An Efficient approach for Detecting Phishing Websites Using Supervised Machine Learning Algorithms

Gentem Varaprasd[#1] K Shalivahana Reddy[*2]

[#]*Assistant Professor in Computer Science and Engineering, Santhiram Engineering College, Nandyal.*

[*]*Assistant Professor in Computer Science and Engineering, Global College of Engineering & Technology, Nandyal.*

***Abstract*—** Phishing is a method of trying to gather personal information using deceptive e-mails and websites. Phishing is a cyber attack that uses disguised email as a weapon. The goal is to trick the email recipient into believing that the message is something they want or need — a request from their bank, for instance, or a note from someone in their company — and to click a link or download an attachment. Phishers use the websites which are visually and semantically similar to those real websites. Machine learning is a powerful tool used to strive against phishing attacks. In this paper discuss an efficient approach to detect phishing websites using supervised machine learning algorithm. In this paper, we propose a classification model in order to classify the phishing attacks. This model comprises of feature extraction from sites and classification of website.

***Keywords*—** Phishing website, Supervised Machine Learning, Decision Tree, Phishing, Phishing Websites, Detection, Machine Learning.

I. INTRODUCTION

In present time, Social networks are common and popular platforms where person interact with other person easily. For communication, share and know to each other is possible by person (users) with help of social networks. In social network platforms, there is huge amount of social and personal data available. So, privacy protection of user becomes more urgent research topics. A lot of privacy violation incidents that caused by phishing attacks and they still work for stealing information in traditional way. An attacker mimic electronic communications by which he get confidential information that provide by user, this social engineering form is phishing. Through emails, such type of communication that tricks users to visit that fraudulent website which is collect passwords, credit card details and confidential information of user

"Phish" is pronounced just like it's spelled, which is to say like the word "fish" — the analogy is of an angler throwing a baited hook out there (the phishing email) and hoping you bite. The term arose in the mid-1990s among hackers aiming to trick AOL users into giving up their login information. The "ph" is part of a tradition of whimsical hacker spelling, and was probably influenced by the term "phreaking," short for "phone phreaking," an early form of hacking that involved playing sound tones into telephone handsets to get free phone calls.

Nearly a third of all breaches in the past year involved phishing, according to the 2019 Verizon Data Breach Investigations Report. For cyber-espionage attacks, that number jumps to 78%. The worst phishing news for 2019 is that its perpetrators are getting much, much better at it thanks to well-produced, off-the-shelf tools and templates.

Phishing is the most unsafe criminal exercises in cyber space. Since most of the users go online to access the services provided by government and financial institutions, there has been a significant increase in phishing attacks for the past few years. Phishers started to earn money and they are doing this as a successful business. Various methods are used by phishers to attack the vulnerable users such as messaging, VOIP, spoofed link and counterfeit websites. It is very easy to create counterfeit websites, which looks like a genuine website in terms of layout and content. Even, the content of these websites would be identical to their legitimate websites. The reason for creating these websites is to get private data from users like account numbers, login id, passwords of debit and credit card, etc. Moreover, attackers ask security questions to answer to posing as a high level security measure providing to users. When users respond to those questions, they get easily trapped into phishing attacks. Many researchers have been going on to prevent phishing attacks by different communities around the world. Phishing attacks can be prevented by detecting the websites and creating awareness to users to identify the phishing websites. Machine learning algorithms have been one of the powerful techniques in detecting phishing websites.

URL[1] is the abbreviation of Uniform Resource Locator, which is the global address of documents and other resources on the World Wide Web. A URL has two main components : (i) protocol

identifier (indicates what protocol to use) (ii) resource name (specifies the IP address or the domain name where the resource is located). The protocol identifier and the resource name are separated by a colon and two forward slashes.

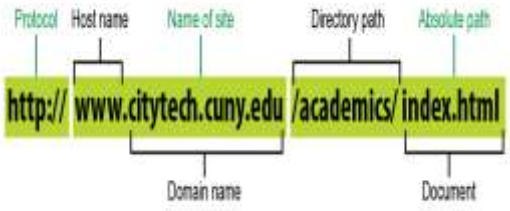

Fig. 1.1

URL Parts

Machine learning[2] is a subset of Artificial Intelligence. Machine learning is estimates the future tasks based on the previous experiences. Machine learning system builds the learning model that learns from experiences of the past to enhance the performance of Intelligence tasks. Machine learning is used in a variety of computational tasks include email spam filtering, recognition of intruders in networks, ranking of web pages, recognizing friend's photo on facebook etc.,

Microsoft Azure[3] platform provides tools for machine learning. In these experiments, the two class boosted decision tree and the two class support vector machine (SVM) were used as spam classifiers. The decision tree is Mainly used in data mining. It has the ability to create a model that foreshows the value of a target variable based on various input variables. The SVM is a supervised learning model that has learning algorithms and the ability to analyze data for classification..

Fig.1.2. A general processing framework for Phishing website detection using Machine Learning

II. RELATED WORK

Another approach by authors [4] proposes feature selection algorithms to decrease the components of dataset to get higher order execution [4]. It also compared with other data mining classification algorithms and results obtained. Dataset for phishing websites was taken from UCI machine learning repository[4]. From the outcomes, it is seen that some classification strategies increment the execution; some of them decline the execution with decreased component. Bayesian Network, Stochastic Gradient Descent (SGD), lazy.K.Star, Randomizable Filtered Classifier, Logistic model tree (LMT) and ID3 (Iterative Dichotomiser)[4] are useful for reduce phishing dataset and Multilayer Perception, JRip, PART, J48[4], Random Forest and Random Tree algorithms are not valuable for the diminished phishing dataset. Lazy. K.Star obtained 97.58% accuracy with 27 reduced features. This study is obtained with the help of WEKA software.

Authors [5]proposed a model with answer for recognize phishing sites by utilizing URL identification strategy utilizing Random Forest algorithm. Show has three stages, namely Parsing, Heuristic Classification of data, Performance Analysis [5]. Parsing is used to analyze feature set. Dataset gathered from Phish tank. Out of 31 features only 8 features are considered for parsing. Random forest method obtained accuracy level of 95%.

Authors [6] proposed a flexible filtering decision module to extract features automatically without any specific expert knowledge of the URL domain using neural network model. In this approach authors used all the characters included in the URL strings and count byte values. They not only count byte values and also overlap parts of neighboring characters by shifting 4-bits. They embed combination information of two characters appearing sequentially and counts how many times each value appears in the original URL string and achieves a 512 dimension vector. Neural network model tested with three optimizers Adam, AdaDelta and SGD. Adam was the best optimizer with accuracy 94.18% than others. Authors also conclude that this model accuracy is higher than the previously proposed complex neural network topology.

In this paper authors [7] made a comparative study to detect malicious URL with classical machine learning technique – logistic regression using bigram, deep learning techniques like convolution neural network (CNN) and CNN long short-term memory (CNN-LSTM)[7] as architecture. The dataset collected from Phishtank, OpenPhish for phishing URLs and dataset MalwareDomainlist, MalwareDomains were collected for malicious URLs. As a result of comparison, CNN-LSTM obtained 98% accuracy. In this paper authors used TensorFlow[7] in conjuction with Keras[7] for deep learning architecture.

Authors [8] in this paper created an extension to Google Chrome to detect phishing websites content

with the help of machine learning algorithms. Dataset UCI-Machine Learning Repository used and 22 features were extracted for this dataset. Algorithms kNN, SVM and Random Forest were chosen for precision, recall,f1-score and accuracy comparison. Random Forest obtained a best score and HTML,JavaScript, CSS[8] used for implementing chrome extension along with python. This extension is having a drawback of declared malicious site list which is increasing every day.

III. CHARACTERISTICS OF PHISHING

The attacker can register any domain name that has not been registered before. This part of URL can be set only once. The phisher can change FreeURL at any time to create a new URL. The reason security defenders struggle to detect phishing domains is because of the unique part of the website domain (the FreeURL). When a domain detected as a fraudulent, it is easy to prevent this domain before an user access to it. Some threat intelligence companies detect and publish fraudulent web pages or IPs as blacklists, thus preventing these harmful assets by others is getting easier. The attacker must intelligently choose the domain names because the aim should be convincing the users, and then setting the FreeURL to make detection difficult. Lets analyse an example given below.

protocol	http://
Domain name	active-userid.com
path	/webapps/89980/
Subdomain item1	com-webappsuserid29348325limited
Subdomain item2	paypal

Fig. 3.1 analysis of URL detection

Although the real domain name is active-userid.com, the attacker tried to make the domain look like paypal.com by adding FreeURL. When users see paypal.com at the beginning of the URL, they can trust the site and connect it, then can share their sensitive information to the this fraudulent site. This is a frequently used method by attackers.

Other methods that are often used by attackers are Cybersquatting and Typosquatting.

Cybersquatting (also known as domain squatting), is registering, trafficking in, or using a domain name with bad faith intent to profit from the goodwill of a trademark belonging to someone else. The cybersquatter may offer selling the domain to a person or company who owns a trademark contained within the name at an inflated price or may use it for fraudulent purposes such as phishing. For example, the name of your company is "abcompany" and you register as abcompany.com. Then phishers can register abcompany.net, abcompany.org, abcompany.biz and they can use it for fraudulent purpose.

Typosquatting, also called URL hijacking, is a form of cybersquatting which relies on mistakes such as typographical errors made by Internet users when inputting a website address into a web browser or based on typographical errors that are hard to notice while quick reading. URLs which are created with Typosquatting looks like a trusted domain. A user may accidentally enter an incorrect website address or click a link which looks like a trusted domain, and in this way, they may visit an alternative website owned by a phisher.

A famous example of Typosquatting is **goggle.com**, an extremely dangerous website. Another similar thing is **yutube.com**, which is similar to **goggle.com** except it targets **Youtube users**. Similarly, **www.airfrance.com** has been typosquatted as **www.arifrance.com**, diverting users to a website peddling discount travel. Some other examples; **paywpal.com**, **microroft.com**, **applle.com**, **appie.com**.

IV. FEATURES USED FOR PHISHING DOMAIN DETECTION

There are a lot of algorithms and a wide variety of data types for phishing detection in the academic literature and commercial products. A phishing URL and the corresponding page have several features which can be differentiated from a malicious URL. For example; an attacker can register long and confusing domain to hide the actual domain name **(**Cybersquatting, Typosquatting). In some cases attackers can use direct IP addresses instead of using the domain name. This type of event is out of our scope, but it can be used for the same purpose. Attackers can also use short domain names which are irrelevant to legitimate brand names and don't have any FreeUrl addition. But these type of web sites are also out of our scope, because they are more relevant to fraudulent domains instead of phishing domains.

Beside URL-Based Features, different kinds of features which are used in machine learning algorithms in the detection process of academic studies are used. Features collected from academic studies for the phishing domain detection with machine learning techniques are grouped as given below.

1. URL-Based Features
2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features

1.URL-Based Features

URL is the first thing to analyse a website to decide whether it is a phishing or not. As we mentioned

before, URLs of phishing domains have some distinctive points. Features which are related to these points are obtained when the URL is processed. Some of URL-Based Features are given below.

- Digit count in the URL
- Total length of URL
- Checking whether the URL is Typosquatted or not. (google.com → goggle.com)
- Checking whether it includes a legitimate brand name or not (apple-icloud-login.com)
- Number of subdomains in URL
- Is Top Level Domain (TLD) one of the commonly used one?

2.Domain-Based Features

The purpose of Phishing Domain Detection is detecting phishing domain names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us. Some useful Domain-Based Features are given below.

- Its domain name or its IP address in blacklists of well-known reputation services?
- How many days passed since the domain was registered?
- Is the registrant name hidden?

3.Page-Based Features

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give information about how much reliable a web site is. Some of Page-Based Features are given below.

- Global Pagerank
- Country Pagerank
- Position at the Alexa Top 1 Million Site

Some Page-Based Features give us information about user activity on target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.

- Estimated Number of Visits for the domain on a daily, weekly, or monthly basis
- Average Pageviews per visit
- Average Visit Duration
- Web traffic share per country
- Count of reference from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

4.Content-Based Features

Obtaining these types of features requires active scan to target domain. Page contents are processed for us to detect whether target domain is used for phishing or not. Some processed information about pages are given below.

- Page Titles
- Meta Tags
- Hidden Text
- Text in the Body
- Images etc.

By analysing these information, we can gather information such as;

- Is it required to login to website
- Website category
- Information about audience profile etc.

All of features explained above are useful for phishing domain detection. In some cases, it may not be useful to use some of these, so there are some limitations for using these features. For example, it may not be logical to use some of the features such as Content-Based Features for the developing fast detection mechanism which is able to analyze the number of domains between 100.000 and 200.000. Another example would be, if we want to analyze new registered domains Page-Based Features is not very useful. Therefore, the features that will be used by the detection mechanism depends on the purpose of the detection mechanism.

V. PHISHING WEBSITE DETECTION

Detecting Phishing Domains is a classification problem, so it means we need labeled data which has samples as phish domains and legitimate domains in the training phase. The dataset which will be used in the training phase is a very important point to build successful detection mechanism. We have to use samples whose classes are precisely known. So it means, the samples which are labeled as phishing must be absolutely detected as phishing. Likewise the samples which are labeled as legitimate must be absolutely detected as legitimate. Otherwise, the system will not work correctly if we use samples that we are not sure about.

A. DECISION TREE

Initially, as we mentioned above, phishing domain is one of the classification problem. So, this means we need labeled instances to build detection mechanism. In this problem we have two classes: **(1)** phishing and **(2)** legitimate.

When we calculate the features that we've selected our needs and purposes, our dataset looks like in figure below. In our examples, we selected 12 features, and we calculated them. Thus we generated a dataset which will be used in training phase of machine learning algorithm.

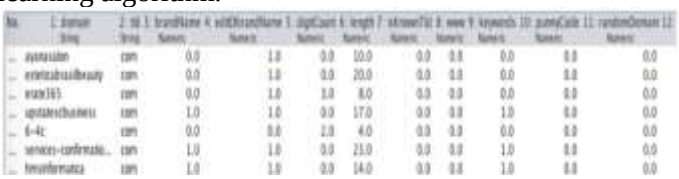

Fig.5.1. training data

A Decision Tree can be considered as an improved nested-if-else structure. Each features will be checked one by one. An example tree model is given below.

Generating a tree is the main structure of detection mechanism. Yellow and elliptical shaped ones represent features and these are called nodes. Green and angular ones represent classes and these are called leaves. The length is checked when an example arrives and then the other features are checked according to the result. When the journey of the samples is completed, the class that a sample belongs to will become clear.

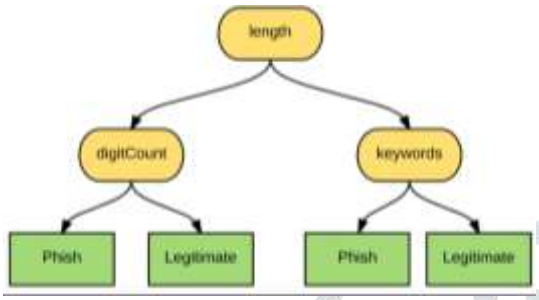

Fig. 5.2 the main structure of detection mechanism

Decision Tree uses a information gain measure which indicates how well a given feature separates the training examples according to their target classification. The name of the method is **Information Gain**. The mathematical equation of information gain method is given below.

$$Gain(S,A) = \underbrace{Entropy(S)}_{\text{original entropy of S}} - \underbrace{\sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)}_{\text{relative entropy of S}}$$

High Gain score means that the feature has a high distinguishing ability. Because of this, the feature which has maximum gain score is selected as the root. **Entropy** is a statistical measure from information theory that characterizes (im-)purity of an arbitrary collection S of examples. The mathematical equation of Entropy is given below.

$$H(S) \equiv \sum_{i=1}^{n} -p_i \log_2 p_i$$

Original Entropy is a constant value, Relative Entropy is changeable. Low Relative Entropy Score means high purity, likewise high Relative Entropy Score means low purity. As we move down the tree, we want to increase the purity, because high purity on the leaf implies high success rate.

In the training phase, dataset is divided into two parts by comparing the feature values. In our example we have 14 samples. "+" sign representing phishing class, and "-" sign representing legitimate class. We divided these samples into two parts according to the length feature. Seven of them settle right, the other seven of them settle left. As shown in the figure below, right part of tree has high purity, so it means low Entropy Score (E), likewise left part of tree has low purity and high Entropy Score (E). All calculations were done according to the equations given above. Information Gain Score about the length feature is 0,151.

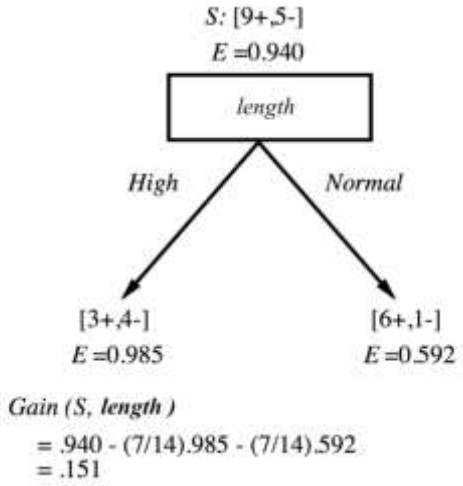

Fig.5.3 calculating Gain

The Decision Tree Algorithm calculates this information for every feature and selects features with maximum Gain scores. To growth the tree, leaves are changed as a node which represents a feature. As the tree grows downwards, all leaves will have high purity. When the tree is big enough, the training process is completed.

The Tree created by selecting the most distinguishing features represents model structure for our detection mechanism. Creating mechanism which has high success rate depends on training dataset. For the generalization of system success, the training set must be consisted of a wide variety of samples taken from a wide variety of data sources. Otherwise, our system may working with high success rate on our dataset, but it can not work successfully on real world data.

B. SVM(Support Vector Machine)

In this paper, a novel method is proposed to detect phishing URL based on SVM. The content representation of proposed system is divided into two categories.

Textual content: "Textual content" in this paper is defined as the terms or words that appear in a given web page, except for the stop words. We first separate the main text content from HTML tags and apply stemming to each word. The function of stemming process is to find out stems rather original world. For ex., 'work', 'works', and 'working' are stemmed into 'work' and considered as the same word.

Visual content: "Visual content" concern to the features with respect to the block regions, layout, overall style including the logos, images and forms. Visual content also can be further specified to the color of the web page background, the font style, the

locations of images, the font size and logos, etc. Moreover the visual content is also user-dependent. On the other hand, let's consider the pixel level web page, whereas an image that enables the total representation of the visual content of the web page.

The phishing approach used in this paper shown in Figure contains the following components.

1. A text classifier using the SVM rules to handle the text content extracted from a given web page.
2. An image classifier using the SVM similarity assessment to handle the pixel level content of a given web page that has been transformed into an image.
3. A SVM approach to estimate the threshold used in classifiers through offline training.
4. A data fusion algorithm to combine the results from the image classifier and the text classifier. The algorithm employs the SVM approach as well.

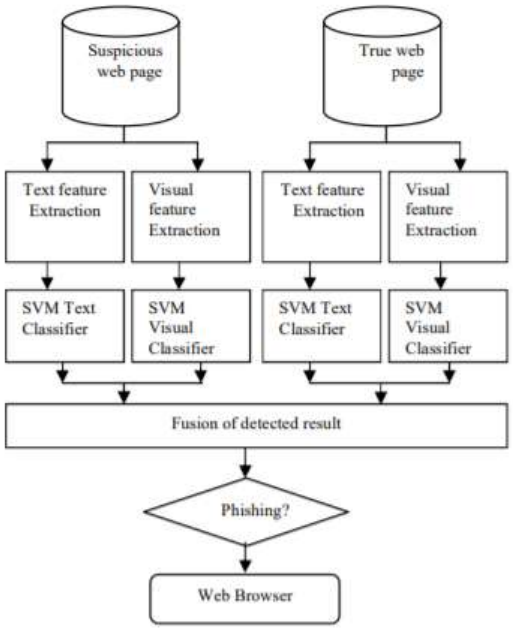

Fig 5.4. Architecture Design for phishing web page detection system

The system includes a training section, which is to estimate the statistics of historical data, and a testing section, which is to examine the incoming testing web pages. The statistics of the web page training set consists of the probabilities that a textual web page belongs to the categories, the matching thresholds of classifiers, and the posterior probability of data fusion. Through the preprocessing, content representations has been done, i.e., visual and textual, are continuously extracted from a given testing web page.

The text classifier is used to classify the given web page into the corresponding category based on the textual features. The image classifier is used to classify the incoming web page into the relevant category based on the visual content. Then the fusion algorithm combines the detection results generated by the two classifiers. The detection results are eventually transmitted to the online users or the web browsers. In preprocessing, first step is to separate HTML tags from the main contexts of an incoming web page. We construct a word vocabulary To form a histogram vector for each web page. This system extracts all the words from a given protected web page and applies stemming to each word. The SVM word-based extraction delivers more discriminative information than stemming-based extraction. But point out that the SVM word-based extraction will largely increase the vocabulary size. In addition, using stemming will build more robustness of detection, because phishers may manipulate the textual content through the change of tense and active to passive.

While stemming for smaller vocabulary detection and robust detection size, to identify similar textual content, we suggest word-based extraction using the SVM. Given a web page, where each component represents the term frequency and n denotes the total number of components in a histogram vector. We explain three points here.

1. We do not extract words from all the web pages in a dataset to construct the vocabulary because phishers use the text from a targeted web page to scam users.
2. For simplicity, we do not use any feature extraction algorithms in the process of vocabulary construction.
3. We do not take the similar web pages into account because the sizes of most phishing web pages are small. In reality, using only text content is insufficient to detect phishing web pages.

This technique usually leads to high FP (false positives), because phishing web pages are mostly similar to the targeted web pages not only in textual content but also in visual content such as layout, logos, and style. In this system, we use the same approach as in using the SVM to measure the visual similarity between an input web page and a secured web page. Firstly, we retrieve the vulnerable web pages and secure web pages from the web. Second, we generate signatures of input webpage, which are used for the calculation of the SVM between them. Each and every web page images are normalized into fixed-size box images. We use these normalized images to generate the signature of each web page. The image classifier is

implemented by setting a threshold, which is later estimated in the subsequent section. If the visual similarity between a input web page and the secure web page exceeds the threshold, it means the web page is classified as phishing.

VI. CONCLUSION

Phishing Website detection plays a critical role for many cyber security applications, and clearly machine learning approaches are a promising direction. In this paper, we discussed detection of Phishing web using supervised machine learning techniques such as Decision tree and Support Vector Machine. The decision tree algorithm generates the rules on the basis of available data. But the number of rule is in large quantity which increases the comparison and detection time. So the support vector machine (SVM) and improved logistic regression is proposed to detect phishing website.

VII. REFERENCES

1. DOYEN SAHOO, CHENGHAO LIU and STEVEN C.H. HOI, "Malicious URL Detection using Machine Learning: A Survey", Vol. 1, No. 1, Article . Publication date: August 2019.
2. GENTEM VARAPRASAD, "A study on Essential of Machine Learning to society and Its types", International Journal of Research and Analytical Reviews (IJRAR), January 2020, Volume 7, Issue 1, pp. 11-16, 2020.
3. Meenu , Sunila godara, "Phishing Detection using Machine Learning Techniques", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, pp. 3821-3829, 2019.
4. M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
5. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.
6. K. Shima et al., "Classification of URL bitstreams using bag of bytes," in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5. 7. A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.
7. A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.