ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 18, Issue.3 July 2024

URL Based Phising Detection Using Ensemble Hybrid Models With Machine Learning

E.ROHINIKUMAR REDDY [1] , Dr.MEERAVALI SHAIK[2]

[1] PG Student, Department of CSE, MALLA REDDY UNIVERSITY,Hyderabad
[2]Professor, Department of CSE,MALLA REDDY UNIVERSITY,Hyderabad

ABSTRACT:

Phishing poses a significant danger to the Internet, since the monetary losses associated with it are increasing. Effective detection of phishing websites relies heavily on feature engineering, which in turn requires a thorough grasp of certain traits. Although including features from several dimensions enhances the effectiveness of the detection approach, it is often time-consuming and requires significant effort. To address these issues, we provide a complete strategy for detecting phishing attempts utilizing an advanced deep learning method known as Multidimensional Feature Phishing Detection (MFPD).By using deep learning techniques, we rapidly categorize the data by extracting distinctive characteristics from the given URL. This is the first phase of the procedure. This method does not need any prior understanding of phishing or assistance from other sources. In the second phase, we generate a set of multidimensional attributes by merging the initial deep learning classification outcome with URL statistics attributes, website code attributes, and webpage text attributes. By using this technique, we may reduce the time it takes to detect by selecting the appropriate threshold. Our method achieves a 98.99% accuracy rate by using a dataset that comprises a large number of both real and fraudulent URLs, with just a 0.59% mistake rate in misidentifying legal URLs as fraudulent. The results of our experiment demonstrate an improvement in detection accuracy via meticulous adjustment of the threshold.

Index Terms: Phishing detection, feature engineering, deep learning, URL analysis, multidimensional features, fast detection, URL statistics, false positive rate, detection accuracy.

I. INTRODUCTION

The Internet has significantly enhanced human civilization, becoming an indispensable infrastructure. However, despite its benefits, major security weaknesses like malware, phishing, and privacy breaches jeopardize customers' financial well-being. According to the Anti-Phishing Working Group (APWG), phishing, defined as fraud that exploits social engineering and technological deception to steal passwords and personal information, can lead to identity theft, privacy invasion, and monetary harm. Kaspersky Lab's data from 2017 reveals that at least 29.4% of user PCs experienced malware-classified online assaults. Furthermore, online antivirus software flagged a total of 199,455,606 distinct URLs as potentially harmful. Additionally, the percentage of recognized financial phishing cases increased from 47.5% to almost 54% in 2017.

Nowadays, phishing is one of the most significant threats to online security. Phishing methods have progressed beyond more traditional channels such as pop-ups, SMS, and email due to the rise of mobile internet and social networks. More contemporary methods are now a part of it, such as spear phishing, phony mobile applications, and phishing QR codes. In addition, more and more phishing attacks are aimed at websites with SSL and HTTPS certificates, hoping to exploit consumers' trust in these security protocols. More and more phishing schemes are popping up, making detection even more of a challenge. Dedicated security researchers and experts have laboriously developed detection techniques to combat phishing attacks. As a safety measure to prevent phishing attempts, most common web browsers have blacklists and whitelists. Google maintains an up-to-date list of potentially harmful websites, and users may verify the security of a URL using the Google Safe Browsing APIs. Despite being fast, user-friendly, and having a low false positive rate, blacklist- and whitelist-based detection systems are not particularly effective because of how slowly they update. To add 47%-83% of phishing websites to blacklists takes around 12 hours, and to include 63% of them takes roughly 2 hours. This delays the process of identifying and blocking new phishing websites.

The detection of phishing websites makes extensive use of machine learning techniques. These strategies use machine learning to assess hazardous URLs and phishing websites by discerning their unique characteristics. Before classification, the majority of contemporary mainstream systems typically gather statistical data from the URL and host, as well as relevant site attributes such as layout, CSS, and content. However, these algorithms often analyze URLs or retrieve data from a singular perspective, disregarding intricate characteristics of phishing sites and sometimes including irrelevant information that reduces detection accuracy.

URL character sequences provide an automatically generated feature that removes the subjective nature of manually selected attributes and does not require any prior knowledge of phishing or assistance from other sources. Extracting connections and semantic information from these sequences is challenging.

Our solution, the Multidimensional Feature Phishing Detection (MFPD) approach, utilizes deep learning to rapidly identify and address these issues. Initially, the distinctive character sequence attributes of the provided URL are extracted and utilized in a deep learning classification model to maximize effectiveness. Specifically, a Convolutional Neural Network (CNN) leverages the URL's letters to identify local correlation features. Phishing sites frequently mimic legitimate URLs by altering or appending characters, thus disrupting the URL's sequential structure.An LSTM (Long Short-Term Memory) network is used to extract relationships and context semantics from the character sequences. The softmax layer is used for the ultimate categorization. Prior understanding of phishing is not required for this first step, CNN-LSTM.

In the second step, we achieve the first classification outcome by combining several characteristics such as URL statistical data, website code features, webpage text features, and multidimensional features. The categorization is performed using XGBoost. Despite the exceptional precision of this multidimensional feature detection method, it results in longer detection times due to the need to extract features from various aspects. On the other hand, the URL character sequence method is more efficient in terms of detection time since it directly scans the URL. To enhance the output judgment of the softmax classifier, we provide a threshold that reduces detection time. This allows us to find a balance between detection accuracy and speed. If the deep learning result reaches or exceeds the threshold, the detection result is generated instantly; otherwise, the second step is executed.

Our primary contributions are as follows:

- We provide a precise explanation of the MFPD approach and analyze the problem of phishing detection.
- We built a real dataset consisting of 1,021,758 phishing URLs retrieved from phishtank.com and 989,021 legitimate URLs retrieved from dmoztools.net.
- We provide a detailed explanation of the MFPD technique and conduct thorough testing on our dataset to demonstrate the efficacy of our approach in terms of speed, accuracy, and false positive rate.
- We propose using a dynamic category decision algorithm (DCDA). To reduce detection time, we tweak the criteria used by the softmax classifier to judge its output and define a threshold.

The paper is organized as follows: Section II contains relevant studies on the identification of phishing websites. The MFPD framework is presented in Section IV. The MFPD process, which includes the combination of CNN-LSTM and multidimensional features, is fully explained. Section V evaluates the effectiveness of our plan. Finally, Section VII concludes the paper and discusses future research directions.

II.PROBLEM STATEMENT

While existing deep learning-based methods for phishing website detection have made significant progress, they still face several limitations. Selvaganapathy et al.'s model, which uses stacked restricted Boltzmann machines and multiple classifiers, improves detection accuracy but involves complex feature selection and classifier combination processes. Bahnsen et al.'s approach, which leverages LSTM for URL character sequence classification, shows better performance than traditional methods like RF but may not fully exploit the potential of combined deep learning architectures. Additionally, these methods often rely on prior knowledge and extensive feature engineering, which can be time-consuming and limit scalability.

Our innovative solution to these problems is to use convolutional neural networks (CNNs) for feature extraction and long short-term memory (LSTM) networks for sequential dependency capture; this method views URLs as sequences of characters.

This method aims to reduce the dependency on manual feature selection, enhance detection accuracy, and improve efficiency. By integrating multidimensional features, including URL statistics, website code features, and webpage text features, our approach seeks to provide a comprehensive and robust phishing detection system.

III.PROPOSED MODEL

Phishing Website Detection Is A Significant Area Of Focus In Current Research, Encompassing Both Traditional And Deep Learning-Based Machine Learning Methods. The Effectiveness Of These Methods Heavily Relies On The Quality Of The Features Extracted. Ongoing Research Efforts Are Primarily Directed Towards Enhancing Feature Extraction And Selection Processes To Enhance Detection Accuracy.Internet Resources Are Predominantly Accessed Via Uniform Resource Locators (Urls), Comprising A Hostname And Freeurl.

Zouina et al. proposed a simple yet effective method for phishing website detection, utilizing only six URL features: size, hyphen count, dot count, numerical character count, presence of an IP address, and a similarity index. Despite its basic feature set, this approach achieves fast detection speeds due to its simplicity. However, the limited amount of experimental data may impact its applicability to a wider range of scenarios.

On the other hand, Le et al. suggested a different approach for identifying phishing websites by employing AROW (Adaptive Regularization of Weights) to extract lexical features from URL strings. This method effectively handles noise in the training data while maintaining high detection accuracy.

We provide a model that improves phishing website identification by integrating deep learning with conventional feature-based methods. The MFPD system uses a quick deep learning technique to combine traditional URL data with information retrieved from website code, content, and character sequences in order to detect phishing attempts. With this connection, we want to increase precision without slowing down detection times. Further, in order to decrease false positives and fine-tune the detection threshold, a Dynamic Category Decision Algorithm (DCDA) is used. Results show that the model is fast, accurate, and has a low false positive rate when tested on a big dataset.

IV.METHODOLOGY

First, we provide the official announcement of phishing website detection in this area. Subsequently, we delve into the overall framework of the Multidimensional Feature Phishing Detection (MFPD) approach and provide its formal definition.

A. Formal Statement of Phishing Website Detection

Phishing website detection is fundamentally concerned with distinguishing between malicious phishing websites and legitimate ones. Formally, given a set UUU comprising all URLs (U={u|u=x,x∈url,i∈N+}U = \{u | u = x, x \in url, i \in \mathbb{N}^+\}U={u|u=x,x∈url,i∈N+}) and denoted as |U|=n|U| = n|U|=n, we define the problem as follows:

Consider two sets: CasC_{as}Cas representing phishing (Cp={c|c=p,p∈phishing}Cp = \{c | c = p, p \in phishing\}Cp={c|c=p,p∈phishing}) and ClC_{l}Cl representing legitimate (Cl={c|c=l,l∈legitimate}Cl = \{c | c = l, l \in legitimate\}Cl={c|c=l,l∈legitimate}). Let uiu_iui represent a suspicious URL. The phishing website detection problem can be formally stated as the task of determining whether uiu_iui belongs to the phishing category (ui∈Cpu_i \in Cpui∈Cp) or the legitimate category (ui∈Clu_i \in Clui∈Cl).

B. Overall Framework of MFPD

The MFPD approach is designed to address the phishing website detection problem through a comprehensive and multidimensional feature-based approach. The framework of MFPD can be outlined as follows:

1. **Character Embedding of URLs**: Initially, MFPD processes the URLs to extract character sequence features. Each URL uiu_iui undergoes character embedding to standardize its length and encode its character sequence into a vector representation. This process aims to

capture the intrinsic characteristics of each URL efficiently.

2. **Extraction of URL Statistical Features**: In addition to character embedding, MFPD extracts statistical features from the URLs. These features include attributes such as URL size, number of hyphens, number of dots, presence of numeric characters, and other relevant metrics. The statistical features provide complementary information to the character embeddings, enhancing the overall feature representation.
3. **Webpage Code and Text Features Extraction**: MFPD also considers features derived from the webpage associated with each URL. This involves analyzing the HTML code structure of the webpage as well as the textual content present on the webpage. By extracting relevant information from both the code and text domains, MFPD aims to capture additional contextual cues for improved detection performance.
4. **Multidimensional Feature Fusion**: The extracted character sequence features, URL statistical features, and webpage-related features are integrated into a comprehensive multidimensional feature representation for each URL. This fusion process results in a feature vector that encapsulates diverse aspects of the URL and its associated webpage content, facilitating robust detection of phishing websites.

Through the amalgamation of character embedding, URL statistical features, and webpage-related features, MFPD leverages a rich feature representation to effectively discriminate between phishing and legitimate URLs. The multidimensional feature-based approach adopted by MFPD enables accurate and efficient phishing website detection.

V.IMPLEMENTATION

A. Data Collection and Indicators

For this research, we utilized data collected from the Internet. First, positive samples were drawn from 1,021,758 URLs in the historical phishing data from the PhishTank website, spanning the years 2014 to 2018. Then, 989,021 legitimate URLs were obtained from the dmoztools.net open catalog to serve as negative samples. The combined dataset, DATA, comprises 2,010,779 URLs.

Due to the transient nature of phishing URLs, feature extraction cannot utilize large data sets effectively. To address this challenge, two datasets—DATA1 and DATA2—were constructed. The first dataset, DATA1, contains 22,445 surviving phishing URLs from the positive samples of DATA and 22,390 randomly selected accessible URLs from the negative samples. The remaining data is included in dataset DATA2.

B. CNN-LSTM Experiment

In the DATA2 experiment, 5-fold cross-validation is used. The CNN-LSTM algorithm's parameters are fine-tuned by comparing the average length of malicious and legitimate webpage samples. With 20 training epochs, we achieve a balanced training time and avoid overfitting.

We compare the CNN-LSTM approach to three classic deep neural networks: RNN, LSTM, and CNN itself. To ensure fairness in comparison, these networks are constructed identically. The structure for the CNN-LSTM algorithm is Input -> Conv -> Maxpool -> LSTM -> Softmax. For the comparison models, the structures are as follows: CNN-CNN is Input -> Conv -> Maxpool -> Conv -> GlobalMaxpool -> Softmax; RNN-RNN is Input -> RNN1 -> RNN2 -> Softmax; LSTM-LSTM is Input -> LSTM1 -> LSTM2 -> Softmax.

The computations are executed on a robust server equipped with 64GB of RAM, an i5 processor, and GTX 1080ti GPUs. This setup ensures that deep learning models can rapidly iterate through massive datasets.

C. Multidimensional Feature Algorithm Analysis

By classifying and extracting features from DATA1 using four ensemble learning methods, we may evaluate the efficacy of the multidimensional feature technique. The findings demonstrate that, when compared to other approaches, XGBoost obtains the best accuracy while minimizing cost, false negative rate (FNR), and false positive rate (FPR).

Additionally, conventional feature extraction techniques are contrasted with the CNN-LSTM multidimensional feature algorithm. Using a multidimensional feature approach greatly enhances accuracy while decreasing

ISSN2454-9940
www.ijsem.org
Vol 18, Issue.3 July 2024

cost, false positive rate, and false negative rate.

Using hybrid features from various sources, MFPD achieves better accuracy and F1 score than other approaches, according to experimental data (Table 5).

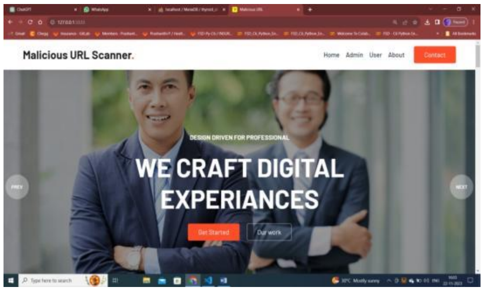
D. Dynamic Category Decision Algorithm Validation

To ensure its effectiveness, the DCDA dynamic category choice algorithm is subjected to 5-fold cross-validation on DATA1. To minimize detection cost and allow speedy and reliable phishing website detection, DCDA seeks to determine the ideal threshold (α).

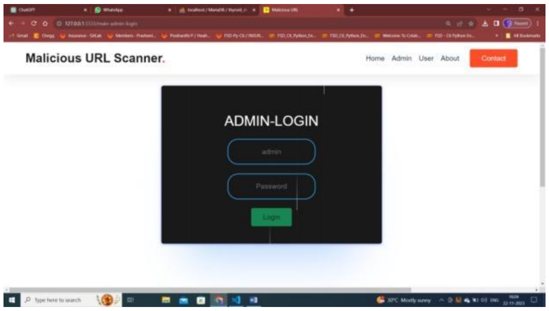
Experimental results indicate that setting the threshold at approximately α=355\alpha = 355α=355 yields stable detection accuracy and low detection cost, almost equivalent to multidimensional feature detection. DCDA significantly reduces the workload by assigning only 28% of websites for multidimensional feature detection at this threshold value.

VI.RESULTS

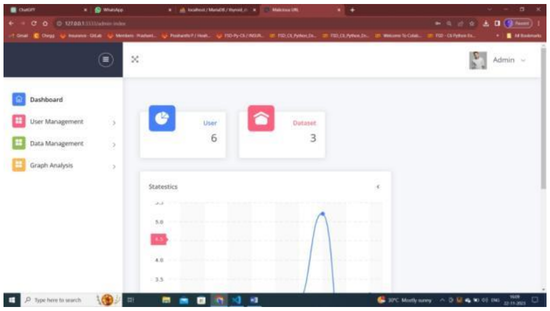
Home Page:



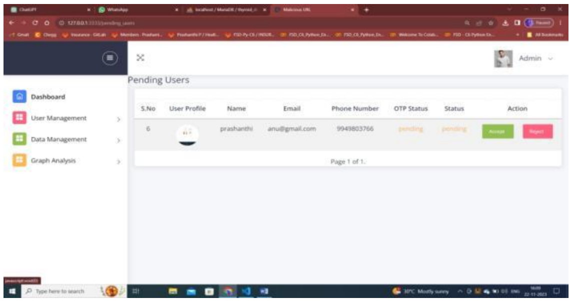
Admin Page:



Admin Dashboard:



Pending Users:

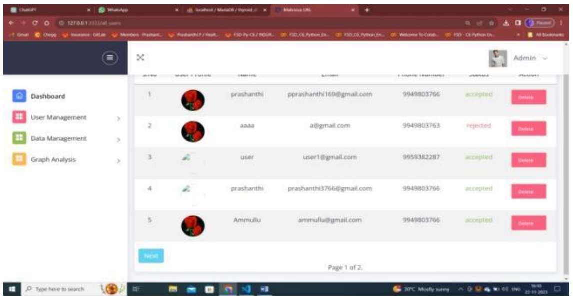


Pending Users

S.No	User Profile	Name	Email	Phone Number	OTP Status	Status	Action
6		prashanthi	anu@gmail.com	9949803766	pending	pending	Accept Reject

Page 1 of 1.

All Users:

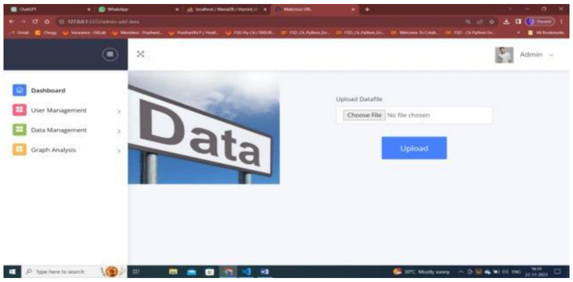


1	prashanthi	pprashanthi169@gmail.com	9949803766	accepted	Delete
2	aaaa	a@gmail.com	9949803763	rejected	Delete
3	user	user1@gmail.com	9959382287	accepted	Delete
4	prashanthi	prashanthi3766@gmail.com	9949803766	accepted	Delete
5	Ammullu	ammullu@gmail.com	9949803766	accepted	Delete

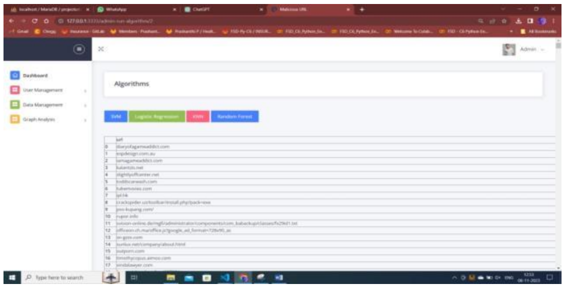
Next

Page 1 of 2.

Upload dataset:



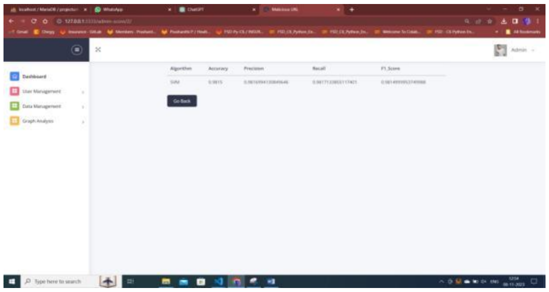
Algorithms Page:-



Algorithms

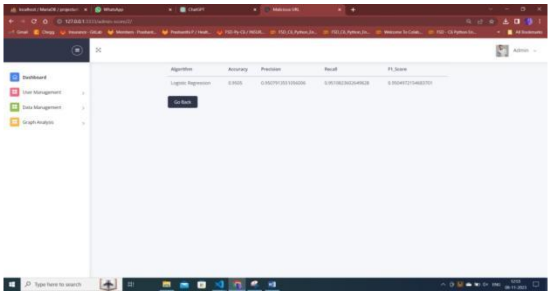
	url
0	diaryofagameaddict.com
1	espdesign.com.au
2	iamagameaddict.com
3	kalantzis.net
4	slightlyoffcenter.net
5	toddscarwash.com
6	tubemoviez.com
7	ipl.hk
8	crackspider.us/toolbar/install.php?pack=exe
9	pos-kupang.com/
10	rupor.info
11	svision-online.de/mgfi/administrator/components/com_babackup/classes/fx29id1.txt
12	officeon.ch.ma/office.js?google_ad_format=728x90_as
13	sn-gzzx.com
14	sunlux.net/company/about.html
15	outporn.com
16	timothycopus.aimoo.com
17	xindalawyer.com

SVM:-



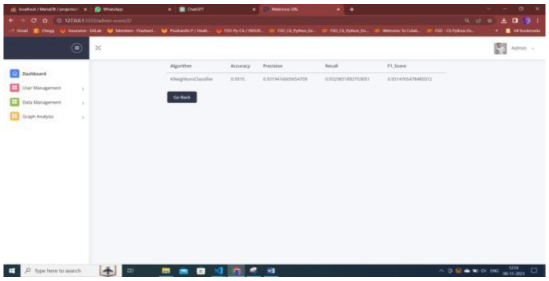
Algorithm	Accuracy	Precision	Recall	F1_Score
SVM	0.9815	0.9816994130849646	0.9817133855117421	0.9814999953749986

Logistic Page:-



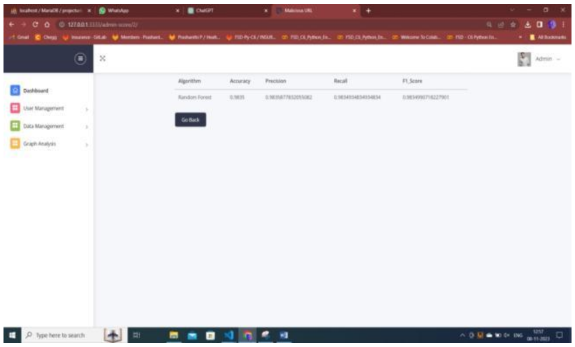
Algorithm	Accuracy	Precision	Recall	F1_Score
Logistic Regression	0.9505	0.9507913551094006	0.9510823602649628	0.9504972154683701

KNN :-



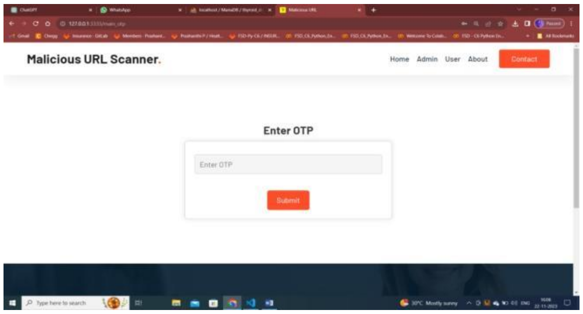
Algorithm	Accuracy	Precision	Recall	F1_Score
KNeighborsClassifier	0.9315	0.9319416605654709	0.9329831892753051	0.9314765478485012

Random Forest:-

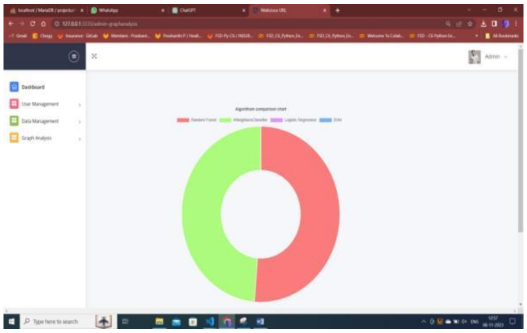


Algorithm	Accuracy	Precision	Recall	F1_Score
Random Forest	0.9835	0.9835877832055082	0.9834934834934834	0.9834990718227901

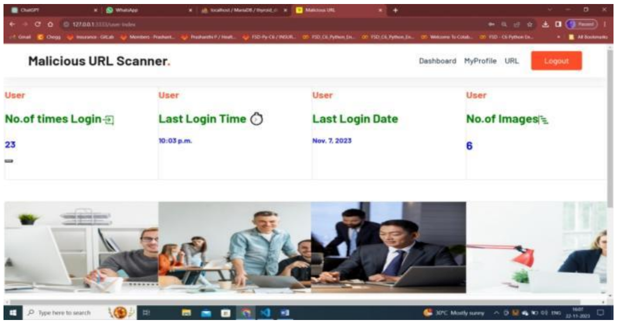
Otp Page:



Graph:-



User dashboard:



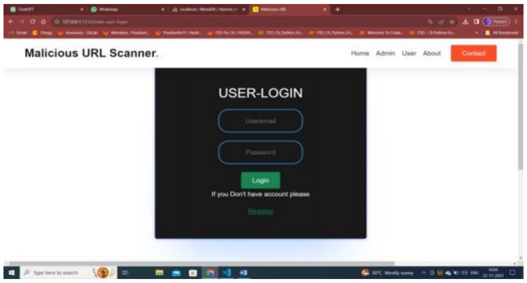
User
No.of times Login
23

User
Last Login Time
10:03 p.m.

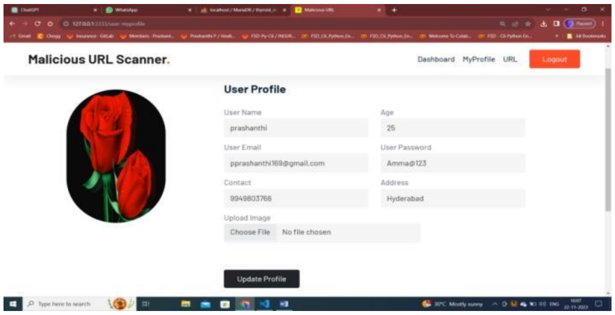
User
Last Login Date
Nov. 7, 2023

User
No.of Images
6

User Page:



Profile Page:



User Profile

User Name: prashanthi

Age: 25

User Email: pprashanthi169@gmail.com

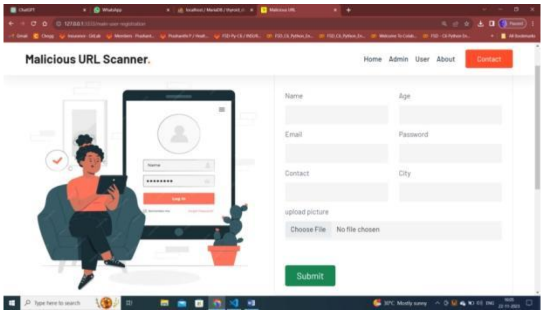
User Password: Amma@123

Contact: 9949803766

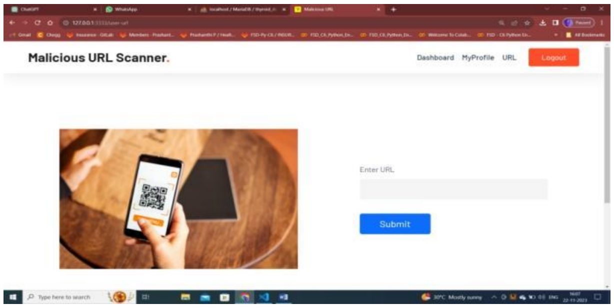
Address: Hyderabad

Upload Image: Choose File No file chosen

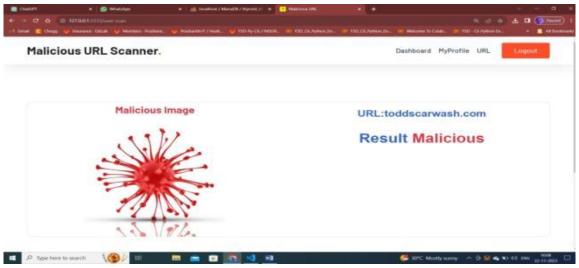
Register Page:



Url Page:



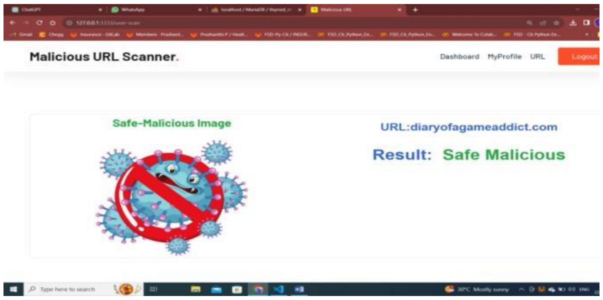
Result Page when bad:



URL:toddscarwash.com

Result Malicious

Result page when good:



URL:diaryofagameaddict.com

Result: Safe Malicious

Aboutus:

Digital Marketing

It is a long established fact that a reader will be distracted by the readable content of a page

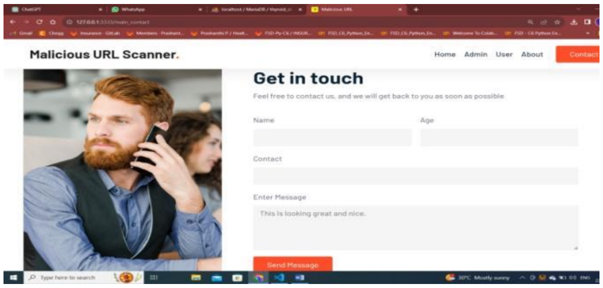
E-mail Marketing

It is a long established fact that a reader will be distracted by the readable content of a page

Buisness Marketing

It is a long established fact that a reader will be distracted by the readable content of a page

Contact us:



Get in touch

Feel free to contact us, and we will get back to you as soon as possible

Name

Age

Contact

Enter Message

This is looking great and nice.

VII.CONCLUSION

The ideal phishing website detection system would respond quickly, accurately, and produce minimal false positives. The proposed approach, Multidimensional Feature Phishing Detection (MFPD), achieves this goal. By combining multidimensional feature detection with a dynamic category selection algorithm, MFPD can quickly and accurately identify phishing attempts using URL character sequences without any prior knowledge. Trials on a dataset with millions of both safe and dangerous URLs showed fast detection rates, low false positive rates, and great accuracy.

In the future, we aim to refine our approach by extracting characteristics from website code and content using deep learning approaches. Additionally, we plan to develop a browser plugin for our method to make it even more widely available and user-friendly.

VIII.REFERENCES

1. APWG. Phishing Attack Trends Report-1Q 2018. Available: https://apwg.org/resources/apwg-reports/, accessed May. 5, 2018.
2. Kaspersky Security Bulletin: Overall statistics for 2017. Available: https://securelist.com/ksb-overall-

statistics-2017/83453/, accessed Jul.12, 2018.

3. A.Ahmad Y, M. Selvakumar, A. Mohammed, A. Mohammed, and A. S. Samer. "TrustQR: A New Technique for the Detection of Phishing Attacks on QR Code," Adv. Sci. Lett., vol. 22, no. 10, pp. 2905-2909, Oct. 2016.
4. C. Inez and F. Baruch. "Setting Priorities in Behavioral Interventions: An Application to Reducing Phishing Risk," Risk Anal., vol. 38, no. 4, pp. 826-838, Apr. 2018.
5. G. Diksha and J. A. Kumar. "Mobile phishing attacks and defense mechanisms: State of art and open research challenges," Comput. Secur., vol. 73, pp. 519-544, Mar. 2018.
6. Google Safe Browsing APIs. Available: https://developers.google.com/safe-browsing/v4/, accessed Oct. 1, 2018.
7. S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. "An Empirical Analysis of Phishing Blacklists," in Proc. 6th Conf. Email Anti-Spam (CEAS'09), Jul. 2009, pp. 59-78.
8. K. Jain and B. B. Gupta. "A novel approach to protect against phishing attacks at client side using auto-updated white-list," Eurasip J. Inf. Secur., vol. 2016, no. 1, May. 2016.
9. M. Zouina and B. Outtaj. "A novel lightweight URL phishing detection system using SVM and similarity index," Human-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.
10. E. Buber, Ö. Demir, and O. K. Sahingoz. "Feature selections for the machine learning based detection of phishing websites," in Proc. IEEE Int. Artif. Intell. Data Process. Symp. (IDAP), Sep. 2017.
11. J. Mao, J. Bian, W. Tian, S. Zhu, W. Tao, A. Li, and Z. Liang. "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," Procedia Comput. Sci., vol. 129, pp. 224-230, Jan. 2018.
12. J. Mao, W. Tian, P. Li, T. Wei, and Z. Liang. "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," IEEE Access, vol.5, no. 99, pp. 17020-17030, Aug. 2017.

13. J. Cao, D. Dong, B. Mao, and T. Wang. "Phishing detection method based on URL features," J. Southeast Univ.-Engl. Ed., vol. 29, no. 2, pp. 134-138, Jun. 2013.
14. S. C. Jeeva and E. B. Rajsingh. "Phishing URL detection-based feature selection to classifiers," Int. J. Electron. Secur. Digit. Forensics, vol. 9, no. 2, pp. 116-131, Jan. 2017.
15. Le, A. Markopoulou, and M. Faloutsos. "PhishDef: URL names say it all," in Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM), Sep. 2010, pp. 191-195.
16. R. Verma and K. Dyer. "On the character of phishing URLs: Accurate and robust statistical learning classifiers," in Proc. 5th ACM Conf. Data Appl. Secur. Priv. (ACM CODASPY), Mar. 2015, pp. 111-122.
17. Y. Li, S. Chu, and R. Xiao. "A pharming attack hybrid detection model based on IP addresses and web content," Optik, vol. 126, no. 2, pp. 234-239, Nov. 2014.
18. G. Xiang and J. Hong. "A hybrid phish detection approach by identity discovery and keyword retrieval," in Proc. Int. Conf. World Wide Web (WWW 2009), Oct. 2009, pp. 571-580.
19. G. Xiang, J. Hong, C. P. Rose, and L. Cranor. "Cantina+: A feature-rich machine learning framework for detecting phishing websites," ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, pp. 21, Sep. 2011.
20. S. Marchal, K. Saari, N. Singh, and N. Asokan. "Know your phish: Novel techniques for detecting phishing sites and their targets," in Proc. IEEE 36th Int. Conf. Distrib. Comput. Syst. (ICDCS), Jun. 2016, pp. 323-333.