2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)
Department of Electrical & Electronics Engineering, Galgotias College of Engineering and Technology, Gr. Noida, India

2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-7281-7741-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICACITE51222.2021.9404714

Survey on Detection and Prevention of Phishing Websites using Machine Learning

1st Malaika Rastogi
Computer Science Department
Galgotias University
Greater Noida, UP, India
malaikarastogi19@gmail.com

2nd Anmol Chhetri
Computer Science Department
Galgotias University
Greater Noida,UP, India
anmol1512@gmail.com

3rd Divyanshu Kumar Singh
Computer Science Department
Galgotias University
Greater Noida,UP, India
divyanshu_singh.scsebtech@galgotiasuniversity.edu.in

4th Gokul Rajan V
Assistant Professor,
School of Computing Science and Engineering,
Galgotias University, UP, India
gokulranjan@galgotiasuniversity.edu.in

***Abstract*—This research paper discusses on the phishing websites Prevention and Detection. A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. Phishing is the most commonly used social engineering and cyber-attack. Through such attacks, the phisher targets naïve online users by tricking them into revealing confidential information, with the purpose of using it fraudulently. In our paper we will be discussing and listing a few of the artificial intelligence models, that will help us to detect these phishing websites so that in the future these data and techniques can be used in machine learning to make our system better and efficient. The problem of phishing is widespread and there is no particular single solution available to effectively reduce all vulnerabilities, so many techniques are often used to reduce certain attacks. Machine learning is a useful tool used to reduce phishing attacks.**

As innovation keeps on developing, phishing strategies began to advance quickly several anti-phishing tools are available and have their own disadvantages. The paper concentrates on basic Machine learning supervised classification techniques to seek out an answer to phishing attacks. The Basic principle of this paper is to execute the frameworks with good efficiency, exactness, and cost-effectively. The task is attained by using 4 ML managed classification models. The four classification models are KNN, Kernel-SVM, Rando Forest Classifier and Decision tree. The supervised classification contains a labeled dataset that is used to train the models. All the four algorithms used: KNN, Kernel-SVM, Rando Forest Classifier and Decision tree are classification models. With machine learning, cybersecurity systems can analyze patterns and learn from them to assist prevent similar attacks and answer changing behavior. It can help users to be more active in preventing threats and respond to active attacks in real-time. So, by using Machine Learning, we could progress towards preventing such attacks.

Keywords – Web service, Cryptography, web Security, Machine Learning, Classification, Clustering.

I. Introduction

Phishing is an increasing sophisticated form of cyber-attacks that is not new to the world, but rather than it has existed among us since 1996, from Usenet newsgroup called AOHELL where it was made [1]. So phishing is basically a method of trying to gather personal information using deceptive e-mails and websites. The goal of phishing is to trick recipient into believing that the received message is something they want or need like something from their bank or note from someone in their company and then to click a link or download an attachment.

In phishing the attacker masquerades as a trusted entity of some kind, often a real or plausibly real person, or a company the victim might do business with [2]. The attacker represents himself as a trustful entity and then by using some of the social engineering, traps the victim which led them to do some actions that may only be in the interest of the attacker, causing loss of the victim. Since now everything is present on internet, from personal information to bank account credentials, there are lot of things that the attacker desire, like Adhar card number, address, mobile number, they all being some basic information, it can be harmful as well by knowing passwords or having some the bank credential which may lead to loss of money or some of the crucial secret information. With advancement in technology, maybe we can protect our self from such attacks. The Phishing URLS can be detected by the concept of machine learning, which may be used further to prevent such attacks [3].

Firstly, machines used to follow instructions given by human, but now humans can train the machine to learn from past data, build a prediction model and act much faster, and this is known as machine learning [4]. It is basically use of tools and technology that can be utilized effectively and efficiently. Machine learning is used to make to make human task easier, faster and simpler just by learning from the past data and working accordingly in the present [5].

Machine language is using data to answer questions, in this using data is referred as Training process, which is using our data to form a creation and fine turning it to predictable model [6]. And this model can be further used to serve our purpose that is to answer the questions. As data is added, models can be improved over time and new predictive model deployed. So, in this way we can provide data to our machine and it will help in prediction and detection of phishing websites [7].

II. Background

Phishing can be termed as fraudulent theft of sensitive information using electronic media fraud to deceive and take users confidential and sensitive information. The phishing scam aims to gain access to sensitive, private information such as passwords, usernames, credit card (CC) details, network credentials, and much more. Criminal attempts to steal sensitive information often occur via email in an attempt to capture sensitive information by contacting another user, such as clicking a malicious link or, downloading a virus attachment [8].

Phishing is nothing more than a con trick where spoofed e-mails and web pages are used as a bait to make people fall in the trap and take advantage of this vulnerability, and hence get some sensitive information. We shouldn't be amazed then by the terminology of phishing which is used to describe the act of sending an email that falsely claims to be from a legitimate organization [9]. We use "ph" rather than "f" in phishing, have we ever thought why we write "ph"? This is because some of the hackers during early times were referred to as "phreaks". Phreaking is a term that describes the action of experimenting with or manipulating communication systems. Phreaks or say hackers don't have much difference. They were always related to each other. The "ph" in phishing was referred to relate phishing attacks with some underground community [10].

According to the records, the very first time "phishing" came into the picture when it was first mention in the Usenet newspaper named as "AOHell` ", on January 02, 1996. Apparently, it was done there too; America Online is where the biggest explosion of what could be a major crime problem [11].

During the time when the (AOL) America Online was referred as the top provider of Internet access, there were many users who logged on frequently to the service every day. Its gaining popularity and usage made it a natural option for the user's having less than pure motives. From the start, hackers and sellers of malicious software have been using this service to communicate with one another. "warez community" was the name of this community. This community was the last to take the initiative to carry out criminal attacks on sensitive information. Since some of the random (CC)credit card numbers that make up the racket are closed, hackers then created most common and permanent set of methods. With the help of AOL- (instant messenger and email systems), they were then able to send messages to users and customers while pretending to be an AOL employee. These messages ask its users to verify accounts or verify payment / contact details. Most people fell into this trick; after all, nothing like this has ever been done. The problem intensified when hackers or so, called phishers introduced AIM accounts over the Internet; such accounts will not be penalized by the AOL TOS department. However, AOL has issued alerts to their emails and instant messenger customers to prevent people from providing sensitive information in these ways.

According to records, Microsoft has encountered some of the newest forms of phishing scams that are evolving so far by 2020 such as pointing email links to Google's fake search results targeting malware-controlled websites, pointing email links to pages that are not hosted by attackers to launch. (404-error page) that can be used as a login page for official sites, (Office-365) company-specific login pages to make it look so realistic that users will think it is real.

III. RELATED WORK

A phishing detection model which is based on a method that involves optimal features' selection and also neural network was proposed. This work involves, feature validity value, including the qualification value, a new index that is generated to assess the impact of factors on finding websites for sensitive identity theft subsequently based on this guide an algorithm is developed to identify relevant features on sensitive identity theft websites.

Fuzzy Rough Set hypothesis is executed as a technique to find the most impactful features from a few standard datasets. The features that are selected are then fed to classifiers for detection of phishing. The three-layered phishing attack detection model known as a "Web Crawler "based Phishing Scam Detector was proposed. It takes as input features the web content, traffic and URL. Based on those features, phishing or non-phishing website classification is made. A detection system was proposed that match the dynamic environment with the phishing websites. This is absolutely a client-side arrangement and doesn't require any third-party help [11].

To give a detailed view of the system (Fig. 1), below are the DFD (Data Flow Diagram) and Use-case Diagram for User as well as Admin

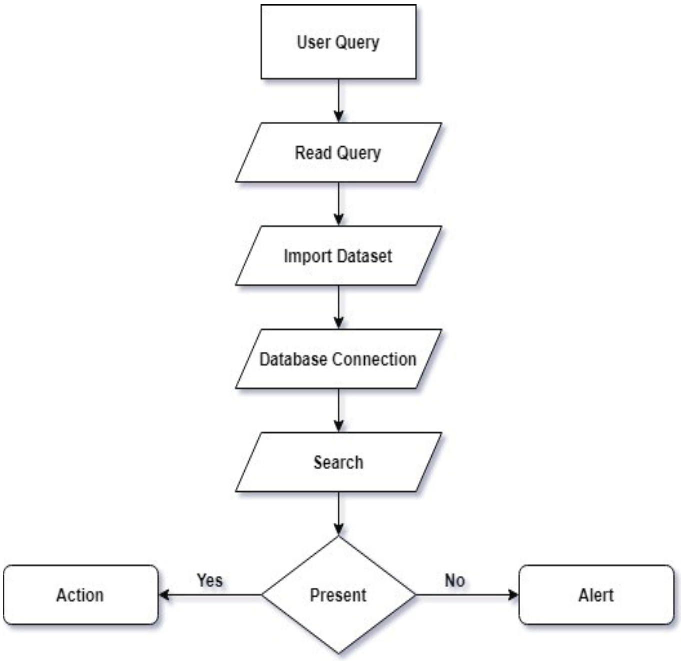

Fig.1. Data Flow Diagram

This diagram represents the graphical notation of the "flow" of data through system of collection of data, describing the process of the model and all its aspects The User's Query is the input to our system.

- Input is collected to process it.
- Phishing URL dataset is imported into the database.
- Database connection is made to perform database operations.
- The features of the given URL such as IP address, URL length is validated.
- If the features are matched with the features of Phishing Website then the user will get the alert.
- Else the user's action will be performed.

IV. DETECTION APPROACH

Because phishing scams deceives and ends up taking advantage of human lack of knowledge and ignorance or even naivety with respect to their experience with electronic communicable channels (e.g. E-Mail, HTTP, etc.), it's not a facile or simple problem to have a permanent solution. All the proposed solutions have one aim that is to minimize the impact and damage of phishing attacks. From a technical perspective, there are usually two nominated solutions to weaken phishing attacks: [12]

a. User's education or knowledge while accessing these websites.
b. Software enhancement

A. The challenges with both of the approaches are:

Non-technical people cannot learn or do no not want to learn, and even if they learn they do not retain their respective knowledge permanently, and because of this reason the training should be made continuous. There is some of the software solutions also present, such as authentication of the user and security warnings, but they are also co-dependent on user behavior and choices while using the web. If user ignores security warnings, the solution is furthermore proved useless.

B. Drawbacks in existing systems

This section focuses on the previous works carried out for detecting phishing websites using Machine learning. A phishing detection model which is based on a method that involves selection of optimal feature and also neural network was brought to practice. This work involves feature validity value, a new index that is generated to check the impact of the features on the detection of infected websites. Then based on it, an algorithm is generated to get the optimal features from infected or harmful websites.

Fuzzy Rough Set hypothesis is executed as a technique to find the most impactful features from a few standard datasets. The features that are selected are then fed to classifiers for the detection of phishing. Web Crawler which was a three-phase phishing attack detection model was proposed as Phishing Attack Detector. It takes in input features like web content, traffic, and URL. Based on those features, phishing or non-phishing website classification is made.

A detection system was proposed that matches the dynamic environment with the phishing websites. This is absolutely a client-side arrangement and doesn't require any third-party help. Parse Tree validation is another technique that is proposed to detect phishing websites. The approach makes use of hyperlinks of the current page by utilizing the Google API and builds a parse tree with intercepted hyperlinks. The parsing starts from the root and follows the Depth-first search algorithm and checks if any intermediate or leaf nodes have the value same as the root.

V. Required Model

This is a supervised machine learning job to do. There exist mainly two types of supervised Machine learning problems, which are known as classification and regression.

This data set comes under classification problem, as the input URL is classified as phishing (1) or legitimate (0). The machine learning models (classification) considered to train the dataset in this notebook is:

- Kernel Support Vector Machines
- Decision Tree
- Random Forest Classifier
- K-Nearest Neighbor

A. Kernel Support Vector Machine

Kernel Support Vector Machine (SVM) which is based on the margin maximization principle is a supervised machine learning algorithm [13]. It can not only be used for regression but also for classification problems, whereas majorly it is used for the classification problem. Some of the features are represented on the n-dimensional plains having the coordinates, and then the SVM model is used to sort the data and perform classification on the basis of finding a hyper-plane or margin that differentiates the two features efficiently. The mathematical function of the decision in SVM is

$$(x1, y1), (x2, y2), \ldots\ldots., (xk, yk); x1\epsilon\ R^d \qquad (1)$$

$$F(x, w, b) = sgn\big((w.xi) + b\big); w\epsilon R^d\ b\epsilon R \qquad (2)$$

Whereas, xi denotes the input, w is the direction of the hyper-plane, b is a threshold and d denoted the number of dimensions.

B. Decision Tree

A very popular data mining technique is a decision tree, which is liked by data miners, analyst, and also in artificial intelligence all over because of its intuitive nature and user-friendly results. So, in decision analysis, the decision tree represents the decision and helps in making the decision. So, in the decision tree, the growth of the tree involves various decisions such as which feature to choose and which condition to follow in order to make a proper data set. This tree is simple to understand and requires less effort from the user for the preparation of the data.[14]

C. Random Forest Classifier

Random Forest Classifier is also a supervised machine learning algorithm. As the name suggests, Forest classifier is a collection of random decision tree together in a classified way. Random Forest Classifier basically builds various decision trees together and further on merges or joins them in order to get a better and classified prediction. This model is quite useful in large databases and also provides an experimental method for detection techniques.[15]

D. K-Nearest Neighbor

K-Nearest Neighbor is also a supervised machine learning algorithm [16]. Amongst the most machine learning algorithms this is the simplest yet widely used classification algorithms. K-NN is used to estimate the occurrence of a data point just by having a look at the surrounding data points. It is basically based on the clustering or collecting of elements that possess the same characteristics or features. KNN can be further explained mathematically, with the help of formula,

$$d(a, b) = \sqrt{\textstyle\sum_{i=1}^{k}}(bi - ai)^2 \qquad (3)$$

Whereas, a and b are the two elements in the k dimensional plain, d= Euclidean distance

VI. Detection Process

To use machine learning, we have to train our system to identify such threats. To do so, we need sample data set of the phishing domains and the legitimate domain of the target. This sample data is required for studying purpose and help the machine make its decision. After collection of data, now it needs to be processed so that it can be used as information.

A. Decision tree

The decision tree can be inferred as nested-if-else block, where each feature will be tested and then moved further. The classes we have as the raw data is phishing domain and legitimate domain, besides this some features are also included in our data set such as domain string, title string, branch name string, length numeric, digit count numeric, www numeric, keyword numeric, second domain name numeric, and many more. Hence the collected data set will be used for detection [17].

Now the decision tree decides which feature to select and which not. This is done by using information gain measure that is specifically used to demonstrate that how efficiently a feature separates training examples followed to their target classification. Information Gain equation is

$$Gain(\,S, A) = \underbrace{Entropy(S)}_{original\ entropy\ of\ S} - \underbrace{\sum_{v\,\epsilon\,values(A)} \frac{|Sv|}{|S|}}_{relative\ entropy\ of\ S} * Entropy(S) \qquad (4)$$

As the measure becomes high, more is the distinguishing ability. The feature that has accomplished highest gain is selected as the root of that tree. The entropy is used to check the purity of the feature, with its equation as

$$H(S) = \sum_{i=1}^{n} - pi\ \log_2 pi \qquad (5)$$

The entropy plays inverse, when it is high means the purity is low whereas when the entropy of a feature is low, it means it is purest. Now in the Decision Tree Algorithm, for each of the feature information gain and purity is calculated and then as the tree grows in downwards direction, all leaves will have higher purity. When the tree is big enough, the training process is said to be completed. Then moving forward, tree is generated by using all these features which are a representation to the model structure for the detection mechanism. By taking a wide variety of data set, we can achieve high success rate and be successful in real world implementation.

B. Support Vector Machine

The Support Vector Machine is amongst the supervised learning algorithms that are found useful in detection of fishing websites [17]. It can effectively classify the malicious URL. This consists of the phases likely known as Training Phase and Testing Phase. In the training phase the support vector machine model is generated, based on which the data is further classified. In the formal testing phase, all the features are extracted or taken from test URL. Then these extracted features are further classified based on the training data set. Figure below shows the combined system architecture of the existing system and the proposed system. The components of these system are briefly described in Fig. 2:

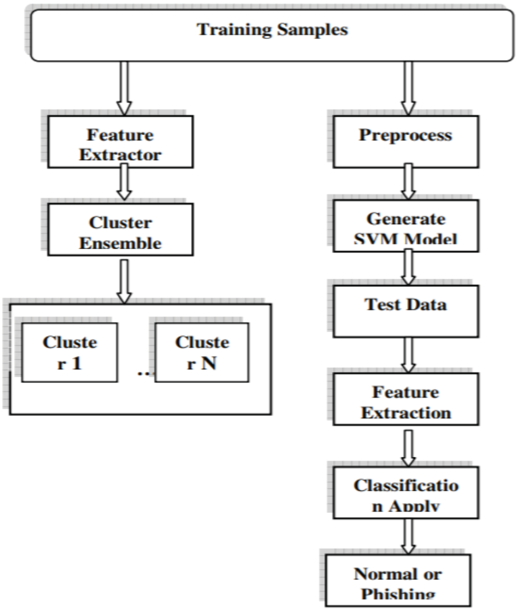

Fig.2. Fig. 2. System Architecture

1) *Feature Extractor:* It is basically used to extract the terms from particular phishing websites which have been collected by our system, then the data is transformed into term-frequency feature vectors. Further these featured vectors are support in the databases.
2) *Cluster Ensemble:* Base clustering solutions are the ones which are generated when different clustering algorithms are applied, that are based on the feature representations. It is useful in combination of various base clustering's. This is used in formation of clusters.
3) *Preprocess:* This is known as the Training phase of the Support Vector Machine. In this the data is collected from internet, which is re-processed, which is further carried out by removing the record which may contain some missing values.
4) *Generate Support Vector Machine Model:* Now the trained data set consist series of URL with many common features in it. Firstly, this data set has been trained using SVM algorithm with the help of the kernel functions which linear, polynomial or sigmoid, that helps in generating the SVM model.
5) *Feature Extraction*: By giving a URL that is test data, the feature extractor withdraws all the features of URL which is based on the above SVM model and furthermore it categorizes in two parts, the first is the safe (normal) website and the second one is the phishing website.

C. Random Forest Classifier

With the aim to efficiently detect phishing websites with high precision and recall, we use this algorithm which carefully selects specific features that represent each URL. The selected features are further used in training and testing phases [18]. To do this we divide this input into four different categories: training input, training output, testing input and testing output. We then use them to teach our Random Forest classifier. Once the classifier has taught to accept the specific features and make the required appropriate decision, we use the testing input to test the classifier. This process is summarized in the figure 3. These predictions are then evaluated based on accuracy, precision, recall and F-score.

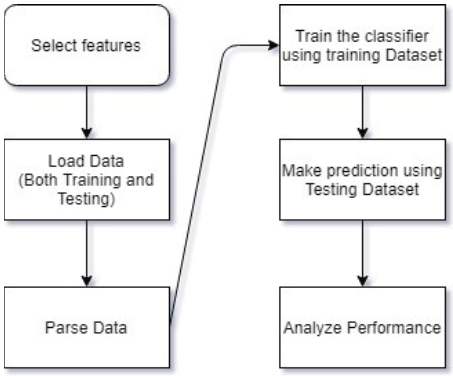
Select features
Load Data (Both Training and Testing)
Parse Data
Train the classifier using training Dataset
Make prediction using Testing Dataset
Analyze Performance

Fig.3. Fig. 3. Processed flowchart for each selected feature set.

VII. Experimental Results

The models listed above need also a mathematical support in order to estimate effectiveness, accuracy and performance. So, each of the models is evaluated and calculated based on True Positive (TP), True Negative (TN), and False Positive (FP) and False Negative (FN) scores respectively. Now we move forward to the calculation, [19]

$$Presision = \frac{TP}{TP+FP} \tag{6}$$

So here precision is the ratio of the True Positive scores to the sum of True Positive and False Positive. Precision is used to find the actual number of infected URLS, from all the URLS that are provided.

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

Recall is now used to find the number of URLS we identified as phishing from the actual number of phishing URL.

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \tag{8}$$

F1 Score is actually average of recall and precision.[20]

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{9}$$

As the name suggest accuracy is the number of correct outputs we get after whole of the process. Whereas here, TP is regarded as the True Positive is the number of rightly classified infected websites. FP is the False Positive which is the count of phishing websites which are classified as legitimate infected websites. TN is the True Negative defined as the count of legitimate websites which are classified as legitimate websites and at last FN is the False Negative which is count of legitimate websites classified as the infected websites.

VIII. Conclusion

The fundamental point mentioned in this paper is to execute the framework with high efficiency, exactness and cost effectively. The task is actualized utilizing 4 Machine Learning managed classification models. The four classification models are K-Nearest Neighbor, Kernel Support Vector Machine, Decision tree and Random Forest Classifier were discussed and analysed in term of the merits and demerits, Performance.

References

[1] Harinahalli Lokesh, G. and BoreGowda, G., “Phishing website detection based on effective machine learning approach,” Journal of Cyber Security Technology, pp.1-14, 2020

[2] Mahajan, R. and Siddavatam, I., “Phishing Website Detection using Machine Learning Algorithms,” International Journal of Computer Applications, 181(23), pp.45-47, 2018.

[3] Harinahalli Lokesh, G. and BoreGowda, G., “Phishing website detection based on effective machine learning approach” Journal of Cyber Security Technology, pp.1-14, 2020.

[4] MIT Technology Review, What is machine learning?. [online] Available at: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/, 2020

[5] Odeh, A., Keshta, I. and Abdelfattah, E., ”Efficient Detection of Phishing Websites Using Multilayer Perceptron, "International Journal of Interactive Mobile Technologies (iJIM), 14(11), p.22. 2020.

[6] Ahmed, k. and Naaz, S., “Detection of Phishing Websites Using Machine Learning Approach,” SSRN Electronic Journal, 2019.

[7] Altaher, A., “Phishing Websites Classification using Hybrid SVM and KNN Approach. International Journal of Advanced Computer Science and Applications, 8(6), 2019.

[8] HR, M., MV, A., S, G. and S, V., “Development of anti-phishing browser based on random forest and rule of extraction framework,” Cybersecurity, 3(1), 2020.

[9] SearchSecurity, What is Phishing? How it Works and How to Prevent it. [online] Available at: https://searchsecurity.techtarget.com/definition/phishing, 2021.

[10] Sahu, K. and K. Shrivastava, S., “Kernel k-Means Clustering for Phishing Website and Malware Categorization. International Journal of Computer Applications, 111(9), pp.20-25, 2015.

[11] YANG, X., YAN, L., YANG, B. and LI, Y., “Phishing Website Detection Using C4.5 Decision Tree” DEStech Transactions on Computer Science and Engineering, (itme), 2017.

[12] Medium, Understanding Random Forest. Available at:https://towardsdatascience.com/understanding-random-forest-58381e0602d2, 2011.

[13] Yaokai, Y. and Rabinovich, M., n.d, “Effective Phishing Detection Using Machine Learning Approach”.

[14] R, I. and Srivastava, T., “K Nearest Neighbor | KNN Algorithm KNN. Available at: <https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>, 2021.

[15] A. J. Park, R. N. Quadari, and H. H. Tsang, “Phishing website detection framework through web scraping and data mining,” In 2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pages 680-684, 2017.

[16] Altaher, A., “Phishing Websites Classification using Hybrid SVM and KNN Approach,” International Journal of Advanced Computer Science and Applications, 8(6), 2017.

[17] KSII Transactions on Internet and Information Systems, Robust URL Phishing Detection Based on Deep Learning. 14(7), 2020.

[18] HR, M., MV, A., S, G. and S, V., “Development of anti-phishing browser based on random forest and rule of extraction framework” Cybersecurity, 3(1), 2020.

[19] Saxena, S., Shrivastava, A. and Birchha, V., “A Proposal on Phishing URL Classification for Web Security,” International Journal of Computer Applications, 178(39), pp.47-49, 2019.

[20] Zouina, M. and Outtaj, B., “A novel lightweight URL phishing detection system using SVM and similarity index,” Human-centric Computing and Information Sciences, 7(1), 2017.