PREDICTOR AND RECOMMENDER SYSTEMS USING AI AND ML

(State Level Symposium on Predictor and Recommender Systems using AI And ML)

BONFRING®
Intellectual Integrity

Dr. P. Devaki
Dr. S. J. Syed Ali Fathima
Ms. S. NithyaRoopa

State Level Symposium on Predictor and Recommender Systems using AI and Ml

Limits of Liability/Disclaimer of Warranty: The authors are solely responsible for the contents of the paper in this volume. The publishers or editors do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are required to communicate such errors to the editors or publishers to avoid discrepancies in future. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Further, reader should be aware that internet website listed in this work may have changed or disappeared between when this was written and when it is read.

Bonfring also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Year of publication 2021

9 788195 162673 >

ISBN 978-81-951626-7-3

Bonfring

309, 2nd Floor, 5th Street Extension
Gandhipuram, Coimbatore-641 012.
Tamilnadu, India.
E-mail: conference.@bonfring.org, Website: www.bonfring.org

DETECTION OF PHISHING WEBSITE

Ramprassana .T, Bharath. S, Balaji .R, Dr. Suganthi .N

Department of Computer Science Engineering, Kumaraguru College of Technology, Coimbatore.

***Abstract*—** Phishing is often performed by email or text chat, respectively. And it also directs users to enter information on a fraudulent site that appears more like a real one [1]. Phishing is one of the threats used by the Social Engineering attack to trick users and manipulate the accessibility of new web security technology [2]. In order to identify and anticipate phishing websites, we suggest an intelligent, versatile and productive framework supported by Machine Learning Technique. We have also provided an algorithm and methods of classification for collecting and categorizing phishing datasets. The phishing website will be identified on the basis of certain essential features such as URL and Domain Identity, and authentication and encryption requirements within the overall phishing detection rate. Our method can use a machine-learning algorithm to detect whether or not the site is a phishing site. This application consistently utilized by many E-commerce enterprises so as to form the entire transaction process secure. The machine learning algorithm used during this process provides improved results compared to other conventional classification algorithms. Users can also buy items online without reluctance with the support of this framework. This mechanism has been used to validate the protection of the website.

***Keywords*—** Phishing websites URL, feature extractions, machine learning algorithm

I. Introduction

1.1 Conceptual Study of the Project

In the area of computer security, phishing is a technically misleading mechanism of pretending to access personal records such like email addresses, passwords, and credit card data, by attempting to portray as a trusted source in internet transactions. We could see the massive negative effect of the Phishing website on the income of the government bodies, Control of public, advertisements campaign and overall brand profile. We analyse previous work in phishing site detection using URL features that have motivated our approach. We are highly precise that Phishing continues to be one of the fast-growing types of internet identity theft scams that inflict short-term and long-term economic harm.

In the year 2020, an unprecedented number of individuals have started operating from home during the time of this global quarantine and are preparing to do so for the near future. The spread of COVID-19 has helped to manage this widespread transition to a remote workplace. In comparison, cyber criminals have also been equipped with a plethora of new techniques to commit fraud and theft. Attacks will increase in sophistication. Cybercriminals would be more constrained in terms of malware distribution methods, according to Kaspersky, as businesses catch up with patching security holes. Attackers are discovering new and creative ways to overcome detection and filtering behaviour. There is nearly 60% of organizations lose data and nearly 50% of organizations have credentials or accounts compromised and so on. According to the Anti-Phishing Working Group's Phishing Activity Trends Report, the average wire-transfer loss from BEC attacks in the second quarter of 2020 was $80,183. This is up from $54,000 in the first quarter.

1.2 Objectives of the Project

The objective of the research is to study and analyse various security issues in phishing attacks on the internet with an attempt to provide a model of security implementation by using some machine learning techniques, which will detect the phishing website and improve the security issues in Internet Computing and to provide benefits to the users. The main objective is to identify and consolidate the vulnerabilities and categorize it. Then it will protect the personal data of the internet users.

1.3 Scope of the Project

- Detect the phishing website
- Notify that the site is malicious
- Sends the phishing website to the blacklist
- Avoids unwanted fake advertisements from the website
- Prevent the personal data from misuse
- Protects our system from cyber attacks

II. Literature Review

Adebowale et al [2] reviewed on the status of the algorithms to classify messages as spam or ham and evaluated several data sets and performance measurements which can be used by all to determine the efficiency of spam filtration. They have compared many different machine learning techniques and their limitations. In general, the literature figure and volume examined indicate that considerable progress has been made in this area and remains. They have the ability to recognize distinctive characteristics of the content of emails. The design of email server and the phases in email processing were discussed. They done a qualitative research in spam filtering using machine learning, deep learning and deep adversarial learning algorithms like Naive bayes classifier, neural networks, rough set classifier, support vector classifiers, decision tree, etc., that can effectively handle the risk of spam emails. This research paper combines a legal and technical solution but wwithout an effective solution. Spam will continue to decrease the value of an efficient communication medium.

Ozgur Koray Sahingoz et al [1] reviewed seven distinct master learning algorithms such as Naive Bayes, KNN, Random Forest, Decision tree algorithm and so on have implemented a phishing website detection to group their list of features into two distinct classes as NLP-based features, mostly human-determined features and word vectors, concentrating on the use of words without any other operation by the URL. Then explain that NLP features work better than word vectors and can improve the accuracy of the Phishing Detection even when using these features together. It detects such active data on the web pages that can be added to the local network blacklist when the phishing of a web page is detected and could be banned from further requests. This paper is mainly based on natural language processing algorithm, so it takes more time for training the datasets and also difficult in nature. According to the experimental and comparative results, with a percentage of 97.98 accuracy rate for phishing URL detection, the Random Forest algorithm with only NLP-based features gave the best performance.

Adebowale et al [3] developed an effective Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm, tested and validated in this paper for the detection of phishing websites and protection-based schemes. The key contribution of the research is the study of hybrid features derived from text, images and frames and then used by a robust ANFIS solutions for spam detection. They used literature review, content-driven methods, approaches to visual similarity and approaches based on heuristics. Training and classification have shown that categorization can be improved. In this experiment, the features used consisted of 35 predictor features. A fair phishing detection rate performed on the basis of criteria is provided in the device detection criteria: search index, URL information, web address bar, image identity, domain identity, source code, JAVA script, page style and layout identity. A set of experiments was performed using 13,000 available datasets. The proposed solution achieves a percentage of 98.3 accuracy. This paper presented an intelligent phishing detection and protection scheme by employing a new approach for the users but something that doesn't guarantee exceptional results.

Faeze Asdaghi and Ali Soleimani [4] discussed web spam is an unethical and illegal way to improve web page visibility by deceiving search engine algorithms throughout this post. Basic principle of their approach is to evaluate the influence of removing a collection of features on a classifier's output instead of a single feature equivalent to the sequential backward collection. They also decided to implement a subset selection function for backward elimination that is called the smart-BT algorithm. They used Naïve bayes classifier method to train and test the dataset within the low storage requirement in low time complexity. Their chi-square attribute can be evaluating the features of extraction from web pages. Their algorithm can be used to finding the unvalued the features and the result can remove it from dataset then leads achieving IBA. The author used Content-based feature to focus on the content of web pages to find whether it is phishing URL or not. The highly non removable subset of the features is defined based on the analysis, and the vector dimensions features are reduced. The key purpose of this study is to understand the effect of the decrease in dimensions on the improvement of the utility of classification, in particular on unbalanced datasets. Behind this algorithm, the basic assumption is sensitivity to the unique behaviours of features, independently and together.

Waleed Ali. [5] discussed the Naive Bayes Network (NB) is a very simple Bayesian network, which includes directed acyclic graphs with only one parent (representing the class label) and some children. NB rejects any connection between the attributes and believes that, given the class name, all the attributes are conditionally independent. NB is based on probability estimates, called posterior probability, to allocate a class to an observed case. The rating decision as an approximation of the class posterior probabilities is expressed in a test example. The most likely class is allocated to the example of the examination. The SVM is one of the most popular and flexible machine-learning techniques in many science and engineering applications. SVM is based on optimizing the margin and thus generating the maximum distance between the separation hyperplane and the instances to reduce the upper limits of the generalization error predicted. In SVM training, some examples of the training data set called support vectors which are similar to the hyperplanes that distinguish the device and provide the most useful information for classification. In order to transform the data in high dimensions, the required kernel function is used to use linear discriminatory functions. A common decision tree that can be utilized for both classification and regression (RF) is Random Forest (RF). RF is an ensemble of multiple decision-making bodies trained separately on a variety of trainings. The classification data is decided by voting among all the qualified decision-making bodies. Thus, Random Forest typically achieves a greater accuracy in classification than a single tree.

Neda Abdelhamid et al [6] reviewed that phishing is considered a type of online threat that is described as the art of spoofing an honest company's website aimed at obtaining private information from users. Blacklist- Whitelist-based methodology, Fuzzy rule-based approaches, Machine learning approaches, CANTINA-based approaches, Image-based approaches, Associative Classification data mining are the most used approaches in this article. Blacklist-Whitelist based approach this define the blacklist method has been generalized to various implementations, one of which is Google Protected Browsing, which uses a collection of predefined phishing URLs to spot fraudulent URLs and insecure website and the whitelist, which is a set of trusted websites, while all other websites are considered untrusted. Fuzzy rule-based approaches this have used a larger set of features to predict websites type based on fuzzy logic. The division of one class line from another class is not well explain by the author. Machine learning methods were used effectively to solve classification problems, and so the authors compared certain widely used machine learning approaches on the topic of email phishing, including SVM, decision trees, and Naïve Bayes and the large data set was used by the authors. CANTINA based approaches this proposed content-driven methodology based on CANTINA approaches also demonstrates the form of websites using term-frequency-inverse-document-frequency and often reviews the content of the website and evaluates if TF-IDF is a phishing URL. TF-IDF measures the word quality of a text by assigning weights and counting their frequency. Image-based methods to classify the form of websites by contrasting phishing pages with visual similarity-based protected sites. The authors collected a limited number of official banking websites, and then carried out the experiments. The results revealed a low error rate.

Yan Ding et al [7] proposed the combination detection approach of Search & Heuristic Rule & Logistic Regression (SHLR) to detect the manipulation technique widely used by phishing websites and increase the filtering performance of valid web pages. Phishing websites and other online spam activities have also seen aware fast growth rates during this time. The preferred principle used in this article is the Search engine-based detection phase, Heuristic rule-based detection phase, LR classifier-based detection phase. The most popular features obtained by the authors are URL-based, host-based, and content-based. The main objectives of this study are to use the header tags content of the webpage as keywords and with the assistance of the web browser, to spontaneously sort legal web pages. As the first method to retrieve the web page's keywords using the frequency and inverse document frequency (TF-IDF) algorithm, the author used the search engine-based detection method. They will decrease the identification time of legal websites and increase the potential for real-time detection. Another methodology is the heuristic rule-based detection process, which will specifically decide if the URL is a phishing website or a legitimate website by comparing and targeting words with string patterns. Finally, to evaluate the type of the remaining pages, the author used logistic regression classifier and DNS extract, a similarity with phishing vocabulary and HTML attributes to enhance the adaptability and accuracy of the detection process.

Xi Xiao et al [8] studied that the constructing blacklists is the conventional approach to minimize such threats and proposed a highly precise CNN-MHSA, a Convolutional Neural Network (CNN), and the MHSA combined approach. By integrating its features and their weights, CNN-MHSA can produce highly-precise detection results for a URL object. They used the graph-based methods to find the victim's website by drawing the relation graph through phishing URLs to find the clue from the phishing URLs of the victim's websites. Machine learning approaches are another technique from the author that can block zero-day attacks, but researchers need to find features manually. This confines the improved performance of the result because no one can assure that the selected features are sufficient to identify malicious sites. Multi-head self-attention (MHSA) is a kind of attention mechanism that is now widely used in machine translation, evaluating the weights to express the different validity for each function, which is believed to be more appropriate in the detection of phishing websites. Convolutional Neural Network (CNN) It is a type of feedforward neural networks whose artificial kernels can respond to not only a single-pixel but also its neighbors. In this paper, we propose a novel and highly precise approach to phishing website detection called CNN-MHSA, which combines Convolutional Neural Network (CNN) and MHSA to improve the accuracy of detection.

Purvi Pujara and Chaudhari [9] clarified that Phishing is a means of extracting personal information from the user via email or website. Authors used the most common approach of detect phishing URL in database and gives the warning if it is phishing otherwise it shows legitimate. The approaches they used are Blacklist method, Heuristic based method and Visual similarity method, Machine Learning based classifiers. Blacklist method they used for easy and faster to implement to bypass the list and necessary to counter the new attack. They used another method is Heuristic based method which find solutions among all possible ones but it does not guarantee and also give accurate algorithm. Visual similarity approach is main feature they use for deceive user by compare the image of legitimate site. In this paper, they performed detailed about phishing website detection. Machine learning based classifier approach it works well in large databases, but fails to identify when attackers use corrupted domains to host their platform. There is no such thing as every single technique that is adequate to detect all Phishing attacks. According to this, they may conclude that tree-based classifiers are better suited to machine learning approaches than others.

The authors Santhana Lakshmi and Vijaya [10] clarified that the supervised learning algorithm like Naive Bayes classifier is intended for use when features within each class are independent of each other but in practice it tends to work well even when the presumption of independence is not true. The data collection used to learn is obtained from PHISHTANK. They have used Identity Extraction, Feature Extraction, Supervised Learning Algorithms, Multi-Layer Perceptron, Decision Tree Induction, Naive Bayes algorithms have been used for modelling the prediction tasks. The method calculates the probability distribution parameters with sampled preparation, as these properties are class-unconditional. Data are divided into two phases The performance evaluated based on mean absolute error, root mean squared error, relative absolute error and root relative squared error. The prediction accuracy is high in case of decision tree induction when compared to other two classification algorithms. The performance of the decision tree is handling both the numerical and categorical data. The later likelihood of the sample belonging to each class is determined for every unseen test sample. The machine learning strategy for prediction and evaluation using cross validation is seen in this paper.

Sheikh Shah Mohammad and Motiur Rahman [11] used three publicly available datasets from UCI repository and Mendeley data which contains some discrete type of attributes. For that they have used of the stack generalization and other important machine learning techniques based on the features to prediction from the meta estimator in two levels and clarified that the Random Forest and Multi-Layer Perceptron (MLP) provide better precision in terms of overall accuracy. Another finding is stacked generalization performs better with multiclass dataset. It provided an average of 97.5% of accuracy with datasets and provided a strong basement to the development of anti-phishing tools and other techniques like Binary classification decision is focused and Confusion Matrix with four attributes namely True Negative, False Negative, False Positive, True Positive are used. Research Gap – Confusion matrix not good for imbalanced data. Good option to reporting results in M-class classification problems because it is possible to observe the relations between the classifier outputs and the true ones.

Joby James et al [12] compared several features using various data mining algorithms like WHOIS, geographic properties and blacklist membership to categorize the websites and identified the phishing website URLs by analyzing the lexical and host-based features in Waikato Environment for Knowledge Analysis (WEKA) and MATLAB. WHOIS properties used to protect the private information of the internet users. They have provided an effective phishing detection method. They can try to identify phishing websites URL's by analyzing their lexical and host-based features and other machine learning techniques they used namely are Geographic Properties and Blacklist Membership. To succeed in this contest, we need algorithms that continually adapt to new examples and features of phishing URLs. Research Gap - This paper used WHOIS protection makes your information private, there have been cases where the registrars release this information the domain owner.

Hossein Shirazi et al [13] authors say that Phishing attacks happens continuously to users, online businesses and institutions it gets their passwords, pin numbers and other Account Details while we use that phishing website. Methodologies used by the authors are Non-Binary Features: Domain Length, URL Length. Binary Length with copyright logo, Page Title and Domain Name Match. In this article, the authors used two main concerns with existing machine learning approaches for phishing detection, the first approach towards the design of only domain name-based features and the second concern is the type of datasets used in the literature that are inadvertently biased based on the website URL. Their approach differs from all previous works in this space as it models the relationship of the domain name to the intent of phishing. With only seven features they were able to achieve a classification rate of 97% with cross-validated data. Research Gap – In this paper they used both binary and non-binary features this make the phishing detection accuracy more in power.

Jian Maoa et al [14] says that, Phishing is a severe and real hazardous threat to internet users. This author utilized Machine Learning Techniques to identify phishing websites with the aid of a hybrid framework based on the combination of SVM and KNN, this author used Machine Learning Techniques to identify phishing websites. Their paper's contribution proposes a conceptual model of techniques for data mining. Their suggested approach has achieved the highest accuracy of 90.04 percent and performs better than the other approaches in our analysis of this article. This research proposed a hybrid approach to categorize the websites as Phishing, Genuine, or Malicious. And they conclude that phishing website detection is an active research area because of its importance to both people and organizations, as phishing websites can cause potential financial losses.

Research Gap – In this paper they used KNN-SVM approach it is slow for large dataset it requires a large amount of time to process and does not perform well in case of overlapped classes.

III. Problem Definition

Using traditional, the web users can check the website manually browse through their knowledge. Only Specialists can identify these types of website immediately, even though the web users are aware of these phishing attacks. We have increasing breakthrough for the users to attain from the phishing or unsecure website with the help of machine learning to predict the phishing URLs. So we have proposed a spam detection of website to identify the unsecured site.

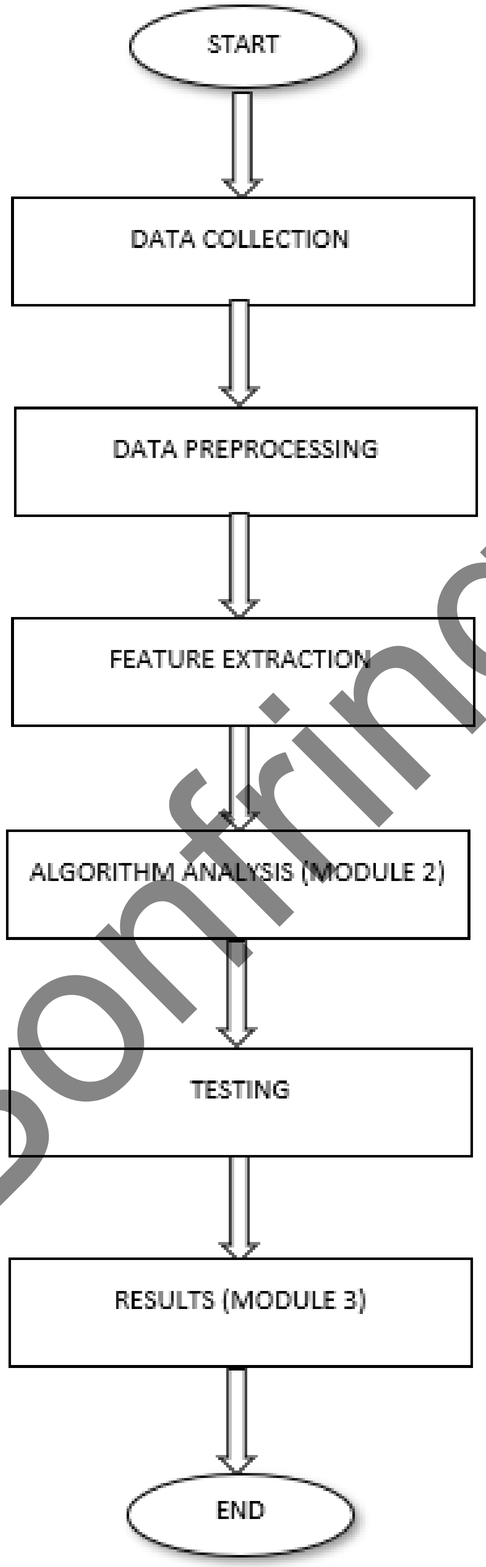
IV. Proposed System

We have proposed a Spam detection of website which has three modules. The first module is Data collection and pre-processing the collection dataset. We perform pre-processing to improve the quality and information of the collected data. After the feature analysis, it can be acquiring the result of phishing website. We can implement the dataset with the best algorithm. To make the best algorithm for finding the output data from the dataset. This Proposed system can make the selected features website into whether the given is phishing or legitimate one.

V. Methodology

In this project we present a Spam detection of website. In the proposed system we have used the dataset to pre-process and data cleaning the missing data and then we take the data to train and test into the proper format dataset. Then we have used the decision tree algorithm and logistic regression algorithm as the classification algorithm with the help of accuracy. This algorithm reduces the time and manual performance by testing and training dataset. Then pre-processing is performed with the algorithm to classify and show the accuracy of the given website. Final module is to detect the given website is whether legal or phishing website.

Flow Diagram



i. *Data Collection:*

In this process, using standard methods to assess the collected data and measure information on targeted variables, we collect the data for our research work and analyses the specific data, even collecting various sources from the different inputs. These are the first step for our project work.

ii. *Data Preprocessing:*

In this process, we used the advanced technique to transform our unprocessed data collection to be processed and converted to an efficient format. We clean up unstandardized data, integrate error data and also reduce missing data, noise data, etc. these make more standardized data.

iii. *Feature Extraction:*

In this process, extract the initial set of measured data and construct the derived features that are intended and non-redeemable, facilitate subsequent learning and generalize the steps. This makes the focus of data better than human.

iv. *Algorithm Analysis:*

In this work, it is a vital part of our project to implement each operation to identify the unknown data that can be used to describe the execution of the deployment. We even estimate the standard of that algorithm for the measurement of the processed data and the extracted features.

VI. Testing

In this task, we validate our work with a view to identifying any errors, gaps or missing any requirement for our use of algorithms as compared to the actual algorithm. We evaluate our initiative with a view to spotting out whether it satisfies all other works.

VII. System Requirements

1. Hardware Requirements:

- Intel Core i3 Processors
- Disk Space 1GB
- RAM: 4 GB or above
- Operating systems: MacOS, Windows 7 or later and Linux

2. Software Requirements:

- Anaconda 2020.07.
- Python 3.7

VIII. Conclusion

In Future we have plans to enhance the website. Now it is a website to find URLs are phishing or not in PCs only. In future we have idea to develop a Mobile Application and set a block options to completely avoid the phishing website. The Application is easy to use in PCs, Mobiles etc.... and find the phishing sites. We are going through a research and case study to avoid and destroy the phishing Sites completely.

References:

[1] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri. (2019), "Machine learning based phishing detection from URLs", Expert systems with Applications. pp 345-357.

[2] M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain. (2019), "Machine learning from email spam filtering", Heliyon Journal.

[3] M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain. (2018), "Intelligent web phishing detection and protection scheme using integrated features of images, frames and text", Expert systems with Applications. pp 300-313.

[4] Faeze Asdaghi, Ali Soleimani. (2018), "An effective feature selection method for web spam detection", Knowledge based System.

[5] Waleed Ali. (2017), "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection", International Journal of Advanced Computer Science and Applications.

[6] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah. "Phishing detection based associative classification data mining", Expert System with Applications. pp 5948-5959.

[7] Yan Ding, Nurbol Luktarhana, Keqin Li,Wushour Slamua. (2019), "A keyword-based combination approach for detecting phishing webpages", Computer & Security. pp 256-275

[8] Xi Xiao, Dianyan Zhang, Guangwu Hu, Yong Jiang, Shutao Xia. (2020), "A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites", Neural Networks. pp 303-312

[9] Purvi Pujara, M. B.Chaudhari. (2019) "Phishing Website Detection using Machine Learning", 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS).

[10] Santhana Lakshmi V Vijaya MS, "Efficient prediction of phishing websites using supervised learning algorithms", International Conference on Communication Technology and System Design. pp798-805.

[11] Sheikh Shah Mohammad Motiur Rahman. (2020), "Phish Stack: Evaluation of Stacked Generalization in Phishing URLs", International Conference on Computational Intelligence and Data Science. pp 2410-2418

[12] Joby James, Sandhya L, Ciza Thomas, "Detection of Phishing URLs Using Machine Learning Techniques", International Conference on Control Communication and Computing (ICCC).

[13] Hossein Shirazi, Bruhadeshwar Bezawada,

[14] Indrakshi Ray. (2018), "Kn0w Thy Doma1n Name: Unbiased Phishing Detection Using Domain Name Based Features", SACMAT '18: Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies. pp 69-75.

[15] Jian Maoa, Jingdong Biana, Wenqian Tiana,b, Shishi Zhua, Tao Weic, Aili Lid, Zhenkai Liange,

[16] "Detecting Phishing Websites via Aggregation Analysis of Page Layouts", 2017 International Conference on Identification, Information and Knowledge in the internet of things.

[17] Altyeb Altaher, "Phishing Websites Classification using Hybrid SVM and KNN Approach". 2017 International Journal of Advanced Computer Science and Applications.