iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33957

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Efficient Phishing Website Detection using Machine Learning Algorithm

Sinduja. S[1], Monisha. S[2], Priya Dharshini. S[3], Sneha. K[4], Vaishnavi. R[5]
[1]*Assistant professor, Department of Computer Science and Engineering, Vivekananda College of Engineering For Women, Namakkal, Tamilnadu.*
[2, 3, 4, 5]*Department of Computer Science and Engineering, Vivekananda College of Engineering For Women*

Abstract: Phishing is defined because the artwork of echoing a internet site of a creditable company proceeding to seize user's private facts along with usernames, passwords and social safety number. Phishing web sites contain loads of cues inside its content-elements in addition to the browser-primarily based totally safety signs supplied alongside the internet site. Several answers had been proposed to address phishing. Nevertheless, there may be no specific magic bullet which can clear up this danger radically. Machine Learning is efficient approach to come across phishing. It additionally eliminates disadvantage of current approach. The proposed technique predicts the URL primarily based totally phishing assaults primarily based totally on functions and additionally offers most accuracy. This technique makes use of uniform resource locator (URL) functions. We recognized functions that phishing web website online URLs contain. The proposed technique employs the ones functions for phishing detection. The proposed machine predicts the URL primarily based totally phishing assaults with most accuracy. Random forest algorithm is used for efficient and accurate classification of phishing web sites and achieves higher end result as compared to current methods.

Keywords: phishing website detection, Random forest, URL based detection, machine learning.

I. INTRODUCTION

Phishing imitates the characteristics and alternatives of emails and makes it seem similar thanks to the very fact the initial one. It looks nearly like that of the legitimate offer. the buyer thinks that this e-mail has return from a true leader or a company. This makes the buyer to forcefully visit the phishing site via the hyperlinks given within the phishing email. This phishing internet sites region unit created to mock the seams of a resourceful web site. The phishes force person to inventory up the private data via giving baleful messages or validate account messages etc. in order that they inventory up the well-liked information which could be utilised by them to misuse it. They devise things like the user isn't left with the opposite selection however to travel to their spoofed internet site. Phishing is that the most venturesome criminal physical activities within the cyber region. Since the utmost of the purchasers logs on to urge admission to the services furnished with the help of presidency and money institutions, there has been a big boom in phishing attacks for the on the far side few years. Phishers commenced to earn money which they struggle this as a thriving business. Some possible solutions to combat phishing were created, as well as specific legislation and technologies. From a technical purpose of read, the detection of phishing usually includes the subsequent categories: detection supported a black list and white list, detection supported Uniform Resource locater (URL) options, detection supported online page, and detection supported machine learning. The anti-phishing manner victimisation blacklist could also be a simple manner, however it cannot notice new phishing websites. The detection on URL is to research the options of URL. The URL of phishing websites could also be terribly just like real websites to the human eye, however they're completely different in information processing. The content-based detection sometimes refers to the detection of phishing sites through the pages of parts, like type data, field names, and resource reference. Uniform Resource Locator (URL) is created to address web pages. The figure below shows relevant parts in the structure of a typical URL.

Figure 1: Characteristics of Phishing Domains

It begins with a protocol wont to access the page. The absolutely qualified name identifies the server United Nations agency hosts the online page. It consists of a registered name (second-level domain) and suffix that we have a tendency to check with as superior domain (TLD). The name portion is forced since it's to be registered with a website name Registrar. a bunch name consists of a subdomain name and a website name. associate degree phisher has full management over the subdomain parts and may set any price thereto. The universal resource locator can also have a path and file elements that, too, is modified by the phisher at can. The subdomain name and path ar absolutely governable by the phisher.

In this paper, we'll target the detection model employing a machine learing framework. the most contributions ar as follows:

1) We gift 2 feature sorts for internet phishing detection: a resourceful feature associate degreed an interaction feature. The first feature is that the direct feature of universal resource locator, as well as special characters in universal resource locator and age of the domain. The interacting feature is that the interaction between websites, as well as in-degree and out-degree of universal resource locator.

2) We introduce DBN to discover internet phishing. we have a tendency to discuss the coaching method of DBN and acquire the acceptable parameters to discover internet phishing.

II. LITERATURE SURVEY

Vahid Shahrivari et.al (2020) presents that the Internet has become a crucial piece of our life, However, It additionally has given freedoms to secretly perform malevolent exercises like Phishing. Phishers attempt to misdirect their casualties by friendly designing or making mockup sites to take data, for example, account ID, username, secret word from people and associations. Albeit numerous strategies have been proposed to distinguish phishing sites, Phishers have developed their techniques to escape from these location strategies. Quite possibly the best techniques for identifying these malevolent exercises is Machine Learning. This is on the grounds that most Phishing assaults have some normal attributes which can be distinguished by AI strategies. In this paper, we thought about the aftereffects of different AI techniques for anticipating phishing sites.

Ping Yi et.al (2018) portrays web administration is one of the key correspondences programming administrations for the Internet. Web phishing is one of numerous security dangers to web administrations on the Internet. Web phishing intends to take private data, for example, usernames, passwords, and Mastercard subtleties, via mimicking an authentic substance. It will prompt data divulgence and property harm. This paper mostly centers around applying a profound learning system to identify phishing sites. This paper first plans two kinds of highlights for web phishing: unique highlights and connection highlights. A recognition model dependent on Deep Belief Networks (DBN) is then introduced. The test utilizing genuine IP streams from ISP (Internet Service Provider) shows that the distinguishing model dependent on DBN can accomplish a roughly 90% genuine positive rate and 0.6% bogus positive rate.

Dželila Mehanović and Jasmin Kevrić (2020) presents security is quite possibly the most real points in the online world. Arrangements of safety dangers are continually refreshed. One of those dangers are phishing sites. In this work, we address the issue of phishing sites grouping. Three classifiers were utilized: K-Nearest Neighbor, Decision Tree and Random Forest with the element determination techniques from Weka. Accomplished precision was 100% and number of highlights was diminished to seven. In addition, when we diminished the quantity of highlights, we diminished opportunity to assemble models as well. Time for Random Forest was diminished from the underlying 2.88s and 3.05s for rate split and 10-crease cross approval to 0.02s and 0.16s separately.

Rishikesh Mahajan and Irfan Siddavatam (2018) talk about Phishing assault is a least complex approach to acquire touchy data from guiltless clients. Point of the phishers is to obtain basic data like username, secret key and ledger subtleties. Network protection people are presently searching for reliable and consistent identification strategies for phishing sites location. This paper manages AI innovation for location of phishing URLs by separating and examining different highlights of authentic and phishing URLs. Choice Tree, arbitrary woodland and Support vector machine calculations are utilized to identify phishing sites. Point of the paper is to recognize phishing URLs just as tight down to best AI calculation by looking at exactness rate, bogus positive and bogus negative pace of every calculation.

Manish Jain et.al (2020) presents phishing is routinely a conventional attack on people by means of causing them to uncover their all out exceptional data the utilization of fake sites. The motivation behind phishing records measure apparatus URLs is to squeeze the individual data like buyer name, passwords and internet banking exchanges. Phishers(attackers) utilizes the sites that rectangular recognition outwardly and semantically the photograph of these genuine sites. As the time keeps on developing, phishing procedures began to advance rapidly and this might be forestalled via practice against phishing components to discover phishing. Machine intending to catch can be an incredible gadget that is consistently utilized toward phishing assaults. This paper studies the capacities utilized for the discovery and location procedures by the utilization of Machine learning.

Sagar Patil et.al (2020) portrays the objective of our task is to execute an AI answer for the issue of identifying phishing and pernicious web joins. The outcome of our venture will be a product item which uses AI calculation to distinguish noxious URLs. Phishing is the method of extricating client certifications and touchy information from clients by taking on the appearance of an authentic site. In phishing, the client is furnished with a mirror site which is indistinguishable from the authentic one yet with noxious code to remove and send client certifications to phishers. Phishing assaults can prompt tremendous monetary misfortunes for clients of banking and monetary administrations. The conventional way to deal with phishing location has been to either to utilize a boycott of known phishing joins or heuristically assess the properties in a suspected phishing page to distinguish the presence of malignant codes. The heuristic capacity depends on experimentation to characterize the limit which is utilized to order pernicious connections from benevolent ones. The disadvantage to this methodology is helpless exactness and low flexibility to new phishing joins. We intend to utilize AI to beat these downsides by carrying out some characterization calculations and looking at the presentation of these calculations on our dataset. We will test calculations, for example, Logistic Regression, SVM, Decision Trees and Neural Networks on a dataset of phishing joins from UCI Machine Learning vault and pick the best model to build up a program module, which can be distributed as a chrome expansion.

III. PROPOSED SYSTEM

The key to success during this issue is to develop rules of thumb to extracting options from websites then utilizing them to predict the kind of internet sites. Machine learning technique is employed to notice the phishing websites by making rules. so as to make a phishing web site sure rules ar accessible in generating the net link of specific web site. exploitation this key plan we have a tendency to ar attending to split the URL into protocol, sub domain, name, and high level domain and file path. Here name plays a serious role; by examination list of parameters with original web site info we will establish the faux web site. additionally to spot with effectives recommendation is enforced. In looking an internet site the URL process is used to spot the initial link instead of faux link. this can end in identification of phishing websites in effective means.

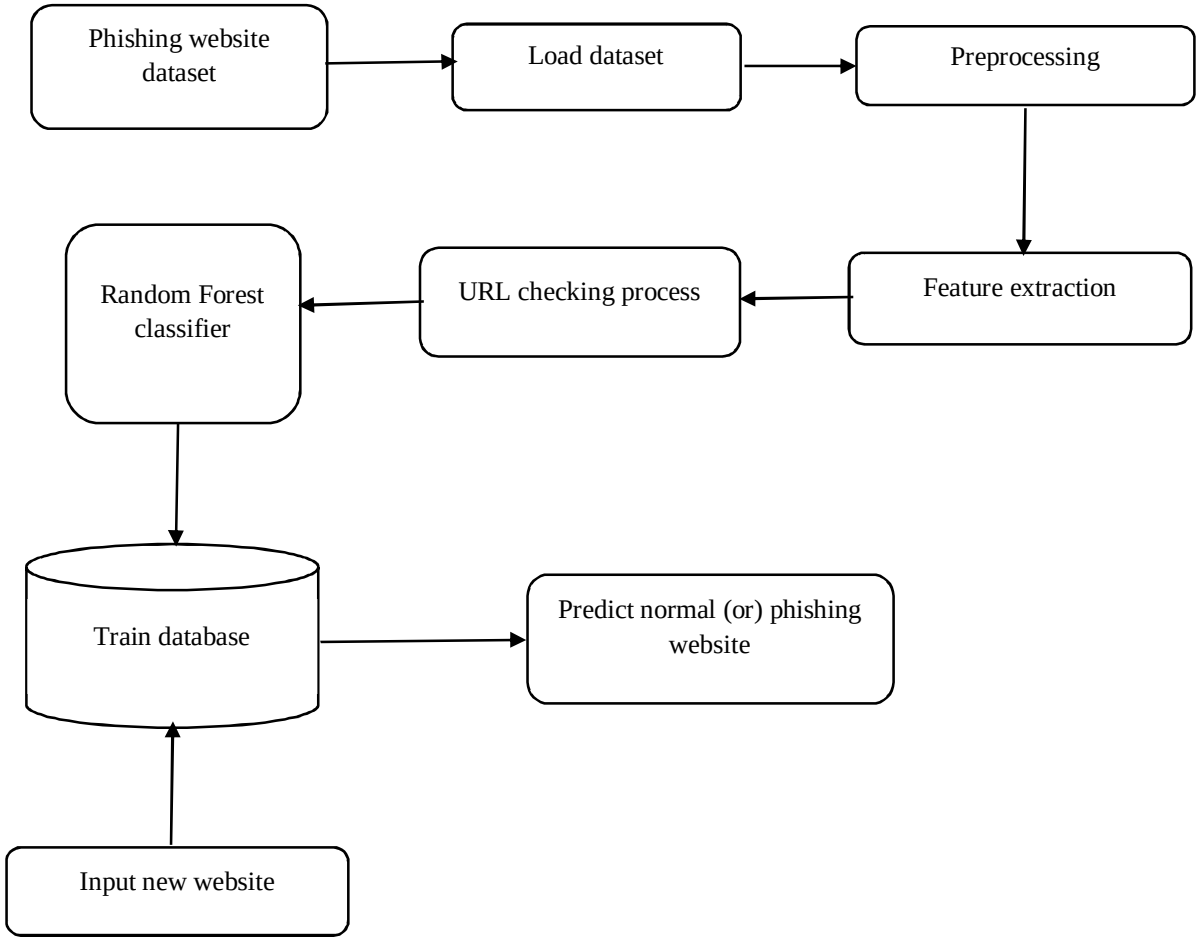


Figure 2: working system of proposed method

A. Implementation Of Proposed Work

1) *Pre-processing Module:* The training data set set is collected from the web. The collected knowledge set is pre-processed. Take away the record, that contains any missing values. Check all the records contains category label (Normal or Phishing).A uniform resource locator consists of some important or insignificant words and a few special characters, that separate some vital parts of the address. Therefore, within the knowledge pre-processing half, firstly, every word is extracted from the uniform resource locator and so they're further to the "word list" to be analyzed within the in progress execution. the most aims of information pre-processing half ar as follows:

a) Detecting the words, which are similar to known brand names,
b) Detecting the keywords in the URL,
c) Detecting the words, which are created with random characters.

2) *Feature Extraction Module:* Feature extraction is filtering out a coaching knowledge set so as to stay attributes / variables that have sensible illustration of the complete coaching dataset. the chosen set attributes sometimes sever as a proportional sample of the population and supply similar performances because the complete coaching dataset's attributes. Feature extraction strategies ar extraordinarily useful in cases once the spatial property of the coaching dataset is massive (very huge numbers of attributes). The spatial property drawback might limit the applicability of looking algorithms on the dataset and thus spatial property reduction becomes close at hand.

In this module the following features are extracted from the URL.

a) IP address - IP address in the domain name of the URL
b) Long URL – Length of the URL
c) URL's having @ symbol – URL contains @symbol
d) Prefix and suffix – domain part has '- '
e) Sub-domain (dots) - dots in domain name
f) Misuse/fake of HTTPs protocol – Not using https protocol
g) Request URL - objects are loaded from a domain other than the URL
h) Server form handler – The server transferred data to another domain.
i) URL of anchor – No of anchor tag in URL page
j) Abnormal URL – No host name in URL
k) Using pop-up window -Usually authenticated sites do not ask users to submit their credentials via a popup window
l) Redirect page – Redirect to suspicious page
m) DNS record – Empty DNS record
n) Hiding the links - change of status bar onMouseOver
o) Website traffic – Determine traffic rate
p) Age of domain- Presence of web site

3) *URL Checking Module:* The uniform resource locator is fragmented into several categorise and checking method is enforced to spot whether or not the user entered web site is legitimate or faux. Few technique and rules square measure delineated during this module. uniform resource locator is split into protocol, sub domain, name, and prime level domain and file path. every fragmented path is compared with set of rules generated within the creation of legitimate web site. This rules comparison and extraction method is finished through {data mining|data methoding} process.

For example,

Proposed Rule:

URL length, 54 $\longrightarrow$ Legitimate
URL length ≥ 54 and ≤ 75 $\longrightarrow$ Suspicious
Otherwise $\longrightarrow$ Phishy

Similarly, many rules are generated for each part in a URL and identification of phishing website is done accurately.

4) *Classification Module:* In this module the test data may be classified traditional or Phishing mistreatment Random forest algorithmic rule. supported generated RF model the take a look at knowledge is assessed. for every universal resource locator extract the options and classify the universal resource locator mistreatment RF. Random forest algorithmic rule could be a supervised classification and regression algorithmic rule. because the name suggests, this algorithmic rule every which way creates a forest with many trees. Generally, the a lot of trees within the forest the a lot of strong the forest feels like. Similarly, within the random forest classifier, the upper the amount of trees within the forest, larger is that the accuracy of the results.

a) *Step 1:* First, start with the selection of random samples from a given dataset.

b) *Step 2:* Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

c) *Step 3:* In this step, voting will be performed for every predicted result.

d) *Step 4:* At last, select the most voted prediction result as the final prediction result.

IV. RESULT AND DISCUSSION

In this section, result of our proposed work is shown and discussed briefly. To detect phishing websites in accurate way one best method is URL based detection. Hence in introduction part URL and its important part and explanation for it were given briefly. Similarly possible URL fraud takes place is also mentioned in our proposed section. By implementing Random forest classifier our work verifies the each and every part in the URL check whether the link is phishing or not.

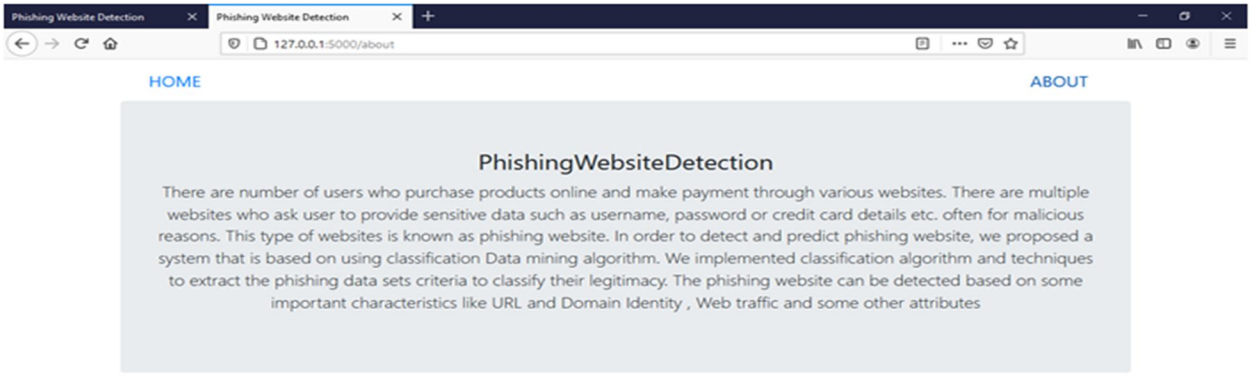


Figure 3: home page of our proposed system

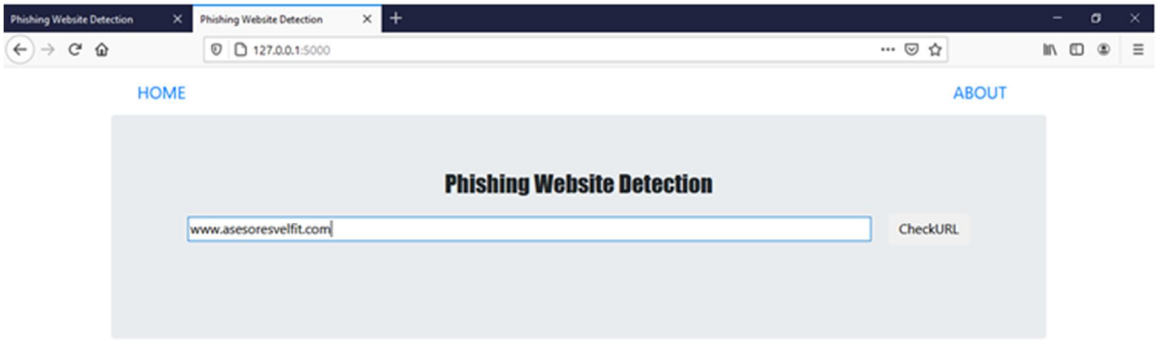


Figure 4: entering URL to check

In above figure it clearly shows entering the URL to check whether it is original or phishing URL. In backend of this process URL will be verified with dataset here each and every part of the URL will be checked with our trained database and predict whether it is original or not.

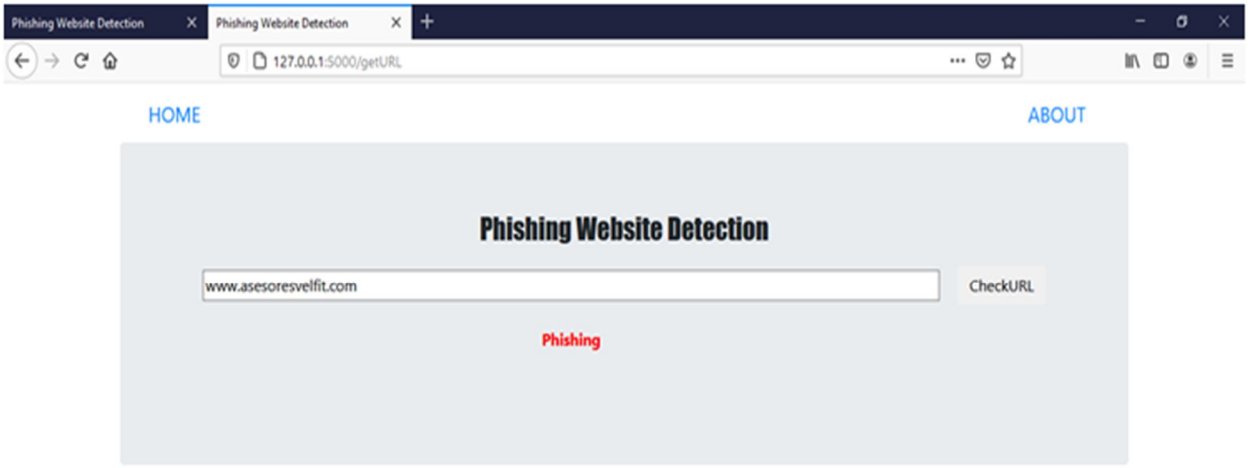


Figure 5: Detect phishing URL

The above graph shows that the entered particular URL is phishing and it intimate.

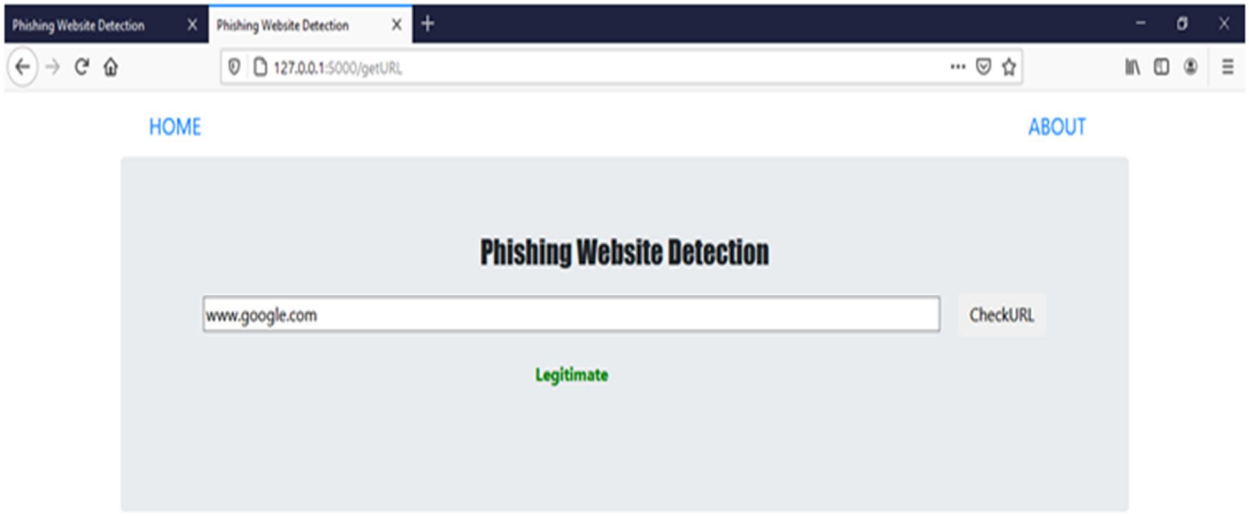


Figure 6: Legitimate website detection

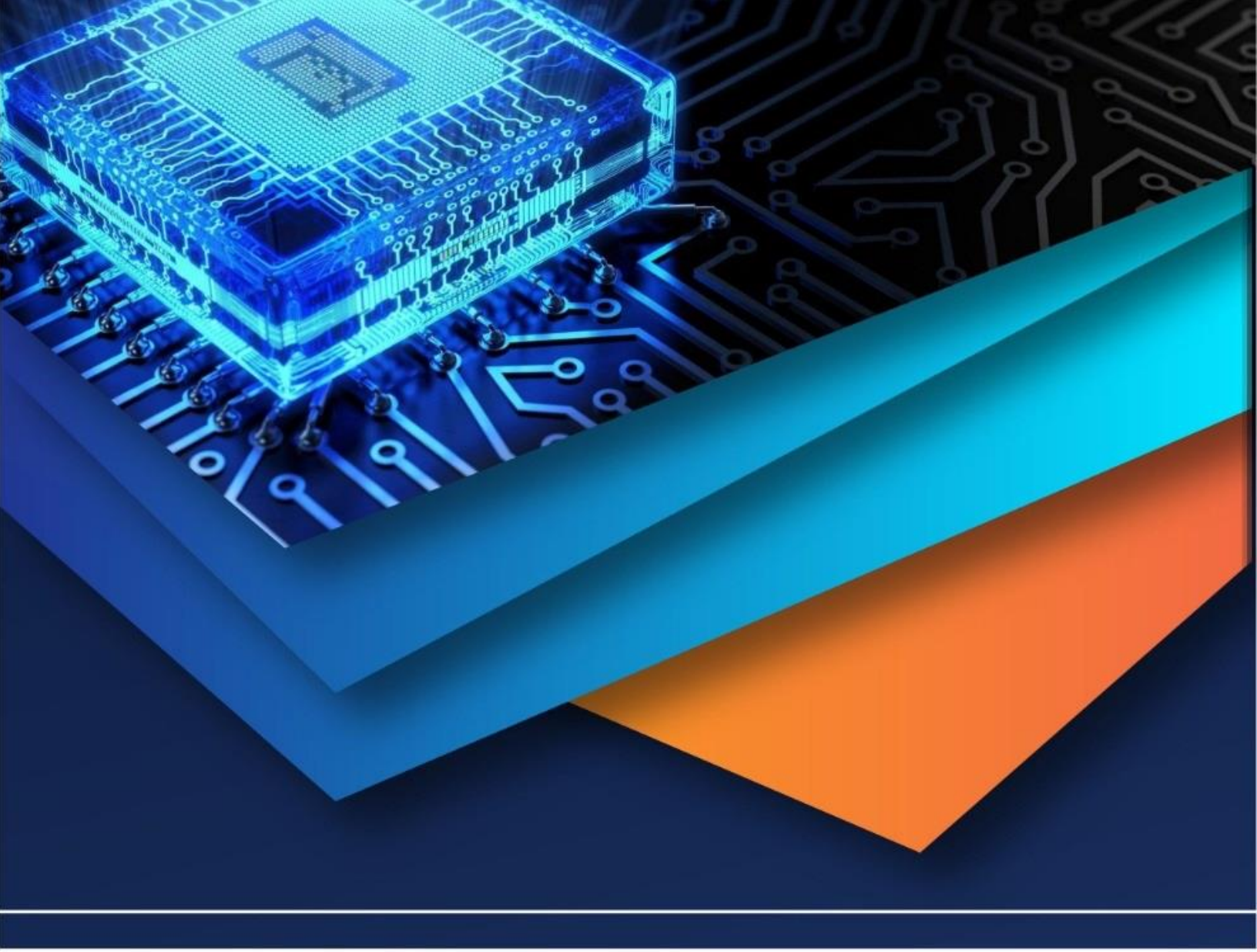
The above graph shows that the mentioned URL is legitimate website by processing through our training phase and it detects accurately by means of Random Forest algorithm.

V. CONCLUSION

The planned System's aim is to implement the detection of the phishing websites mistreatment data processing. This task is going to be done by extracting the options of the web site via address once the user visits it. The obtained options can act as check knowledge for the model. Random Forest formula is often accustomed train the planned model. The most task of this technique is to observe the phishing web site and alert the user beforehand thus on forestall the users from obtaining their credentials victimized. If any user still needs to proceed, it is often done at their own risk. This paper investigates the matter of computing device categorization i.e., traditional or Phishing. This paper presents the supervised machine learning approach RF is employed to class's phishing and malware sites. This paper extracts varied numbers of options from the address. The RF formula achieved high classification accuracy for analyzing similar knowledge elements to those of rule-based heuristic techniques.

REFERENCES

[1] Vahid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi, "Phishing Detection Using Machine Learning Techniques" arXiv:2009.11116v1 [cs.CR] 20 Sep 2020.
[2] Ping Yi, Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, and Ting Zhu, "Web Phishing Detection Using a Deep Learning Framework" Wireless Communications and Mobile Computing Volume 2018 |Article ID 4678746.
[3] Dželila Mehanović*, Jasmin Kevrić, "Phishing Website Detection Using Machine Learning Classifiers Optimized by Feature Selection" international information and engineering technology association.
[4] Rishikesh Mahajan and Irfan Siddavatam, "Phishing Website Detection using Machine Learning Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 23, October 2018.
[5] Manish Jain, Kanishk Rattan, Divya Sharma, Kriti Goel, Nidhi Gupta, "Phishing Website Detection System Using Machine Learning" International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 05 | May 2020.
[6] Sagar Patil, Yogesh Shetye, Nilesh Shendage, "Detecting Phishing Websites Using Machine Learning" International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 02 | Feb 2020.
[7] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," 2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015, pp. 769–770, 2015.
[8] Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.
[9] Lakshmi, V. S., & Vijaya, M. S. (2012). Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering, 30, 798- 805.
[10] P. Yi, T. Zhu, Q. Zhang, Y. Wu, and L. Pan, "Puppet attack: A denial of service attack in advanced metering infrastructure network," Journal of Network and Computer Applications, vol. 59, no. 1, pp. 325–332, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY