International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 259-266 www.ijaem.net ISSN: 2395-5252

Investigating the Solutions of Phishing Detection Using ML Algorithm

Sonam Malviya

Submitted: 02-01-2022 Revised: 09-01-2022 Accepted: 12-01-2022

ABSTRACT—The online presence of various services motivates us to study and design new approaches for new generation online security threats. In this context, this paper is focused on studying the phishing attack and also involves a proposal to enhance the performance of existing Machine Learning (ML) based phishing attack detection. In this context the paper first includes the detailed study of data mining techniques and methods, further study includes the understanding the phishing, the attack deployment method, the root causes of the phishing attack, and the recent and popular anti-phishing techniques. Further, a review has been carried out for investigating the different available solutions of phishing detection. Finally based on the study a proposal has been offered for designing an accurate and efficient machine learning model for identifying the phishing URL patterns. In near future, the proposed model has been implemented and their performance has been compared with an existing machine learning model.

Keywords—machine learning, data mining, online security, review, proposal, phishing URL classification.

I. INTRODUCTION

Phishing is an act by which the attacker tries to misguide the internet user to recover their personal and confidential information. This information is used to perform financial fraud. In order to detect and prevent phishing attacks a significant amount of techniques have been contributed to the literature. Where two techniques are most popular first is black listing the URLs and maintain a large database to match the attack and the second is based on Machine Learning (ML) techniques. However, the list-based techniques are easy to implement but we required a huge resource which increases the implementation cost and the computational complexity. But the ML approaches are adaptive and can learn from the historical data and also adopt new knowledge. Additionally, these techniques are efficient and can accurately identify malicious patterns.

Thus in this paper, we are focused on investigating the data mining and ML techniques. Basically, the proposed work is motivated by a research article where S. C. Jeeva et al. [1] focuses on discriminating the features that can discriminate between two URLs in terms of legitimate and phishing. The extracted significant features are used with Apriori algorithms to construct the rules. The rules are inferred to underline the features. Additionally, some recent contribution on phishing detection has also been investigated. Further, the identified issues in the current system have been discussed and dedicated to design and implement an efficient and accurate phishing URL classification system. In this context, a basic architecture of the proposed model and the possible algorithms as a component are mentioned in the proposed model. Finally, the conclusion of the work carried out in this paper has been reported.

II. DATA MINING TECHNIQUE

The development of IT has generated a large amount of data in various fields. Data mining has given rise to store data and manipulates previously stored data for the decision-making process. Now we believe that information leads to power and success, and thanks to new technologies that are able tocollect data in bulk. Now we started collecting and storing all sorts of data and enhancing the computational power to sort the valuable information. These collections of data are very rapidly growing. Data mining involves the use of tools for data analysis to identify the unknown and valid patterns, as well as relationships. These tools can include statistical models, ML methods, and a set of algorithms. Data mining is more than collecting and managing data [2].

Contributing factors include the computerization in scientific, governmental, and management, and advanced tools to online instrumentation in manufacturing and shopping, and remote sensing. In addition, the use of the WWW as a system has flooded us with huge data. This growth has required new and innovative techniques that can assist us to turn raw data into useful knowledge. The

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 259-266 www.ijaem.net ISSN: 2395-5252

objective is to identify valid, novel, useful, and meaningful relations and consequences [3].

Data mining also referred to as knowledge discovery from data (KDD), automates the extraction and representation of knowledge captured from data sources. It is a set of process or computer learning techniques to analyze and extract knowledge. The purpose is to identify trends in data and can be defined as discovering meaningful correlation, patterns, and trends, using statistical, ML, artificial intelligence (AI), and visualization. Industries that are taking advantage may include medical, aerospace, etc. [4]. It is not only limited with a particular industrybut it requires technologies and enthusiasm to investigate the possibility of hidden information. The following figure 1 shows steps in a knowledge discovery process. The process comprises of a few steps as follows [5]:

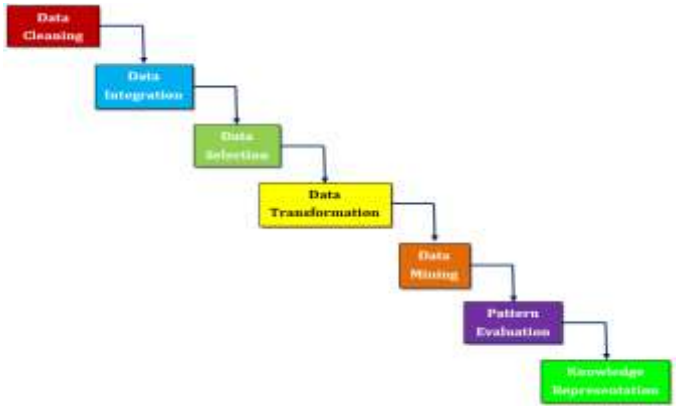

Figure 1.1: KDD Steps

- **Data Cleaning:** the noiseand irrelevant data from input data has removed.
- **Data Integration:**in this step different data sources, and heterogeneous nature of data is combined.
- **Data Selection: in this step the** key and essential data are identified for examination.
- **Data Transformation:** in this stage the chose information is changed into proper configurations.
- **Data Mining:** in this stage learning methods are applied to separate examples.
- **Pattern Evaluation:** In this progression, stringently fascinating examples addressing information are recognized.
- **Knowledge Representation:** in this stage found information is outwardly addressed. This progression utilizes representation procedures to comprehend and decipher the outcomes.

A. Classification of Data Mining System

Data mining systems can be categorized according to the following [6]:

- **According to the data sources:** The type of data handled such as multimedia data, spatial data, time-series, Web, text, etc.
- **According to database involved:**The data model involved such as relational, object oriented, data warehouse, transactional, etc.
- **According to the knowledge discovered:** The kind of knowledge discovers, discrimination, association, clustering,classification, etc. The systems offering several functionalities together.
- **According to techniques used:** The data analysis approach used such as ML, optimization, statistics, visualization, and others.
- **The user interaction with the mining:** such as query-driven, interactive exploratory systems, or autonomous systems.

B. Data Mining Functions

The type of consequences required is depending upon employment of the mining technique. The data mining techniques and the type of pattern obtained is briefly described as follows [7]:

Characterization: It is the conclusion of features ofthe objects as characteristic. The data relevant to a class are retrieved using query and passes through a module to extract the meaningful representation of the data.

Discrimination: This produces discriminate rules and is comparison of the features of objects in reference of two classes target class and evaluating class.

Association Analysis: The rate of items found together in transactions is the aim of association rule mining.The technique is usages a support that filters the frequent items. The confidence is atype of probability that measure how an item emerges in a transaction with other item. That is also supportive to association rule generation.

Categorization: It utilizes provided class names to arrange the articles. Characterization regularly utilizes a preparation set where articles are related

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 259-266 www.ijaem.net ISSN: 2395-5252

with realized class names. The calculation gains from the preparation set and construct a model. The model is utilized to group new articles.

Forecast: There are two sorts of expectations: one can attempt to foresee some inaccessible qualities or drifts, or anticipate a class mark. The second is order. When a model is constructed dependent on a preparation set, the class mark of an article can be anticipated.

Clustering: it is the association of information, in contrast to order, here class marks are obscure and the calculation finds classes. It is additionally called solo learning since it isn't directed by class names.

Outlier Analysis: Outliers are information components that can't be gathered in a bunch. Otherwise called special cases, they are vital to recognize.

III. UNDERSTANDING OF PHISHING

Phishing is a luring technique used by phishing attackers to exploiting the personal details of a user. It is a type of fraud that happens when a noxious Web website mimics an authentic to procure touchy data. There are a few enemies of phishing strategies for identifying phishing endeavors in messages and content on sites. Phishers consistently concoct new techniques to bypass the software and techniques. Phishing is a social engineering technique that is used to bypass technical controls to mitigate security risks. People are the weakest link. Phishing capitalizes on this weakness and exploits human nature to gain access to a system. In the beginning, phishers were usually acting alone. Early phishers desiring data to cause mischief and to make long-distance phone calls. Phishing attacks became more professional, organized, and systematic [8].

With the increasing online financial services and e-commerce, the focus of phishing attacks turned to consumers of online banks, retailers, and service providers such as PayPal. The phishing is usually online deployed using e-banks, Chatting, Messaging, and Emails. The phisher poses as an employee of an organization and then deceives the consumers into sending their information. Phishers create fake websites to increase the success rate. Phishers register dozens of domain names that look like a famous brand. Victims, who enter one of these websites by mistake, may believe that the website is real, and use their account [8].

A. Overview

Attackers are intruding to the network and collect the confidential data. Attacks may either be active attacks or passive attacks. It is a passive attack. It is a threat in social media also. Phishing messages contain links to an infected website. The link directs the user to the infected website to enter the information, so the hacker will use the information. It is normally difficult to know the website is actual or spoofed [9]. Let's take Facebook as an example, attacker creating a page that perfectly looks like Facebook but in a different URL that pretends to be legit. When a user lands on the page, the user might think it is the real Facebook page. So individuals who don't discover the page is dubious may enter their username, secret key and the data would be shipped off the programmer. It is negatively impacting all the sectors and also increases day by day. Such attacks make it possible for the adversary to orchestrate. The goals of phishing are to carry out fraudulent transactions illegally [10]. In phishing attacks the following steps are used:

- Creating a fake web site looks exactly same as legitimate Web site
- Then send link of the web site to a large amount of target users by the name of legitimate companies, trying to visit the web sites.
- Victims visit the web site and input its information
- Then perform the financial fraud such as unauthorized money transfer.

B. Phishing AttacksTypes

The different types of phishing attacks are [11]:

- **Deceptive Phishing:**It is a message that required confirms information about account, requesting users to re-enter information, account charges, new services offer requiring action, and others. Such kind of message is sent toa number of people and then attacker will wait to be get react someone to signina fake website.
- **Malware-Based Phishing:** This is a software, which is executing on users' device. It can be an email attachment, or downloadable file for a small and medium businesses that are not always keep their software up to date.
- **Key loggers and Screen loggers:** This is also a malware which tracks the keyboard input and the relevant information and send it to the hackers. They use browsers program and run when the system is started as drivers.
- **Session Hijacking:** This will monitor the user'sactivitieswhen user sign in to the account or do transaction. Then use information to perform actions which is confidential.
- **Data Theft:** Sensitive data will be taken by the victims without knowledge of the user. This information is such as passwords, credit card information, personal information, or confidential information by intercepting communications, legal data, records, etc., to cause damage to challengers.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 259-266 www.ijaem.net ISSN: 2395-5252

- **DNS-Based Phishing:** This phishing will modify the hosts file. The hackers will use bogus address and message will be sent from the fake website. Users will enter the personal information and be hacked.
- **Search Engine Phishing:** Phishers will create web pages for fake products, and wait to index by search engines, and then trap the customers to enter confidential information as a part of an order, sign-up, or balance transfer.

C. Anti-phishing Techniques

This section explains the different popular anti-phishing techniques.

1. Attribute Based Technique

This anti phishing technique employs different checks such as certificate, image, and URL. It checks whether an URL is new. Image comparison demonstrates the similarity or difference among the legitimate and suspected webpage. The testament legitimacy and validity are additionally checked with a confided in power. The legitimacy is its capacity to recognize the new and obscure assaults. The plan additionally identifies all the more bogus sites. Be that as it may, this strategy is costly. Staggered checking is slow consequently creating a high setback for identification [12].

2. Genetic Algorithm-based Techniques

The genetic algorithm based technique works to find out traces in phishing webpage to classify as illegitimate. The algorithm evolves rules based on include determination. These standards separate among unique and phished website pages. The benefit is the identification of phishing messages and noxious connection location. The hindrance lies in the mind boggling rules. The likelihood of bogus up-sides and the precisionhave to be considered [13].

3. Character-based Approach

This approach is based on the URL structure. A URL is defined as $< \text{a herf} = \text{hidden destination} > \text{visibletext} </\text{a} >$. The information of URLs is used to classify. The IP is in dotted format is classified as suspicious. Secondly, the URL with visible text corresponding to other destinations can be a phishing URL. The merit is that it can detect known and unknown both attacks. The disadvantage is false positives. This approach may fail for many cases [14].

4. Content Based Approach

The phishing webpage has the small lifecycle. Therefore has a low page rank. So attacker prepares a website and waits to be rank. In this scenario this technique decides the legitimacy of webpage. The merit is zero day attacks detection and less misclassification rate. The demeritis,it is time consuming for producing results and dependency on page rank algorithm [15].

5. Identity Based Approach

Some character of a site is utilized for avoidance of the assault is known as Identity based methodology. The method might utilize shared validation, where a client and a site commonly confirm one another.

6. Fuzzy Logic based Approach

Fuzzy logic has been suggests a promising unusual measures for operational risks [16]. These techniques present information to risk managers to assess and rank phishingwebsite risks. The benefit enables us to process, define, and compute relationships by mathematics.

D. Phishing URL

Location of phishing URLs has become troublesome because of the development of phishing and endeavors to keep away from moderation by boycotts. Cybercrime has made it conceivable to have crusades with more limited lifecycles, to sidestep boycott [17]. Getting data by persuading the clients to uncover their usernames, passwords, charge card, and so forth by imagining as a trusty source is known as phishing. It is an offense that objectives both social designing and specialized stunts to take data and is a type of wholesale fraud. Phishing sites are influencing people and monetary associations, to a genuine danger. Each URL has this normal punctuation-

$$< \text{protocol} >://< \text{hostname} >< \text{path} >$$

The $< \text{protocol} >$ part used to fetch the resource hosted in server using different protocols like http, https and FTP.The $< \text{hostname} >$ part is describes the web server. The hostname is the domain name and associated with postfix like co.in or com and others.The $< \text{path} >$ is similar to the path of a file in computer. It contains different marks like slashes, dots, dashes, etc.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 259-266 www.ijaem.net ISSN: 2395-5252

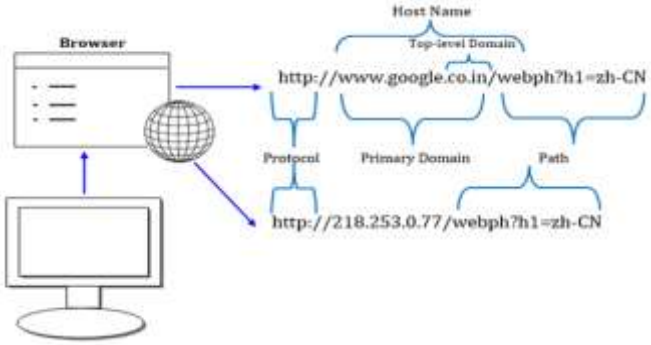

Figure 2 Example of a URL [18]

A phishing URL is developed with malicious intentions to distribute malwares, to attacks or control the consequences. The technical specialization of attackers is increasing to assemble sustainableinfrastructures. Botnets are the basicsof hosting phishing sites. Attacker lures the target user to click on a URL. It is usually modified URL. The black list techniques can used to provide protection [19].

IV. LITERATURE REVIEW

A significant contribution has been done for designing an efficient and accurate mode for phishing content analysis. Various approaches have been proposed for analyzing the website data or URLs to keep safe from phishing attack. In this section, describes recent studies done for identifying phishing.

H. Y. A. Abutair et al. [20] introduce a Case-Based Reasoning System also known asCBR-PDS. The system is adaptive and dynamic to detect new phishing attacks with a small data set to train classifier. The different scenarios of experiments on a 572 URLs have been carried out. The results show the accuracy of 95.62% and can work with a small feature set.

R. Gowtham et al. [21] presented a technique to overcome difficulties in phishing websitesdetecting, and alsoidentifies the target. The method bunches the domain from URLs having an immediate or aberrant relationship with the malicious website page. The domains assembled from the direct related website pages are contrasted and domains from indirect related pages to discover the Domain set. On applying the Target Identification algorithm, we find the target domain. Then, at that point play out an outsider DNS query of the dubious space and the objective area and on contrasting the authenticity.

R. S. Rao et al [22] carried out a work forPhishShield.This technique used for analysis of URL and Web Content. It accepts URL as an info and yields the situation with URL. The heuristics used to recognize phishing. It can recognize party time phishing assaults which boycott unfit to distinguish and quicker than visual assessment techniques. The accuracy for PhishShield is found 96.57%.Additionally itcan identify phishing web sites with less misclassification rates.

Table 1 Review Summary

Reference	Methods	Research area	Dataset
[20]	Case-Based Reasoning Phishing Detection System (CBR-PDS)	detect new phishing attacks	572 phishing and legitimate URLs
[21]	Target Identification (TID) algorithm	overcomes difficulties in detecting phishing websites and identifies the phishing target	third-party DNS look up
[22]	desktop application called PhishShield	Heuristics used to detect phishing.	URL and Website Content of phishing page
[23]	WEKA tool	data mining classification approach to detect malware behavior	a real case study data set.
[24]	sequence mining algorithm, All-Nearest-Neighbor (ANN)	recognize new, unseen malicious executables	Malware sample dataset
[25]	confidence weighted	content based phishing	URL Dataset

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 259-266 www.ijaem.net ISSN: 2395-5252

[26]	classification Phishing-Alarm, to detect phishing attacks	URL detection algorithm to quantify the suspiciousness ratings of Web pages based on visual appearance	real-world websites

MonireNorouzi et al. [23] presented an approach to classify malware behavior. The different classification techniques on the feature and behavior of malware have been applied. A dynamic analysis has also presented to identify the features. An algorithm is also developed for converting behavior of malware to WEKA based format. To show the performance a real case study based data set is obtained. The results show the approach is efficient and able to be use forbehavioral classification of malwares.

To protect users, a line of defence is anti-malware. The Anti-malwares softwares are using signature-based techniques. These techniques are not much effective for new kind of malware behaviour. Thus, **Y. Fan et al. [24]** proposed an algorithm to discover malicious patterns using sequential pattern mining.The instruction sequences aretaken out from the sample dataset then All-Nearest-Neighbor (ANN) classifier is used for malware classification. The developed framework composed of the sequential pattern mining and ANN classifier to detect new unseen malwares. Experiments on a real data are performed to show that framework outperforms other data mining methods.

Aaron Blum et al. [25] explores the use of confidence weighted classification technique to classify the content based phishing URL. Additionally they investigate the dynamics, extensibility of system, and different types of phishing domains. The system is able of detecting the phishing threats earlier and can provide security for zero hour threats.

Jian Mao et al. [36] presents Phishing-Alarm, to detect phishing attacks. They also present an algorithm to quantify the ratings of Web pages in terms of suspiciousness. That rating is determined using visual appearance of Web pages. The CSS is used to define a page layout. As page elements do not have the same influence as method based on rating on weighted page similarity. The evaluation shows the approach is correct and accurate and reflects low computational overhead.

V. PROPOSED WORK

The motivation of the proposed work is taken from the research article [1]. In this paper a machine intelligence based technique has implemented. The given technique works on 14 features of URL using association rule mining. To extend this work the following problems are addressed:

1. The given work usages the apriori algorithm for classification. The apriori algorithm usage candidate set which is a resource consuming process
2. The technique is a rule based technique, therefore the rule based technique needs a significant amount of comparisons

In order to resolve the addressedissues the following solution is proposed as described in figure 3.

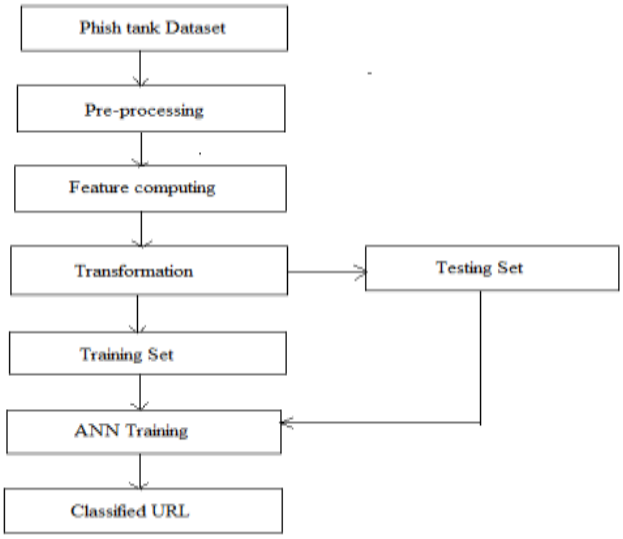

Figure 3 proposed system

In this diagram the proposed machine learning based phishing URL detection technique has been presented. The system accepts initial input in terms of learning samples; the proposed system utilizes the Phish tank dataset in this experiment. That phish tank dataset contains a number of different attributes, among them not all the attributes are going to be used. Therefore the pre-processing of the dataset is carried out. Using the pre-processing techniques we are just extracting the phishing URLs to learn additionally other remaining not usable attributes are removed from initial dataset. After extracting the URLs from the dataset the features from the URLs are computed. The calculated features of the URLs are further transformed into the two dimensional vector.

This 2D vector is further being used for experiment and classification for identifying the malicious or phishing URL. Thus for experimentation purpose we sub-divided the dataset into the two parts, first part is termed as the training set which consist of 70% of entire data sample and second part is called here as the testing dataset which consist of 30% of the samples to validate the model performance.After splitting of the dataset of pre-processed data samples the Artificial Neural Network (ANN) has been used to perform training of the supervised learning model.After training the model is able to recognize the similar featured URLs, thus to validate the trained ANN model we apply the test dataset for classifying the URL patterns. The ANN model returns the two types of class labels on which the model takes training. These classes are malicious (phishing) and legitimate (secure) URL. Finally based on the test sample classification results the performance of given system is evaluated in terms of accuracy and other parameters.

VI.CONCLUSIONS

The proposed work is aimed to study the data mining and machine learning techniques to solve the real world problem. In this context the problem of phishing URL classification has been proposed for exploration. Therefore a detailed study on phishing nature, deployment techniques, their types and the different solutions for phishing identification has been reported in this paper. In addition, to recognize the phishing URLs, emails, and web pages the recent development and research has also been explored in this paper. Finally some basic issue in phishing URL classification on existing approach has been addressed and for resolving the issues an machine learning model has been presented. This model will be implemented in near future;additionally a comparative study has been proposed to see the impact of changes in performance.

REFERENCES

[1] S. C. Jeeva, E. B. Rajsingh, "Intelligent phishing url detection using association rule mining." Human-centric Computing and Information Sciences 6.1 (2016): pp. 1-19.

[2] U. Fayyad, G. P. Shapiro, P. Smyth, "From data mining to knowledge discovery in databases", AI magazine 17, No. 3 (1996): 37.

[3] O. R. Zaïane, “Chapter I: Introduction to Data Mining”, CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta

[4] B. M. Ramageri, “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering, Volume 1 Number 4, pp. 301-305

[5] S. Sumathi, S. N. Sivanandam, "Introduction to data mining principles." Introduction to data mining and its applications (2006): 1-20.

[6] M.H. Dunham, “Data mining introductory and advanced topics”, Upper Saddle River, NJ: Pearson Education, New Delhi, 2003. Print. ISBN: 81-7758-785-4, 2006

[7] D. Watson, T. Holz, S. Mueller, “Know your enemy: Phishing, behind the scenes of Phishing attacks”, The Honeynet Project & Research Alliance (2005)

[8] T. Jagatic, N. Johnson, M. Jakobsson, F. Menczer, “Social Phishing”, Community. ACM, Vol. 50, No. 10 (pp. 94-100) (2007)

[9] E. Kirda, C. Kruegel, “Protecting Users Against Phishing Attacks with AntiPhish”, Computer Software and Applications Conference, 2005. 29th Annual International (Volume: 1)

[10] H. Tout, W. Hafner, “Phishpin: An identity-based anti-phishing approach”, in proceedings of international conference on computational science and engineering, Vancouver, BC, pages 347-352, 2009

[11] V. Suganya, “A Review on Phishing Attacks and Various Anti Phishing Techniques”, International Journal of Computer Applications, Volume 139 – No.1, April 2016

[12] V. Shreeram, M Suban, P Shanthi, K Manjula, “Anti-phishing detection of phishing attacks using genetic algorithm”, Proceedings of the International Conference on Communication Control and Computing Technology, pp. 447-450, 2010

[13] J. Chen, C. X. Guo, “Online Detection and Prevention of Phishing Attacks”, Proceeding of the First International Conference on Communication and Networking in China, Beijing, pp. 1-7, 2007

[14] M. Dunlop, S. Groat, D. Shelly, “Goldpolish: Using Images for Content-based Phishing Analysis, In Proceedings of the Fifth

International Conference on Internet Monitoring and Protection, Barcelona, pp. 123-128, 2010

[15] S. Shah, "Measuring Operational Risks using Fuzzy Logic Modeling," Article, Towers Perrin, JULY 2003

[16] H. Tout, W. Hafner "Phishpin: An identity based anti-phishing approach", in proceedings of international conference on computational science and engineering, Vancouver, BC, pp. 347-352, 2009

[17] D. Sahoo, C. Liu, S. CH Hoi, "Malicious URL detection using machine learning: A survey", arXiv preprint arXiv: 1701.07179 (2017)

[18] T. Gundel, "Phishing and Internet Banking Security, Technical Security report", IBM Crypto Competence Center

[19] S. Garera, N. Provos, M. Chew, A. D. Rubin, "A framework for detection and measurement of phishing attacks", In Proceedings of the 2007 ACM workshop on Recurring malcode, pp. 1-8

[20] H. Y. A. Abutair, A. Belghith, "Using Case-Based Reasoning for Phishing Detection", 8th International Conference on Ambient Systems, Networks and Technologies, Procedia Computer Science 109C (2017) pp. 281–288

[21] R. Gowtham, Dr. I. Krishnamurthi, K. S. S. Kumar, "An efficacious method for detecting phishing webpage through Target Domain Identification", Decision Support Systems November 30, 2013

[22] R. S. Rao, S. T. Ali, "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach", Eleventh International Multi-Conference on Information Processing-2015, Procedia Computer Science 54, pp.147 – 156

[23] N., Monire, A. Souri, M. S. Zamini, "A data mining classification approach for behavioral malware detection." Journal of Computer Networks and Communications (2016): 1.

[24] Y. Fan, Y. Ye, L. Chen, "Malicious sequential pattern mining for automatic malware detection." Expert Systems with Applications 52 (2016), pp. 16-25.

[25] A. Blum, B. Wardman, T. Solorio, G. Warner, "Lexical feature based phishing URL detection using online learning", In Proceedings of the 3rd ACM workshop on Artificial intelligence and security, pp. 54-60, 2010.

[26] J. Mao, W. Tian, P. Li, T. Wei, Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity", IEEE Access 5 (2017): pp. 17020-17030c.