International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Anti-Phishing Technology Using Machine Learning Approach

Achu Thomas Philip[1], Jain James[2], Nisha Mohan P.M.[3]

[1,2]B. Tech Student, Computer Science and Engineering, APJ Abdul Kalam Technological University, Kerala, India
[3]Asst. Professor, Computer Science and Engineering, Mount Zion College of Engineering, Kadammanitta, Kerala, India

---

Abstract - *Phishing is a social engineering attack that aims at exploiting the weakness found in the system at the user's end. Phishing attacks are the most common type of attacks leveraging social engineering techniques. Attackers use emails, social media to trick victims into providing sensitive information or visiting malicious URL (Uniform Resource Locator) in the attempt to compromise their systems. For individuals, this includes unauthorized purchases, the stealing of funds or identifies theft. An organization succumbing to such an attack typically sustains severe financial losses in addition to declining market share, reputation and consumer trust. It refers to exploiting weakness on the user side, which is vulnerable to such attacks. The phishing problem is huge and there does not exist only one solution to minimize all vulnerabilities effectively, thus multiple techniques are implemented. In this paper, we discuss Random Forest Machine Learning approach for detecting phishing websites. First step is to extract various features of URL such as domain license, elements, idiosyncrasies; then checking legitimacy of website by predicting the result. We make use of Machine Learning techniques and algorithms for evaluation of these different features of URL and websites. In this paper, an overview about these approaches is presented.*

Key Words: **Phishing, Anti-phishing, Machine Learning, Random Forest, Phishing Detection.**

1. INTRODUCTION

The term "phishing" was first recorded in a Usenet newsgroup called AOHell on 2nd January 1996, to describe the theft of users' credentials on America Online (AOL) by a group of hackers and since then, the scale and sophistication of phishing attacks have been on increase with huge financial and reputational damages on online users. Phishing is a social engineering attack that aims at exploiting the weakness found in the system at the user's end. For example, a system may be technically secure enough for password theft but the unaware user may leak his/her password when the attacker sends a false update password request through forged (phished) website. Phishing is metaphorically similar to fishing in the water, but instead of trying to catch a fish, attackers try to steal consumer's personal information. When a user opens a fake web page and enters the username and protected password, the credentials of the user are acquired by the attacker which can be used for malicious purposes. Phishing websites look very similar in appearance to their corresponding legitimate websites to attract large number of Internet users. For addressing this issue, a layer of protection must be added on the user side to address this problem. A phishing attack is when a criminal sends an email or the URL pretending to be someone or something he's not, in order to get sensitive information out of the victim. The victim in regard to his/her curiosity or a sense of urgency, they enter the details, like a username, password, or credit card number, they are likely to acquiesce. The recent example of a Gmail phishing scam that targeted around 1 billion Gmail users worldwide. Phishing is a technique used by hackers or attackers to trick the users into entering their sensitive credentials such as usernames, passwords, and credit cards details into a non-genuine entity such as a website. In this type of attack, unauthentic entities disguise themselves as genuine and trustworthy entities. Thus users are misled by the look and feel of the fake website which is almost identical to the legitimate one. Generally, attackers use banking and payment sites, social media sites and E-Commerce sites to lure potential victims. In 2016, a variety of changes in spam flows with an increase in the number of malicious mass E-mails containing links to phishing sites was observed. Until recently, PhishTank has verified and Validated 2,668,949 websites as phishing sites. Hence, phishing has now become the leading delivery vehicle for ransom ware and other malware. So, there is a grave need for developing a very dynamic anti-phishing solution. An organization succumbing to such an attack typically sustains severe financial losses in addition to declining market share, reputation, and consumer trust. Depending on scope, a phishing attempt might escalate into a security incident from which a business will have a difficult time recovering.

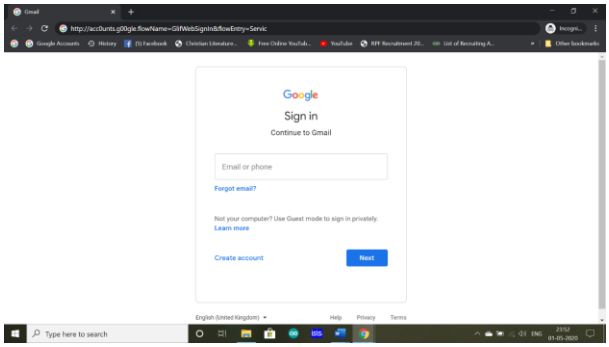


Fig 1: *Gmail Phishing Scam URL*

The Fig. 1 looks exactly like a Gmail sign-in form, the URL is slightly changed, but it's not filling in this form would give the attacker full access to the victim's Gmail account. The kind of theft and fraud that could take place by just acquiring the details of someone's or some organizations' account couldn't really be imagined. All other accounts are controlled by the Gmail account. That could be a huge threat. Microsoft Outlook fraud is the second-most targeted and Google drive being the third. Other targets are Facebook, bank logins and Pay-tm, Pay-pal, etc.

Anti-Phishing: All those intentations to prevent spam attacks via internet can be termed as anti-phishing. An anti-phishing service is a technological service that helps prevent unauthorized access to secure and sensitive information. Anti-phishing services protect various types of data in diverse ways across a variety of platforms.

Machine Learning: It is a field of artificial intelligence and it has ability to learn without explicitly programmed. Various machine learning techniques are supervised learning, unsupervised learning and reinforcement learning.

Random forest: It is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

2. RELATED WORKS

Websites can be categorised using sophisticated techniques in light of specific features such as, URL length, prefix, suffix, sub-domain, and so forth. Authors of [3] created distinctive learning bases utilising space understanding to recognise phishing sites and real sites. Lately, there have been different studies for acquiring automated rules to separate genuine and phishing sites utilising statistical analysis [2, 12, 13]. For example, [4] and [5] characterised various intelligently derived rules in light of different website features by using frequency counting of websites gathered from various sources, including PhishTank and Yahoo directory (Yahoo, 2011). Advancements in rules for decision making have been developed in [2] in which the authors utilised a computational intelligence method on a bigger phishing dataset gathered from numerous sources.

R. B. Basnet and A. H. Sung [1] introduced a broader definition of website phishing. This tells how it can be used to trick a user into visiting a new website and revealing their username, password and other sensitive information. It discusses about how meta data is a very useful tool in determining whether a URL is phishing or not. Meta data can be made available about a URL from the various search engines like Google, Yahoo, Bing, etc. This meta data can help in differentiating a phishing URL from a non-phishing one. It proposed the use of Logistic Regression as a classifier. But it can be failed to work when there are a massive number of features and are not necessarily linear in nature.

Ankit Kumar Jain et al. [6] presented a comprehensive analysis of all the Phishing attacks known, along with their resulting consequences. Moreover, it also provides a very useful insight over the various machine learning based approaches for phishing detection with the help of a comparative study. This opens up various viewpoints in terms of finding more efficient solutions with help of machine learning in near future.

R. Aravindhan et al. [7] proposed a list based anti-phishing approach, which has two types: (i)Black list and (ii)White list. In black list some online databases such as phish tank provides list of phishing websites. In white list the user manually builds a white list by adding the trusted website to the white list. In heuristics based anti-phishing approach the characteristics are determined such that it reflects the nature of the website accurately, machine learning techniques is used to find the phishing.

Another heuristic features detection method by authors [08] explains about the feature of URL such as Primary-Domain, Sub-Domain, Path-Domain and ranking of website such as PageRank, AlexaRank, AlexReputation to identify the phishing websites. Dataset used from PhishTank and experimental is splitted into 6 phases through MYSQL, PHP with 10 testing datasets. The proposed model contains two phases. In Phase I site features were extracted and in Phase II six values of heuristic are calculated. According to authors, if heuristic value is nearest to one, the site is considered as legitimate and if it is nearest to zero then the site is doubted as phishing site. Root Mean Square Error (RMSE) is used to calculate accuracy and obtained 97% accuracy.

Naghmeh Moradpoor et al. [9] proposed a neural network-based model for detection and classification of phishing emails. It uses real benign emails from "SpamAssassin" dataset and real phishing emails from "Phishcorpus" dataset. Python and MATLAB is used to measure the accuracy, true-positive rate, false positive-rate, network performance, and error histogram.

Another attempt to accurately classify websites based on features was conducted in [4]. The authors manually categorised features into six criteria and then loaded them into an environment for analysis on WEKA [10]. During which various experiments ran using four classification algorithms against 1006 instances from PhishTank. The evaluation measure to determine the applicability of these features was the classification accuracy. The outcomes uncovered that decision tree algorithms detected on average, 83% of the phishing sites. Accordingly, the authors' proposed that with appropriate pre-processing the results would improve.

In this paper authors [11] proposed a phishing detection model to detect the phishing performance effectively by using mining the semantic features of word embedding, semantic feature and multi-scale statistical features in Chinese web pages. Eleven features were extracted and categorized into five classes to acquire statistical features of web pages. AdaBoost, Bagging, Random Forest and SMO are used to implement learning and testing the model. Legitimate URLs dataset obtained from DirectIndustry web guides and phishing data was obtained from Anti-Phishing Alliance of China. According to study, only semantic features well identified the phishing sites with high detection efficiency and fusion model achieved the best performance detection. This model is unique to Chinese web pages and it has dependency in certain language.

3. PROPOSED SYSTEM

Here we proposed a new method of anti-phishing technology. The Anti-Phishing Technology using Machine Learning Approach is a mechanism that is proposed in order to ensure high security. In this mechanism we deal with the URLs (Uniform Resource Locaters) and the URL (Uniform Resource Locater) check with machine learning technique and predict whether it is phishing website or not. Here we create a web browser for browsing. Each time we browse a site the corresponding URL (Uniform Resource Locater) of site will be checked with machine learning technique. If the predicted result will negative then that site will be blocked. Then we can't access the site from that system.

Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally, we block the website in our system. Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers. Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. The misleading sites are intended to emulate the look of a genuine organization site page.

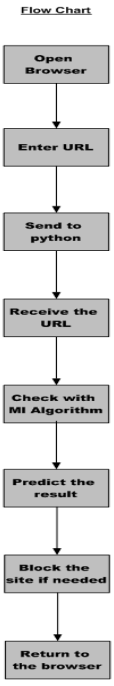


Fig 3: *Flowchart / Architecture of Proposed System*

We proposing a system that detects the Malicious URLs websites hosting phishing, spam etc. by using Machine Learning. The system should be useful in preventing online frauds leading to leakage of important and private user data.

Out of all the previous work, only the blacklist and whitelist are implemented which has a drawback of not being updated in long time. The basic idea of our proposed solution is the blocking of URL (Uniform

Resource Locater) in the system properties. Above given block diagram briefly illustrates the working of our proposed system and it has the following algorithm:

1. START
2. The URL -that to be checked whether it is phished or not- is entered into the newly developed browser.
3. The security issues related to the URL is checked using the machine learning approach (Random Forest).
4. If the result is "not phished" ; go-to step.10
5. If the result is "phished " ; go-to step.6
6. The user can make a decision whether to perform blocking or not.
7. If decision = "No" ; go-to step.10
8. Else decision = "Yes " ; go-to step.9
9. Blocking of the URL in the HOST of the user system is performed and go-to step.11
10. Browsing is allowed to continue
11. Stop the process

The above given algorithm briefly explains the working concept behind the proposed system. In step.1, the URL that is to be checked whether it is phished or not is fed to the web browser we have created. In step.2, the evaluation of the data variables present in the URL is carried out using the machine learning approaches. Then, a result is generated indicating the authenticity of checked URL. If the result reveals that it is not phished then, the browsing is allowed to continue and else the result indicate that the URL is phished, user can take a decision regarding the prevention(step.5). If the decision is not to block, user can continue to access the web data on that resource locater (the system data are vulnerable to hackers). Else the decision made by the user is to block, then control goes to system properties and mark that URL as a spam in the host (the system data are safe). Note that access to the blocked site is denied not only from our browser (we created) but also from other browsers too and this is the main advantage of the system proposed.

Working of Random Forest Algorithm

Step1: First, start with the selection of random samples from a given dataset.

Step2: Algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step3: In this step, probability will be performed for every predicted result.

Step 4: At last, select the most probability prediction (average) result as the final prediction result.

3.1 Use case Diagram

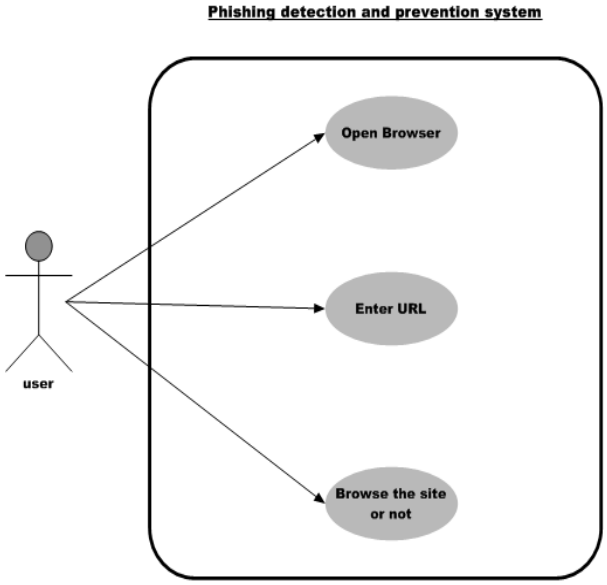


Fig 3.1: *Use case Diagram of Developed System*

Detecting and preventing the Phishing Websites using Machine Learning Approach is a mechanism that is proposed in order to ensure high security. In this mechanism we deal with the URLs and the URL check with machine learning technique and predict is it is phishing website or not. Here we create a web browser for browsing. Each time we browse a site the corresponding URL of site will be checked with machine learning technique. If the predicted result will negative then that site will be blocked. Then we can't access the site from that system. The main advantage of our system is that we can't access the blocked site not only from our browser but also from other browsers too.

3.2 Machine Learning Steps

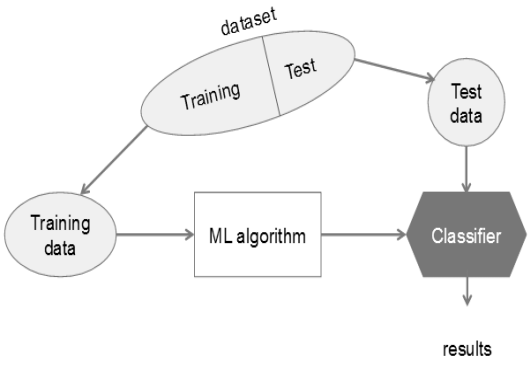


Fig 3.2: *Machine Learning Steps*

a) First time normal train with RF algorithm
b) Then test manually by input some URLs
c) Analyse the result
d) Train the same dataset using some hyper parameters in RF algorithm

e) Then again test manually by input some URLs
f) Analyse the result
g) By compare with old and new result we concluded that accuracy was improved little more.

3.3 Modules

USER: In this module user enter a URL for browse a site, that URL be the input for our machine learning algorithm (RANDOM FOREST). Based on the predicted result of the algorithm a user can access the site. If the entered site is phishing site then the user can't access it.

SYSTEM *(Process modules):*

- **Dataset Collection**: Text dataset of phishing site details are collected in the first process module. A data set is a collection of numbers or values that relate to a particular subject. Here it is the symbols, letters, variables used in the particular uniform resource locator which are pre-identified as of phishing nature.
- **Pre-processing**: In pre-processing step, the text data is tokenized and processed using natural language techniques like stemming, postaging, tf-idf etc. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. The arrangement of all these distinct values of strings or numbers into related group of values is made true by pre-processing stage.
- **Feature Extraction**: After pre-processing stage, it is necessary to extract the text features. Thereby text feature extraction is done at this module. The system cannot understand the text data so we convert the text data into numerical values. We can convert string arrays, character vectors, and cell arrays of character vectors to numeric values. Text can represent hexadecimal or binary values, though when you convert them to numbers they are stored as decimal values. This stage works over a random extraction of the observations from the dataset and a random extraction of the features. Not every tree sees all the features or all the observations, and this guarantees that the trees are de-correlated and therefore less prone to over-fitting. Each tree is also a sequence of yes-no questions based on a single or combination of features. At each node (this is at each question), the three divides the dataset into 2 buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, the importance of each feature is derived from how "pure" each of the buckets is.
- **Training**: Here we train the machine learning Classifier. We use the random forest classifier to classify the entered site into phishing site or not. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The reason that the random forest model works so well is 'a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.'
- **Testing**: In this phase the classifier predict a new URL of website is phishing site or not. The model has now been trained to learn the relationships between the features and the targets. The next step is figuring out how good the model is! To do this we make predictions on the test features (the model is never allowed to see the test answers). We then compare the predictions to the known answers. i.e., We expect there to be some relationship between all the features and the target value, and the model's job is to learn this relationship during training. Then, when it comes time to evaluate the model, we ask it to make predictions on a testing set where it only has access to the features (not the answers)! We can compare these predictions to the true value to judge how accurate the proposed system is.

4. APPLICATION

The popularity of applications on social networking websites has increased a great deal this year. This has led to a new wave of phishing attacks targeting the users of these applications. Symantec has examined phishing websites exploiting three major social networking brands. The fake websites display attractive offers on the social networking applications to lure end users. Some of the applications that the phishing sites were based on are:

i. **Social networking on mobile** – Due to the rise in the number of users accessing the Internet through smart phones, social networking websites have expanded their services on smart phones, including messaging, chatting, photo viewing, etc. This increase in users has opened more doors to attackers because there are now more potential victims. Hence, attackers have created phishing websites on social networking brands claiming to provide these services on smart phones.
ii. **Live chat** – In November, Symantec observed that five percent of the targeted applications were on live chat, and among them adult sex

chat was the most common target. The phishing attacks show fake offers of free sex chat to lure end users into entering their login credentials.

iii. **Blogging** – Phishing websites that attacked blogging in social networking comprised 23 percent of all targeted applications. Various attractive blog topics are used in the login pages of the phishing site as a means to con end users. Pornographic material is one of the most common topics observed in these phishing attempts.

iv. **Gaming** - In 2009, gaming has become an increasingly popular aspect of social networking. Symantec evaluated gaming and found that it comprised 13 percent of the targeted applications. Gaming applications in social networking generally require various kinds of credit points to progress to higher levels of the game. Some of these credit points typically require online payment. The phishing websites trick users by providing fake offers of free credit points on these gaming applications.

5. RESULT

Our simulation results indicate that the efficiency of phishing identification after training with the training dataset by improving the accuracy by mean AUC score. In particular, we study the phishing website problem and propose an identification architecture (random forest approach). We train the system using with RF algorithm and checked the Mean AUC Score.

Table 5: Observations

	Precision	Recall	F1 score	Support
-1	0.95	0.92	0.94	1653
1	0.94	0.96	0.95	1996
avg/total	0.95	0.94	0.94	3649

All AUC Scores:

[0.99209681 0.99164624 0.99099854 0.99193778 0.98909853

0.98114067 0.9723529 0.96853327 0.98689045 0.98632184]

Mean AUC Score (Random Forest) = 0.9851017031833058

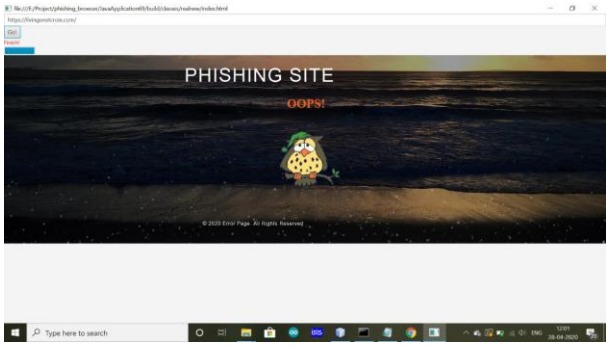
- Then test manually by input some URLs
- Analyse the result
- Train the same dataset using some hyper parameters in RF algorithm
- Then again test manually by input some urls
- Analyse the result
- By compare with old and new result we concluded that accuracy was improved little more.

[0.99264014 0.99238669 0.9920322 0.99302114 0.98981248

0.98367844 0.97576199 0.97456446 0.98938933 0.98701515]

Mean AUC Score (Random Forest) = 0.987030202601052

From the above mean accuracy result we concluded that the hyper-parameters can influence the accuracy of a model.



Input (URL to be checked phished or not) is given to the new browser we have created. Then, URL passed to the back-end where the machine learning steps are carried out. A prediction is generated indicating phished or legitimate. If the output is “1”, phishing and if it is “-1”, legitimate. Thereby user can choose decision regarding the blocking of URL. If it is to proceed, the occlude is done successfully.

6. CONCLUSION

The proposed system enables the internet users to have a safe browsing and safe transactions. Its helps users to save their important private details that should not be leaked. Providing our proposed system to users in the form of extension makes the process of delivering our system much easier. The results points to the efficiency that can be achieved using Random Forest approach (machine learning algorithm). Phished URLs are completely blocked from the System in which other browsers cannot access the page in future. A particular challenge in this domain is that criminals are constantly making new strategies to counter our defence measures. To succeed in this context, we need algorithms that continually adapt to new examples and features of phishing URL’s. And thus we use online learning algorithms. This new system can be designed to avail maximum accuracy. Using different approaches altogether will enhance the accuracy of the system, providing an efficient protection system. The drawback of this system is detecting of some minimal false positive and false negative results. These drawbacks can be eliminated by introducing much richer feature to feed to the machine learning algorithm that would result in much higher accuracy.

REFERENCES

[1] R. B. Basnet, A. H. Sung, "Mining web to detect phishing URLs", Proceedings of the International Conference on Machine Learning and Applications, vol. 1, pp. 568-573, Dec 2012.

[2] Abdelhamid N., Thabtah F., Ayesh A. (2014) Phishing detection based associative classification data mining. Expert systems with Applications Journal. 41 (2014) 5948-5959.

[3] Mohammad, R. M., Thabtah, F. & McCluskey, L. (2013) Predicting Phishing Websites using Neural Network trained with Back Propagation. Las Vigas, World Congress in Computer Science, Computer Engineering, and Applied Computing, pp. 682-686.

[4] Aburrous M.., Hossain M., Dahal K.P. and Thabtah F. (2010) Experimental Case Studies for Investigating E- Banking Phishing Techniques and Attack Strategies. Journal of Cognitive Computation, Springer Verlag, 2 (3): 242-253.

[5] Mohammad R., Thabtah F., McCluskey L., (2014B) Intelligent Rule based Phishing Websites Classification. Journal of Information Security (2), 1-17. ISSN 17518709. IET.

[6] Jain, Ankit Kumar, and B. B. Gupta. "Comparative analysis of features based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016, pp. 2125-2130.

[7] R.Aravindhan, Dr.R.Shanmugalakshmi, Certain Investigation on Web Application Security: Phishing Detection and Phishing Target Discovery, January 2016.

[8] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "A novel approach for phishing detection using URL-based heuristic," 2014 Int. Conf. Comput. Manag. Telecommun. ComManTel 2014, pp. 298–303, 2014.

[9] Naghmeh Moradpoor, Employing Machine Learning Techniques for Detection and Classification of Phishing Emails, July 2017.

[10] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[11] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017.

[12] Mohammad R., Thabtah F., McCluskey L., (2014A) Predicting Phishing Websites based on Self-Structuring Neural Network. Journal of Neural Computing and Applications, 25 (2). pp. 443-458. ISSN 0941-0643. Springer.

[13] Qabajeh I., Thabtah F., Chiclana F. (2015) Dynamic Classification Rules Data Mining Method. Journal of Management Analytics. Volume 2, Issue 3, pages 233-253. Wiley

BIOGRAPHIES

Achu Thomas Philip, currently pursuing B.Tech degree in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala, India at Mount Zion College of Engineering, Kadammanitta, Kerala, India. His primary research interests are in Artificial Intelligence (Machine Learning oriented programming), Robotics and Cyber Security.

Jain James, currently pursuing B.Tech degree in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala, India at Mount Zion College of Engineering, Kadammanitta, Kerala, India. His primary research interests are in Artificial Intelligence (Machine Learning oriented programming), Internet of Things, and Cyber Security.

Nisha Mohan P.M. received the M.Tech degree in Communication and Networking from MS University, Tirunelveli, India in 2013. She is currently working as Assistant Professor in the Department of Computer science and Engineering at Mount Zion College of Engineering, Kadammanitta, Kerala, India. Her primary research interests are in Cloud Computing, Image Processing and Cyber Security.