Phishing Websites Classification using Hybrid SVM and KNN Approach

Nishat Khan
Pune vidhyarthi Griha College of Engineering, Pune

***Abstract*—Phishing is a potential web threat that includes mimicking official websites to trick users by stealing their important information such as username and password related to financial systems. The attackers use social engineering techniques like email, SMS and malware to fraud the users. Due to the potential financial losses caused by phishing, it is essential to find effective approaches for phishing websites detection. This paper proposes a hybrid approach for classifying the websites as Phishing, Legitimate or Suspicious websites, the proposed approach intelligently combines the K-nearest neighbors (KNN) algorithm with the Support Vector Machine (SVM) algorithm in two stages. Firstly, the KNN was utilized as a robust to noisy data and effective classifier. Secondly, the SVM is employed as a powerful classifier. The proposed approach integrates the simplicity of KNN with the effectiveness of SVM. The experimental results show that the proposed hybrid approach achieved the highest accuracy of 90.04% when compared with other approaches.**

Keywords—Information security; phishing websites; support vector machine; K-nearest neighbors

I. Introduction

Phishing is a serious threat that is potentially dangerous to the internet users. Phishing is a sort of semantic attack and attacks are targeting social or financial achievements [1]. In phishing attack, attackers attempt to trick and take money from the users of the internet by sending them e-mails instead of using malware software. First a fake web site which looks like the legitimate website is made by the attacker. Then users are requested to access the fake website and the attacker takes their money and important information. Phishing attacks become advanced continually as attackers find innovative methods and adapt their policies consequently. The most common method for phishing is e-mail. Phishing e-mails utilize different strategies to deceive the internet users into releasing their personal information such as account number, passwords and usernames. For example, requesting the user to validate his bank account information by providing his bank information to be compromised. The growing complexity of these approaches makes it difficult to protect users from phishing attacks [1]. The Anti-Phishing Working Group (APWG) is a group which gathers the phishing data from several sources, stated that phishing attacks continue spreading, as there were 69,533 unique phishing websites counted in December 2016, 80% of theses phishing attacks were targeting the online payment divisions [2].

Heuristic approach for phishing website detection [3]-[6] explores the content, structure and URL of phishing websites, finds the phishing websites features and develop an approach to identify phishing sites. The advantages of this method are the high speed and it generates less false negative or false positive. An attacker can avoid the filter and can achieve his target and get the user credentials when he understands the heuristic method strategy. The approach for phishing website detection based on visual similarity [7], [8] matches the suspicious website visual object such as text and images with original domain visual object. If the matching is less than definite threshold it is considered as legitimate site else as phishing. This method is not fast if matched with heuristic approach since it matches the suspicious website with the visual contents of all the legitimate websites.

The fast evolution of advanced phishing methods developed by professional attackers enabled the new phishers to build phishing websites using phishing software tools, which are offered in the internet. Therefore, the use of traditional anti-phishing approaches is not enough for efficient detection of phishing websites [8]. To limit the growing of phishing attacks, there is a need for effective and efficient solution. In this paper, a hybrid approach based on the combination of SVM and KNN is proposed to detect phishing webpages. The contributions of our paper are as follows:

1) To explore the potential of data mining techniques for phishing websites classification.

2) To propose a hybrid KNN-SVM approach for phishing websites classification

3) To validate and evaluate the performance of the proposed hybrid KNN-SVM approach for phishing websites classification and compare it with existing works.

The rest of the paper is structured as: Section 2 shows the related work. Section 3 explains the proposed hybrid approach for phishing websites classification. Section 4 presents the experimental results. Finally, Section 5 concludes our paper.

II. Related Work

Several research efforts were conducted to safe the users from the phishing websites. There are several approaches for phishing websites detection including machine learning [9], blacklists, visual similarity calculation [7] and classification of URL feature and domain name exploration. Our paper belongs to the first area. However, we also present a brief overview of the other fields to provide background to our research. In [1], Purkait provided a comprehensive phishing detection literature which introduced a complete review about the phishing detection techniques.

Google Safe Browsing [10] utilizes a blacklist phishing detection approach. The suspected uniform resource locator (URL) is matched in the blacklist to check its existence, if it is found in the blacklist the suspected URL is categorized as phishing website, else it will be categorized as valid website. The main drawback in this method is that phishing websites that are not exist in blacklist are not identified. The phishing websites appears for the first time is not presented in blacklist are named Zero day phishing sites. This approach could raise the false negative percentage. Cantina is proposed by Zhang *et al*. In [11], it is an approach for phishing website detecti depends on the text in the website. The Term Frequenc Inverse Document Frequency (TF-IDF) method is employ on the text in website to identify the phishing attacks. The top five terms based on the TF-IDF is sent to search unit and compared with the results obtained by searching unit using the suspicious link. Their results show that the proposed approach detected 89% of the phishing sites. This approach cannot detect the phishing websites when the text in the website is substituted with images.

In [12] the authors proposed an approach for phishing website detection. The authors have carefully chosen six structural-features of the website, and then the Support-Vector-Machine algorithm is utilized to decide if the website is phishing website or legitimate website. The accuracy of the classification of this approach was 84%. However, this approach mentioned significant features that can potentially help in defining the legitimate website. Moreover, it is not affected by prior familiarity of the user about computer security techniques. Aburrous *et al.* [13] proposed an intelligent phishing detection approach based on Fuzzy data-mining algorithms, they used 27 features for phishing website detection and achieved an accuracy of 83.7%. However, the features used in their proposed approach are inadequate. Xiang and Hong in [14] proposed an approach for phishing websites detection based on linear classifier. Their approach employed the Domain Object Model and Hyper-Text Mark-up language with 10 attributes to classify phishing sites. Their results attained an accuracy of 89%.

BaitAlarm [15] utilizes the visual features contrast to categorize legitimate and phishing websites. Phishing attacks usually employ similar designs to reproduce the layout of the real website so authors employed Cascading Style Sheets (CSS) for classifying phishing websites. The authors considered a valid site and matched with many phishing sites signifying the necessity of whitelist. The drawback of BaitAlarm is that calculation cost of CSS style and matching with the records of whitelist is excessively high.

III. The Proposed Hybrid KNN-SVM Approach for Phishing Websites Classification

In this section, we present the techniques that have been employed in this paper for the classification of the phishing websites.

A. *K-Nearest Neighbors (KNN)*

KNN is an effective supervised learning method for many problems including security techniques [16]. K-nearest neighbor is based on the clustering of the elements that have the same characteristics; it decides the class category of a test example based on its k neighbor that is near to it. The value of k in the KNN depends on the size of dataset and the type of the classification problem [17]. Fig. 1 shows that KNN classifies the target based on its neighbors.

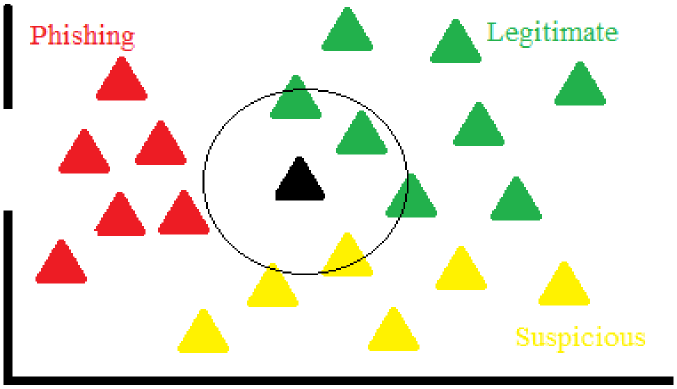

Fig. 1. A k-nearest neighbor (KNN) classifier

KNN is explained as follows:

Find the nearest elements from the test data *a* to training data K based on euclidean distance to calculate the distance. For two elements in k dimensional space, *a* = [*a*1, *a*2, …, *ak*] and y = [*b*1, *b*2, …, *b*k], the Euclidean distance based on the two elements can be computed by using (1) :

$$d(a,b) = \sqrt{\sum_{i=1}^{k}(b_i - a_i)^2} \tag{1}$$

After collecting the k-nearest neighbors, the majority of the k-nearest neighbors will be considered as a class for the test data.

B. *Support Vector Machine (SVM)*

SVM is a machine learning technique based on supervised learning and appropriate to both regression and classification [18]. The SVM is considered a modern technique achieving fast acceptance due to the good results achieved in a many fields of data mining problems, based on its solid foundation in statistical learning theory. SVM is a classification technique based on the statistical learning, which successfully utilized in many applications of nonlinear classification and large datasets and issues [19]. SVM classifiers employ the hyper-plane to isolate categories. Every hyper-plane is determined by its direction (w), the precise position in space or a threshold is (b), (xi) denotes the input array of constituent N and indicates the category. A set of the training cases are shown in (2) and (3).

$$(x1, y1), (x2, y2), \dots, (xk, yk); xi\epsilon R^d \tag{2}$$

K represents the number of training dataset and d denotes the number of the dimensions of the input dataset. The function of decision is specified as follows:

$$f(x,w,b) = sgn\big((w.xi) + b\big), w\epsilon\, R^d,\ \ \mathrm{b} \in R \tag{3}$$

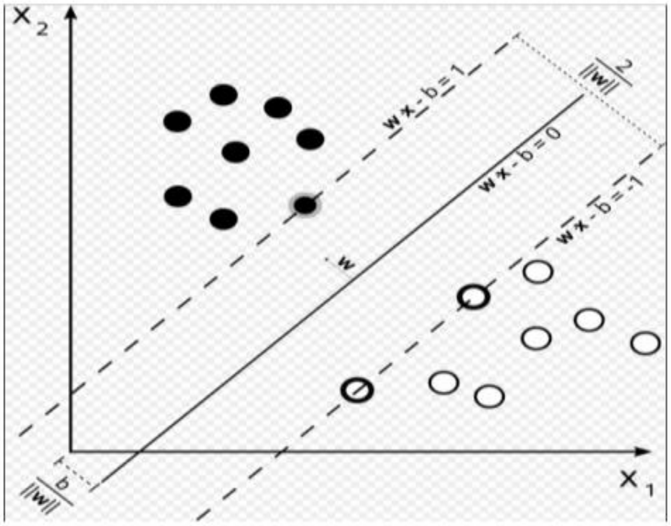

Fig. 2. SVM for phishing websites classification

One of the advantages of employing the SVM for training the system is its ability to work with multi-dimensional data. SVM is a classifier that takes a given labeled training data as input and outputs an optimal hyperplane which classifies new examples. SVM makes a hyperplane between data sets by maximizing the margin as shown in Fig. 2

C. The Proposed Hybrid KNN-SVM Approach for Phishing Website Detection

This paper proposes an integration of nearest neighbor classifier and support vector machine for phishing website detection. The proposed KNN-SVM hybrid classification approach can be used effectively for phishing website detection with low computational complexity in the training and detection stage. The lower computational complexity property is gained from KNN classification approach that does not require construction of a feature space. KNN algorithm has been used in the proposed hybrid approach KNN-SVM as the first step in the phishing website detection, and then the SVM method is employed in the second stage as a classification engine of this hybrid model. Fig. 3 shows the proposed KNN-SVM hybrid approach for phishing website detection.

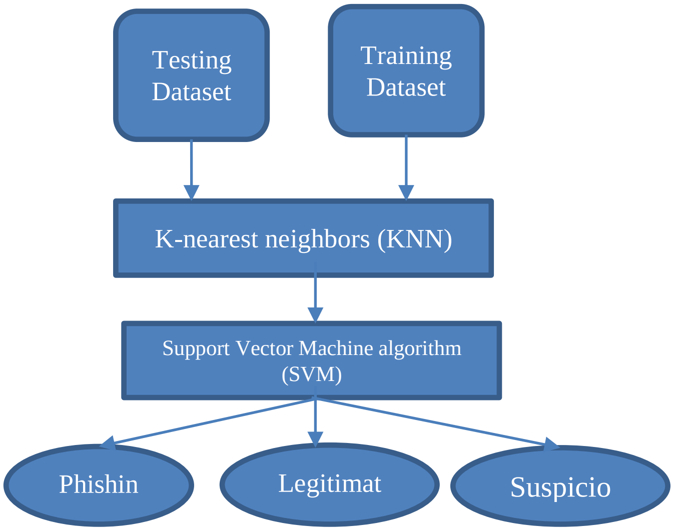
Testing Dataset
Training Dataset
K-nearest neighbors (KNN)
Support Vector Machine algorithm (SVM)
Phishin
Legitimat
Suspicio

Fig. 3. The proposed hybrid approach for phishing website detection

IV. Experimental Results and Discussion

In our study, we used a dataset prepared by Abdelhamid *et al*. [20]. The dataset contains more than 1353 samples collected from different sources, each sample record consists of nine different features plus the class which is Phishing, Legitimate or Suspicious website. The used features in the dataset are explained in Table 1.

To evaluate the effectiveness of the proposed hybrid KNN-SVM approach for phishing website detection, two classification metrics were considered, the Accuracy and Recall.

Accuracy: Accuracy is a well-known measure for the classification assessment. It is known as the proportion of correctly classified samples to the total number of samples, whereas the error rate uses wrongly classified rather than correctly. Equation (4) shows the mathematical formula for accuracy.

$$Accuracy = \frac{(TN+TP)}{(TN+FP)+(TP+FN)} \quad (4)$$

TABLE I. Description of the Features Used in the Dataset

Feature	**Description**
IP Address	The existence of IP address in the URL domain name indicates that someone is attempting to access the personal information.
URL Length	Phishers redirect the user’s submitted information by hiding the suspicious part of the URL to differentiate between legitimate and phishing URLs based on the URL length, Mohammad *et al* in [4] proposed that if the length of URL is more than 54 symbols the URL is classified phishy.
Pop Up Window	Generally, genuine sites do not request users to send their private information through secondary window.
Request URL	A webpage commonly contains characters, audio files, videos and images. Usually, these entities are uploaded in the webpage using the similar server hosting the webpage. When the entities are loaded from other domain, the webpage is possibly suspicious.
web_traffic	The traffic rate in Legitimate websites is usually high because they are used frequently. The traffic rate of phishing websites is usually low because they used rarely.
Fake HTTPs protocol	Using HTTPs protocol to transfer Important information indicates that the user is surely linked with a true website. Phishers could employ a false HTTPs protocol to trick the users.
URL of anchor	Like the URL, and differs in the links in the webpage which may possibly connect to another domain unlike the domain entered in the URL address bar.
Server Form Handler (SFH)	When the user send the data, the webpage will send the data to the server for processing. Usually, the information is handled from the same domain that hosting the website. Phishers try to send the information to fake domain.

Where, True Positive (TP): The number of phishing websites correctly classified as phishing websites.

False Positive (FP): The number of phishing websites classified as legitimate websites.

True Negative (TN): The number of legitimate websites classified as legitimate websites.

False Negative (FN): The number of legitimate websites classified as phishing websites.

Recall: Recall is the ratio of correct items that were selected. Equation (5) presents the mathematical formula for recall.

$$Recall = \frac{TP}{(TP+FN)} \tag{5}$$

The experiments conducted using the training and testing datasets to assess the performance of the proposed hybrid KNN-SVM approach for phishing websites detection. The total dataset is divided into two parts: 20% for testing and 80% for training. The proposed hybrid KNN-SVM approach combines the advantages of both SVM and KNN classifiers. The experiments compared the performance of proposed hybrid KNN-SVM approach with that of basic SVM, Naïve Bayes, Neural Network, Decision tree and basic KNN.

As shown in Fig. 4, the proposed hybrid KNN-SVM approach achieved the highest accuracy of 90.04% and outperform the other machine learning classifiers, which indicates the advantage of the proposed hybrid KNN-SVM approach. The performance of the proposed KNN-SVM is better than those of the SVM (83.76%) and KNN (87.45%), separately. Consequently, the proposed KNN-SVM can better classify the phishing websites than a single classifier. Also, Fig. 4 shows that the other machine learning classifiers including Naïve Bayes, Neural Network, DT and KNN achieved lower accuracy of 83.10%, 87.08%, 83.76%, 86.72%, 87.45%, respectively.

Fig. 5 demonstrates the recall results of the proposed hybrid KNN-SVM outperform the all other machine leaning classifiers, namely the Naïve Bayes, Neural Network, SVM, DT and KNN. The recall is calculated for the three classes: phishing, legitimate and suspicious, then the average recall for all the classes is presented. The proposed KNN-SVM achieved the highest average recall of 89.08%, while the Naïve Bayes, Neural Network, SVM, DT and KNN achieved an average recall of 64.41%, 80.65%, 58.56%, 64.89% and 84.78%, respectively as shown in Table 2.

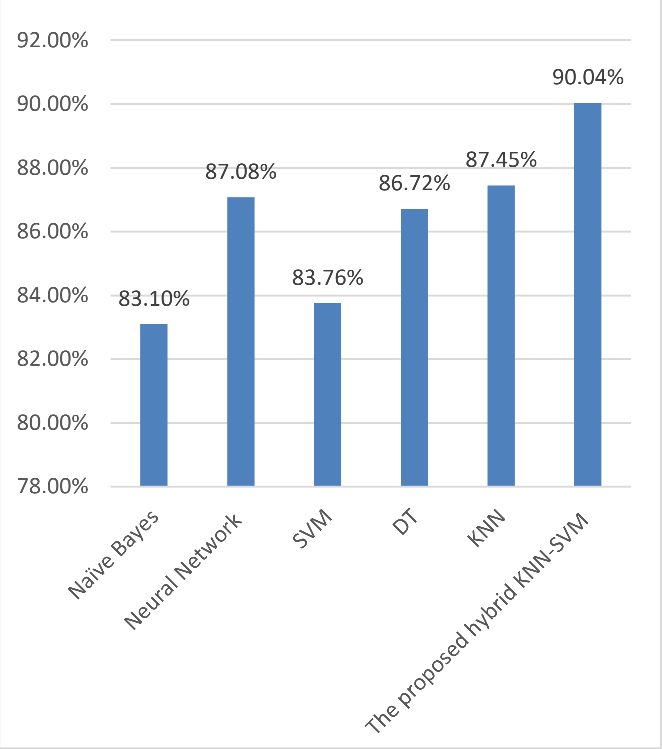

Fig. 4. Performance comparison between the proposed hybrid KNN-SVM and other approaches in terms of accuracy

TABLE II. COMPARISON OF THE RECALL ACHIEVED BY THE PROPOSED HYBRID KNN-SVM AND OTHER DATA MINING CLASSIFIERS

Used approach	**Suspicious**	**Legitimate**	**Phishing**	**AVG**
Naïve Bayes	17.65%	85.29%	90.30%	64.41%
Neural Network	66.67%	83.96%	91.33%	80.65%
SVM	0.00%	83.02%	92.67%	58.56%
DT	13.33%	88.68%	92.67%	64.89%
KNN	80.00%	83.02%	91.33%	84.78%
The proposed hybrid KNN-SVM	86.67%	90.57%	90.00%	89.08%

TABLE III. PERFORMANCE COMPARISON BETWEEN THE PROPOSED HYBRID KNN-SVM AND EXISTING APPROACHES

The approach	**Accuracy**
SVM based approach by Pan etal. [12]	84%
An intelligent phishing detection approach based on Fuzzy data-mining algorithms by Aburrous *et al* [13]	83.7%.
Text based approach By Zhang etal. [11]	89 %
phishing websites detection based on linear classifier by Xiang and Hong in [14]	89 %
The proposed hybrid KNN-SVM	90.04%

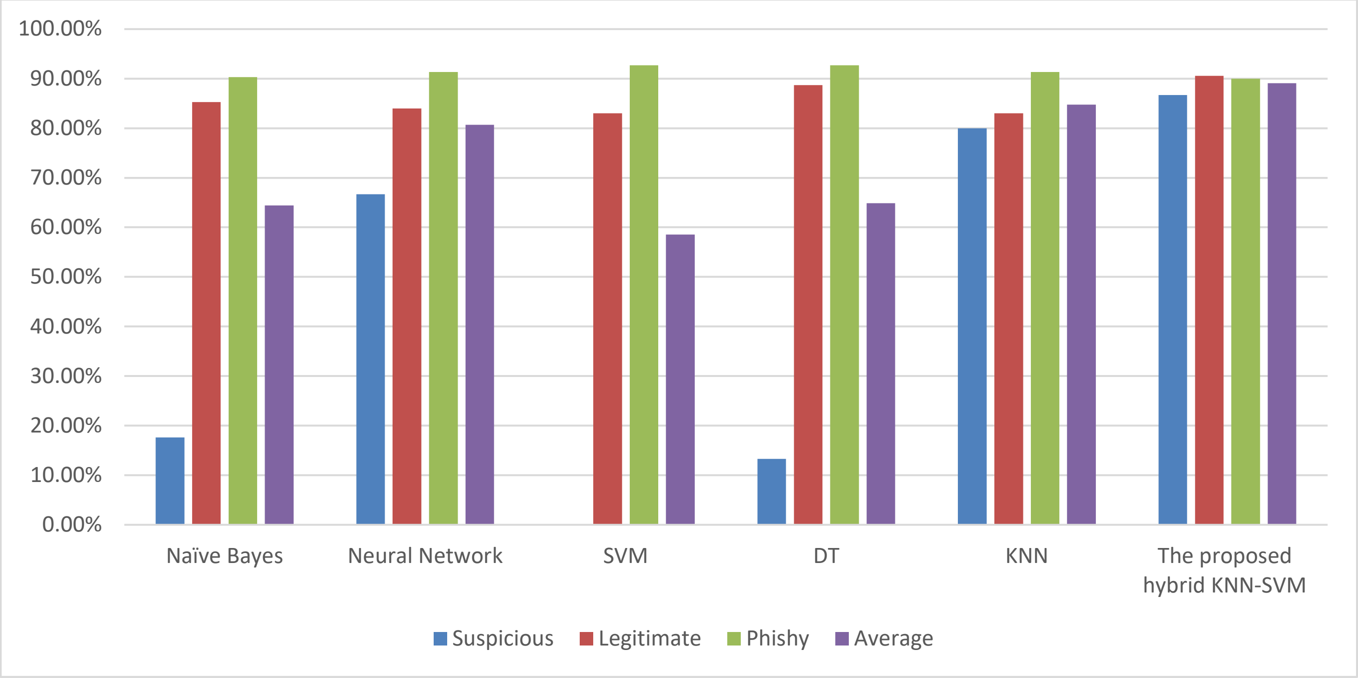

Fig. 5. Performance comparison between the proposed hybrid KNN-SVM and other approaches in terms of recall

The proposed hybrid approach is also compared with other related approaches as show in Table 3, it is clear form Table 3 that the proposed hybrid KNN-SVM approach achieved the highest accuracy of 90.04% and performs better than the other approaches. Among the listed related works, the intelligent phishing detection approach based on Fuzzy data-mining algorithms proposed by Aburrous *et al* [13] achieved the lowest accuracy of 83.7%. Text based approach proposed by Zhang *et al*. [11] and the phishing websites detection based on linear classifier proposed by Xiang and Hong [14], achieved performance close to the performance of our proposed hybrid KNN-SVM.

V. Conclusion

Detection of Phishing websites is an active research area due to its significant importance for both individuals and organizations, because phishing websites can cause potential financial loses. Artificial Intelligence techniques have been used successfully in many fields and it offer potential possibility to classify the fishing websites. This paper proposed a hybrid approach for classifying the websites as Phishing, Legitimate or Suspicious, the proposed approach combines the K-nearest neighbors (KNN) algorithm with the support vector machine algorithm (SVM). The proposed approach integrates the effectiveness and simplicity of KNN with the powerful of SVM. Thus, the proposed hybrid KNN-SVM gains the advantages of combining KNN with SVM and avoids their own drawbacks when they used separately. The experimental results show that the proposed hybrid approach achieved an accuracy of 90.04%. For the future, we will consider more advanced data mining techniques for the classification of the phishing websites.

References

[1] Purkait, Swapan. “Phishing counter measures and their effectiveness–literature review.” Information Management & Computer Security 20.5 , pp.382-420,2012.

[2] Anti-Phishing Working Group,2016. Phishing Activity Trends Report—4th Quarter 2016. http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

[3] Joshi, Yogesh, Samir Saklikar, Debabrata Das, and Subir Saha. “PhishGuard: a browser plug-in for protection from phishing.” In Internet Multimedia Services Architecture and Applications, 2008. IMSAA 2008. 2nd International Conference on, pp. 1-6. IEEE, 2008.

[4] Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. “An assessment of features related to phishing websites using an automated technique.” In Internet Technology And Secured Transactions, 2012 International Conference for, pp. 492-497. IEEE, 2012.

[5] ALmomani, Ammar, Tat-Chee Wan, Ahmad Manasrah, Altyeb Altaher, Eman Almomani, Karim Al-Saedi, Ahmad Alnajjar, and Sureswaran Ramadass. “A survey of learning based techniques of phishing email filtering.” International Journal of Digital Content Technology and its Applications 6, no. 18 (2012): 119..

[6] Chou, Neil, Robert Ledesma, Yuka Teraguchi, and John C. Mitchell. “Client-Side Defense Against Web-Based Identity Theft.” In NDSS. 2004.

[7] Fu, Anthony Y., Liu Wenyin, and Xiaotie Deng. “Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD).” IEEE transactions on dependable and secure computing 3, no. 4 (2006).

[8] Rao, Routhu Srinivasa, and Syed Taqi Ali. “A computer vision technique to detect phishing attacks.” In Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on, pp. 596-601. IEEE, 2015.

[9] Whittaker, Colin, Brian Ryner, and Marria Nazif. “Large-Scale Automatic Classification of Phishing Pages.” In NDSS, vol. 10, p. 2010. 2010.

[10] Safe Browsing API – Google Developer, [Online] Available at https://developers.google.com/safe-browsing/

[11] Zhang, Yue, Jason I. Hong, and Lorrie F. Cranor. “Cantina: a content-based approach to detecting phishing web sites.” In Proceedings of the 16th international conference on World Wide Web, pp. 639-648. ACM, 2007.

[12] Pan, Ying, and Xuhua Ding. “Anomaly based web phishing page detection.” In Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual, pp. 381-392. IEEE, 2006.

[13] Aburrous, Maher, M. Alamgir Hossain, Fadi Thabatah, and Keshav Dahal. “Intelligent phishing website detection system using fuzzy techniques.” In Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on, pp. 1-6. IEEE, 2008.

[14] Xiang, Guang, and Jason I. Hong. “A hybrid phish detection approach by identity discovery and keywords retrieval.” In Proceedings of the 18th international conference on World wide web, pp. 571-580. ACM, 2009.

[15] Mao, Jian, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang. “BaitAlarm: detecting phishing sites using similarity in fundamental visual features.” In Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference on, pp. 790-795. IEEE, 2013.

[16] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. “A detailed analysis of the KDD CUP 99 data set.” In Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on, pp. 1-6. IEEE, 2009.

[17] Bremner, David, Erik Demaine, Jeff Erickson, John Iacono, Stefan Langerman, Pat Morin, and Godfried Toussaint. “Output-sensitive algorithms for computing nearest-neighbour decision boundaries.” Discrete & Computational Geometry 33, no. 4 , pp.593-604, 2005.

[18] Suthaharan, Shan. “Support Vector Machine.” In Machine Learning Models and Algorithms for Big Data Classification, pp. 207-235, 2016.

[19] Hussain, Hanaa, Khaled Benkrid, and HÜSEYİN ŞEKER. “Novel dynamic partial reconfiguration implementations of the support vector machine classifier on FPGA.” Turkish Journal of Electrical Engineering & Computer Sciences 24, no. 5 ,pp. 3371-3387, 2016

[20] Abdelhamid, Neda, Aladdin Ayesh, and Fadi Thabtah. “Phishing detection based associative classification data mining.” Expert Systems with Applications 41, no. 13, pp.5948-5959, 2014.