IJIEMR Transactions, online available on 24th Dec 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 12

DOI: 10.48047/IJIEMR/V11/ISSUE 12/42

Title **Phishing URL Detection using Hybrid Ensemble Model**

Volume 11, ISSUE 12, Pages: 304-310

Paper Authors

Adepu Vedansh, Padigapati Soumith Reddy

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Phishing URL Detection using Hybrid Ensemble Model

Adepu Vedansh, Vedanshadepu9999@gmail.com
Padigapati Soumith Reddy, soumithreddypadigapati@gmail.com

ABSTRACT: Recently, it has been reported that consumers have lost money after unintentionally doing transactions using links provided by strangers. People can be tricked in a number of methods, including through email, SMS, calls, phoney websites, and even in-person. Phishing assaults are what are used to trick or scam individuals. Hence, in this undertaking, we will focus on one of the ways of leading a phishing assault, in particular, by utilizing a malevolent URL or site. Since these URLs are composed so near lawful URLs, it very well may be hard to discern whether somebody has visited a substantial URL or not. The qualifications of the individual visiting the connection will rapidly be given to the assailant on the off chance that there is no component set up for impeding or erasing these malignant URLs. Malignant URLs can be communicated in private or public. The target of our examination is to make an machine learning based model that guides in deciding if a URL is protected to use. The objective of this undertaking is to find destructive URLs and make a solid machine learning model for recognizing risky and dependable URLs.

Keywords *– Classification, phishing, URL, ensemble model.*

1. INTRODUCTION

Phishing continues to be a significant cause of security difficulties and the bulk of cyber-attacks in the current environment. The 2021 Cybersecurity Threat Trends study from Cisco found that 86 percent of businesses have at least one employee who has clicked on a phishing link. Over 90% of data breaches, according to the company's study, are the result of phishing. Every year, phishing assaults against clients cost American businesses $2 billion. The significant goal of this undertaking is to apply machine learning methods to distinguish hazardous URLs and caution clients of potential phishing dangers. Various techniques might be utilized to decide whether a phishing URL is certifiable or false. One methodology is to deny the URL and update it at whatever point another unsafe URL is found. Another is heuristic-based recognition, which might distinguish party time phishing attacks and contains characteristics that have been found in genuine world phishing endeavors. Be that as it may, the characteristics are not generally destined to be available in such assaults, and the bogus positive rate for location is very high. With a 98 percent accuracy rate, the Deep Learning approach is used; however, because of its intricate models, this approach has the drawback of requiring a very big dataset. Through their hidden layers, convolutional neural networks were used to recognise traits. We will have a lot of characteristics to identify because our dataset is so large, and doing so will help us find new URLs. Albeit the quantity of characteristics utilized was less than ten inferable from their little dataset, a cross breed strategy was utilized to recognize phishing URLs. At the point when another URL is provided that doesn't match any of the standards they are perceiving, this system might have constraints. In this venture, we'll utilize a half and half group model that consolidates MLP, SVM, Choice Tree, and Irregular Woodland to characterize a URL as genuine or phishing. The disadvantage of the boycott system is that it can't distinguish party time phishing assaults, which might be recognized utilizing a heuristic technique. A heuristic-based procedure's key disadvantage is that a chunk of time must pass to carry out. To improve the model's ability to distinguish phishing URLs, we'll incorporate HTML and JavaScript-based capacities.

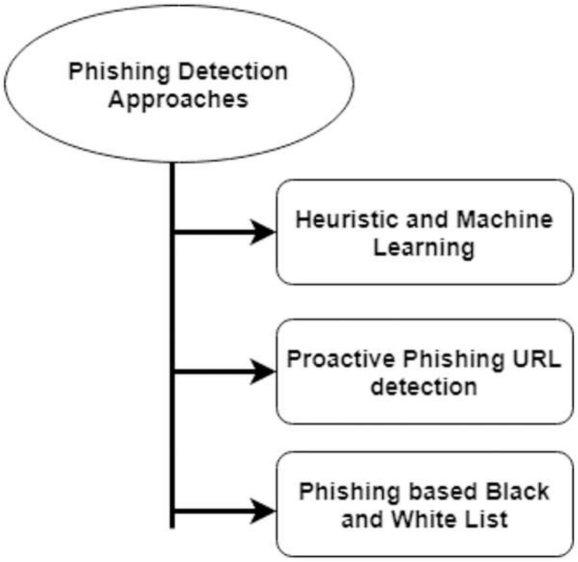

Phishing Detection Approaches
Heuristic and Machine Learning
Proactive Phishing URL detection
Phishing based Black and White List

Fig.1: Example figure

Due of how simple it is to develop a phoney website that closely resembles a real website, phishing is becoming a top worry for security researchers. In spite of the fact that specialists can detect deceitful sites, not everything clients can, and thus, certain individuals succumb to phishing tricks. The assailant's essential objective is to get login data for financial balances. Organizations in the US lose US$2 billion yearly because of their clients succumbing to phishing [1]. As per the third Microsoft Registering More secure List Report, which was distributed in February 2014, the yearly worldwide impact of phishing may reach $5 billion [2]. Since clients don't know about phishing attacks, they are climbing to a higher level. It is exceptionally difficult to battle phishing assaults since they go after client weaknesses, yet it is vital to improve phishing discovery strategies. The "boycott" strategy, which is the standard method for identifying phishing sites, includes adding boycotted URLs and Internet Protocol (IP) locations to the antivirus information base. Aggressors change the URL to appear to be true by obscurity and numerous other direct ways, for example, quick transition, in which intermediaries are consequently developed to have the site, algorithmic creation of new URLs, and so on, to avoid boycotts. This technique's essential defect is its powerlessness to distinguish party time phishing assaults. Heuristic-based identification, which considers characteristics that have been seen to exist in real phishing assaults, is equipped for spotting party time phishing assaults. In any case, the attributes are not generally destined to be available in such assaults, and the misleading positive rate for recognition is exceptionally high.

2. LITERATURE REVIEW

Phishing Websites Detection using Machine Learning Techniques:

Internet and cloud innovation upgrades as of late have essentially expanded electronic exchange, or shopper to-customer online exchanges. The assets of an organization are hurt by this development, which licenses unlawful admittance to delicate data about clients. One notable attack that hoodwinks clients into getting to hazardous substance and surrendering their data is phishing. Most phishing sites utilize a similar site point of interaction and universal resource location(URL) as the real sites. There have been a few suggested strategies for distinguishing phishing sites, including boycotts, heuristics, and so on. Be that as it may, the quantity of casualties is rising dramatically because of deficient security frameworks. Phishing attacks are bound to prevail over the Web due to its unknown and unregulated nature. Existing examination exhibits that the phishing discovery framework's presentation is compelled. An insightful technique is expected to shield customers against digital assaults. In this paper, the creator put out an machine learning based URL distinguishing proof calculation. To recognize phishing URLs, a repetitive brain network procedure is utilized. With 7900 pernicious and 5800 veritable destinations, analysts tried the proposed technique. The consequences of the examinations show that the recommended technique performs more actually in recognizing noxious URLs than additional ongoing strategies.

Phishing Website Detection using Machine Learning Algorithms:

The simplest technique for getting touchy data from accidental individuals is through a phishing assault. The objective of phishers is to get vital information, for example, login, secret key, and financial balance data. Individuals working in digital protection are currently looking for solid and steady strategies for recognizing phishing sites. To recognize legitimate and phishing URLs, this article utilizes machine learninginnovation. It concentrates and investigations numerous parts of the two sorts of URLs. Calculations, for example, CDecision Tree, Random Forest, and Support Vector Machine are utilized to recognize phishing sites. By assessing every calculation's exactness rate, misleading positive and bogus negative rates, the review intends to recognize phishing URLs and distinguish the best machine learning strategy.

Phishing URL Detection: A Machine Learning and Web Mining-based Approach:

Over the past ten years, the use and growth of online transactions have accelerated. Phishing assaults are on the rise as a result of cybercriminals' rising sophistication. Phishing, malware, and spam are rapidly spreading due to the World Wide Web's ongoing development. This study suggests a feature-based method for categorising URLs as phishing or non-phishing. By understanding the structure of URLs, several URL properties may be used. Two alternative algorithms have been used to categorise URLs. An effective classifier that determines whether or not a particular URL is phishing is created using the Random Forest machine learning method. Additionally, an unique method for detecting phishing URLs has been suggested, which mines the publicly accessible material on the URLs.

Detecting Phishing Websites, a Heuristic Approach:

Phishing is a website fraud tactic used to hunt down and steal the private data of internet users. Using social engineering strategies including SMS, voice, email, websites, and malware, the hacker deceives the user. To identify different phishing assaults, a number of methods have been developed and put into practise, including the use of blacklists and whitelists. In this review, we propose a work area program called PhishSaver that focuses on the phishing page's URL and site content. Using a desktop programme called PhishSaver, we try to identify phishing websites. To identify various phishing assaults, PhishSaver combines a blacklist with a variety of heuristic characteristics. We utilised the Google safe browsing blacklist, which is part of the GOOGLE API SERVICES, since Google updates and maintains this list on a regular basis. PhishSaver may also be used as a daemon process, which enables it to identify phishing attempts as a user browses the internet in real time. PhishSaver accepts a URL as input and returns information about whether it is a real or phishing website. The criteria for phishing detection include copyright content, title content, zero links in the HTML body, footer links with null value, and website identity. PhishSaver is faster than visual based appraisal moves toward that are utilized to recognize phishing and is fit for distinguishing party time phishing attacks that might not have been boycotted. We see that PhishSaver has a lower false negative and false positive rate and a greater accuracy rate across a wider spectrum of phishing assaults.

Phishing URL Detection: A novel hybrid Approach using Long Short-Term Memory and Gated Recurrent Units:

Phishing is one of the earliest forms of cyberattack, and it typically takes the form of disguised URLs to deceive individuals into disclosing personal information for the purpose of the attacker's malicious objectives. It is one of the simplest methods for persuading individuals to provide personal information, such as credit card data. The attackers use the URLs of malicious websites that appear to be legitimate websites in the majority of phishing attempts to successfully breach data. Consequently, it is essential to differentiate between malicious and benign URLs. In addition to three non-hybrid deep learning models—CNN (1D), LSTM, and GRU—this study suggests four hybrid deep learning models—GRU-LSTM, LSTM-LSTM, BI (GRU)-LSTM, and BI (LSTM)-LSTM. The results showed that the BI (GRU)-LSTM model performed best, with F1-Score values of 93.91%, 93.94%, and 93.38% for accuracy, precision, and recall, respectively. Therefore, the primary objective of this research is to assess the phishing URL identification accuracy, precision, recall, and f1 score of hybrid deep learning methods.

3. METHODOLOGY

In this review, we utilize a half and half outfit model to build the accuracy of phishing URL discovery. Two unmistakable techniques for group learning are alluded to as "packing" and "helping." The notable outfit learning model arbitrary timberland might be tracked down in the sacking classification. AdaBoost is a notable gathering learning model that has a place with the helping class. The helping models utilize the whole dataset, yet the sacking models just utilize a subset of it. Our model, otherwise called a cross breed troupe model, is comprised of a gathering of frail understudies who are united to exhibit their consolidated strength. The frail understudies vote on the URL class, still up in the air by the model. We will be using a variety of classifiers, creating a heterogeneous collection of models. You can increase accuracy by including more weak pupils.

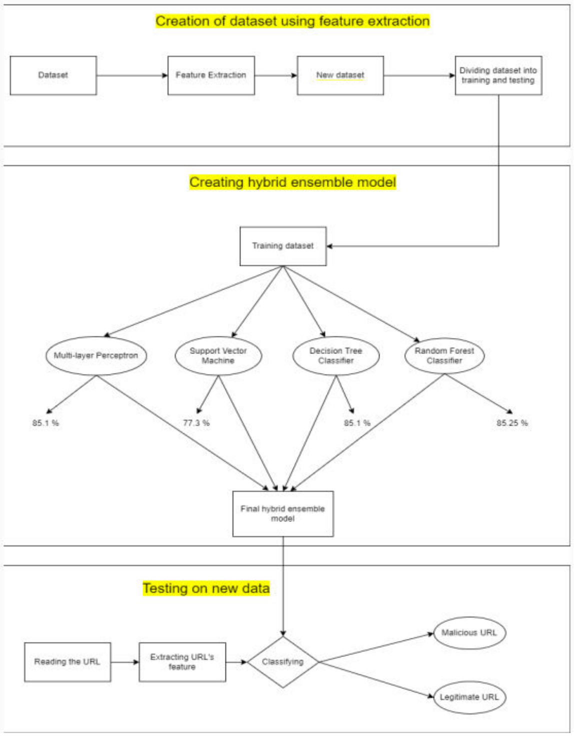


Fig.2: System architecture

MODULES:

The accompanying modules were made to do the previously mentioned project.

1) Dataset: A total of 20,000 genuine and malicious URLs make up the dataset under consideration.

2) Getting features out: URLs from the dataset for various characteristics return either 0 or 1, depending on the circumstances. A csv file is created from the tabulated values that have been

returned.

3) Dividing the dataset into parts for testing and training: The training and test portions of the dataset are distributed in varying proportions.

4) Ensemble hybrid model: The classifiers are applied to the dataset, and the corresponding accuracy is calculated.

In this work, we define a set of models a variable number of times to create weak learners. Finally, the Max Voting Classifier approach is employed, and the ensemble model's final class prediction will be the one that has been primarily predicted by the weak learners.

4. IMPLEMENTATION

MLP:

One more counterfeit brain network strategy with a few layers is the multi-layer perceptron (MLP). Obviously direct issues can be tended to in a solitary perceptron, however non-straight models are not appropriate to it. MLP can be utilized to determine these difficult issues. A feedforward fake brain network that delivers a bunch of results from a bunch of information sources is known as a multi-layer perceptron (MLP). A coordinated diagram interfacing the information and result layers of a MLP is comprised of many layers of information hubs. Backpropagation is utilized by MLP to prepare the organization.

SVM:

A regulated AI approach called Support Vector Machine (SVM) is used for both grouping and relapse. In spite of the fact that we frequently allude to relapse concerns, order is the most proper term. Finding a hyperplane in a N-layered space that plainly characterizes the information focuses is the objective of the SVM technique.

DECISION TREE:

A decision tree is a diagram that utilization the spreading way to deal with show every expected outcome for a specific information. Drawing choice trees manually, utilizing a designs instrument, or utilizing expert programming are choices. At the point when a gathering needs to pursue a choice, choice trees can assist with concentrating the discussion.

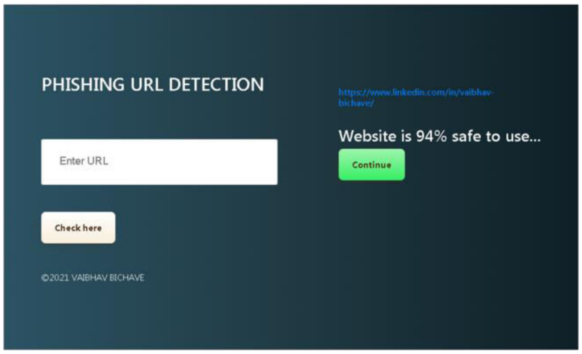
RANDOM FOREST:

The Random Forest Algorithm, a well-known method for supervised machine learning, is used to solve problems with classification and regression. There are many different kinds of trees in a forest, and the more trees there are, the stronger the forest will be.

HYBRID ENSEMBLE MODEL:

In programming languages like C++, a hybrid algorithm is one that mixes two or more different methods to solve the same issue. The hybrid algorithm either selects one (based on the data) or alternates between them throughout the procedure.

5. EXPERIMENTAL RESULTS



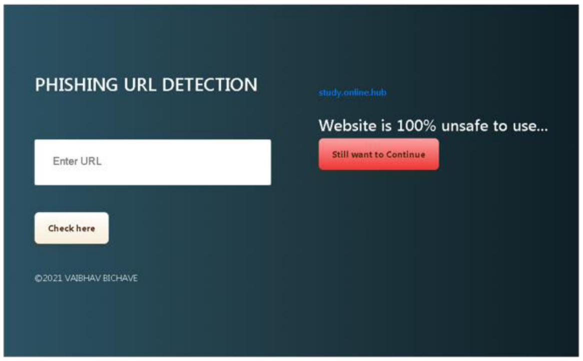
PHISHING URL DETECTION

https://www.linkedin.com/in/vaibhav-bichave/

Website is 94% safe to use...

©2021 VAIBHAV BICHAVE

Fig.3: Output screen



PHISHING URL DETECTION

study.online.hub

Website is 100% unsafe to use...

©2021 VAIBHAV BICHAVE

Fig.4: Output screen

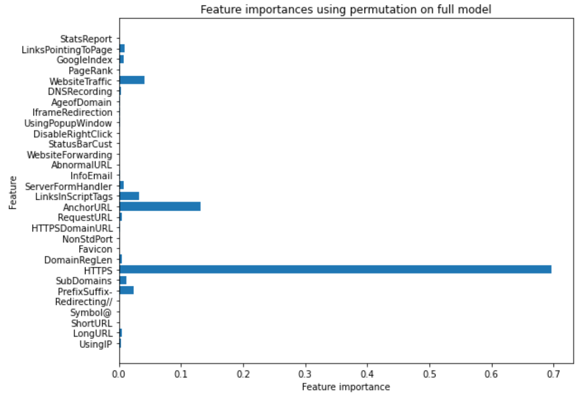


Fig.5: Output screen

6. CONCLUSION

The primary benefit of this study is that this model can be used as a browser extension to determine whether the website we are currently viewing is safe or harmful. This could be beneficial to users in order to stop malware from getting into their devices. We are currently 85.37 percent accurate. The precision score is 86.65%. The recall score is 83.95%. Choosing which weak learners should be incorporated into the hybrid architecture was a challenging decision to make while developing the hybrid model. As we navigate the internet, the present work may be assembled and distributed as a browser plugin that will automatically determine if a site is harmful or safe to visit. Additionally, this model may be improved by utilising a variety of deep learning approaches to boost its overall accuracy.

REFERENCES

[1] Mr. Kondeti Prem Sai Swaroop1, Ms. Konka Renuka Chowdary2, Ms. S. Kavishri 3 Phishing Websites Detection using Machine Learning Techniques International Research Journal of Engineering and Technology (IRJET)

[2] Mahajan, Rishikesh & Siddavatam, Irfan. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications. 181. 45-47. 10.5120/ijca2018918026.

[3] Mahajan, Rishikesh & Siddavatam, Irfan. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications. 181. 45-47. 10.5120/ijca2018918026.

[4] Bhagyashree E. Sananse Tanuja K. Sarode, Phishing URL Detection: A Machine Learning and Web Mining-based Approach International Journal of Computer Applications (0975 – 8887)

[5] Suman Bhattacharyya1 , Chetan kumar Pal2 , Praveen kumar Pandey3 Detecting Phishing Websites, a Heuristic Approach International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455- 7137

[6] B.A.S. Dilhara (2021) Phishing URL Detection: A novel hybrid Approach using Long Short-Term Memory and Gated Recurrent Units International Journal of Computer Applications (0975 – 8887)

[7] Tomas RASYMAS, Laurynas DOVYDAITIS.(2020). Detection of Phishing URLs by Using Deep Learning Approach and Multiple Features Combinations Baltic J. Modern Computing, Vol. 8 (2020), No. 3, 471-483

[8] Ray, K.S., Kusshwaha, R. (2021). Detection of Malicious URLs Using Deep Learning Approach. In: Chakraborty, M., Singh, M., Balas, V.E., Mukhopadhyay, I. (eds) The "Essence" of Network Security: An End-toEnd Panorama. Lecture Notes in Networks and Systems, vol 163. Springer, Singapore.

[9] Luong Anh Tuan Nguyen, Huu Khuong Nguyen, and Ba Lam To.(2016).An Efficient Approach Based on Neuro-Fuzzy for Phishing Detection .Journal of Automation and Control Engineering Vol. 4.

[10] Ashritha Jain R,Chaithra Kulal ,Mrs. Mangala Kini,Deekshitha S .(2019).A Review Paper on Detection of Phishing Websites using Machine Learning.International Journal of Engineering Research & Technology (IJERT).

[11] Arun Kulkarni1 , Leonard L. Brown, III2.(2019).Phishing Websites Detection using Machine Learning.International Journal of Advanced Computer Science and Applications, Vol. 10.

[12] Mehanović, D., Kevrić, J. (2020). Phishing website detection using machine learning classifiers optimized by feature selection. Traitement du Signal, Vol. 37, No. 4, pp. 563-569. https://doi.org/10.18280/ts.370403

[13] S. Mercy Shalinie,Ming Hour Yang,Raja Meenakshi U. Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions, IET Network

[14] Taha, Altyeb. (2017). Phishing Websites Classification using Hybrid SVM and KNN Approach. International Journal of Advanced Computer Science and Applications. 8. 10.14569/IJACSA.2017.080611.

[15] Anindita Khade, Dr. Subhash K Shinde, 2013, Detection of Phishing Websites Using Data Mining Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 02, Issue 12 (December 2013),