International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 5, Sep- Oct 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

An Efficient Method of Predicting Phishing Websites Using Machine Learning Algorithm

Pavithra[1],Dr.E.Uma[2]

[1] Student, College of Engineering, Guindy

[2]Assistant Professor, College ofEngineering, Guindy

ABSTRACT

Phishing assault is currently a major risk to individuals' everyday life and systems administration conditions. By masking illicit URLs as genuine ones, assailants can initiate clients to visit the phishing URLs to get private data and different advantages. Powerful techniques for recognizing phishing sites are desperately expected to lighten the dangers presented by phishing assaults. As the dynamic taking in ability from huge informational collections, the neural system is generally used to distinguish the phishing assaults. Be that as it may, in the phase of preparing informational collections, numerous pointless and little impact highlights will trap the neural system model into the issue of over-fitting. This issue ordinarily causes the prepared model that can't adequately recognize phishing sites. So as to ease this issue, this paper proposes OFS-NN, a compelling phishing sites location model dependent on the ideal component choice technique and neural system. In the proposed hybrid intelligent phishing website prediction approaches, the most influential features and the optimal weights of website features are heuristically identified with the genetic algorithm (GA) to help in increasing the accuracy of phishing website prediction. Accordingly, the website features selected and weighted by the GA are utilized to train DNNs to accurately predict the phishing websites. The methodology can lessen the discovery time for setting a limit. Testing on a dataset containing a great many phishing URLs and real URLs, the exactness arrives at 98.99%, and the bogus positive rate is just 0.59%. By sensibly modifying the edge, the test results show that the recognition effectiveness can be improved.

INTRODUCTION

Phishing is presently one of the quickest developing digital assaults. Phishing assaults fundamentally use social building and innovation duplicity to get client protection data. The most widely recognized method for phishing assaults is to send an unlawful connection to the client and actuate the client to click. The clients are then deceived by entering private data without their affirmation. At present, phishing assaults of the time show up in PCs and versatile stages. Hide thermore, phishing assaults are developing quickly. – The Anti-Phishing Alliance of China (APAC) has detailed that, at a month ago of 2018, there is an aggregate of 435193 phishing sites identified [1]. Successful techniques for recognizing the phishing sites are critically expected to lighten the dangers presented by phishing assaults.

1.1 PHISHING WEBSITES

As a rule, the potential estimations of the phishing sites location are often delegated a two-way issue: the "phishing sites" or the "real sites". Since phishing assaults as a rule take points of interest of clients' imprudent practices or symptom on utilizing organizing devices, it's a troublesome issue to be for all time settled

[2]. Targeting relieving the chance of phishing assaults, various methodologies square measure projected to organize and teach finish purchasers to understand and acknowledge phishing URLs [3], [4]. These methodologies turn out results somewhat by intermittently causation messages to caution finish purchasers with potential phishing dangers. Be that because it could, they still smitten by the users' behaviors and information of utilizing the underlying systems [5].

Due to high accuracy and potency, the software system based mostly automatic methodology is wide want to find phishing attacks. At present, the automated phishing detection ways are often classified into four categories: the blacklist and whitelist ways, the heuristics ways, the visual similarity meth- ods and also the machine learning ways [6]. Through recording antecedently detected phishing or legal URLs, IP addresses and keywords, the tactic of constructing black and white lists will effectively forestall phishing attacks [7]. thanks to tiny employment on analyzing the content of internet sites, this methodology has its advantage of requiring tiny resources on the underly- ing systems. However, since databases for storing black and white lists square measure created supported the antecedently detected URLs, this methodology has the problem in managing freshly emerged phishing attacks.

1.2. DETECTION

The heuristic phishing detection approaches can be taken as the extension of black and white lists [8]. The heuristic approaches are usually based on assigned signatures for identified phishing attacks. Through scanning websites for the assigned signatures, this kind of approaches raises a warning if malicious behaviors are found [9]. Because of the ability of detecting newly emerged URLs, the heuristic approach exhibits better performance than the method of blacklist and whitelist. However, due to complicated nature of phishing attacks and the time-consuming heuristic tests, this method tends to have higher false positive on phishing detection than the blacklist and whitelist method.

Through visually comparing the suspicious website with legitimate target, the visual similarity method can also achieve the similar accuracy as the blacklist and whitelist technique [10]. This method is based on catching snapshots of websites' appearances in the web browsers and sorting the acquired snapshots. However, these works may incur very high time and space costs.

Given the active learning capability from massive data sets, the machine learning method is widely used to detect the phishing websites. This method usually uses feature selection algorithm to extract a sensitive features vector that could help distinguish phishing and legitimate websites at first. Then, based on the extracted features, the underlying machine learn- ing classifiers are trained to detect the phishing websites [11]. The classifiers are usually constructed based on the neural network model, the support vector machine model, the Naïve Bayes model, and so on. The machine learning method is accurate in phishing websites detection. Meanwhile, it also has the ability to adapt to newly emerged phishing websites. The key to the success of this method is to acquire highly qualified features from phishing URLs and their relevant websites [12]. However, improper selection of sensitive features will make the underlying classifier not able to precisely detect the phishing websites. Meanwhile, some useless or small impact features will cause machine learning methods falling into the problem of over-fitting.

This paper proposes OFS-NN, effective phishing attacks detection model based on the optimal feature selection and neural network. Under this model, the new FVV index is firstly

defined to evaluate the impact of sensitive features on phishing website detection. Then, based on the FVV index, an algorithm is designed to select optimal features from phishing URLs and their relevant websites. Finally, the selected features are used to train the underlying neural network to build a classifier to detect phishing attacks. Generally speaking, the contributions of this paper are listed as follows:

(1) Defines a new index—FVV. In order to better evaluate the impact of a selected sensitive feature on detecting phishing attacks, this paper presents the FVV index. The new FVV is defined by combining the positive and negative features of URLs. Through calculating the FVV values, some useless or small impact features can be eliminated to improve the performance of the entire model.

(2) Designs an optimal feature selection algorithm. This algorithm calculates the FFV values of all features of the input URLs and their relevant websites at first. Then, a threshold is set to select sensitive features to construct an optimal feature vector. Through this algorithm, many useless and small influence features are pruned. Due to no disturbance from these redundant features, the over-fitting problem of the underlying neural network is alleviated. Meanwhile, this algorithm is also able to reduce the time cost of the process of phishing websites detection.

(3) Presents the OFS-NN model. Through the selected sensitive features and a large number of experimental analyses, the optimal structure of the neural net- work is trained and constructed as the final classifier of the OFS-NN model. This model is able to accurately detect many types of phishing attacks. Benefits from the powerful learning and fitting capabilities of the neural network, OFS-NN exhibits better performance than many existing systems in phishing website detection.

Z

2.RELATED WORKS

Notwithstanding the technique for preparing and instructing end clients, programming based programmed recognition strategy is the most widely recognized approach to facilitate the danger of phishing assaults.

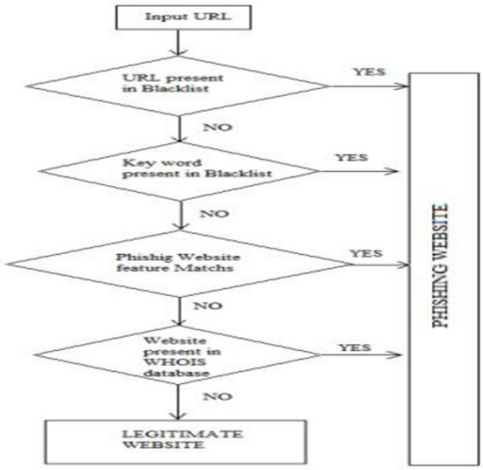
As the most immediate technique, the highly contrasting records can moderate phishing assaults with generally low asset consumptions [14]. The Google Safe Browsing API is a representative work of boycott that is continually kept up by Google Inc. [15]. Through speaking with the established APIs, customer applications can check whether the objective URL is in the boycotts. In view of the whitelist, Kang [16] proposes a technique to distinguish phishing sites. This strategy sorts out client access to the site by distinguishing URL comparability. It manages nearby and DNS parodying assaults by looking at DNS enquiry results. Sharifi and Siadati [17] proposes a boycott generator technique for identifying phishing sites. This technique decides if it is a phishing site by coordinating the area name of the site and Google's indexed lists. Han [18] sandboxes live phishing toolboxes by estimating the effect of boycotting administrations on phishing sites toward the starting when these administrations are introduced. So as to moderate the dangers of recently rose phishing URLs, Lee [19] proposes the Phish Track system for naturally refreshing the boycott of phishing destinations. The high contrast records devour little assets on the fundamental frameworks. Be that as it may, it can't appropriately manage the recently rose phishing assaults [20]. So as to alleviate this deficiency, as the work that is displayed right now, high contrast records are typically working in a joint effort with different techniques [21].

Because of dynamic learning capacity, AI meth-ods are broadly concentrated to identify the phishing sites. Bar [22] is a phishing assaults identification model based on 27 delicate highlights separated from URLs and

their significant sites. This model utilizes the TF-IDF calculation to distinguish phishing assaults. This calculation can distinguish numerous sorts of phishing assaults however to the detriment of consuming much time cost of the hidden frameworks. Then, some legitimate sites are accounted for as phishing ones [23].Bar [12] is the continuation work of Cantina. Com-pared with Cantina, CANTINA includes 10 additional highlights. In the meantime, the phishing location TF-IDF calculation is supplanted with SMV. Through these enhancements, deficiencies of Cantina are alleviated. In any case, the new CANTINA has a tight scope of uses [24].

As a programmed phishing identification model, Phish Storm [25] is actualized as an interface between informal communication devices and email servers or HTTP intermediaries. Right now, irregular woodland classifier is prepared by separating 12 URL significant highlights. Be that as it may, because of little inclusion of touchy highlights, it can't identify numerous sorts of phishing assaults. Phish Shield [26] is actualized as heuristic phishing distinguish ing apparatus running on PCs.This apparatus can distinguish phishing assaults that are planned by changing substance of sites. As a result of lacking dynamic taking in limit from phishing assaults, it can't appropriately manage assaults that persistently change their touchy highlights. The PhiDMA [27] model is actualized as a multi-channel. It comprises of five degrees of channels: the auto-redesign whitelist layer, the URL highlight layer, the lexical mark layer, the string coordinating layer and the availability rating examination layer. This model recognizes phishing sites dependent on openness mistakes (known blunder, likely blunder and potential blunder) correlation. Be that as it may, in certain sites, the model has come up short on the capacity to experience all the three mistakes. In the meantime, this model just gets 92.72% phishing system identification precision. The Off-the-Hook [28] model is worked as modules to specific programs. Through joining boycott, AI technique and 210 highlights, this model can recognize numerous phishing assaults. Be that as it may, an excessive number of delicate highlights genuinely debase the exhibition of this model.

The AI technique can precisely manage recently developed phishing assaults. Be that as it may, it depends a lot on the chose delicate highlights. Actually, many existing AI strategies utilize the gathered touchy highlights



to preparing the fundamental classifier. How-ever, they accomplish little work on pruning out pointless and little effect highlights. The two sorts of highlights will expand the outstanding burden of the classifier. To this end, this paper characterizes the new FVV record to get ideal delicate highlights. Benefits from the chose highlights, the prepared ideal neural system classifier (the center of the OFS-NN model) shows preferable execution over many existing models.

3.SYSTEM ARCHITECHTURE

Configuration is a multi-step that centers around information structure programming design, procedural subtleties, technique and so on… and interface among modules. The plan strategy additionally interprets the necessities

into introduction of programming that can be gotten to for greatness before coding starts. PC programming configuration change constantly as novel strategies; improved investigation and fringe understanding advanced. Programming proposition is at moderately essential stage in its insurgency.

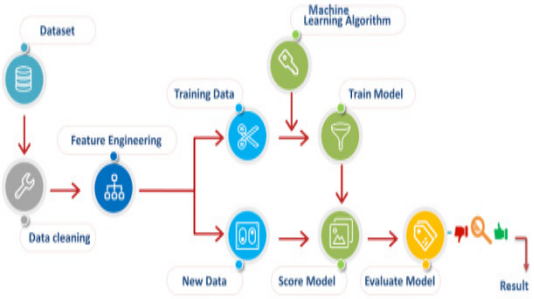
Along these lines, programming structure approach comes up short on the profundity, adaptability and quantitative nature that are generally connected with progressively traditional designing orders. Anyway, techniques for programming plans do leave, criteria for structure characteristics are existing and structure documentation can be applied.

REQUIREMENTS:

The product prerequisites detail is created at the finish of the investigation task. The capacity and execution assigned to programming as a major aspect of framework building are refined by setting up a total data depiction as utilitarian portrayal of framework conduct, a sign of execution prerequisites and plan limitations, fitting approval criteria.

4.METHODOLOGY

Right now, first characterize the conventional explanation of phishing site discovery, at that point portray the general system of the methodology MFPD and its proper definition.



4.1. Data Collection and Pre-processing

Information cleaning: Fill in missing qualities, smooth uproarious information, recognize or expel anomalies, and resolve irregularities. Information change may incorporate smoothing, collection, speculation, change which improves the nature of the information. Thus determination incorporates a few techniques or capacities which permit us to choose the valuable information for our framework. Information Collection is one of the most significant errands in building an AI model. It is the social event of assignment related data dependent on some focused on factors to examine and create some significant result .Notwithstanding, a portion of the information might be uproarious, for example may contain erroneous qualities, deficient qualities or off base qualities. Subsequently, it is must to process the information before breaking down it and going to the outcomes. Information pre-handling should be possible by information cleaning, information change, information determination.

4.2 Data Sets

Data sets can hold information as records to be used by a program running on the system. Data sets are also used to store information needed by applications or the operating system itself, such as source programs, macro libraries, or system variables or parameters.

Subsequent to finding the best calculation we utilize that calculation for finding the spam locales. At that point we are going to give an info that thecalculation discovers the information is spam or not.

4.2.1 Feature Selection

Highlight determination is one in every of the essential errands which might be utilised once building AI models. Highlight alternative is critical in lightweight of the very fact that selecting right highlights would facilitate construct models of upper exactness in addition as facilitate accomplish targets known with building a lot of simple models, decrease overfitting so on. springing up next ar a little of the procedures that may well be utilised for embody choice:

Filter ways that helps in selecting highlights

captivated with the results of measurable tests.

Pearson's correlation
Linear discriminant analysis (LDA)
Analysis of Variance (ANOVA)
Chi-square tests

Wrapper strategies that helps in feature choice by employing a set of options and crucial the model accuracy. the subsequent ar a number of the algorithms used:

Forward choice
Backward elimination
algorithmic feature elimination

Regularization techniques that penalizes one or a lot of options fitly to come back up with most significant options. the subsequent ar a number of the algorithms used:

LASSO (L1) regularization
Ridge (L2) regularization.

4.2.2 Clustering

Clustering are tied in with discovering normal groupings of information and a name related with every one of these groupings (bunches). A portion of the basic model incorporates client division, item includes distinguishing proof for item guide. A portion of coming up next are basic ML techniques:

- Mean-move (Higher precision)
- Hierarchical grouping
- K-implies

4.2.3 Multivariate quering

It is tied in with questioning or finding comparable articles. A portion of the accompanying ML techniques could be utilized for such issues:

- Nearest neighbors
- Range Search
- Farthest neighbors

4.2.4 Density estimation

Density estimation issues are connected with discovering probability or recurrence of articles. In likelihood and insights, thickness estimation is the development of a gauge, in light of watched information, of an inconspicuous hidden likelihood thickness work. A portion of the accompanying ML techniques could be utilized for unravelling thickness estimation assignments:

- Kernel thickness estimation (Higher exactness)
- Mixture of Gaussians
- Density estimation tree

4.2.5 Dimension reduction

Dimension reductionisthe way toward diminishing the quantity of arbitrary factors viable, and can be partitioned into include choice and highlight extraction. Following are some of ML techniques that could be utilized for measurement decrease:

- Manifold learning/KPCA (Higher exactness)

- Principal segment investigation
- Independent segment investigation
- Gaussian graphical models
- Non-negative lattice factorization
- Compressed detecting

4.2.6 Testing and matching

Testing and matchingerrands identifies with looking at informational indexes. Following are a portion of the strategies that could be utilized for such sort of issues:

- Minimum spreading over tree
- Bipartite cross-coordinating
- N-point connection

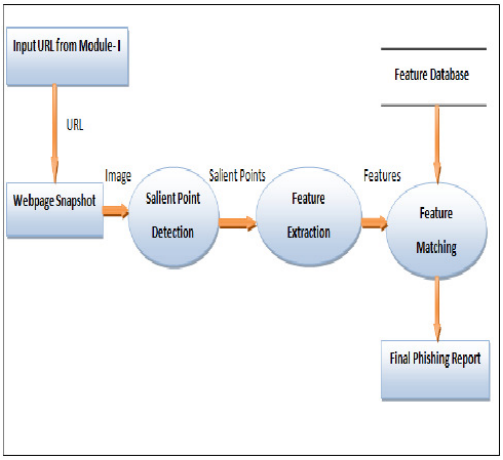
4.3 CLASSIFICATION

Classification tasks is simply related with predicting a category of a data (discrete variables). One of the most common examples is predicting whether or not an email if spam or ham. Some of the common use cases could be found in the area of healthcare such as whether a person is suffering from a particular disease or not. It also has its application in financial use cases such as determining whether a transaction is fraud or not. The ML methods such as following could be applied to solve

classification tasks:
Kernel discriminant analysis *(Higher accuracy)*
K-Nearest Neighbors *(Higher accuracy)*
Artificial neural networks (ANN) *(Higher accuracy)*
Support vector machine (SVM) *(Higher accuracy)*
Random forests *(Higher accuracy)*
Decision trees
Boosted trees
Logistic regression
naive Bayes
Deep learning

SupervisedLearning:

These calculations are prepared utilizing named models, in various situations, as an information where the ideal result is as of now known. A hardware, for example, could have information focuses, for example, "F" and "R" where "F" speaks to "fizzled" and "R" speaks to "runs". A learning calculation will get a lot of info directions alongside the relating exact results. The learning calculation will at that point contrast the genuine result and the precise result and banner a blunder, if there is any error. Utilizing various strategies, for example, relapse, arrangement, inclination boosting, and expectation, regulated learning utilizes various examples to proactively anticipate the estimations of a name on extra unlabelled information. This technique is ordinarily utilized in zones where recorded information is utilized to foresee occasions that are probably going to happen later on. For example, foresee when a charge card exchange is probably going to be deceitful or anticipate which protection clients are probably going to document their cases.



Unsupervised Learning:

This strategy for ML discovers its application in territories were information has no authentic names. Here, the framework won't be given the "right answer" and the calculation ought to distinguish what is being appeared. The principle point here is to break down the information and distinguish an example and structure inside the accessible informational collection. Value-based information fills in as a decent wellspring of informational collection for unaided learning. For example, this kind of learning distinguishes client fragments with comparable qualities and afterward lets the business to treat them correspondingly in showcasing efforts. Thus, it can likewise distinguish qualities that separate client sections from each other. Either ways, it is tied in with distinguishing a comparative structure in the accessible informational collection. In addition, these calculations can likewise recognize exceptions in the accessible informational collections.

A portion of the generally utilized systems of unaided learning are -

- k-implies bunching
- self-arranging maps
- value disintegration
- mapping of closest neighbour

Semi-supervised Learning:

This sort of learning is utilized and applied to a similar sort of situations where regulated

learning is material. In any case, one must note that this method utilizes both unlabelled and marked information for preparing. In a perfect world, a little arrangement of marked information, alongside an enormous volume of unlabelled information is utilized, as it takes less time, cash and endeavours to procure unlabelled information. This sort of AI is frequently utilized with techniques, for example, relapse, grouping and expectation. Organizations that generally think that its difficult to meet the significant expenses related with named preparing process settle on semi-administered learning.

Reinforcement Learning:

This is for the most part utilized in route, apply autonomy and gaming. Activities that yield the best rewards are recognized by calculations that utilization experimentation strategies. There are three significant segments in support learning, to be specific, the specialist, the activities and the earth. The specialist right now the leader, the activities are what an operator does, and the earth is whatever a specialist associates with. The primary point right now learning is to choose the activities that boost the prize, inside a predefined time. By following a decent arrangement, the specialist can accomplish the objective quicker.

4.4PREDICTION:

Consider the case of a bank processing the likelihood of any of advance candidates blaming the advance reimbursement. To register the likelihood of the shortcoming, the framework will initially need to order the accessible information in specific gatherings. It is depicted by a lot of rules recommended by the experts.

When we do the grouping, according to require we can figure the likelihood. These likelihood calculations can register over all divisions for changed purposes

4.5 EXTRACTION:

Data Extraction (IE) is another utilization of AI. It is the way toward separating organized data from unstructured information. For instance pages, articles, websites, business reports, and messages. The social database keeps up the yield created by the data extraction.

The procedure of extraction accepts contribution as a lot of archives and creates an organized information. This yield is in a condensed structure, for example, an exceed expectations sheet and table in a social database.

These days extraction is turning into a key in the huge information industry.

As we realize that the gigantic volume of information is getting created out of which the majority of the information is unstructured. The primary key test is taking care of unstructured information. Presently change of unstructured information to organized structure dependent on some example with the goal that the equivalent can put away in RDBMS.

Aside from this in current days information assortment system is additionally getting change. Prior we gathered information in clusters like End-of-Day (EOD), however now business needs the information when it is getting created, for example continuously.

4.6REGRESSION:

Regression tasks mainly deal with estimation of numerical values (continuous variables). Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

Kernel regression *(Higher accuracy)*

Gaussian process regression *(Higher accuracy)*

Regression trees

Linear regression

Support vector regression

LASSO

We can apply Machine figuring out how to

relapse too.

Accept that x= x1, x2, x3, … xn are the info factors and y is the result variable. Right now, can utilize AI innovation to deliver the yield (y) based on the info factors (x). You can utilize a model to communicate the connection between different parameters as underneath: Y=g(x) where g is a capacity that relies upon explicit attributes of the model. In relapse, we can utilize the guideline of AI to enhance the parameters. To cut the estimate mistake and ascertain the nearest conceivable result.

We can likewise utilize Machine learning for work advancement. We can decide to modify the contributions to show signs of improvement model. This gives a better than ever model to work with. This is known as reaction surface structure. Along these lines, this was about Machine Learning Applications. Expectation you like our clarification.

5.EXPERIMENTATIONFOR MODEL VALIDATION:

Now, conducts five-overlap cross approval on DATA1 to demonstrate the legitimacy of the dynamic classification choice calculation DCDA. The key of DCDA is to locate the ideal edge α so it can rapidly recognize phishing sites with high precision and low identification cost. The analysis results are appeared in Fig. 18 and Fig. 19. At the point when an edge of around α=355, the identification precision and the discovery cost will in general be steady, arriving at 98.88% and 4.56, which is practically identical to the multidimensional element recognition. The most significant job of DCDA is constant discovery. Fig. 20 shows that as the limit expands, the normal number of sites that CNN-LSTM is liable for recognizing bit by bit diminishes, and the quantity of sites that the multidimensional component recognition is answerable for identifying slowly increments. At the point when the edge is roughly α=355, just 28% of the sites need to experience the multidimensional component location, which significantly diminishes the remaining task at hand.

It is found in the test that the time taken by the CNN-LSTM calculation to foresee the whole test set doesn't surpass 10 s. The multidimensional element calculation has a normal location time of 3.5 s for per URL because of the need to separate the code highlights of the site page, the content highlights of the site page, and the WHOIS data in the URL highlights. For comfort, we set the normal length of CNN-LSTM site location to 0.5 s and the quantity of information tests from DATA1 to 10 000. The exploratory outcome is appeared in Fig. 21. As the limit expands, the normal identification time of the DCDA calculation changes straightly. At the point when the edge =355 □, the normal recognition time is short of what one-portion of the multidimensional component identification. In synopsis, thinking about recognition exactness, identification cost and location time, the limit of the DCDA calculation is set to 355, which ensures fitting precision and discovery cost of phishing site location and altogether lessens the location time.

CONCLUSION:

It is notable that a decent phishing site location approach ought to have great ongoing execution while guaranteeing great exactness and a low bogus positive rate. Under the control of a powerful class choice calculation, the URL character arrangement without phishing earlier information guarantees the location speed, and the multidimensional include location guarantees the recognition precision. We direct a progression of investigations on a dataset containing a large number of phishing and authentic URLs. This section describes the proposed model of phishing attack detection. The proposed model focuses on identifying the phishing attack based on checking phishing websites features, Blacklist and WHOIS database. According to

few selected features can be used to differentiate between legitimate and spoofed web pages. These selected features are many such as URLs, domain identity.Thus, SVM algorithm, randomforest, decision tree algorithm is used to effectively identify this phishing websites. Inaddition, we plan to approach into a plugin for embedding into a web browser.

REFRENCES:

[1] (2016). PhishMe Q1 2016 Malware Review. [Online]. Available: https://phishme.com/project/phishme-q1-2016-malware-review/

[2] A. Belabed, E. Aimeur, and A. Chikh, ‘‘A personalized whitelist approach for phishing webpage detection,’’ in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.

[3] Y. Cao, W. Han, and Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc. 4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60.

[4] T.-C. Chen, S. Dick, and J. Miller, ‘‘Detecting visually similar Web pages: Application to phishing detection,’’ ACM Trans. Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.

[5] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, ‘‘Clientside defense against Web-based identity theft,’’ in Proc. 11th Annu. Netw. Distrib. Syst. Security Symp. (NDSS), 2004, pp. 1–16

[6] C. Inc. (Aug. 2016). Couldmark Toolbar. [Online]. Available: http://www.cloudmark.com/desktop/ie-toolbar

[7] J. Corbetta, L. Invernizzi, C. Kruegel, and G. Vigna, ‘‘Eyes of a human, eyes of a program: Leveraging different views of the Web for analysis and detection,’’ in Proceedings of Research in Attacks, Intrusions and Defenses (RAID). Gothenburg, Sweden: Springer, 2014.

[8] X. Deng, G. Huang, and A. Y. Fu, ‘‘An antiphishing strategy based on visual similarity assessment,’’ Internet Comput., vol. 10, no. 2, pp. 58–65, 2006.

[9] Z. Dong, K. Kane, and L. J. Camp, ‘‘Phishing in smooth waters: The state of banking certificates in the US,’’ in Proc. Res. Conf. Commun., Inf. Internet Policy (TPRC), 2014, p. 16.

[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] A. 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 statisticalal 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 G and J. Hong, “A hybrid phish detection approach by identity discovery and keywords 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 featurerich machine learning framework for detecting phishing web sites,” 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.

[21] R. Patil, B. D. Dhamdhere, K. S. Dhonde and R. G. Chinchwade, “A hybrid model to

detect phishing-sites using clustering and Bayesian approach," in Proc. IEEE Int. Conf. Convergence technol. (I2CT), Apr. 2014, pp. 1-5.

[22] M. Arab and M. K. Sohrabi, "Proposing a new clustering method to detect phishing websites," Turk. J. Electr. Eng. Comput. Sci., vol. 15, no. 1, pp. 92-95, Jun. 2015.

[23] A. Shinde, A. Pandey, R. Pawar and V. Gangule, "Clustering and Bayesian Approach-based Model for Detection of Phishing," Int. J. Comput. Appl., vol. 118, no. 24, pp. 30-33, May. 2015.

[24] X. Zhang, Z. Yan, H. Li and G. Geng, "Research of phishing detection technology," Chin. J. Netw. Inf. Secur., vol. 3, no. 7, pp. 724, Jul. 2017.

[25] J. Ma, L. K. Saul, S. Savage and G. M. Voelker, "Beyond blacklists: learning to detect malicious web sites from suspicious URLs," in Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD09), Jan. 2009, pp. 1245-1254.

[26] R. M. Mohammad, F. Thabtah and L. Mccluskey, "Predicting phishing websites based on self-structuring neural network," Neural Comput. Appl., vol. 25, no. 2, pp. 443-458, Aug. 2014.

[27] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecommun. Syst., vol. 68, no. 4, Aug. 2018.

[28] J. Zhang, Y. Ou, D. Li and Y. Xin, "A Prior-based Transfer Learning Method for the Phishing Detection," J. Netw., vol. 7, no. 8, pp. 12011207, Aug. 2012.

[29] L. Chang, X. Deng, M. Zhou, Z. Wu, Y. Yuan, S. Yang and H. Wang, "Convolutional Neural Networks in Image Understanding," Acta. Automatica. Sinica., vol. 42, no. 9, pp. 1300-1312, Jul. 2016. [30] Z. C. Lipton, J. Berkowitz and C. Elkan, "A critical review of recurrent neural networks for sequence learning," arXiv.1506.00019, Oct. 2015.

[31] S. G. Selvaganapathy, M. Nivaashini and H. P. Natarajan, "Deep belief network-based detection and categorization of malicious URLs," Inf. Secur. J. Glo. Perspect., vol. 27, no. 3, pp. 145-161, Apr. 2018.

[32] A. C. Bahnse, E. C. Bohorquez, S. Villegas, J. Vargas and F. A. González, "Classifying phishing URLs using recurrent neural networks," in Proc. IEEE APWG Sym. Electron. Res. (eCrime), Jun. 2017, pp. 1-8.

[33] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," in Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS'15), Dec. 2015, pp. 649-657.

[34] Y. Xiao and K. Cho, "Efficient character-level document classification by combining convolution and recurrent layers," arXiv.1602.00367, Feb. 2016.

[35] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Improving neural networks by preventing coadaptation of feature detectors," arXiv.1207.0580, Jul. 2012.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929-1958, Jun. 2014. [37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv.1412.6980, Dec. 2014.

[38] A open-content directory of World Wide Web links. [Online]. Available: https://dmoztools.net/, accessed Oct. 12, 2018.

[39] . Jo, I., Jung, E., Yeom, H.Y.: You're not who you claim to be: website identity check for phishing detection. In: 2010 Proceedings of 19th International Conference on Computer Communications and Networks, Aug 2010.

[40] Abunadi, A., Akanbi, O., Zainal, A.: Feature extraction process: a phishing detection approach. In: Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on, pp. 331–335. IEEE (2013)

[41] Shrestha, N., Kharel, R.K., Britt, J.,

Hasan, R.: High-performance classification of phishing URLs using a multi-modal approach with MapReduce. In: 2015 IEEE World Congress on Services (SERVICES), pp. 206–212. IEEE (2015).
[42] Prasad, B.R., Agarwal, S.: Comparative study of big data computing and storage tools: a review. Int. J. Database Theory Appl. 9(1), 45–66 (2016).
[43] Li, B., Sun, R., Fang, X., Luo, X., Chang, W.: Emergent challenges and IPDS for anti-phishing attack. In: 2014 International Conference on IT Convergence and Security (ICITCS), pp. 1–4. IEEE (2014).
[44] Blum, A., Wardman, B., Solorio, T., Warner, G.: Lexical feature based phishing URL detection using online learning. In: Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security, pp. 54–60. ACM (2010)
[45] Zhuang, W., Jiang, Q., Xiong, T.: An intelligent anti-phishing strategy model for phishing website detection. In: 2012 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 51–56. IEEE (2012)

[37]S.Neelakandan, and D.Paulraj,''A Gradient Boosted Decision Tree based Sentiment classification of Twitter Data".

