IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE DETECTION USING MACHINE LEARNING

Mr. S.Bosu Babu [1], Fouziya Farheen [2], B.Ram Deepak [3], S.Nirupam Kumar [4], Sagar Chanchlani [5]

[1] Assistant Professor, Dept. of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

[2,3,4,5] Students, Dept. of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

ABSTRACT

The Internet's advancement has drawn attention to network security, as a secure network environment is fundamental for the Internet's fast and healthy growth. Cybercriminals employ phishing, a malicious act of deceiving users into clicking on phishing links, stealing their information, and using it to fake logins and steal funds. Network security is an iterative issue of attack and defense, and phishing and its detection technology continually evolve. Blacklists and whitelists are traditional methods for identifying phishing links but fail to identify new ones, which necessitates predicting whether a new link is a phishing website and improving the prediction's accuracy. Machine learning has emerged as a critical tool in predicting phishing websites, with this paper offering system learning technology for the detection of phishing URLs via extracting and studying diverse functions of valid and phishing URLs. KNN Classifier, Random Forest, and Support Vector Machine algorithms are used to locate phishing websites. The paper aims to detect phishing URLs as well as narrow them down to the fine algorithm that gets to know the set of rules with the aid of using evaluating the accuracy rate.

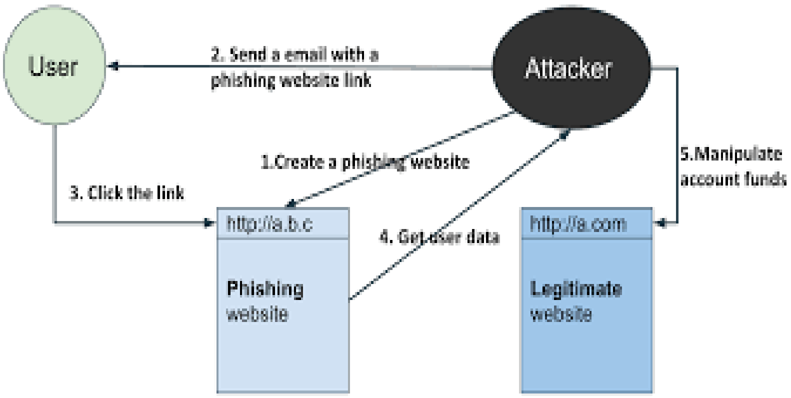
Keywords: Phishing, Phishing website detection, Machine Learning, KNN, SVM, Random Forest

1. INTRODUCTION

The Internet has become an integral part of people's lives and is now essential. According to the January 2021 global digital population report, 4.66 billion people, which is 59.5% of the world's population, are active Internet users, with 92.6% of users connecting via smartphones. The Internet has revolutionized the way people live and work by changing communication, shopping, chatting, and office work. With the advent of the COVID-19 pandemic, many traditional industries have shifted to online services, such as catering and retail, and cybercriminals are taking advantage of this situation by stealing sensitive data, such as usernames, passwords, and credit card information. Cybersecurity problems are divided into several categories, including man-in-the-middle, denial-of-service attacks, phishing, and malware. Phishing is one of the most common attacks and involves deceiving users into providing sensitive personal information. The number of phishing attacks increased significantly in 2020, with cybercriminals using phishing to obtain sensitive personal information and fraudulently applying for government subsidies.

2. LITERATURE SURVEY

Phishing is a prevalent form of cyberattack that involves sending fraudulent emails or messages to trick recipients into visiting fake websites and providing sensitive information, including usernames, passwords, and credit card numbers, for financial gain.



The phishing life cycle, illustrated in the figure, begins with the creation of a fake website that closely resembles a legitimate one. Cybercriminals often use tactics like misspelled URLs and imitation of web content, such as logos and text, to deceive users into thinking that they are accessing a legitimate website. The attackers usually target form submission pages that require sensitive information, such as login, payment, and find password pages.

The next step in the phishing process involves sending emails or messages that coerce recipients into clicking on a link to the fake website. Cybercriminals may use various methods, such as SMS, voice messages, QR codes, and spoof mobile applications, to spread false information and trick users into clicking on a link. They often use social engineering techniques, such as psychological manipulation, to deceive users.

Once users click on the link, they are directed to the fake website where cybercriminals collect their personal information. The fake website looks like the real website and includes similar logos, user interfaces, and content, making it challenging for users to detect that they are on a fake website. Attackers typically target login, reset password, payment, and renewal personal information pages to collect sensitive data.

In the final step, cybercriminals use the user's real information to steal their account funds. Some attackers may also use stolen data for other criminal activities. Phishing methods have evolved with the growth of the Internet, and attackers now target various forms of online payment phishing. According to the 2020 Internet Crime Report, phishing scams accounted for approximately 30% of cybercrime complaints, causing more than USD 54 million in losses.

Therefore, users need to be cautious and able to distinguish between real and fake web pages. They can use visual tools to identify phishing websites and protect their sensitive information from cybercriminals.

3. DATASET

We have considered **phishing.csv** dataset from Kaggle which has 11054 instances and 30 features.
Dataset description:
There are four forms of functions that we will extract from the URL.

- Address Bar-based Features
- Abnormal Based Features
- HTML and JavaScript-based Features
- Domain-based Features

ADDRESS BAR-BASED FEATURES:

1. Using the IP address: In cases where the IP address is substituted for the domain name in a URL, it can indicate an attempt to illicitly obtain personal information from users. It is not uncommon for the IP address to be converted into hexadecimal code as well.
2. Long URL to Hide the Suspicious Part: To obscure the suspicious part of a URL in the address bar, scammers may employ lengthy web addresses. Research findings indicate that URLs that are 54 characters or longer are categorized as phishing attempts.
3. Using URL Shortening Services "TinyURL": URL shortening is a technique used on the internet to create a much shorter version of a URL that still directs users to the intended webpage. This is achieved through an "HTTP Redirect" on a short domain name that links to a webpage with a long URL.
4. URL's having "@" Symbol: When the URL contains the "@" symbol, the browser usually disregards all characters before the "@" symbol and directs to the actual address that comes after it.
5. Redirecting using "//": If a URL contains "//" within its path, it signifies that the user will be directed to a different website. The position where "//" appears is a crucial factor. If the URL begins with "HTTP", then "//" should appear in the sixth position. However, if the URL uses "HTTPS", then "//" should appear in the seventh position.
6. Adding Prefix or Suffix Separated by (-) to the Domain: Legitimate URLs seldom incorporate the dash symbol. Fraudulent entities, on the other hand, typically append prefixes or suffixes separated by hyphens to the domain name in an attempt to create the impression of a genuine website to users.
7. Sub Domain and Multi Sub Domains: Suppose we possess a link, for instance, http://www.hud.ac.uk/students/. A domain name may comprise country-code top-level domains (ccTLD), which in this example is "uk". The term "ac" is a shorthand for "academic", and when combined with "ac.uk" is known as a second-level domain (SLD), whereas "hud" represents the actual domain name. To establish a guideline for extracting this characteristic, we must initially exclude the subdomain (www.) from the URL. After that, we must remove the ccTLD, if present. Finally, we must count the remaining dots in the URL. If the number of dots surpasses one, we must categorize the URL as "Suspicious" since it has a single subdomain. If the dots are greater than two, it is classified as "Phishing" since it will have multiple subdomains. However, if the URL has no subdomains, we will label the feature as "Legitimate".
8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer): Although HTTPS is an essential element in establishing a website's credibility, it is not enough on its own. To ensure a website's authenticity, one should also evaluate the certificate linked with HTTPS, taking into account the certificate issuer's trustworthiness and the certificate's age. A trustworthy certificate must be at least two years old.
9. Domain Registration Length: Considering the short lifespan of phishing websites, we infer that reputable domains are often purchased in advance for several years. The longest duration for which fraudulent domains were employed was just one year.
10. Favicon: A favicon is a graphical icon that is linked to a particular webpage. Several user agents, including graphical browsers and newsreaders, display the favicon in the address bar as a visual cue of the website's identity. If the favicon is obtained from a domain that differs from the one displayed in the address bar, it is highly probable that the webpage is a Phishing attempt.
11. Using Non-Standard Port: This attribute is beneficial in verifying the availability of a specific service (e.g., HTTP) on a particular server. In order to regulate intrusions, it is preferable to open only the necessary ports. By default, several firewalls, Proxy, and Network Address Translation (NAT) servers block all or most of the ports and only allow the selected ones to be opened. If all the ports are left open, phishers can operate nearly any service they desire, putting user information at risk.
12. The Existence of "HTTPS" Token in the Domain Part of the URL: To deceive users, phishers may include the "HTTPS" token in the domain section of a URL. For an instance, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/ can be used.

ABNORMAL BASED FEATURES:

13. Request URL: The Request URL attribute verifies if the external elements such as images, videos, and sounds present on a webpage are loaded from another domain. In genuine webpages, both the webpage address and the majority of the embedded objects share the same domain.
14. URL of Anchor: An anchor is a feature represented by the tag in HTML. It is analyzed in the same way as the "Request URL" attribute. However, for this feature, the focus is on whether the tags and the website have different domain names. This is similar to the "Request URL" feature. Additionally, if the anchor does not link to any webpage, e.g.:
    <a href="#">
    <a href="#content">
    <a href="#skip">
    <a href="JavaScript ::void(0)">
    , then it's far considered legitimate.
15. Links in <Meta>, <Script> and <Link> tags: Legitimate websites often employ <Meta>, <Script>, and <Link> tags to provide metadata, client-side scripts, and retrieve other web resources, respectively. These tags are usually associated with the same domain as the webpage.
16. Server Form Handler (SFH): SFHs that consist of an empty string or "about:blank" are considered suspicious as there is no proper action taken upon the submitted information. Moreover, if the domain name in SFHs is distinct from the domain name of the webpage, this is an indication that the webpage is suspicious as the submitted information is generally not handled by external domains.
17. Submitting Information to Email: The purpose of a web form is to enable users to input their personal details, which are then forwarded to a server for processing. However, a phisher can redirect this information to their personal email. This can be achieved using a server-side script language, such as the "mail()" function in PHP, or a client-side function like the "mailto:" function.
18. Abnormal URL: The identity of a legitimate website can usually be identified from its URL, which can be obtained from the WHOIS database.

HTML AND JAVASCRIPT-BASED FEATURES:

19. Website Forwarding: Phishing websites can be differentiated from legitimate ones based on the number of times they have been redirected. A legitimate website usually has a maximum of one redirection, whereas a phishing website typically has at least four. Therefore, the number of redirections is an important factor in determining the authenticity of a website.
20. Status Bar Customization: JavaScript can be utilized by phishers to display a fake URL in the status bar, which can mislead users. In order to identify this deceptive tactic, it is necessary to inspect the webpage's source code, specifically the "onMouseOver" event, to determine if it modifies the status bar.
21. Disabling Right Click: Phishers may prevent users from viewing and saving the webpage source code by using JavaScript to disable the right-click function. This method is considered similar to the use of "onMouseOver" to conceal the link. To identify this technique, the webpage's source code should be examined for the event "event.button==2" to determine if the right-click functionality has been disabled.
22. Using Pop-up Window: It is uncommon for a legitimate website to request that users provide their personal information through a pop-up window. Nevertheless, some genuine websites employ this feature to alert users about fraudulent activities or to display a welcome message. However, these pop-up windows do not request any personal information to be entered by the user.
23. IFrame Redirection: IFrame is an HTML element that enables the display of another webpage within the current webpage. Phishers can exploit the "iframe" tag by making it invisible, without any frame borders. To achieve this, they utilize the "frameBorder" attribute, which triggers the browser to produce a visual demarcation.

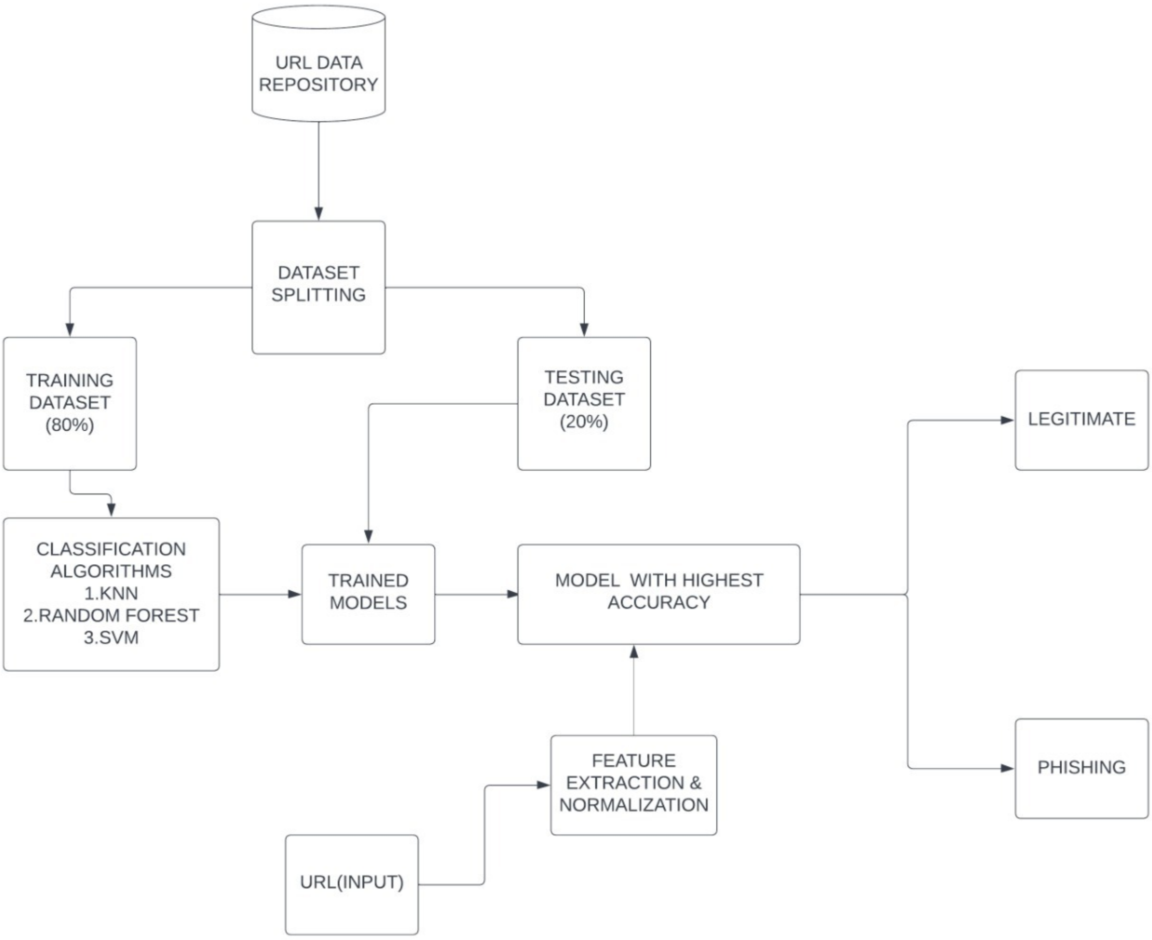
DOMAIN-BASED FEATURES:

24. Age of domain: According to the Whois 2005 database, it is possible to obtain this characteristic. Phishing websites usually have a brief lifespan, while authentic domains have a minimum age requirement of six months.
25. DNS Record: Phishing websites can be identified in two ways: either the claimed identity is not registered in the WHOIS database (Whois 2005), or there are no records found for the hostname (Pan

and Ding 2006). If the DNS record is either empty or not found, the website is considered "Phishing." On the other hand, if there is a DNS record, the website is labeled as "Legitimate."

26. Website Traffic: The popularity of a website can be determined by the number of visitors and pages they visit. Nonetheless, due to their short lifespan, phishing websites may not be acknowledged by the Alexa database (Alexa the Web Information Company., 1996). Websites that are considered legitimate are usually ranked within the top 100,000. A domain is classified as "Phishing" if it has no traffic or is not recognized by Alexa. Conversely, if a domain is recognized by Alexa, it is considered "Suspicious."
27. PageRank: The PageRank is a numerical value between "0" and "1" that aims to assess a webpage's importance on the internet. The higher the PageRank value, the more significant the webpage is. Typically, 95% of phishing webpages have no PageRank value, while the remaining 5% may have a PageRank value of up to "0.2".
28. Google Index: This function examines whether or not an internet site is in Google's index. When a site is indexed by Google, it is displayed in search results (Webmaster resources, 2014). Usually, phishing webpages are merely accessible for a short period and as a result, many phishing webpages may not be found on the Google index.
29. Number of Links Pointing to Page: The number of links that lead to a webpage is indicative of its level of legitimacy, including those originating from the same domain (Dean, 2014). 98% of phishing items had no links pointing to them due to their short lifespan. Conversely, legitimate websites typically have at least two external links pointing to them.
30. Statistical-Reports Based Feature: Several organizations such as PhishTank and StopBadware publish statistical reports on phishing websites periodically, with some reports being monthly and others quarterly. Our study focused on utilizing two types of statistical reports from PhishTank, namely "Top 10 Domains" and "Top 10 IPs". Additionally, we referred to the "Top 50" IP addresses report from StopBadware.

4. SYSTEM ARCHITECTURE



- Data splitting: Dataset splitting is used to make our model generalized. This splitting is performed with a ratio of 90:10.
  - • The 90% portion of the dataset is utilized for model training.
  - • The 10% portion of the dataset is utilized for model testing evaluations.
- Classification of Algorithms:
  1. KNN: It is a powerful non-parametric classification technique that can predict outcomes by identifying similar data points based on their distance from the target and its nearest neighbors. By using distance metrics like the Euclidean distance for continuous data and the Hamming distance for discrete values, k-NN can efficiently calculate the proximity between data points. Since it does not have a training process, the algorithm can quickly adapt to new data, making it ideal for online learning scenarios. While it may not be suitable for real-time applications due to its longer prediction time, k-NN has proven to be effective in various fields such as recommendation systems, image recognition, and anomaly detection. Overall, the k-NN algorithm is a valuable tool in machine learning, offering simple yet powerful solutions for classification tasks.
  2. Random Forest: It is a popular choice in machine learning. One of its key advantages is its ability to handle high-dimensional datasets with a large number of features. It can also deal with missing data, making it a flexible and robust method for handling real-world problems. Moreover, random forests are resilient to outliers and noise in the data, which can be a significant challenge in other machine learning algorithms. Additionally, the ensemble nature of random forests helps to reduce the variance and overfitting, leading to better generalization performance. Random forests also provide feature importance rankings, which can be useful in identifying the most important variables that contribute to the prediction. Overall, random forests are a versatile and effective machine learning technique that can be applied to a wide range of classification and regression problems.

3. SVM: It is a robust machine learning algorithm that operates by plotting each data point in n-dimensional space. SVM constructs a hyperplane to classify two classes of data, which is essentially a separating line. The algorithm identifies the support vectors, the closest points to the hyperplane. It then draws a line connecting these points and constructs a separating line that is perpendicular and bisects this connecting line. To ensure accurate classification, SVM seeks to maximize the margin, which is the distance between the hyperplane and support vectors. However, it may not be possible to separate complex or non-linear data. To address this challenge, SVM employs a kernel trick that transforms the lower-dimensional space into a higher-dimensional space.

- Feature extraction: Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data.
- Normalization: The goal of normalization is to transform features to be on a similar scale. This improves the performance and training stability of the model.
- Model training: It is used to run the input data through the algorithm to correlate the processed output against the sample output.
- Model testing: It is referred to as the process where the performance of a fully trained model is evaluated on a testing set.
- Predict: It refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data.

5. MODULES DIVISION

- Feature Extraction
- Normalization
- Applying Algorithm
- Connecting to GUI(User Interface)

6. METHODOLOGY

- We consider **phishing.csv** dataset from Kaggle which has 11054 instances and 30 features namely UsingIP,LongURL,ShortURL,Symbol@,Redirecting//,PrefixSuffix-,SubDomains,HTTPS,AgeOfDomain,LinksInScriptTags,LinksPointingToPage, etc.
- The dataset taken is already normalized (i.e its features are on similar scale).
- Using train_test_split function imported from sklearn.model_selection, we split the total data into 90% training, 10% testing data.
- First, we train the KNN, Random Forest, SVM (Support Vector Machine) models using the training data.
- Secondly, these trained models are tested using the testing data and the model which produces the highest accuracy is considered.
- Now features are extracted from the URL (input entered in the GUI) and normalized and this is given as input to the selected model i.e the model with the highest accuracy.
- It gives the result to what extent the entered link is safe or unsafe.

7. INPUT

The input given to the models is a vector of normalized features while the input that is entered in the GUI is a URL in string format from which the features are extracted and normalized and given as input for model for predicting whether the URL is phishing or legitimate one.

8. EXPECTED OUTPUT

A real-time system requires a phishing website detection model that can accurately identify such websites while minimizing both false warnings and computational time, which is exactly what this model achieves.

9. CONCLUSION

The objective of this paper is to improve the detection of phishing websites through the application of machine learning techniques. By employing the random forest algorithm, we were able to achieve a detection accuracy of 96.9% with a low false positive rate. Furthermore, our findings indicate that the classifiers perform better when trained with a larger dataset.

10. FUTURE SCOPE

The promising results of this project suggest that there is a future scope for further research in the field of phishing website detection using machine learning technology. The project can be extended to include the testing of other machine learning algorithms and feature selection techniques. Additionally, there is potential to investigate the use of more complex data preprocessing techniques to further enhance the accuracy of the detection method. Moreover, the developed model can be integrated into real-world applications to provide users with an effective defense against phishing attacks.

11. REFERENCES

- https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector?select=phishing.csv
- https://www.kaggle.com/code/akashkr/phishing-url-eda-and-modelling
- https://www.mdpi.com/2504-4990/3/3/34