Detection of phishing attacks using Random Forest Algorithm

1Mrs.R. Sathya, 2U.Sai Kiran, 3J.V.Chowdary, 4B.Kasi Kumar Reddy

1Assistant Professor, 2,3,4Student, 1,2,3,4Department of Computer Science and Engineering,

1,2,3,4SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

1sathyaraman76@gmail.com, 2kiran98118@gmail.com, 3adithya.chowdary810@gmail.com, 4kasikumar214@gmail.com.

**Abstract: The internet was a great resource to everybody in 21st century. after the deveolpment of the internet the criminal elements also updated to the information age. the criminal elements have been responsible to the biggest data breaches in the history of the internet. the forces are using several methods to infiltrate the common people information and data. one of the main method is using phishing methods by using the malicious websites. these locales have been created.In order to stop these crimes and protect the civilians from visiting these malicious websites we propose a machine learning based protection system.the Websites are catogorised into 3 types Benign, spam and Malicious the ML system will classify the websites into three categories and our system will investigates the UNIFORM RESOURCE LOCATOR(URL) without getting to the core of the malignant web locales. it will kill the run-time inactivity of the websites wekeep the clients and general public safe. by using the Machine-Learning Techniques, our program occomplishes best execution on slimplificationThe inclusion contrasted and boycutting the general internet.**

Index Terms—Machine Learning, Random forest algorithm, Natural language processing.

**Key Words**:MACHINE LEARNING,RAndom Forest Algorithm AND natural language processing.

# I.INTRODUCTION

As you know the internet is one of the great inventions and it keeps on expanding into every sector from messaging/connecting people

everything like emails, phonecalls, Banks, security, weapon systems and even governments. Governments and Big Coorperations hire a lot of people to protect their data but for regular guys the data breaches are really common when an data breach occurs the person's Identity passwords, personal information like fingerprints and other intimate details, financial information like bank account details, passwords, credit/debt card numbers and their pins ..ETC. To get to all this information criminals usually use a technique called as Phishing. The phishing attacks were generally targeted at indivudual sensitive information. Phishing assaults generally take two forms.

1. the first type of attack is to deceive and make then to reveal their own sensitive information by generally pretending to be an

noteworthy and big entity like Multinational coorperations or any social network who need real information.

2. by injecting a trojan or an malware into the HOST's System and get access to their sensitive information.

The specific types of the Trojan and Malware used in the phishing attacks are subject to extreme research. the new types trojans and malwares. were invented as the cyber security is implementing new protocols. the malware is multipyling like rats a new malware has been created and put on the internet every 8 seconds so the malware and trojan system were not discussed at this time. we will focus on first type of Phishing Attack.

II.RELATED WORK:

Generally in this case we will use a neural Network. but, the neural network model may be a but premature and under-fit the training data-set.on the other hand we could tailor the NN Model to accomodate the training dataset but, that may cause overfitting of the network. one of the way to ovoid the overfitting of the network is to restructure the neuralnet itself.we may have to change some parameters or maybe use even more neurons to either the hidden layer or even the addition of new layer. but to many hidden neurons will definetely overfit the dataset and even without the overfitting the model cannot be improved after a certain restructing of neuralnet but the error rate is fairly specific and the error rate will be set by the model designer that value may or may not be unreachable so they specify the minimal value but the error rate should always must have a room for improvement. the nn model is fairly used for these kinds of projects and the cybersecurity issues. but still ther is a room for a lot of important improvements

# III.PROPOSED SYSTEM

The lexical features are generally based on the observations we noted on several URL's of the malicious websites. the url's of many of these

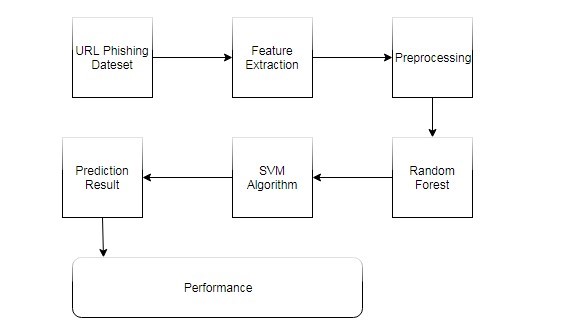
shady sites were different when compared with a legitimate websites.after analyzing few Lexical features we have been able to capture the properties for the classification system.In our classification system the URL consists of two parts, the host name and the path from these two we will extract a couple of words and we will eliminate all the symbols the strings are eliminated from taking symbols like('/','?','=','.','-',':' and ',').e.t.c, If the phishing websites have longer URL like more Levels, more Tokens in either Domain or Path it is easier to track. generally some of the phishing or malware websites are designed in a specific way so that they fool the user that they are benign websites by using popular names in the world as tokens other then using new unknown names as tokens. most malicious and malware websites use the IP adressesdirectly so that it can pass itself as a benign one and also they will use several words like confirm,login, signin, account, card details, age, verification,...e.t.c for their suggestive qualities.we constantly check websites for these suggestive words and as you know generally malicious and spam websites are generally less well known as benign websites and we can also assume that from hosting space we can identify some of these sites ex:nigeria email scam due to all these reasons like site popularity and traffic rankings we can consider these things as a criteria also the site popularity and the traffic rnkings data can be found in alexa..

Advantages:

1. the one advantage is that all the URL present in the Dataset are orderly marked and labelled

2. we have used two supervised learning algorithms such as random forest algorithm and a support vector machine to train the model using the SCIKIT-Learn libraby

ARCHITECTURE



MODULES

FEATURE EXTRACTION : One of the most Basic problems faced in these kinds of projects is that the training datasets are far too unreliable. but, in cybersecurity we are facing we have lot of sources about phishing, malware and trojans as the data mining techniques have been improving constantly,the datasets that are published are far to old or they are not at all published. maybe, one of the reasons is that there are no definative set of rules or characteristics of Phishing sites. it is very difficult or maybe impossible to create a dataset that covers all the possible characteristics. but,we tried to include all the sound and effective techniques used in predicting a malicious websites and also we have included some of the experimental new features to the effective ones and we think it will be good enough to predict the sites.we will be updating the features as we go along.

DATA PREPROCESSING : In this stage we will process the collected dataset of URL's mostly our dataset contains a collection of all BENIGN, SPAM and MALICIOUS URL'snow these urls are carefully analysed in the dataset so there are no false values or repeated or unmarked set of URL's then the dataset is handed over to the Future extractor. the future extractor will interface with dataset and extract the feature values from the dataset and predefined URL based features like IP adress,URL Length, suffix and prefix and characteristics of subdomain, no of '/' and protocol of HTTPS the value of the following features are based on set of different conditions.

IP ADDRESS : If the url contains an ip address that url will be given the value as 1 otherwise it will be zero.

URL LENGTH : If the url lenght is more than 54 the value of 0 will be assigned otherwise the value will be 1. Suspicious Characters : if some or any suspicious characters are present in the url then the value of that url will be 1 or it will be 0. Suffix and Prefix : If the both suffix and prefix are present in an Url the that url value will be 1 otherwise it will be 0.

DATA SPLITTING : It involves two steps Firstly, it tests the dataset whis is also known as test dataset then the train dataset will takes place which will train the data. the test dataset will cuntribute to 20% and the Train dataset will contribute to 80% of the total dataset.

FEATURE EXTRACTION OF URL : the data extracted from the the url is vital for detecting fake sites from the real one because phishing sites cannot copy the same url's as the legitimate(Benign) sites without changing it even for a little bit.

URL's Structures : an url is an virtual address for a website or an web locale for which the people can use to find their specific content. it functions as a real world address for people but in a digital world on a network. the URL structure is as follows.

There are several components in URL like primary domain, subdomain, protocol, top level domain(TLD) and Path domain. some of these parts such as primary domain, subdomain and top level domain are collectively known as Domain. the protocol is also called as a communication protocol which is used to exchange information between two or more different connected devices for ex : FTP, HTTP,HTTPS..e.t.c. there are several types of sub-domain the sub-domain depends on the types of services and differnt type used by the domain. In the domain hierarchy the TLD occupies the highest position in hierarchy Architecture for ex:- .com, .net, .gov.in, .in ...e.t.c . these are some of the features used in the common phishing sites and techniques.

URL FEATURES :

IN order to detect the malicious/spam or Phishing sites we need data. the data can be extracted from the URL's are generally called as URL features. the URL features that are extracted are

IP ADDRESS : we check if the IP address is present in the url if it is present then the URL will be suspected with phishing.

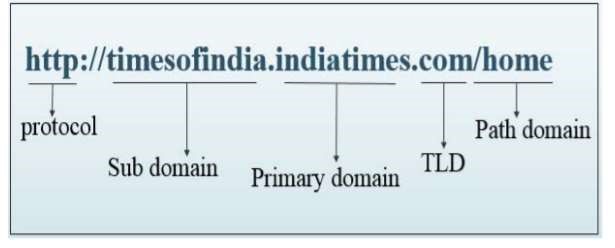
URL Length : to make sure the URL is passable as a legitimate one. the Phishers make the urls longer and with more digits. if the URL length is more then 54 words or digits then the url is suspected for being Malicious.

Suspicious Characters : In almost every URL there are many special chracters,digits and symbols. for example @ or \_ and mostly the general network browsers ignore these symbols. but some sequences of these symbols means the URL is a mailcious Website.

# Suffixes and Prefixes : Mostly benign website URL the prefix and suffix are barely used but in other sites the prefix and suffix are used quite commonly so having an suffix and prefix in an url is definetely an huge red flag.

# FEATURE EXTRACTION FROM URL

The highlights of URL assume a crucial job in recognizing counterfeit locales since phishers can't duplicate the specific URLs of real destinations. Structure of the URL A URL is a web address which is a reference to a web asset. The URL is utilized to indicate the area of the web asset on the system. The structure of URL is as per the following: The URL is made of a few parts, for example, convention, subdomain, essential area, TLD (top level space) and way area. Subdomain, essential area, and TLD are as one called as space. The convention alludes to a correspondence convention for trading data between data gadgets; e.g., HTTP, FTP, HTTPS, and so on. The subdomain is an auxiliary space given to the area and has different sorts relying upon the administrations gave by the area page. The TLD is the space in the most elevated situation in the area name chain of importance design; e.g., .com, .net, and so on. Every one of these highlights are utilized for discovery of phishing destinations.



**Features of URL**

So as to identify ill-conceived locales, different highlights of the URL are removed from the submitted URL. Following are the highlights that are extricated from the URL:

• IP address: It checks if the URL comprise of an IP address. In the event that IP address is available, at that point the URL is suspected as phishing.

• Length of URL: To shroud the dubious piece of URL, phishers generally utilize long URLs. This component figures the length of the URLs. On the off chance that the length surpasses 54, at that point the URL is suspected as phishing.

• Suspicious Characters: Many URLs comprise of images, for example, '@' or '\_'. When'@' image is utilized, the program will in general disregard everything present in the URL before the '@'symbol.

• Prefix and addition: Suffixes or prefixes are not really utilized in real URL. Subsequently this component checks if prefixes and postfixes are available or not.

# TECHNIQUES AND ALGORITHMS

Random Forest :

the Random forest algorithm is definetely not a brand new algorithm it has been used for a long time. it is a supervised basic machine learning algorithm and it is based on statistics and ensamble learning. we use the random forest because of its ensamble learning techniques. in this technique all the different algorithms or a same algorithm will be connected or rejoined several times over and over again to make the predection model more powerful and much more accurate. this algorithm is combining the multiple algorithms of the same type again and again for example MULtiple Descision Trees which resulted random forest algorithm in the name because trees makes the forest. this algorithm is mostly used for Regression and Classification Sector of the programing. There are several basic steps involved in using the algorithm.

1.firstly, we will pick a X number of random data records from the dataset.

2.Then, we will build and program the random forest algorithm using the X number of data Records.

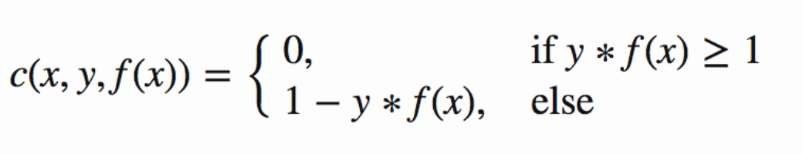
3. then, we will repeat step 1 and 2 and build more and more trees until we are satisfied with the resultant number of trees.

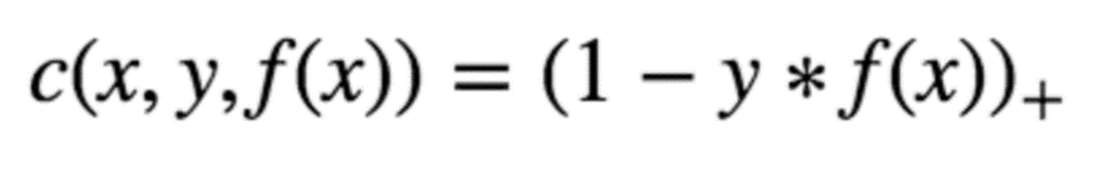
4. As we go in the classification end of the algorithm. we have already created a forest with trees then we feed the data records as the new data gets feeded into the forest the trees will come up with a solution and the majority vote will win and produce the result yes or no on a website.

Var^frf(x)=p(x)..

# **Support Vector Machine Algorithm**

The support vector machine(SVM) algorithm is generally used to find the hyper-equations in an X-dimensional space where x is the number of data-records in the datasets and the svm distintly classifies the Hyperplane datapoints. the SVM Algorithm generally try to balance and maximize the margin between Datapoints and the following Hyperplane affets there are two functions both the loss funtion and profit funtion. the loss funtion will maximize the hinge loss margin. the hinge loss is really instructive in the svm algorithm.the 0 value is placed if the original value has the similar to the predicted value. if they are not the same sign. we will have to calculate the hinge loss value. we also change and add some kind of regulation parameters to the following cost function. the regulation parameter is to maintain objective to balance the regulation between margin loss and maximization.





Hinge loss function (function on left can be represented as a function on the right)

# TensorFlow Architecture

1. Data Preprocessing

2. building an Efficient Model

3. training and the estimation of the model.

The name TensorFlow came into exsistance because it takes a multidimensional array or row as an standard set of input these arrays

are also called as tensors. you can construct a graph with a kind of 0.7 of the operations. you can perform the operations on the and the input inserted on one end will give accurate result on the other end as output the system in the middle will perform several operations. the tensors act as a perfect input processing systems and perform multiple operations through a list of specific sets of conditions and give out the output fairly accurately so the TensorFlow is the most used piece of software in the category of data science and machine learning. This is the reason it is called TensorFlow in light of the fact that the tensor goes in it courses through a rundown of activities, and afterward it comes out the opposite side.

**IV. SYSTEM REQUIREMENTS**

Python 2.7

Anaconda Navigator

Python's standard library:

Pandas

Numpy

Sklearn

tkMessageBox

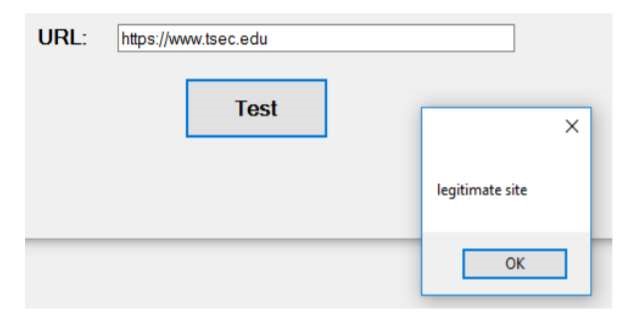
Matplotlib

Dataset

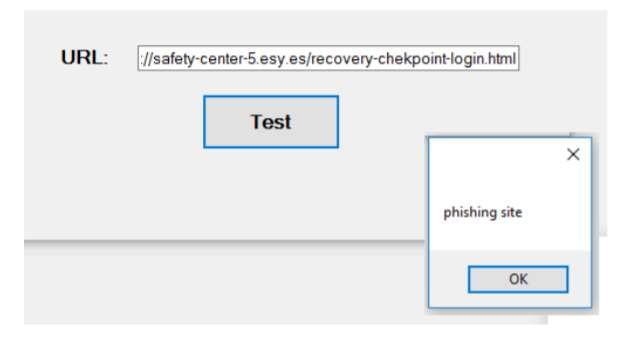
We are going to use a Famous Graphical User Interface(GUI) known as the Anaconda Navigator. the Anaconda GUI helps us to manage the environments, Packages and Channels without using complicated command lines and we have easy access to differnt types of development softwares at the same place which will make the development os applications that much easier. we can download your required packages from Anaconda Cloud or the Anaconda Repository. the application is available in several operating systems like windows, mac, linux . e.t.c..

**V. RESULT**

Our classifier was used to test a significantly large Dataset. But as for our c4.5 classifier can analyse more than the 3000 URL and Phishing sites. there are also some Legitimate website urls to increase the efficiency and accuracy of the classifier. the random forest algorithm we will split3 subsets of the dataset with 1000 URL's each.



http://safetycenter-5.esy.es/recovery-chekpoint login.html is fed as an input to the system.



**VI. CONCLUSION**

After this entire charade we know that the phishing is a major problem in the current day and age. the work we have done in our current project will help an ordinary men and women with no knowledge about the phishing attacks. it is also been stressed enough that these kinds of phishing and these other kinds of attacks should be properly countered and eradicated before it causes any permenant and irreversable damage. Our project is just a prototype which can be improved and applied to counter all the other kinds of cyber-attacks. This cyber security tools and work should always be improved to make sure that internet is safe for everyone including a large group of ordinary people.

# **VII. FUTURE WORK**

As for the future of this work goes we can give sove very practical and some of the very technical solutions by updating with the changing technologies and with a significantly large and very diversified number of datasets. we can use several new technolgies like machine learning, deep learning combined with artificial intelligence where all of these technoligies can be fully automated to operate some of the basic cyber security technologies like an efficient Spam filters and several email and ad-blockers. using these technologies we can eradicate another major incident such as "RANSOMWARE" trojan. which took the world by strom.Currently we are making huge progress in the User-phishing interaction model. this model is currently under development in this study a large number of human interactions and interviews is condected with a diversified group of people with and without any technical knowledge about networks and its security. after analyzing all their logical, rational, irrational, emotional resposes to some of the Legitimate and malicious websites and how are phishers are luring the unsuspecting people into their trap. after all these data has been collected with practical application of cognitive responses the results are looking very promising. we have tested most of the Phish Limiter and Spam Filter Softwares both theriotically, experimentally and practically in the aspects of accuracy and safety they provide is decresing day by day since new trojans and malwares have been ntroduced everyday. we have done careful and deep analysis of SDN flow through deep packet inspections and to avoid any petential dangers or hazards. we have observered that the two modes present in Phish-Limiter SF and FI which are resposible for detecting and countering the attacks have been failing miserabily in detecting and stoping the Phishing attacks before it has done some serious damage to the end user. but, using our GENI and Phishing datasets and testing them using theriotical, experimental and practical real world testing of our combined Phish-Limiter we have been able to Detect and Stop

the attacks on our end user with an effective and efficient accuracy of 98.38%. as the new technologies keep on emerging there always will be a room for improvement.

# **REFERENCES**

1. APAC. (Dec. 2018). Fishing Website Processing Bulletin. Accessed: 2019.[Online]. Available: http://www.apac.cn/gzdt/
2. M. Khonji, Y. Iraqi, and A. Jones, ‘‘Phishing detection: A literature survey,’’ IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091–2121, 4th Quart., 2013.
3. A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson, ‘‘Nudges for privacy and security: Understanding and assisting users’ choices online,’’ ACM Computing Surv., vol. 50, no. 3, 2017, Art. no. 44.
4. M. M. Moreno-Fernández, F. Blanco, P. Garaizar, and H. Matute, ‘‘Fishing for phishers. Improving Internet users’ sensitivity to visual deception cues to prevent electronic fraud,’’ Comput. Hum. Behav., vol. 69, pp. 421–436, Apr. 2017.
5. M. Junger, L. Montoya, and F.-J. Overink, ‘‘Priming and warnings are not effective to prevent social engineering attacks,’’ Comput. Hum. Behav., vol. 66, pp. 75–87, Jan. 2017.
6. E.-S. M. El-Alfy, ‘‘Detection of phishing websites based on probabilistic neural networks and K-medoids clustering,’’ Comput. J., vol. 60, no. 12, pp. 1745–1759, 2017.
7. C. Huang, S. Hao, L. Invernizzi, Y. Fang, C. Kruegel, and G. Vigna, ‘‘Gossip: Automatically identifying malicious domains from mailing list discussions,’’ in Proc. ACM Asia Conf. Comput. Commun. Secur. (ASIA CCS), Abu Dhabi, United Arab Emirates, Apr. 2017, pp. 494–505.
8. F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof, and M. Köppen, ‘‘Detecting malicious URLs using machine learning techniques,’’ in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), Dec. 2016, pp. 1–8.
9. J. Saxe, R. Harang, C. Wild, and H. Sanders, ‘‘A deep learning approach to fast, format-agnostic detection of malicious Web content,’’ in Proc. IEEE Symp. Secur. Privacy Workshops (SPW), San Francisco, CA , USA, Aug. 2018, pp. 8–14.
10. L. Wu, X. Du, and J. Wu, ‘‘Effective defense schemes for phishing attacks on mobile computing platforms,’’ IEEE Trans. Veh. Technol., vol. 65, no. 8, pp. 6678–6691, Aug. 2019.
11. R. Gowtham and I. Krishnamurthi, ‘‘A comprehensive and efficacious architecture for detecting phishing webpages,’’ Comput. Secur., vol. 40, pp. 23–37, 2014.
12. G. Xiang, J. Hong, C. P. Rosé, and L. Cranor, ‘‘CANTINA+: A featurerich machine learning framework for detecting phishing Web sites,’’ ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, 2011, Art. no. 21.
13. E. Zhu, C. Ye, D. Liu, F. Liu, F. Wang, and X. Li, ‘‘An effective neural network phishing detection model based on optimal feature selection,’’ in Proc. 16th IEEE Int. Symp. Parallel Distrib. Process. Appl. (ISPA), Melbourne, NSW, Australia, Dec. 2018, pp. 781–787.
14. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, ‘‘An empirical analysis of phishing blacklists,’’ in Proc. 6th Conf. Email Anti-Spam (CEAS), Mountain View, CA, USA,

Jul. 2009, pp. 1–20. VOLUME 7, 2019 73283 E. Zhu et al.: OFSNN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network

1. GitHub. Implementation for the Usage of Google Safe Browsing APIs (v4). Accessed: 2019. [Online]. Available: https://github.com/google/ safebrowsing
2. J. Kang and D. Lee, ‘‘Advanced white list approach for preventing access to phishing sites,’’ in Proc. Int. Conf. Converg. Inf. Technol. (ICCIT), Gyeongju, South Korea, Nov. 2007, pp. 491–496.
3. M. Sharifi and S. Siadati, ‘‘A phishing sites blacklist generator,’’ in Proc. 6th ACS/IEEE Int. Conf. Comput. Syst. Appl. (AICCSA), Doha, Qatar, Mar./Apr. 2008, pp. 840–843.
4. X. Han, N. Kheir, and D. Balzarotti, ‘‘PhishEye: Live monitoring of sandboxed phishing kits,’’ in Proc. 23rd ACM Conf. Comput. Commun. Secur. (CCS), Vienna, Austria, Oct. 2016, pp. 1402–1413.
5. L.-H. Lee, K.-C. Lee, H.-H. Chen, and Y.-H. Tseng, ‘‘POSTER: Proactive blacklist update for anti-phishing,’’ in Proc.

[20] A. Aleroud and L. Zhou, ‘‘Phishing environments, techniques, and countermeasures: A survey,’’ Comput. Secur., vol. 68, pp. 160– 196, Jul. 2017.