Research Article

A Novel Framework For Malicious URL Detection Using Hybrid Model

Srisailapu D Vara Prasad[1], Dr. K. Rajasekhara Rao[2]

[1]Research Scholar, Department of Computer Science & Engineering, Dr. Y.S.R. ANU College of Engineering & Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh-522510, India.
& Assistant Professor, Department of Computer Science & Engineering, School of Technology, Gandhi Institute of Technology and Management (Deemed to be University), Hyderabad, Telangana-502329, India.
Email: prasad.phdanu@gmail.com. Orcid Id: 0000-0002-5112-5909

[2]Professor, Department of Computer Science & Engineering & Director URCE, Vijayawada, Andhra Pradesh-521109, India. Email: krr_it@yahoo.co.in

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: Web searching is a process where a software program searches the database and gathers information related to the specified terms. Search is done with the help of search engines like Google, Bing, etc. Generally, the search engines offer search engines through which users are allowed to search for the required content via World Wide Web (WWW). The WWW includes massive amounts of data sites where it gets difficult to obtain the relevant data. These days, web users experience problems with information overload and sink because the amount of information and number of users has grown significantly and rapidly. In general, clients can enter keywords in the search engines to get the appropriate data. The search engines return a list of URLs as a result in a ranked order. Clients will most often choose the first link as the most relevant link for the requested data. However, sometimes it so happens that users may click the top-ranked URLs, which may not be the legitimate URL, and by doing so, users' data will be stolen by third-party attackers. Many scientific studies show that the number of methods is based on machine learning to detect malicious URLs. In this paper, we have proposed a hybrid model to see the given URL is a phishing URL or not.

Keywords: Phishing URL, Hybrid model, search engines, legitimate URL

1. Introduction

Nowadays, malicious URLs are a common threat to institutions, easy-going networks, and net banking. Existing systems have focused on two-fold prominence, i.e., either URL is friendly or malicious. Close-to no composting is solved that revolves across the popularity of malicious URLs along with their assault forms. This way, it becomes essential to comprehend the strike kind and typify a possible countermeasure. It proposes a strategy to describe your website page determined by URL, content and visible characteristics of their WebPages are executing the staggered classifier to dictate your website pages. We propose a Hybrid model (3.1, 3.2, 3.3) to check the given URL is legitimate URL or not. Within this work, we propose 30 fresh brand features of spam, antivirus, and malware URLs. The dual and multi-class data set is manufactured utilizing 49935 malicious and kindhearted URLs. We actualize a demo model to check the URL, whether it is hurtful, malware, or phished when the URL properties are reached the any of the groupings; we send a programmed alarm to the analyzer for better improvement of the pursuit.

With the drastically quick and dangerous development of information out there over the net, World Wide Web has become a vigorous stage to store, scatter and recover information likewise to mine supportive details. Inferable from the properties of the vast, differing, dynamic, and unstructured nature of web information, web information examination has experienced many difficulties, as quantifiable, mixed media framework and fleeting issues, and so on. Subsequently, web clients are everlastingly suffocating in a "sea" of data and confronting the matter of data over-burden once associating with the Web. Regularly, the following issues are typically brought up in web-related investigations and applications.

To discover detailed data on the Web, clients commonly either peruse web records straightforwardly or utilize a web crawler to get the ideal data. When a client uses a web crawler to discover data, the individual in question ordinarily enters one or numerous watchwords as an inquiry; at that point, the program restores a posting of reviewed pages identified with the given query. As it may, there usually are two significant contemplations identified with the inquiry-based web search. The essential drawback is low exactitude, which is brought about by many inapplicable pages got by the internet searcher. The following drawback is an intense review attributed to the deficiency of ordering all pages possible on the Web. It causes the issue in finding the un-indexed data that is genuinely pertinent. Step-by-step instructions to understand extra pages relevant to the question, along these lines, are transforming into a popular theme in web information the executives in a decade ago. Finding required data: Most web indexes act in

an exceedingly question-activated methodology principally on a premise of a catchphrase or numerous watchwords entered. Commonly the outcomes given back by the web index don't explicitly coordinate what a client needs in light of the very certainty of the presence of the similitude. In elective words, the semantics of web information is never contemplated inside the setting of web search.

2. Related work

This article is focused on the analysis and study of their malicious URLs and proposes methodologies for discovering the malicious URLs of the web applications. The rise in internet usage has brought dramatic changes in commerce and communicating. The rise of the Web offers new opportunities and challenges for the industry, government, e-commerce, healthcare, education, and the public. At the moment, the user ratio of web programs accomplishes an enormous increase in speed. Individual life is dependent upon the Web.

Nevertheless, the malicious strike spoils online usage. Academic, financial, and e-commerce organizations are severely affected by these attacks. The spiteful sites are accessed through malicious URLs. The principal problem with an internet application is detecting malicious URLs. A summary of URL detection techniques, static approaches, dynamic approaches, hybrid approaches, the scope of the issue, diagnosis of classification methods, data analysis, and challenges in detection methods are addressed in this section.

Gupta, Renu, et al. (2016) world is held with massive data, and searching is the most ordinary task on the Web. As the data accessible on Web is expanding, it is hard to gain pertinent data on Web. The client arrives at a query for recovering obligatory data from www, and a considerable number of web pages are obtained. These website pages or search lists contain relevant pages and irrelevant exploration consequences in response to a client's query. For this matter, an effective Page Ranking algorithm is required. Google utilizes an exceptionally fundamental algorithm called the Page Rank algorithm, which uses web structure mining and has a few drawbacks. This paper examines various improvements of Page Rank, which utilizes web content mining for efficient ranking. Relative qualities and restrictions of a few algorithms are investigated to discover the extent of research.

Caraballo, Willis E. Polanco (2013)the quickened development of Social Networks has permitted electronic commerce (e-commerce) solutions for increment considerably more and turns out to be more prevalent. The outstanding criteria for these answers for being effective are the ease of use, the availability, and the content relevance of the Websites joined with the data produced inside the social networks. This theory's motivation is to perform research about the combination of Web mining advancements with e-commerce solutions arrangements in social networks systems as a potential engine for enhancing sales. The composition of this study is as per the following. The initial segment depicts the diverse stages of Web mining and its pertinence for growing better Web based business arrangements from the informal communities' point of view. The second part describes the commitments of Web mining to the grouping of the information in the Web and how it executes multi-classifier techniques. The third piece of this examination introduces customization in e-commerce solutions to address the diversity of the customer's preferences acquired in the social networks systems.

Due to the advantage of new communication technologies, online-banking, e-commerce, and social networking applications, including businesses, promote growth. The use of the World Wide Web is growing day to day basis. By using the Internet, most of the time, malicious software, known as malware or attacks, will be promoted. More than half the world's population to the Internet to facilitate the distribution of malicious content on the Web due to the bad actors has become a standard technology. New security threats, rapid changes in technology, and new IT security experts, due to a lack of significant limitations of traditional security management techniques, are becoming more serious. Identification of the malicious URL in the hot area of information security always is an important task. Drive-by downloads phishing and social engineering, and spam using the most popular attacks poisoned URL. Malware download by visiting the URL is called drive-through-down. A drive-by-down attack is usually made by exploiting plug-in attacks or by inserting malicious code into JavaScript. Theft and social engineering attacks protect users from sensitive information that may affect actual web pages. Phishing can be used for advertising or spam messages charity.

Every year, many of the problems that lead to this kind of attack. So the primary concern promptly to detect such a malicious URL. In this chapter, we mainly discuss the various methods proposed by different authors to detect malicious URLs. In this chapter, the other techniques in machine learning used to see adversarial URL focus on purpose. The characteristic feature of the representation of the stage classification of machine learning methods is described in detail. Types of the various elements of machine learning techniques are well represented in the extraction phase. The paper gives a thorough description of the methods used to identify a wide variety of machine learning.

3. Proposed methodology

We propose a Hybrid model to break down the web URL. Proposed model is arranged into e steps. Basically URLs and the kind of physical attack dependent on multi-class characterization. In this work, we propose 30 recent features of spam, phishing, and malware URLs. The double and multi-class dataset is developed utilizing 49935 malicious and kindhearted URLs. It comprises 26041 friendly and 23894 malicious URLs containing 11297 malware, 8976 phishing, and 3621 spam URLs. We actualize a demo model to check the URL, whether cruel, malware or phished. When the URL properties are reached the any of the groupings, we send a programmed alarm to the analyzer to improve the pursuit. In the wake of breaking down the URL, it will secretive looked through URL substance to image. In the wake of changing over the image, we separate the highlights and applying the perceptual image hashing framework utilizing which we can distinguish bonafide and non-true website pages using prepared and test datasets and remove image choices dependent on four phases as they are the transition phase, the feature extraction phase, the size of the stage and the stage of compression and encryption. In the Transformation stage, the info image experience uncommon or potentially recurrence change to make all extricated choices depend either on the estimations of image pixels or on the image recurrence coefficients. The information extraction stage of extracting the image from the image to highlight options and obtain an uninterrupted hash vector allows the seizure of the image hashing framework. When the outbreak of the consistent vector hash, the hash it isolated vector quantization step is calculated. Presently the discrete hash vector is changed over to parallel perceptual hash string in the third stage. Finally, the reduced horizontal eclipse string hash, compression, and encryption at the stage briefly and scrambled hash of the last seizure. Exploratory outcomes execution and improving the discovery rate exactness by contrasting the current models like AI models, such as SVM and RFC.

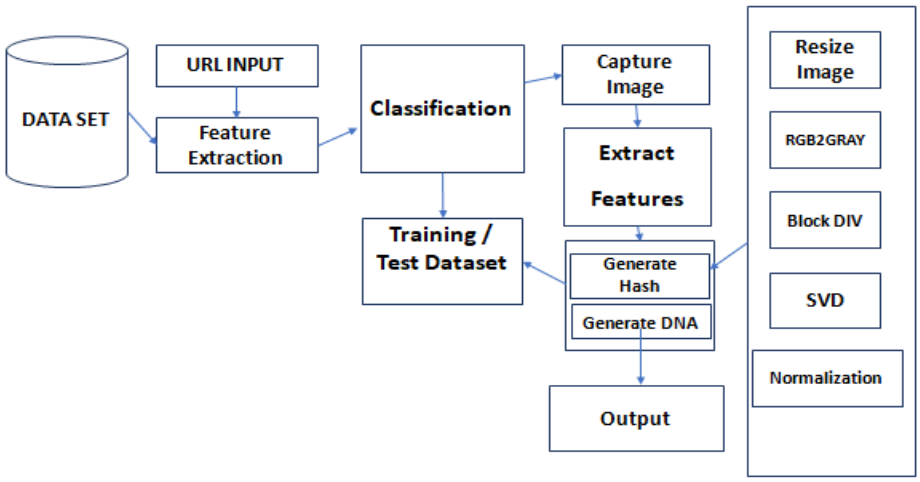

Figure 1: Proposed Hybrid model

3.1 Extract the features from the given URL and classify using random forest algorithm

URL is the first thing to analyze 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 URLBased Features are given below.

having_IP_Address
URL_Length
Shortining_Service
having_At_Symbol
double_slash_redirecting
Prefix_Suffix
having_Sub_Domain
SSLfinal_State
Domain_registeration_length
Favicon
port
HTTPS_token
Request_URL
URL_of_Anchor

Links_in_tags
SFH
Submitting_to_email
Abnormal_URL
Redirect
on_mouseover
RightClick
popUpWidnow
Iframe
age_of_domain
DNSRecord
web_traffic
Page_Rank
Google_Index
Links_pointing_to_pageStatistical_report

Most of the existing works used only structural URL features, but in our work, we are also adding content and visual features for more accurate classification. In order to classify the URL we have implemented Random Forest algorithm which is given below

Algorithm 1 Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
function RandomForest(S, F)
    H ← ∅
    for i ∈ 1, ..., B do
        S^(i) ← A bootstrap sample from S
        h_i ← RandomizedTreeLearn(S^(i), F)
        H ← H ∪ {h_i}
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    At each node:
        f ← very small subset of F
        Split on best feature in f
    return The learned tree
end function
```

Step 1: First, start with the selection of random samples from a given dataset.
Step 2: Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
Step 3: In this step, voting will be performed for every predicted result.
Step 4: At last, select the most voted prediction result as the final prediction result.

Some Factors/properties considered for classification a) Using the IP Address b) Long URL to Hide the Suspicious Part c) Using URL Shortening Services "TinyURL" d) Adding Prefix or Suffix Separated by (-) to the Domain e) Sub Domain and Multi Sub Domains f) HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer) g) Domain Registration Length.

3.2 Capturing the image of the web page by visiting the URL and Generation of perceptual hash for the URL page: Content and visual features include colors, shapes, different orientations, and text. To extract visual elements, we are capturing the webpage by visiting the URL and generate a perceptual hash value.Image hashing or perceptual hashing is the process of examining the contents of an image.Constructing a hash value that uniquely identifies an input image based on the contents of an image. The pipeline consists of the following steps perceptual image hashing system. The transitional phase, the feature extraction phase, the sizing phase, and phase compression and

encryption. In the transformation phase, the input image is unique, and frequency transition experiences as all collected options create either image pixel values or image frequency coefficients. The feature extraction phase of the input image from the image capture options and continuous hash vector allows the system to obtain a perceptual image hashing.

Perceptual Hash Algorithm:

Convert the image to grayscale.
Resize the image into 32x32.
Apply DCT which results in the high frequency pixels to be in the upper left corner.
Crop the upper 8x8 pixels.
Compute the median of all the pixel values.
Compute the bits: Each bit is simply set based on whether the color value is above or below the median.
Construct the hash: Set the 64-bits in to a 64-bit integer.
Calculate the hamming distance between stored image and URL image.

```
phash_row_col(image,size=8):
width=32, height=32
image=image.resize(width, height)
grays = get_grays(image, width, width)
RoI=DCT(grays)
RoI=RoI.crop(size,size)
med=median(RoI)
for y in range(size):
for x in range(size):
if RoI[x][y] > median:
k=1
Hash=Hash<<1|k
else
k=0
Hash=Hash<<1|k
return Hash
```

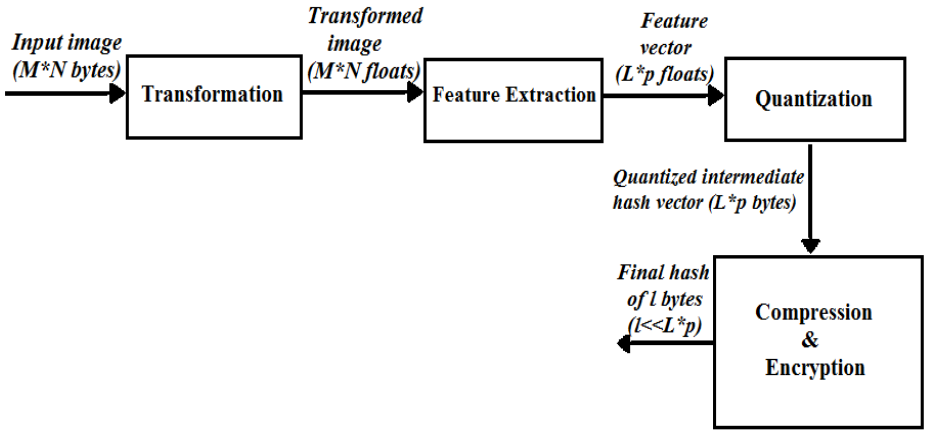

Figure 2: Pipeline Stages of perceptual Image Hashing System

Convert the image to grayscale. Resize the image into 32x32. Apply DCT, which results in the high-frequency pixels being in the upper left corner. Crop the upper 8x8 pixels. Compute the median of all the pixel values. Calculate bits, whether above or below the intermediate value of each bit, will be set based on it. Build a hash, a set of 64-bit to 64-bit integers. Calculate the distance between the image and the URL hamming stored image.

3.3 Developing a DNA sequence for web pages for authentication purpose

A compelling strategy for DNA sequencing algorithm is proposed in the area of Image Cryptography. DNA sequencing in an image processing approach includes background correction of images, resolution enhancing, base detection, and sequence interpretation. The idea is changed over to DNA code and partitions into sections. The fragments are arranged in a database that acts as a cipher. The extracted features are provided with a signature generated using the DNA sequencing technique. This is done by considering natural DNA sequences as the primary

keys. The proposed methodology introduces an approach to storing and retrieving DNA sequences. A perceptual hash function is generated to identify the similarity among the digital images adapted for DNA sequences.

Algorithm DNA Sequence Generation:

1. Input the plain image P(m,n) where m,n are image dimensions of rows and columns, respectively.
2. Generate the key sequence K
3. Splitting the RGB image into R, G, B components, and transform the decomposed matrixes of R, G, B to binary matrices R(m, n ×8), G(m, n×8) and B(m, n×8), then encode respectively in accordance with the chosen DNA encoding rule ruleenc, Table 2[4], and get three DNA sequence matrices Pr(m, n×4), Pg(m, n × 4) and Pb(m, n × 4).
4. Transform K to binary sequence Kb, then generate the matrix Mk (m, n×8) by repeating Kb, t times, where t = m×n×8/32 .Encode Mk with the same encoding rule and get Mke.
5. According to DNA XOR operation, Table 1[4], do: Pr' = Pr XOR Mke, Pg' =Pg XOR Mke and Pb' = Pb XOR Mke.
6. Join encoded Pr', Pg' and Pb' to get encoded RGB.
7. According to DNA XOR operation, do: XOR column wise on encoded RGB to get DNA sequence.

Table 1: One type of XOR operation for DNA Sequence

XOR	A	G	C	T
A	A	G	C	T
G	G	A	T	C
C	C	T	A	G
T	T	C	G	A

Table 2: Eight types of DNA map rules

	1	2	3	4	5	6	7	8
A	00	00	01	01	10	10	11	11
G	11	11	10	10	01	01	00	00
C	11	10	00	11	00	11	01	10
T	10	01	11	00	11	00	10	01

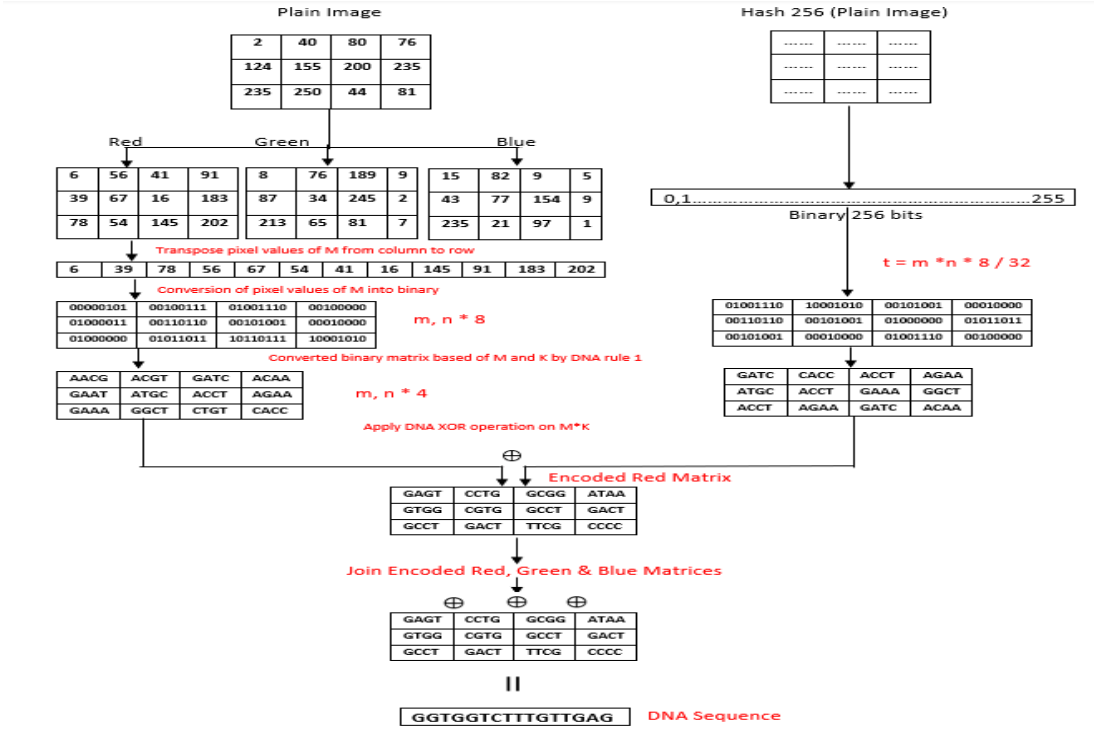

Figure 3: DNA Sequence generation

The experimental results are obtained as follows which shows on comparing the legitimate DNA sequence and search sequence the result for the required web page is positive if they are matched if not the result is unmatched. This can be identified in the following table as per the experimental tests performed.

Table 3: Test Results using DNA sequence for web pages

Sno	Trained legitimate Domain DNA	Search Sequence	Match	Result
1	CGAGACAGCGA GCATATGCAGGA AGCGGCAGGAA TAAGGAAAAGC AGC	CGAGACAGCGAGCA TATGCAGGAAGCGG CAGGAATAAGGAAA AGCAGC	Yes	Match
2	AAGCTCGGGAG GTGGCCAGGCGG CAGGAAGGCGC ACCCCCCCAGCA ATCCGCGCGCCG GGACAGAATGCC	CTCCTGACTTTCCTC GCTTGGTGGTTTGA GTGGACCTCCCAGG CCAGTGCCGGGCCC CTCATAGGAGAGG	No	No Match
3	CTGCAGGAACTT CTTCTGGAAGAC CTTCTCCTCCTG CAAATAAAACCT CACCCATGAATG CTCACGCAAG	CTGCAGGAACTTCT TCTGGAAGACCTTC TCCTCCTGCAAATA AAACCTCACCCATG AATGCTCACGCAAG	Yes	Match

4. Performance Evaluation and Results

In this work, we have spam, phishing, and malware feature 30 features of URL used. 49935 URL using malicious and benign, double, and multi-class status has been developed. It contains malware 11297, 8976, and 3621 fishing-friendly spam URL containing 26041 and 23894 contain the URL is malicious. We actualize a demo model to check the URL, whether evil, malware, or phishing. After verifying URL, Content, and Visual features, we send a programmed alarm to the analyzer to improve the pursuit. Dataset used is Kaggle Phishing Dataset, Phishing URLs from Phishtank; we have implemented proposed model and verified the results. We compared proposed model with existing algorithms like SVM and LR and proved that our model gave better results than them.

Table 4: Performance of the classifiers

Algorithm	**Accuracy**	**Precision**	**F1**
SVM	88.5	88.5	88.7
LR	90.1	90.1	90
HYBRID MODEL	95.8	95.6	95.6

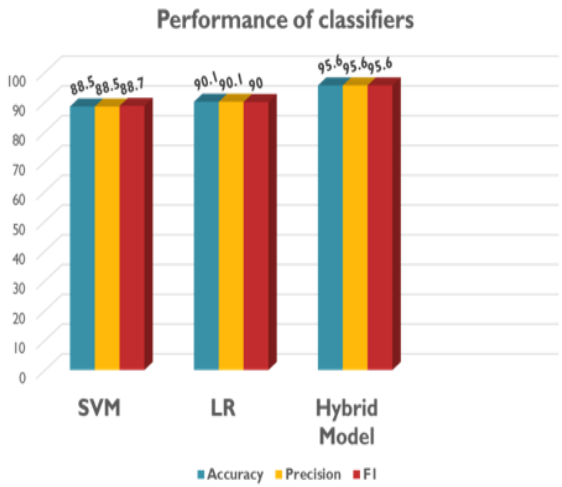

Figure 4: Performance of the classifiers for a hybrid model

In this work, we proposed a hybrid model for improving the accuracy in detecting Phishing URLs. It involves verifying URL features and using a perceptual image hashing system to identify authentic and non-authentic web pages using trained and test datasets.

This work provides a hybrid model to classify the web pages and improves the search engine efficiency to generate DNA sequences for a web page layout. A snapshot of the web page is extracted for which the DNA sequence is generated. The feature extraction is done using the SURF algorithm, which gives the best compression results by utilizing the hessian matrix, detection, and localization to discover the exciting points of a given image and extract visual features efficiently. The experimental results are obtained accurately based on the tests performed. One could get the desired web page by determining the matched and unmatched products acquired by comparing DNA sequence with search sequence. The work also involves generating a DNA sequence for a web page layout. The proposed hybrid model, which analyzes URL, Content, Visual features, produced better results than existing classifier algorithms like SVM, LR, etc.

5. Conclusion

This work provides a hybrid model to classify the web pages and improves the search engine efficiency to generate DNA sequences for a web page layout. A snapshot of the web page is extracted for which the DNA sequence is generated. The experimental results are obtained accurately based on the tests performed. One could get the desired web page by determining the matched and unmatched products acquired by comparing DNA sequence with search sequence.

Despite the tremendous progress achieved over the past few years, extensive studies and using machine learning to detect malicious URLs remains the most challenging open issue. The feature extraction and representation of the future directions for more effective learning, predictive train models, are more powerful machine learning algorithms. In this work, we have improved the accuracy of detecting phishing URLs. In the future, we want to develop an add-on or plug-in software to a web browser that will check the URLs clicked by the user and give an alert message by verifying URL, Content, and Visual Features.

6. References

1. Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi, —"Malicious URL Detection using Machine Learning": A Survey‖, IEEE, 2017
2. Dharmaraj Rajaram Patil and JB Patil. 2015. "Survey on Malicious Web Pages Detection Techniques." International Journal of u-and e-Service, Science and Technology (2015)
3. LLC OpenDNS. 2016. PhishTank: "An anti-phishing site." Online: https://www,PhishTank. Com (2016).
4. R. Guesmi, M. A. B. Farah, A. Kachori, M. Samet "A novel chaos-based image encryption using DNA sequence operation and Secure Hash Algorithm SHA-2," 2016.
5. Rose, Daniel E., and Danny Levinson. "Understanding user goals in web search." proceedings of the 13th international conference on World Wide Web, ACM, 2004

6. A.Naga Venkata Sunil, Anjali Sardana “A PageRank Based Detection Technique for Phishing Web Sites" IEEE Symposium on Computers and Informatics, 2012.
7. Gaurav Varshney, Manoj Misra and Pradeep K. Atrey “A survey and classification of web phishing detection schemes," SECURITY AND COMMUNICATION NETWORKS, Security Comm. Networks 2016.
8. Michael Bendersky, W. Bruce Croft, YanleiDiao “Quality-Biased Ranking of Web Documents”ACM, 2011
9. David G. Lowe “Distinctive Image Features from Scale-Invariant Keypoints," 2004
10. Brin, Sergey, and Lawrence Page. “Reprint of The anatomy of a large-scale hypertextual web search engine.” Computer networks 56.18 (2012): 3825-3833.
11. Broder, Andrei. “Taxonomy of web search.”,ACM Sigur forum. Vol. 36. No. 2 ACM, 2002
12. Silverstein, Craig, et al. “Analysis of a huge web search engine query log.”, ACM SIGIR Forum. Vol. 33. No. 1. ACM, 1999
13. Zhang, Qingjiu, and Shiliang Sun. “Multiple-view multiple-learner active learning.” Pattern Recognition 43.9 (2010): 3113-3119.
14. Taskar, Ben, Pieter Abbeel, and Daphne Koller. “Discriminative probabilistic models for relational data.” Proceedings of the Eighteenth Conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 2002
15. Hanneke, Steve. “Activated learning: Transforming passive to active with improved label complexity.” The Journal of Machine Learning Research 13.1 (2012): 1469-1587.