Phishing Fraud Detection

Aditya Mache[1], Ashish Gade[2], Shreeyash Dhole[3], Nilima Kulkarni[4]

[1-4]*Department of Computer Science and Engineering, MIT School of Engineering, MIT Arts Design and Technology University, Pune, Maharashtra, India*

---***---

Abstract *In the modern era of information technology being connected on a social media platform and mail services have been an abundant process in the life of human being; along with the virtues of instant connectivity and exchange of information via an internet platform, some common social engineering attacks are been carried out by the evil-minded people namely called as hackers. Web attacks are the major part of cybercrime in which criminal uses internet services or URLs of related or similar identity as a mediator for resembling the legitimate website with a motive to steal some personal information about user entity that is not been publicly available and use them this information for personal benefit to gain access to the social media accounts or to access the bank account for laundering the money or to gain profit in any term. This website could be extremely dangerous for both the user end and the service provider. Thus, to detain the user from getting fraud and detect various phishing websites a proper proactive data analysis is abundant so that by using the analyzed data, the internet services can be more secure and reliable to transact with. This paper proposes a vital method of implementing technologies for extracting features and classification. A Comparative approach of analysis has also been discussed in the paper where various machine-learning techniques and their performance are evaluated, extorted and classified. There are also some suggestions of methods that could be implemented to increase the efficiency of the system. The best performing method is used as a backend for efficient working of a web Plug-in.*

Keywords Phishing attacks, fraud, scam, detention, hackers, EDA algorithm- Booster, intrusion detection, Address bar, Domain name, HTML JavaScript.

1.INTRODUCTION

Phishing nowadays can be so powerful that it can mislead a human to perform a cybercrime too. Humans are been habitual of Searching for information or purchasing a product from the online market because it is a very easy task nowadays. Also, many people are been addicted to surfing on the internet so it has been a boon in the life of human beings, but along with this some bad intentions people make use of their evil mind and try to exploit the users of the internet for their benefit and make the internet a curse. Attacks in which sensitive information of an entity is retrieved by implementing the social- engineering methods for personal benefit are described as Phishing. In this type of attack entities like emails and sensitive text messages seem to come from a reputable source. Their main aim is an attempt to get the users credentials like passwords, account numbers, Social security numbers, answers to private security questions, etc. As per the report, https://www.antiphishing.org/resources/apwg-reports/ submitted by APWG phishing trend in the year 2018 in which the accounts of phishing found in the initial days of the financial year 2018 was 4,96,578 to a total number of phishing sites detected in the latter half was 3,71,519 thus the number of attacks is rising rapidly. The Attackers use many of such cybercrime attacks each day and is often successful in this attack due to lack of information by the USER using the resources.

The Research paper is structured in the below format; in which section 2, we have summarized information about phishing and try to elaborate some questions like, what is phishing? what is the technical perspective of phishing? how it is held? We have also discussed some of the related work and existing projects for the detection of phishing URLs. Section 3 consists of all the information about the proposed system including its detailed description, the purpose of the system, its architecture, Feature extraction parameters, Implementation of the resultant output filesystem. The resultant output of our experiment is analyzed in section 4. The conclusion is described in section 5.

1.1 Social-Engineering attack (Phishing)

Social engineering attack mainly deals with the human psychology and susceptibility to manipulate and unleash a sensitive data from a victim or break security measures that allow an attacker to access the network. Phishing is one of the most preferred and used methodologies to trap the victim. In these Phishing attacks, the attacker delivers a phony transmission that seems to be designated through a legitimate source. The purpose against performing such an attack is to gain secret information including the credit card details, password of an account or to install an untrusted source file to the victim's computer as a means to gain access to the system. This attack is a prevalent kind of cyber-attack where each user of the internet must learn to bulwark themselves. Fig(1) describes the step-by-step method used by the attacker. In which, the attack begins with an unauthentically bogus email or some form of transmission that is intended to attract a victim. The message seems to come from a reliable source in this form of attack. If the victim is attracted by the appealed offer, the victim is more likely to lead some sensitive details on a scammed platform. A malware file or a trojan file of the virus is often attached with such emails and the if victim downloads such file to the device the virus starts to destroy the system.

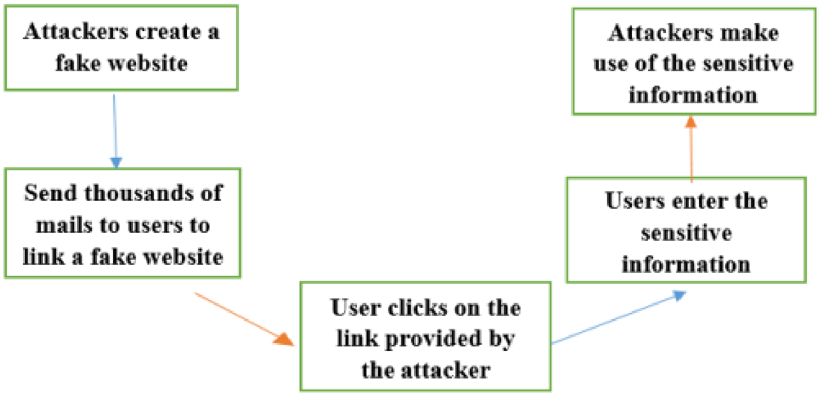

Attackers create a fake website
Send thousands of mails to users to link a fake website
User clicks on the link provided by the attacker
Users enter the sensitive information
Attackers make use of the sensitive information

Fig (1) Method used by an attacker for Phishing

Assailants get to benefit financially from obtaining the credit card details or other personal information of their victims [5]. Phishing emails are also sent to obtain authentic information or other data about staff to use them in a sophisticated assault on a specific organization. Phishing is a popular starting point for cyberattacks such as Advanced Perpetual Threats and ransomware. In a phishing attack, information gathering is a very crucial step also called reconnaissance in which the attacker collects context knowledge of targets' personal and job backgrounds, intrigues, or social activities by using social information tools, like gregarious networks including LinkedIn, Facebook However Twitter. Hackers may use this information to classify prospective victims' names, work titles, and email addresses, as well as the knowledge about denominations of important workers at the workplace. This data will then be used to compose a trustworthy e-mail or text message. Criminals can also be an advanced sedulous threat group, for example, typically begin with a malicious connection or attachment in an e-mail. The most common susceptibility or clickable phishing environments in this form of attack have been described as Facebook victuals. If phishing assailants are created, they are often used to spread false information. A victim is usually given a however if the high seems to come from Kennedy individual, the agency [17]. Malevolent file injection, which involves phishing malware, or connections to malevolent websites is used to carry out the attack.fig (2) In any situation, the aim is to guide the target to a malevolent site where they can download malevolent software or be duped into sharing personal and financial information, and also in some cases the device used by the victim person is compromised and damaged.

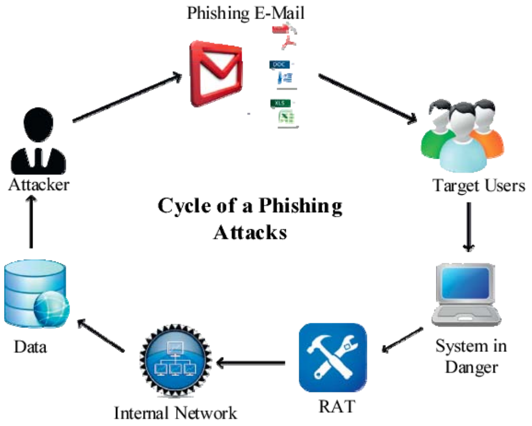


Fig (2) Phishing attack cycle

2. Literature Survey

To detect phishing attempts, several techniques have been suggested. The accuracy rate of the experiment performed by

Paper Title	Description	Draw-back
1. Muhammet Baykara et.al., Sept 2018	Examining the context of mail with the help of parameters like sentimental keywords, leverage offer, grammatical mistakes, etc the nature of mail is identified and segregated as spam.	A phishing attack is not a platform-based attack thus, implementing a solution based on only one technology is not the complete solution. Mail can be defamed by the attacker by spoofing it to be legitimate.
2. Athulya AA, Praveen K et.al., 2020	A hybrid cull technique named desultory is used to designate the web-page and listing approach is used where a Website is designated as BLACK-LIST, WHITE-LIST	The listing approach cannot mitigate a Zero-Day attack. As each day many new phishing sites are being developed by the attackers.
3. Moruf akin adebowale,khan T.Liwin ,M.A Hussain et.al., 2018	A Neuro-Fuzzy interference System based scheme is used to integrate the feature of text, image, and frames of a web-phishing detection and protection are examined.	The accuracy rate of the proposed solution decrease as the number of sites increases in the experiment. Also along, With the different datasets provided the performance varies of the solution.
4. Cassandra Cross, Rosalie Gillet et.al., 2020	ELaboration of current cognizance of business email compromise for fraud is mentioned.	No solution to the discussed problem is provided in the paper.
5. Jyoti chikara, Ritu Dahiya,Neha Garg et.al., 2013	The paper has expounded the vigilance and inculcation about phishing quandaries and phishing and anti-phishing techniques.	The solution in this paper has a time constraint delay i.e. the performing time required by the solution is high.

6. Adam. Lakshmi, M. Purushotham Reddy et.al., 2021	Hyperlink present in the source code of the HTML page is actively monitored for identification of a phishing site. in total 30 parameters are been used and a supervised machine learning approach is proposed.	Zero-Hour attacks cannot be mitigated by using this technique.
7. M. Noushad Rahim, K.P. Mohamed Basheer et.al., 2021	The proper study of gap analysis of conventional antiphishing techniques and challenges faced by the Machine Learning based approach is discussed.	The performance of the Machine Learning Algorithm varies based on the characteristics and dataset used in the testing model
8. Priya Saravanana, Selvakumar Subramanian et.al., 2020	In this paper extraction by various features from the collection of phishing and legitimate website obtain from PhishTank and starting point directory service is used. The Framework provides the ideal solution for an individual website.	The proposed algorithm has a time delay exceeding more than 5 sec per Url which is a huge time delay.
9. Ms. Sophiya Shikalgar Dr. S. D. Sawarkar Mrs.Swati Narwane et.al., 2019	A Hybrid approach is proposed in the paper where the Machine Learning technique of Random Forest and Navies Bayes approach is implemented.	This is an excellent proposed system but, could be much improvised if a structured dataset of phishing is used, and also adding more features such as Lexical analysis and content-based Feature can assure more accuracy. The Accuracy rate of XGboost is 85.5%
10) Ayam el Assael, Shahryar Baki, Avishai Das, Rakesh M Verma et.al., 2019	A new system named PHISHBENCH is introduced in which the evaluation and comparison of existing features under the identical experimental condition is proposed.	A statistical and quantitative approach is proposed which alone cannot be a solution for each framework.
11) Mehmet Korkmaz et.al., 2021	A phishing site is identified by using a voting method. The ML model used in this experiment is LR, K-nearest Neighborhood, XgBoost, Random forest.	The Accuracy rate of the XGBoost algorithm in this experiment is 80.25 and the time taken is 95.18sec.

2.1 Related Work

NAME	DESCRIPTION
The Google Safe Browsing	The Strategy includes the utilization of blocklists of web pages for the Hacker's identification of spamming assaults. The primary inadequacy of this methodology is its powerlessness to distinguish phishing URLs that are absent inside the blocklist where may be indicated by the

	bogus positive way.
Phish Net	The methodology beats issues identified by the blocklist. It had different significant advances like the generation of URL varieties comparative with the initial ones which develop the boycott additionally as an information structure that appoints every point to the web-dependent in closeness along with previous web pages.
Phish Guard	This performs a calculation URL based on the external appearance of the website pages to rank suspect locations. The calculation is based on the tests like Whitelist and Blacklist, shortened URL, pattern matching, DNS test.
Spoof Guard	This strategy compares phishing manifestations to suspicious sites to see if they're genuine or phishing sites. A weight is applied to each of these heuristics depending on how close the sites are clustered. On the off chance that the all-out score of the phishing side effects recorded above surpasses the edge, it is named a phishing site else, an authentic site.

3. About the System

In this project, Artificial intelligence models are used for automation in detecting Phishing sites. Working of these models depends on the Pre-requisite function that is the Feature Extraction Module where discrimination of data into legitimate sites and Phishing sites is derived from a collection of websites in a CSV (comma-separated values) file. The overall perspective of this project mainly focuses on increasing the average accuracy rate of ML and proactive monitoring with warning of a webpage with help of embedding integration of web plug-ins that detects the site to be legitimate or fraudulent before redirection of URL and defies the user from forgery. In machine re-training alone it is not enough to overcome new attacks, new features and strategies are needed to stop the attack from deceptive detection systems. Thus continuous proactive monitoring should be done on the activities which can be achieved by using a web extension that can detect the URL nature.

3.1 Reasons for a phishing attack project?

One of the most important modes of correspondence is e-mail. Increased spam emails result in traffic jams, decreased competitiveness, and phishing, and this is a somber quandary in the information world. Per year, the volume of spam emails increases. As a result, URL identification is a critical and crucial step in the prevention of an attack.

Phishing can be very dangerous. Its severity could be so high e.g., intrusion of the backdoor file in a system to gain access to user documents and also to use the system for personal gain. One might not be aware that such criminal activities are also being conducted by evil people; the lack of knowledge is the main reason for the attack to be successful. Thus, creating awareness about the attack and its importance is not suitable for all users. Hence an automated system is thus crucial for checking whether the recipient's website is Legitimate or it is a kind of Bait received for catching a Whale.No single solution has been introduced to give the total immunity from the attack held by phishing.

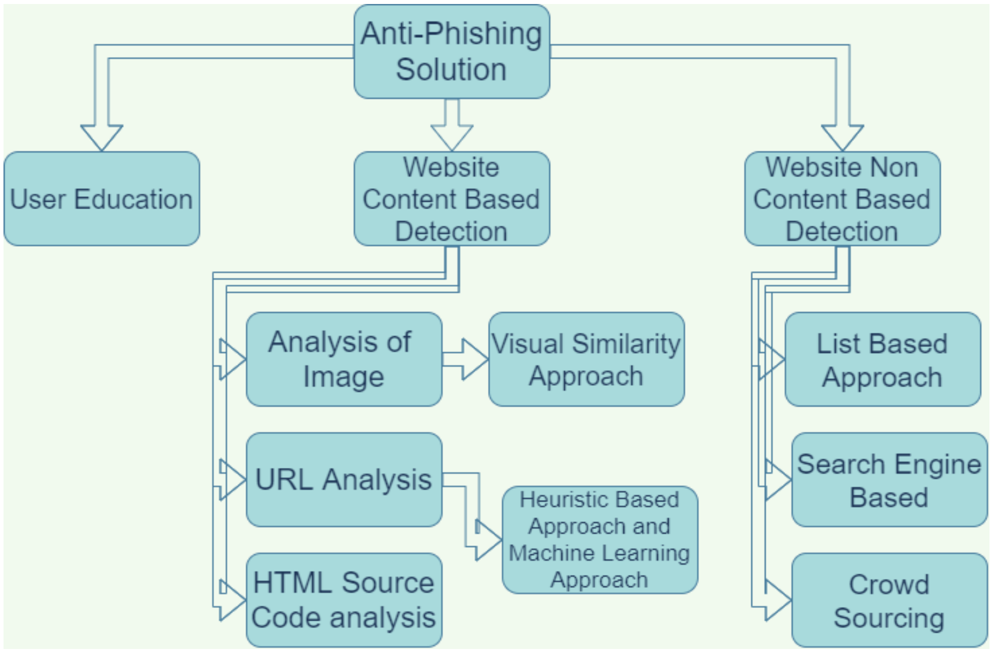

Anti-Phishing Solution
User Education
Website Content Based Detection
Website Non Content Based Detection
Analysis of Image
Visual Similarity Approach
URL Analysis
Heuristic Based Approach and Machine Learning Approach
HTML Source Code analysis
List Based Approach
Search Engine Based
Crowd Sourcing

Fig (3) Preventive measures for Phishing

3.2 System Architecture

The system architecture of a proposed system for the detection of Phishing URLs focuses on the correct selection of Features based on which the ML will perform its operation.

The system design includes the prevention of URL redirection and pop-up a window to indicate the results to the user.

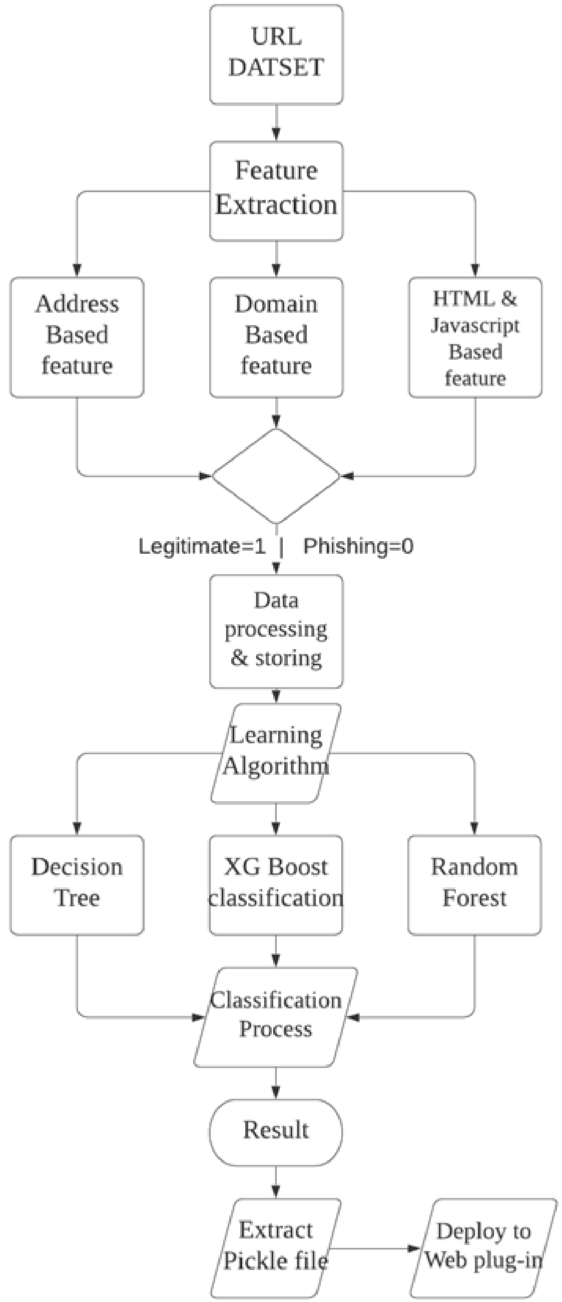


Fig (4) System Block Diagram

3.3 System Model

The system is based on an analysis of the feature extraction performed on the open-source and user-entered database. Extraction of Key factor includes:

3.3.1 Address Bar Based Features

a) The domain of URL- Extraction of Domain present in the URL by matching with the URL-parse after "WWW."

b) IP Address in URL: -If an IP address is present there might be a possibility that an attacker is trying to exploit it. As a result, if an IP address is used instead of a domain name, it should indicate spoofing.

c) The existence of the '@' sign in the URL is verified. When you use the "@" symbol in a URL, the browser ignores anything before the "@" symbol, and the actual address always comes after the "@" symbol.

d) URL Length- Calculating the URL's length. Phishers will mask the suspicious aspect of a URL in the address bar by using a long URL. If the URL is longer than or equivalent to 54 characters, it is considered phishing. Otherwise, it is considered valid.

e) Depth of URL- Computes the depth of the URL. This feature calculates the number of sub-pages in the given URL based on the '/’. The value of the feature is numerical based on the URL.

f) URL redirection "/”: - The presence of the character "//" in the URL indicates that the user will be forwarded to another website. The “//” in the URL's position is calculated. We discovered that if the URL begins with “HTTP,” the “//” should be placed in the sixth row, however, if the URL, uses "HTTPS," the "//" should appear in the seventh point.

g) "HTTP/HTTPS" in Domain Name-Checks for the presence of "HTTP/HTTPS. To deceive users, phishers can add the “HTTPS” token to the domain part of a URL.

h) Shortening URL: - It is a method of reducing the length of a URL while still directing it to the desired webpage on the "World Wide Web." This is achieved by using an “HTTP Redirect” on a short domain name that points to a long URL website.

i) Prefix or Suffix "-" in Domain-Checking the presence of '-' in the domain part of URL. In legitimate URLs, the dash symbol is seldom used. Phishers also apply prefixes or suffixes to domain names distinguished by (-) to make people believe they are communicating with a genuine website.

3.3.2 Domain-Based Features

j) Web Traffic - All features measure the total number of visits gained by the site including the visits on many pages is measured and accordingly the popularity of a website is determined. In the case of a phishing site, the living period is of short time thus it may not be recognized by the Alexa database.

k) Tenure of Domain-All features could be obtained from the WHOIS database The minimum age of the legitimate domain is considered to be 12 months for this project. Age here is nothing but the difference between creation and expiration time

l) Expiration Tenure for Domain-This feature could be extracted through WHOIS DB. Thus, the remaining domain time is calculated by finding the difference in expiration tenure & current time. The end period considered for the legitimate domain is less than 6 months.

3.3.3 HTML, JavaScript-based Features

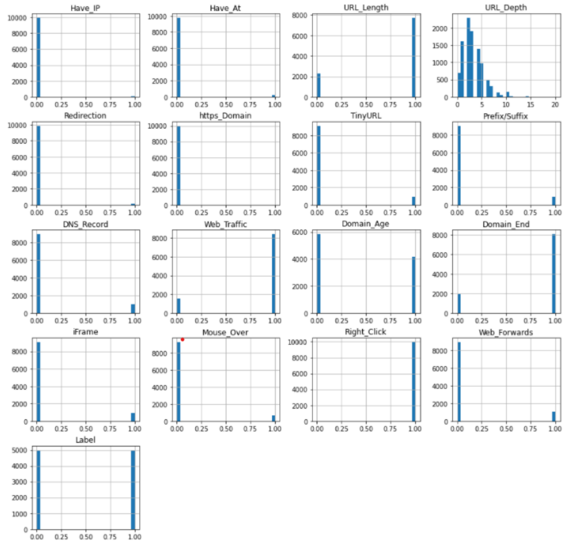
m) IFrame Redirect- The HTML tag IFrame helps you to insert another website into the one you're now using. Phishers will use the "iframe" attribute to make the frame invisible. In such cases, attackers make use of the "box boundary" feature, which allows browsers to replicate the original page.

n) Status Bar -Phishers will fool users into seeing a bogus URL in the status bar by using JavaScript. We'll need to delve through the webpage source code, especially the "on Mouseover" example, to see if it modifies the status bar, to get this feature.

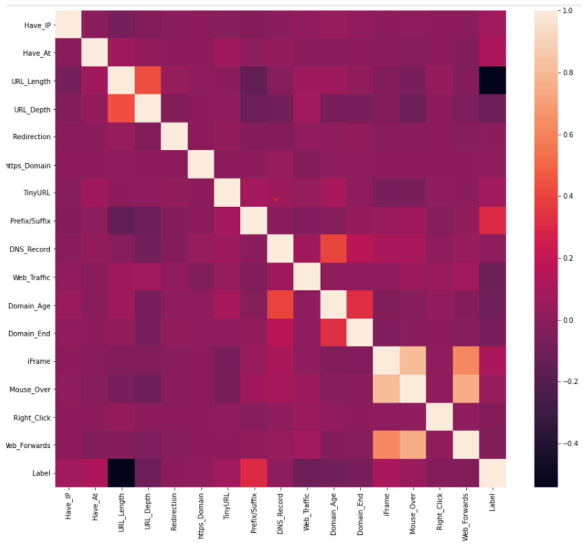
o) Disabling Right Click: - Phishers disable the right-click button using JavaScript, blocking users from reading and saving a webpage's source code. This feature is done similarly to "Hiding the Link with on Mouseover." Thus, we'll check the webpage source code for the event "event. Button==2" and see whether the right-click is disabled for this feature.

p) Website Forwarding-The number of times a page has been diverted is a parameter that separates phishing websites from legitimate ones. We discovered that legal websites were only routed once in our dataset. Phishing websites with this functionality, on the other hand, have been diverted at least four times.

The Extraction includes the allocation of Binary numbers to the feature, 1 if the feature is associating with Phishing URLs or 0 if not.



Fig(5) Bar diagram of Feature Report.



Fig(6) Heat Map of Feature Report.

A Collection of this Dataset is then further used for the implication of Task on Supervised Machine Learning Classification and regression are the two main categories of supervised machine learning problems. The classification problem applies to this data collection. For the Working of the model, the dataset is divided into two sets: Training and Testing dataset. In, this project the 85:15 ratio is maintained which includes 85% of Train data and 15 % of the Testing dataset.

Experiments conducted by using Machine Algorithm result in different accuracy measures. The approach is to maximize the resultant accuracy output thus changes have been made with the parameters of the algorithm where train and test set is split into x and y attribute and keeping the maximum depth at which the tree should grow is kept equal to 10 also changes made with the Learning rate make the algorithm to adapt the problem. Thus, comparative accuracy helps in the identification of the best outcome results.

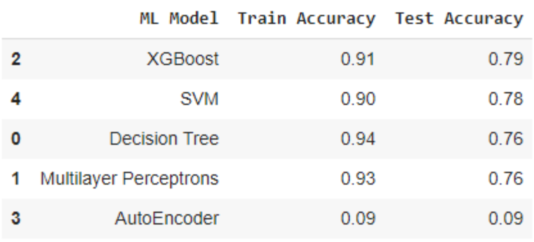
Accuracy = True Positive + True Negative /True Positive +False Negative +True Negative False Positive x 100%

Further then, this model is used for the Web Plug-in as a backend. In this Plug-in, 70% of code is based on JavaScript in which the overall page feature extracting features are been measured. The resultant best-performed model from the previous section is then used to evaluate the CSV values of the URL. If parameters are indicated towards the Phishing site the resultant output is represented by a pop-up before visiting the URL.

4. Result

Because of the increasing proliferation of phishing attacks, various methods of avoidance have evolved. Existing server and anti-spam filtering clients are used to detect different aspects of spam messages, but since the firewall is just a wall of security with tiny gaps or vulnerabilities, the attacker might be able to get through it.

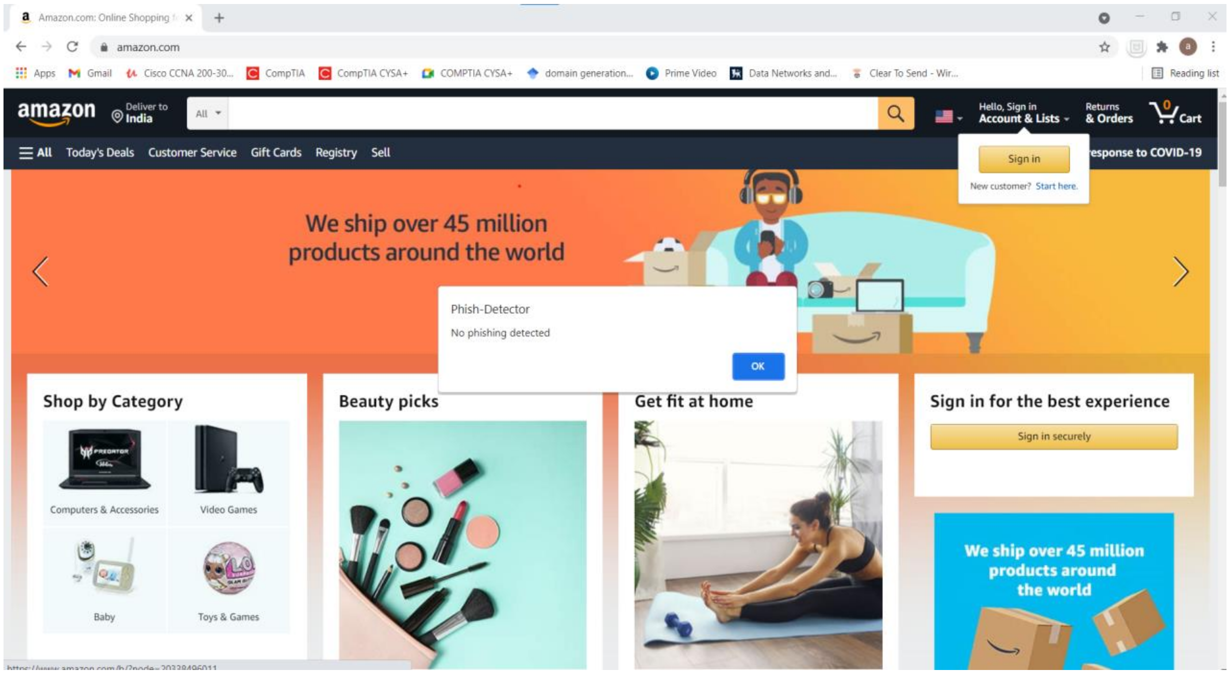
In this experiment, Several useful tricks have been developed in which a model decides whether a dataset containing websites is true or fake, thanks to the integration of Machine Learning. To get the best results selection of feature and classifier is very crucial.



	ML Model	Train Accuracy	Test Accuracy
2	XGBoost	0.91	0.79
4	SVM	0.90	0.78
0	Decision Tree	0.94	0.76
1	Multilayer Perceptrons	0.93	0.76
3	AutoEncoder	0.09	0.09

Fig(7). Results of accuracy rate.

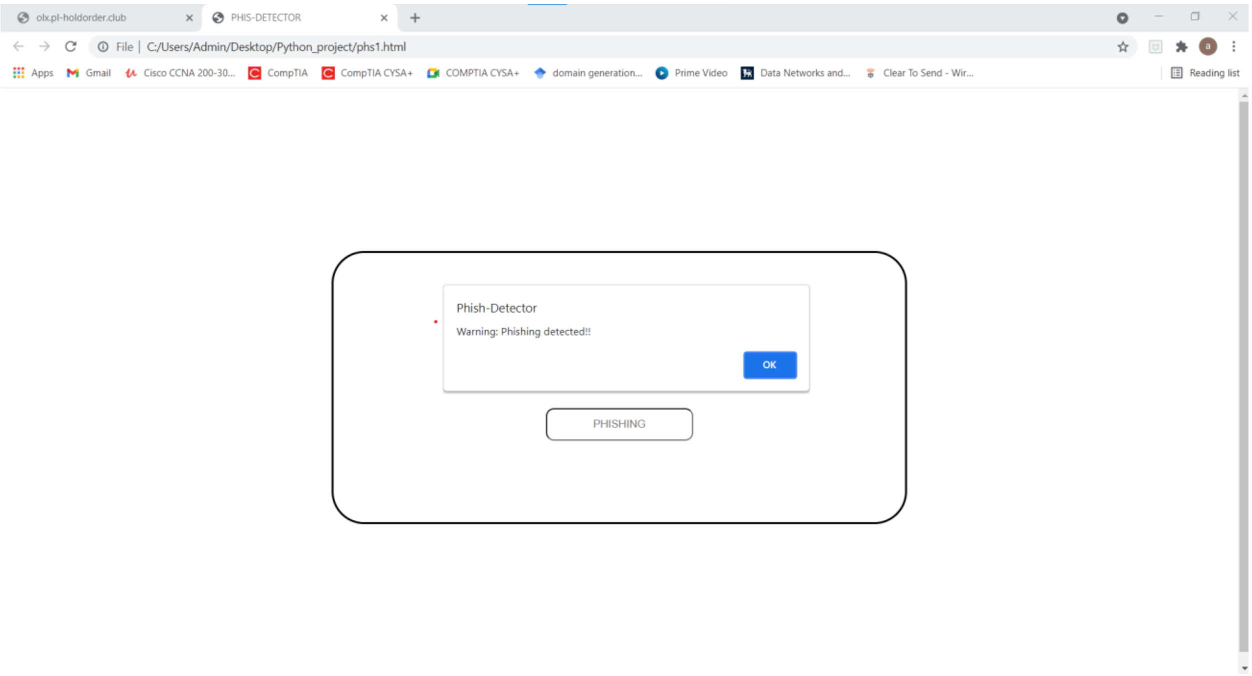
In collusion with this experiment, demonstrations on various Machine learning algorithms, XGBoost Classifier suggest better results. The Accuracy of the model is exceeded and the resultant output is 91%. An experiment can be additionally targeted to gain maximum precise outcomes on relegation along with artificial neural networks. The application of this model is extracted in the form of a pickle file and is used as a backend in the web extension as a method to indicate the user to avoid Phishing.



Phish-Detector

No phishing detected

Fig(8) The pop-up indication of Legitimate Website



Phish-Detector

Warning: Phishing detected!!

Fig(9) Pop-up indication of a phishing website.

5. Conclusion

This project aims to solve the issue of unawareness of the nature of websites and identify it by using datasets created specifically for this purpose and to train machine learning models to detect phishing websites. The dataset, which contains both phishing and benevolent URLs of websites, is used to generate the necessary URL and website content-based functionality. The performance standard of each model is estimated and compared. The Implementation of this proposed system will avoid the phishing site to compromise a user by identifying the attack and notifying the same. Henceforth, to avoid such an attack the preventive measure is taken into account as a troublesome job within the system security domain. A good identification system is now able to detect phishing attacks with a limited number of false positives. Data analysis and heuristics, machine learning, and deep learning algorithms are some of the guiding techniques discussed in this article. While heuristic and data analysis methods have low False Positive rates and high computational costs, they are better at identifying phishing attacks. As opposed to other approaches, ML procedures provide the most straightforward outcomes. Some machine learning algorithms can detect TP up to 91 % of the time. As malicious URLs are generated daily, attackers employ tactics to dupe users and change URLs to assail. The web Plug-in is an astonishing method that indicates a warning message to the user and avoids the attack.

6. References

1) Muhammet Baykara, Zahit Ziya Gure, NC “Detection of Phishing Attacks” 2018 6th International Symposium on Digital Forensic and Security (ISDFS) Firat University, Elazig, Turkey.
2) Athulya A.A, Praveen K. “Towards the Detection of Phishing Attacks” 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)
3) More akin Adebowale, Khin T.Lwin, M.A.Hossain “Intelligent phishing detection scheme using deep learning” (convolutional neural network (CNN) and the long short-term memory (LSTM)). (04 Aug 2018)
4) Cassandra cross, Rosalie Gillet “Exploiting Trust for Financial gain: an overview” on Business email Compromise fraud. (22 April 2020)

5) M somewhat “Efficient Deep Learning Technique for the Detection of Phishing Website”(27 June 2020)

6) Krishnamurthy“Chrome anti-phishing 43, Firefox anti-phishing 75”

7) Anu Yadav and Jatin Gemini “The Security threat in Cyber World – cybercrime as PHISHING” p- ISSN: 2393-9907; e-ISSN: 2393-9915 (April-June, 2017).

8) Phirashisha, Syiemlieh, Golden, Mary Khongsit1, Usha Mary Sharma, Bobby Sharma “Phishing-An Analysis on the Types, Causes, Preventive Measures And Case Studies in the Current Situation” e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 01-08 IOSR Journal of Computer Engineering (IOSR-JCE).

9) Ike Vayansky and Sathish Kumar “Phishing – challenges and solutions “ (January 2018).

10) Ram Basnet, Srinivas Mukkamala, and Andrew H. Sung ”Detection of Phishing Attacks: A Machine Learning Approach” New Mexico Tech, New Mexico 87801, USA

11) Amruta Deshmukh1, Sachin Mahabale2, Kalyani Ghanwat3, Asiya Sayyed “WEB PHISH DETECTION AN EVOLUTIONARY APPROACH” Department Of Computer Science and Engineering, Zeal Education Society’s, DCOER, Pune, Maharashtra, India.

12) Abdulghani Ali Ahmed, Nik Quosthoni Sunaidi “Malicious Website Detection: A Review” Faculty of Computer Systems & Software Engineering University, Pahang, Malaysia (February 01, 2018)

13) Moruf akin Adebowale, Khin T.Lwin, M.A.Hossain “Intelligent phishing detection scheme using the deep learning” (convolutional neural network (CNN) and the long short-term memory (LSTM)). (04 Aug 2018)

14) Jyoti Chhikara, Ritu Dahiya, Neha Garg, Monika Rani “Phishing & Anti-Phishing Techniques” ISSN: 2277128X

15) Priya Saravanana, Selvakumar Subramanian. “A Framework for Detecting Phishing Websites using GA-based Feature Selection and ARTMAP based Website Classification”.

16) Ayam el Assael, Shahryar Baki, Avishai Das, Rakesh M Verma “An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs.”

17) M. Noushad Rahim and K.P. Mohamed Basheer “A survey on anti-phishing techniques: From conventional methods to machine learning.”

18) Victor E. Adeyemo, Abdullateef O. Balogun.“Ensemble-Based Logistic Model Trees for Website Phishing Detection”

19) Mehmet Korkmaz, Emre Kocyigit, Ozgur Koray Sahingoz, and Banu Diri “Deep Neural Network-based Phishing Classification on a High-Risk URL Dataset.”

20) Ms. Sophiya Shikalgar Dr. S. D. Sawarkar Mrs.Swati Narwane. “Detection of URL-based Phishing Attacks using Machine Learning”

21) Tanusree Sharma1, Priscilla Ferronato2, and Masooda Bashir. “Phishing Email Detection Method: Leveraging Data Across Different Organizations”

22) Andr´e Bergholz, Gerhard Paaß, Frank Reichartz, Siehyun Strobel,Jeong-Ho Chang“Improved Phishing Detection using Model-Based Features”

23) Adam. Lakshmi ·M. Purushotham Reddy, Chukka Santhaiah, · U. Janardhan Reddy.“Smart Phishing Detection in Web Pages using Supervised Deep Learning Classification and Optimization Technique”

24) Bireswar Banik and Abhijit Sarma.“Phishing URL detection system based on URL features using SVM”

25) Abdulhamit Subasia, Emir Kremicb“Comparison of Adaboost with MultiBoosting for Phishing Website Detection”

26) Routh Srinivasa Rao, Alwyn Roshan Pais.“Two-level filtering however to detect phishing sites using lightweight visual similarity approach.”

27) Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. “PhishNet: Predictive Blacklisting to Detect Phishing Attacks.”(Oct 2010)

28) Aditya Mache, Ashish Gade, Shreyash Dhole, Nilima Kulkarni.“Phishing Fraud Detection-A Review” (Jan 2021)