page and Google's Page Rank quality recommendations, are among the features. Without accesstothesameURLsandfeaturesasourapproach,it'sdifficultto conduct a direct comparison.

In this research, Yong et al. created a novel approach for detecting phishing websites that focuses on detecting a URL which has been demonstrated to be an accurate and efficient way of detection. To offer you a better idea, our new capsule-based neural network is divided into several parallel components. One method involves removing shallowcharacteristicsfromURLs.Theothertwo,ontheotherhand,constructaccurate feature representations of URLs and use shallow features to evaluate URL legitimacy. The final output of our system is calculated by adding the outputs of all divisions. Extensive testing on a dataset collected from the Internet indicate that our system can compete with other cutting-edge detection methods while consuming a fair amount of time.Forphishingdetection,VahidShahrivarietal.used machinelearningapproaches. They used the logistic regression classification method, KNN, Adaboost algorithm, SVM, ANN and random forest. They found random forest algorithm provided good accuracy. Dr. G. Ravi Kumar used a variety of machine learning methods to detect phishingassaults.For

improvedresults,theyusedNLPtools.Theywereabletoachieve high accuracy using a Support Vector Machine and data that had been pre-processed usingNLPapproaches.AmaniAlsalemetal.trieddifferent machinelearningmodelfor phishing detection but was able to achieve more accuracy in random forest. Hossein et al. created the "Fresh-Phish"open-source framework. This system can be used to build machine-learning data for phishing websites. They used a smaller feature set and built thequeryinPython.Theycreateabig,labelleddatasetandtestseveralmachine-learning classifiers on it. Using machine-learning classifiers, this analysis yields very high accuracy.Thesestudieslookathowlongittakestotrainamodel.X.Zhangsuggesteda phishing detection model based on mining the semantic characteristics of word embedding,semanticfeature,andmulti-scalestatisticalfeaturesinChinesewebpagesto detect phishing performance successfully. To obtain statistical aspects of web pages, elevenfeatureswereretrievedanddividedintofiveclasses .Toobtainstatisticalaspects of web pages, eleven featureswere retrieved and divided into five classes. To learn and evaluate the model, AdaBoost, Bagging, Random Forest, and SMO are utilized. The legitimateURLsdatasetcamefromDirectIndustryonline guides,andthephishingdata came from China's Anti-Phishing Alliance. With novel methodologies, M. Aydin approaches a framework for extracting characteristics that is versatile and straightforward. Phish Tank provides data, and Google provides authentic URLs. C# programmingandRprogrammingwereutilizedtoobtainthetextattributes.Thedataset and third-party service providers yielded a total of 133 features. The feature selection approaches of CFS subset based and Consistency subset-based feature selection were employedandexaminedwiththeWEKA tool. TheperformanceoftheNaveBayesand Sequential Minimal Optimization (SMO) algorithms was evaluated, and the author prefers SMO to NB for phishing detection.

EXISTING SYSTEM

Anti-phishing strategies are broadly categorized into two main approaches: user education and technical defenses. While educating internet users (netizens) about phishing risks remains important, technical countermeasures have become increasingly crucial given the sophistication and frequency of modern phishing attacks. This paper focuses primarily on reviewing technical defense mechanisms proposed in recent years to combat phishing threats more effectively.One of the most efficient technical defenses involves identifying and blocking phishing websites before users can interact with them. Phishing websites are designed to mimic legitimate websites and deceive users into revealing sensitive information such as login credentials, credit card numbers, or personal identity details. Detecting such websites in real-time is a vital step in disrupting the attack cycle.

With the advancement of machine learning (ML) technologies, researchers have developed a wide array of ML-based techniques to identify phishing websites. These techniques analyze multiple features of URLs, website content, and behavior to predict whether a site is malicious. Supervised learning models like Random Forest, Support Vector Machines (SVM), and deep learning models have shown high levels of accuracy in distinguishing phishing websites from legitimate ones. Feature extraction, dataset quality, and continuous learning from evolving threats contribute significantly to improving prediction performance.

This paper surveys these recent machine learning approaches, evaluates their effectiveness, and highlights their potential for deployment in real-time environments. The emphasis is on real-world applicability, scalability, and minimizing false positives to ensure user safety without compromising experience.

Disadvantages of the Existing Technical Defense Systems:

1. **High False Positive Rates:** Many ML-based systems may incorrectly classify legitimate websites as phishing, disrupting the user experience and trust in the detection mechanism.
2. **Dependence on Training Data:** The effectiveness of ML models heavily relies on the quality and freshness of training data. Outdated or biased datasets can lead to poor detection of new phishing patterns.
3. **Inability to Detect Zero-Day Phishing Attacks:** Most ML models require prior knowledge of phishing patterns. Detecting entirely new (zero-day) phishing websites remains a challenge due to the lack of historical data.
4. **Adversarial Evasion Techniques:** Cyber attackers constantly evolve their tactics to bypass detection algorithms. Slight modifications in phishing websites may fool even robust ML classifiers.
5. **Resource Intensive:** Deep learning and ensemble models can require substantial computational resources, making them difficult to deploy in low-resource or real-time IoT and mobile environments.

PROPOSEDSYSTEM

One of the most frequent and dangerous forms of phishing attacks is spear phishing, where cybercriminals impersonate trusted entities such as well-known institutions, domains, or organizations to deceive victims into disclosing sensitive information. Unlike generic phishing attacks that target a broad audience, spear phishing is highly targeted and personalized, making it more convincing and effective.

Attackers craft emails that appear legitimate by incorporating personal details about the victim, such as their full name, job title, company name, or recent activities. These emails often contain malicious URLs or attachments that, when clicked, lead to fake login pages or initiate malware downloads. The primary objective is to steal confidential data like login credentials, passwords, banking details, or credit card information.

A specialized form of spear phishing is known as whaling, where the attacker targets high-profile individuals within an organization, such as the CEO, CFO, or other top executives. These attacks are designed with extreme precision, as executives hold access to sensitive corporate resources. In a whaling attempt, the attacker may send an email that appears to be from a known associate or internal department, embedding a malicious URL that infects the target's system or redirects them to a counterfeit login portal.

These forms of phishing are especially dangerous in enterprise environments as they can lead to massive data breaches, financial fraud, and loss of corporate secrets.

Advantages of Spear Phishing and Whaling (from the attacker's perspective)

1. **Low Level of Customization in Some Cases**: While some spear phishing emails are highly personalized, many others lack deep customization. This makes it easier for attackers to scale their campaigns without the need for individual targeting.
2. **Use of Basic Social Information**: Even without deep personalization, attackers may still include social cues such as company names or positions scraped from public profile. This creates a false sense of legitimacy, increasing the chances that a victim will interact with the message.
3. **Exploitation of Human Behavior**: Attackers capitalize on urgency, curiosity, or trust. Messages may include prompts like "click to accept a friend invitation" or urgent financial requests, which victims may act on without thinking critically.
4. **Generic Information Still Effective**: Spear phishing messages may not always contain direct personal references, but general terms such as "your account" or "your company's security

team" can still convince recipients to click the link.

5. **Access to Broader Networks**: By targeting a high-profile individual and compromising their credentials, attackers can gain access to internal systems and use that access to further exploit co-workers or the organization as a whole.

RESULTS

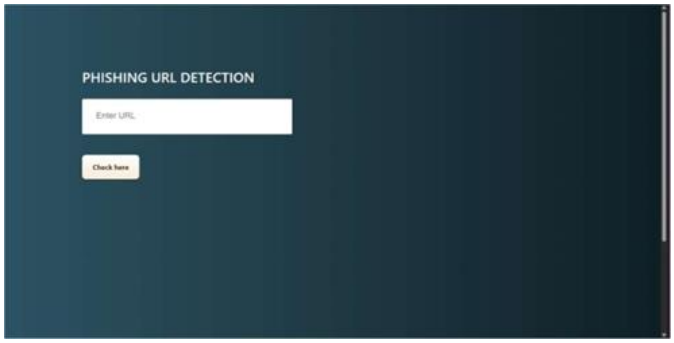

Fig : Home Screen

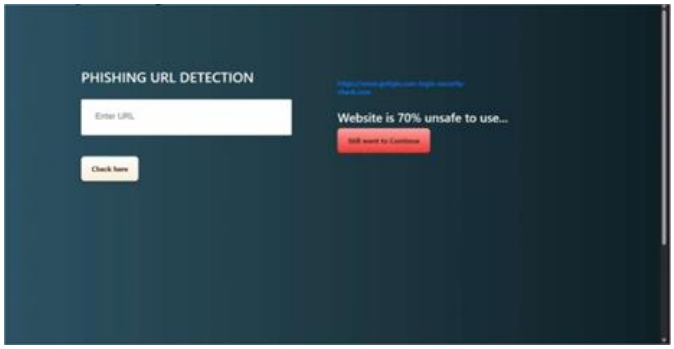

Fig : Output Screen

CONCLUSION

This survey presented various algorithms and approaches to detect phishing websites by severalresearchersinMachineLearning.Onreviewingthepapers,wecametoaconclusion that most of the work done by using familiar machine learning algorithms like Naïve Bayesian, SVM, Decision Tree and Random Forest. Some authorsproposed a new system likePhishScoreandPhishCheckerfordetection.Thecombinationsoffeatureswithregards to accuracy, precision, recall etc. were used. Experimentally successful techniques in detectingphishingwebsiteURLsweresummarizedasphishing websites increases day by day,somefeaturesmaybeincludedorreplaced with new ones to detect them.

References

[1]. 'APWG | Unifying The Global Response To Cybercrime' (n.d.) available: https://apwg.org/

[2]. 14 Types of Phishing Attacks That IT Administrators Should Watch For [online] (2021)https://www.blog.syscloud.com,available: https://www.blog.syscloud.comtypes-o f-phishing/

[3]. Lakshmanarao, A., Rao, P.S.P., Krishna, M.M.B. (2021) 'Phishing website detection using novel machine learning fusion approach', in 2021 International ConferenceonArtificialIntelligenceandSmartSystems(ICAIS),Presentedatthe2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 1164– 1169

[4]. H.Chapla,R.KotakandM.Joiser,"AMachineLearningApproachforURLBased Web Phishing Using Fuzzy Logic as Classifier", 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 383-388, 2019, July

[5]. Vaishnavi, D., Suwetha, S., Jinila, Y.B., Subhashini, R., Shyry, S.P. (2021) 'A ComparativeAnalysisofMachineLearningAlgorithmsonMaliciousURLPrediction', in 2021 5th International Conference on Intelligent Computing and ControlSystems (ICICCS),Presentedatthe20215thInternationalConferenceonIntelligentComputing and Control Systems (ICICCS), 1398–1402

[6]. Microsoft,MicrosoftConsumersafety report. https://news.microsoft.com/ensg/2014/02/11/microsoft-consumersafety-index-reveals- impact-of-poor-online-safety-behaviours-in-singapore/sm.001xdu50tlxsej410r11kqvks u4nz.

[7]. Internal Revenue Service, IRS E-mail Schemes. Available at

https://www.irs.gov/uac/newsroom/consumers-warnedof-new-surge-in-irs-email-schem es-during-2016-tax-season-tax-industry-also-targeted.

[8]. Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S. (2007), A comparison of machine learning techniques for phishing detection. Proceedingsof theAnti-phishingWorkingGroups2ndAnnualECrime Researchers Summit on - ECrime '07.doi:10.1145/1299015.1299021.

[9]. E.,B.,K.,T.(2015).,PhishingURLDetection:AMac hineLearningandWeb Mining-based Approach. International Journal ofComputerApplications,123(13),46-50.doi:10.5120/ijca2015905665.

[10]. WangWei-Hong,LVYin-Jun,CHENHui-Bing,FANGZhao-Lin.,AStatic Malicious Javascript Detection Using SVM, InProceedings of the 2nd International Conference on Computer Science and Electrical Engineering (ICCSEE 2013).