IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11708

Detecting Phishing Attacks Using Hybrid Learning

Shreetej Sharma[1], Darshan M[1], Prof. Usha C.R[2]

[1]Student, Department of Information Science & Engineering, Global Academy of Technology, Bengaluru, India

[2]Assistant Professor, Department of Information Science & Engineering,

Global Academy of Technology, Bengaluru, India

Abstract: Phishing is a type of cyber-attack in which attackers open deceptive websites that look similar to renowned and legitimate websites in order to steal the user's sensitive information. Numerous ordinary human life activities, including electronic banking, social networks, ecommerce, and so on, have been transferred to cyberspace as a result of the fast growth of communication technologies and worldwide networking. The internet's anonymous, accessible, and unmanaged architecture provides an ideal platform for cyber-attacks. The world has seen an increase in the number of Phishing attacks in the past years. It poses a serious threat to the privacy of individuals and can be used to cause financial theft, and identity theft, and can also cause disruption of an organization. It is of prime importance to detect such attacks and eliminate the risk of being a victim of a phishing attack.

Keywords: AWPG, Hybrid Learning, Link Guard Algorithm, Phishing attack, Website,

1. INTRODUCTION

Phishing is the criminally fraudulent process of attempting to gain sensitive information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in electronic communications. Phishing is a type of cyber-attack in which fraudulent websites are used to steal sensitive user information such as credit card numbers, account login credentials, and so on. Phishing poses direct risks through the use of stolen credentials and indirect risks to institutions that conduct business online through the erosion of customer confidence. Phishing can cause anything from a loss of email access to significant financial loss. Phishing attacks are divided into two types: deceptive phishing and malware phishing.

- Deceptive phishing is any attack in which fraudsters impersonate a legitimate company in order to steal personal information or login credentials from victims. In order to intimidate recipients into following the attackers' instructions, those emails frequently use threats and a sense of urgency.
- Malware-based phishing occurs when an attacker ascribes a malicious computer program masked as helpful to emails, websites, and other electronic documents on the internet.

Phishing attack can be implemented in different ways such as follows:

- **Email to email:** When someone receives an email requesting sensitive information to be sent to the sender.
- **Browser-to-website:** When some misspelled a legitimate web address on a browser and then referred to a phishing website that has a semantic similarity to the legitimate web address.
- **Email-to-website:** When someone receives an email embedded with phishing web address.
- **Website-to-website:** When some clicks on phishing website through a search engine or an online advert.

The detection and mitigation of phishing attacks is a great challenge due to the complexity of current phishing attacks. Machine learning techniques such as J48, Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB) and Artificial Neural Network (ANN) were used to detect the phishing attacks. However, getting good-quality training data is one of the biggest problems in machine learning because data labelling can be a tedious and expensive one. We will be concentrating on browser-to-website and email-to-website phishing attacks in this paper. One must always double-check the URL before opening it to avoid phishing attacks. In this paper, a Hybrid Learning technique called Link Guard is used to classify an URL as legitimate or phishy.

2. LITERATURE SURVEY

For detecting phishing websites, a semi-supervised learning approach known as Transductive Support Vector Machine (TSVM) [4] was proposed. Web page features were initially extracted in order to counteract the disadvantage of phishing detection using Document Object Model technology (DOM). It included the spatial relationship between the subgraphs

DOI: 10.17148/IJARCCE.2022.11708

as well as the colour and grey histograms. The characteristics of sensitive information were examined using a page analysis based on DOM objects. To train classifiers that took into account the distribution information implicitly represented in the significant number of unlabelled samples, TSVM was introduced. The classifier, however, needs to be more accurate.

To detect phishing attacks in internet banking, a new rule-based method [5] has been proposed. To identify the website, this method used two novel feature sets. Four features to assess the page resource identity and four features to ascertain the access protocol of page resource elements are included in the feature sets. An approximate string-matching algorithm was used in the feature sets to determine the relationship between the content and URL of a page. To distinguish between legitimate and phishing web pages, the Support Vector Machine (SVM) was used to process the feature sets. SVM, however, frequently needs a lot of memory to classify web pages.

For phishing detection in Iranian e-banking, a fuzzy-rough hybrid system [6] was introduced. The fuzzy-rough hybrid system combined rough sets-based data mining with fuzzy logic. The dimensionality of the data was decreased using a rough attribute reducer. The input data was then transformed into linguistic variables using fuzzy logic. It produced the rules that are used to identify phishing websites. But the effectiveness of phishing detection is significantly impacted by the membership function used in fuzzy logic.

For the purpose of identifying phishing websites, a brand-new fast associative classification algorithm (FACA) [7] was put forth. The FACA worked well for supervised learning. It combined combining classification and association rule mining process. FACA used a technique for vertical mining that is Finding every common item set and a new one is a challenge. An estimation technique was used to categorise the websites as websites that are trustworthy and phishing websites. Though, the method's computational complexity is high.

For the purpose of identifying phishing websites, a brand-new fast associative classification algorithm (FACA) [7] was put forth. The FACA worked well for supervised learning. It combined combining classification and association rule mining process. FACA used a technique for vertical mining that is Finding every common item set and a new one is a challenge. An estimation technique was used to categorise the websites as websites that are trustworthy and phishing websites. Though, the method's computational complexity is high.

A novel framework [9] based on dynamic evolving networks and reinforcement learning was proposed for the detection of online phishing emails. The phishing attacks were caught by this framework in the online mode. A Feature Evaluation and Reduction (FEaR) algorithm was created in the novel framework to explore the new behaviour as to rank a chosen list of features. The FEaR algorithm extracted significant features from the subsequent email while dynamically adjusting the number of them. The classification model was built around a neural network (NN), and a dynamic NN using reinforcement learning (DENNuRL) was created to let the NN evolve dynamically and create the best NN capable of addressing the phishing attack issue. To improve the offline dataset, more datasets will be added.

Based on Optimal Feature Selection and Neural Network, an effective phishing detection model [10] was proposed. (OFS-NN). This model addressed the issue of overfitting in neural network as a result of numerous pointless and minor influences the neural network's features A Feature in OFS-NN FVV, or Validity Value, was developed to analyse the impact. Detection of sensitive features on phishing websites. Then, based on the new FVV index, a formula was created to choose the best options from the phishing websites. A neural network was trained with a subset of features to recognise website phishing scams. However, as more features are added, the efficiency of the classifier.

3. METHODOLOGY

The approach is divided into two parts, and each part's output is an input to the next part as shown in the proposed framework-

The first part is based on data collection, processing of data sets, and URLs feature extraction. We consider different heuristic features in the structure of URLs, ranging from a generic social engineering feature, lexical feature in the URL, multiple alphabets, and phishing target brand name. The feature vector is constructed with important features to model our classifiers. The second part is based on the classification of data set using link guard to evaluate our approach.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11708

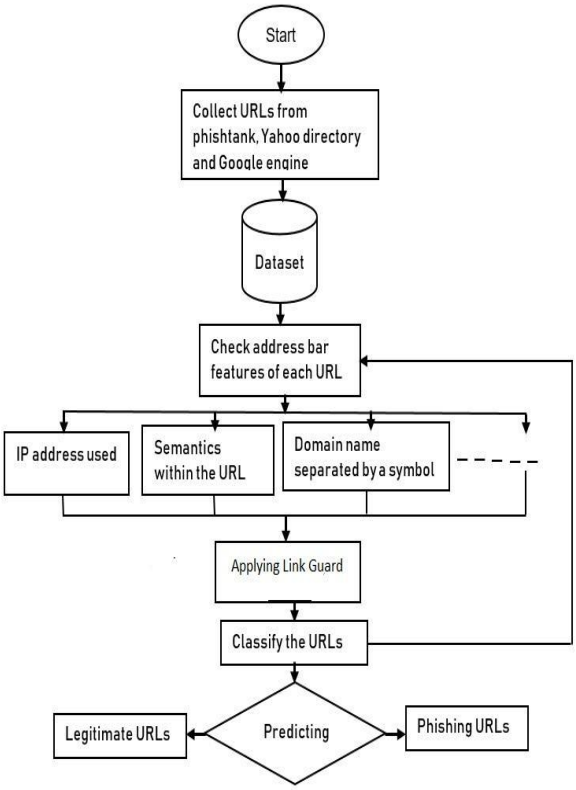

Figure 1: Flowchart of proposed model

4. PROPOSED MODEL

The proposed model is designed to classify the weblinks provided by the user as legitimate or phishy using a Hybrid Approach. The user inputs URL, which is processed and required features are extracted from it, which is then analysed by our developed model which uses Random Forest and IBK algorithms for classifying an URL as legitimate or phishy based on its extracted features.

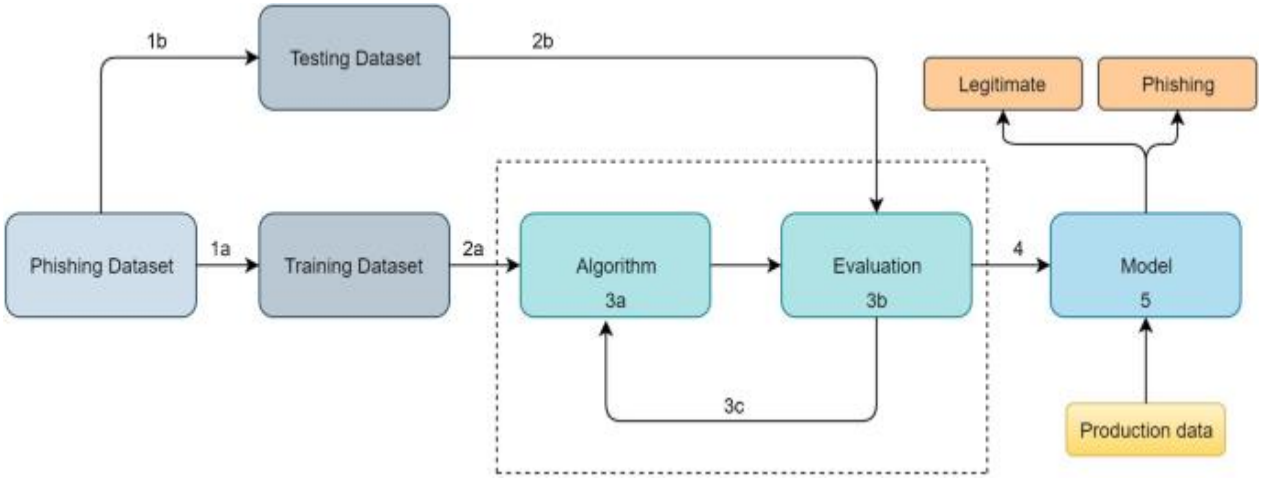

Figure 2: Basic Structure of proposed model

The model is trained using datasets obtained from different open-source repositories, like Phishtank, APWG etc... Random Forest and IBK algorithms provide efficient classification and our model also classifies live URLs.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11708

5. IMPLEMENTATION

5.1 Domain and Architecture

1. Cyber-Security- Cyber Security is the practice of Protecting computers, mobile devices, Servers, electronic Systems, networks, and data from malicious attacks. It is also known as Information Security (INFOSEC) or Information Assurance (IA), System Security. The first cyber malware virus developed was pure of innocent mistakes. But cybersecurity has evolved rapidly because of the impeccable increase in the cybercrime law field on the Web.

Advantages of Cyber Security:

i. Cyber security safeguard business: The most significant benefit is that the best in IT security cyber security solutions can give your company comprehensive digital protection. This allows flexibility of accessing the internet by the staff and ensuring the safety and protection from possible threats and risks.

ii. Protects Personal Information: In this age of a digitally-driven world, one of the most valuable commodities is personal information. If a virus is able to collect personal information about your employees or customers, it is quite likely that it will be sold or used to steal their money.

iii. Protects and Enhances Productivity: Viruses infecting your systems and network will result in functioning resulting in the almost impossibility of further working. In effect, this will cause downtime in work for your staff and wastage additionally bringing the entire company to a halt.

iv. Prevents crashing of websites: If you're a small business, you're probably hosting your own website. If your system is infected, there's a good risk your website will be forced to go down. This means that not only will you incur losses due to missed transactions, but you will also run the risk of losing trust from your clients, and some viruses may cause long-term damages to your systems.

v. Support Your IT Professional: Typically, a good security system equips your organization and employees with the best tools, techniques, and assistance in combating cyber-attacks and criminals.

2. Hybrid Learning- HYBRID MACHINE LEARNING is a progress of the MACHINE LEARNING work process that perfectly unites different computations, processes, or procedures from equivalent or different spaces of data or areas of usage fully intended to enhance each other. As no single cap fits all heads, no single MACHINE LEARNING procedure is appropriate for all issues. A couple of strategies that are extraordinary in managing boisterous data anyway may not be prepared for dealing with high-layered input space. Some others could scale pretty well on high-layered input space anyway may not be good for managing sparse data. These conditions are a fair motivation to apply HYBRID MACHINE LEARNING to enhance the contender procedures and use one to overcome the deficiency of the others. The open doors for the hybridization of standard MACHINE LEARNING methodologies are ceaseless, and this ought to be workable for every single one to collect new combination models in different ways.

3. Random-Forest is a well-known supervised learning machine learning algorithm. It can be applied to ML problems involving both classification and regression. It is based on the idea of ensemble learning, which is a method of combining various classifiers to address complex issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the given dataset and averages them to increase the dataset's predictive accuracy. Instead of relying on a single decision tree, the random forest uses predictions from each tree and predicts the result based on the votes of the majority of predictions. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4. IBK- K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint. The algorithm's learning is:

✓ Instance-based learning: Here we do not learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data.

✓ Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance.

✓ Non -Parametric: In KNN, there is no predefined form of the mapping function.

5.2 Requirement Planning

Tools-

1) Anaconda Spyder Anaconda Navigator could be a desktop graphical computer programme (GUI) included in Anaconda® distribution that permits you to launch applications and simply manage condo packages, environments, and channels without using command-line commands. Navigator can seek for packages on Anaconda Cloud or in an exceedingly local Anaconda Repository. so as to run, many scientific packages depend upon specific versions of other packages. Data scientists often use multiple versions of the many packages and use multiple environments to separate

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.11708

these different versions. The command-line program condo is both a package manager and an environment manager. This helps data scientists make sure that each version of every package has all the dependencies it requires and works correctly. Navigator is a straightforward, point-and click thanks to work with packages and environments with no need to type condo commands in a very terminal window. It is accustomed find the packages required, install them in an environment, run the packages, and update them – all inside Navigator. Spyder may be a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a novel combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the information exploration, interactive execution, deep inspection, and delightful visualization capabilities of a scientific package. Furthermore, Spyder offers built-in integration with many popular scientific packages, including NumPy, SciPy, Pandas, I Python, QtConsole, Matplotlib, SymPy, and more [18].

2) VS Code the Visual Studio integrated development environment could be a creative pad that you just can use to edit, debug, and build code, then publish an app. An integrated development environment (IDE) may be a feature-rich program that may be used for several aspects of software development. Over and above the quality editor and debugger that almost all IDEs provide, Visual Studio includes compilers, code completion tools, graphical designers, and plenty of more features to ease the software development process. Visual Studio offers a set of tools that enable you to simply create cloud-enabled applications powered by Microsoft Azure which may be wont to configure, build, debug, package, and deploy applications and services on Microsoft Azure directly from the IDE [19].

Libraries

1) Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.
2) Sklearn was initially developed by David Cournapeau as a Google summer of code project in 2007. Later, in 2010, Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and Vincent Michel, from FIRCA (French Institute for Research in Computer Science and Automation), took this project at another level and made the first public release (v0.1 beta) on 1st Feb. 2010. Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy SciPy and Matplotlib.
3) Beautiful Soup is a Python library that is used for web scraping purposes to pull the data out of HTML and XML files. It creates a parse tree from page source code that can be used to extract data in a hierarchical and more readable manner.
4) NumPy could be a Python package. It stands for 'Numerical Python'. it's a library consisting of multidimensional array objects and a group of routines for processing of array. Numeric, the ancestor of jumpy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created numpy package by incorporating the features of Numarray into Numeric package. There are many contributors to the present open source project. Using NumPy, a developer can perform Mathematical and logical operations on arrays, Fourier transforms and routines for shape manipulation and Operations associated with algebra.
5) Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution. Pandas is built on top of two core Python libraries—matplotlib for data visualization and NumPy for mathematical operations. Pandas acts as a wrapper over these libraries, allowing you to access many of matplotlib's and NumPy's methods with less code. For instance, pandas.plot() combines multiple matplotlib methods into a single method, enabling you to plot a chart in a few lines.

6. RESULT

The users are provided an interface where they may type in the URL they intend to verify. The outcome is based on the extracted characteristics from the input URL as well as the open-source repository URL datasets. After examining each of these factors, the trained model gives the user a result: Legitimate or Phishy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11708

Figure 3:User Interface

Input URL's	Expected Result	Actual Result	Correctness
www.google.com	Legitimate	Legitimate	Yes
https://www.irctc.co.in/	Legitimate	Legitimate	Yes
https://www.cricbuzz.com/	Legitimate	Legitimate	Yes
https://www.rbi.org.in/	Legitimate	Legitimate	Yes
www.pakgovakdfn.in	Phishing	Phishing	Yes
http://www.internetbanking@login	Phishing	Phishing	Yes
www.indloanfree.com	Phishing	Phishing	Yes
http://www.paypal.com.attacker. site/login	Phishing	Phishing	Yes

7. CONCLUSION

Phishing detection on social network platforms is thought to be a modern, expanding technique that focuses on achieving greater detection accuracy levels. The Link Guard Algorithm is a novel classification approach presented in this paper that is highly effective in detecting the authenticity of an URL and informs the user to avoid any unintentional data leak. Additionally, it protects users against harmful or unwanted links in Web sites and Instant messaging, ensuring safe user presence online.

REFERENCES:

1. R. S. Rao, and S. T. Ali, "PhishShield: a desktop application to detect phishing webpages through heuristic approach," Procedia Comput. Sci. vol. 54, 2015, pp. 147-156.
2. O. A. Akanbi, I. S. Amiri and E. Fazeldehkordi, "Cyber security: A machine learning approach to phishing," 2014
3. T. Chin, K. Xiong, and C. Hu, "Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking," IEEE Access, vol. 6, 2018, pp. 42516-42531.
4. Y. Li, R. Xiao, J. Feng, and L. Zhao, "A semi-supervised learning approach for detection of phishing webpages," Optik, vol. 124, no. 23, 2013, pp. 6027-6033
5. M. Moghimi, and A. Y. Varjani, "New rule-based phishing detection method," Expert syst. appl., vol. 53, 2016, pp. 231-242.
6. G. A. Montazer, and S. ArabYarmohammadi, "Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system," Appl. Soft Comput., vol. 35, 2015, pp. 482-492.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11708

7. W. E. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," Appl. Soft Comput., vol. 48, 2016, pp. 729-734.
8. K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," Inf. Sci., vol. 484, 2019, pp.153-166
9. S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," Deci. Support Syst., vol. 107, 2018, pp. 88-102.
10. E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, "OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network," IEEE Access, vol. 7, 2019, pp. 73271- 73284.