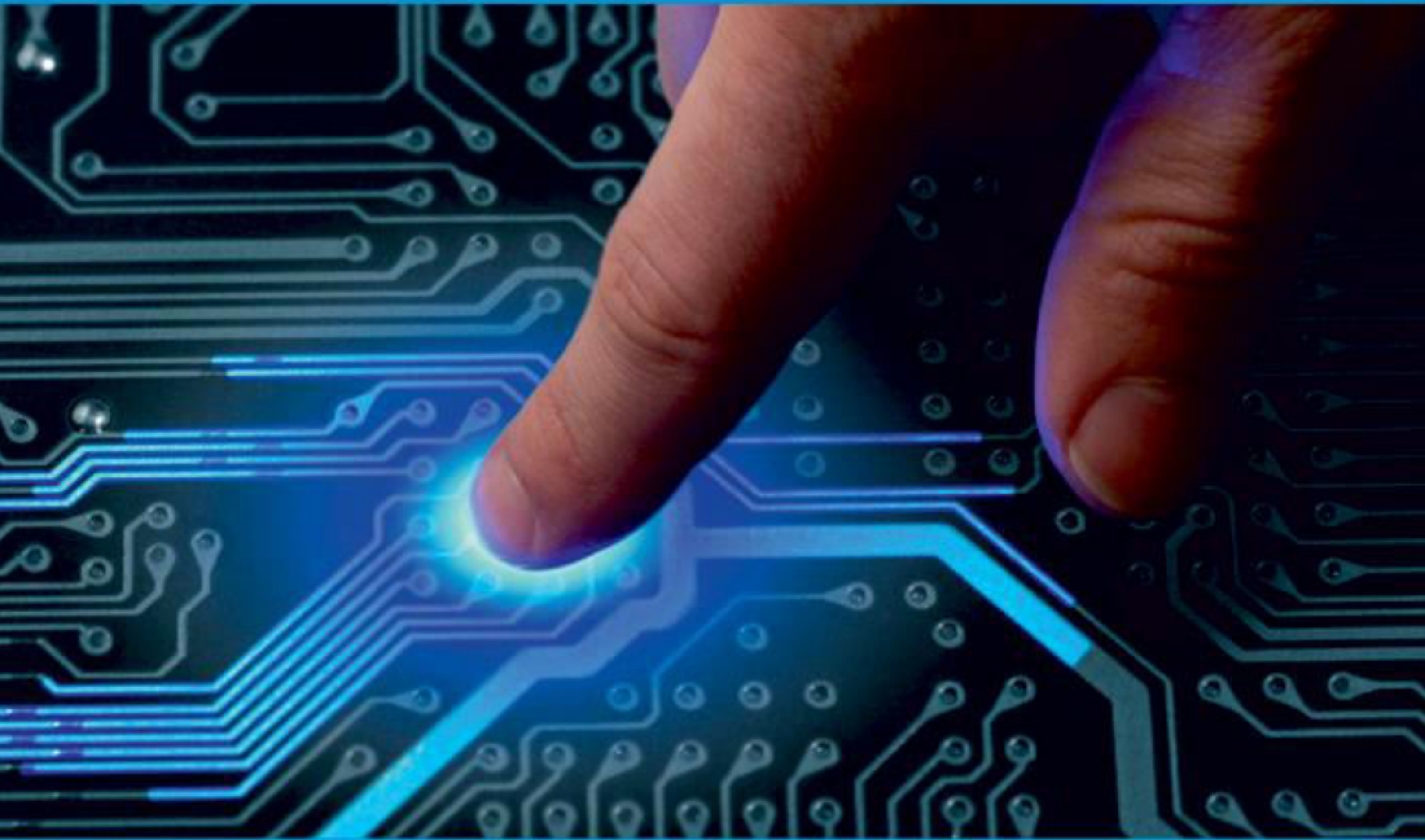
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| DOI: 10.15680/IJIRCCE.2023.1104108 |

Phishing URL Detection Using Machine Learning

Rohit Kakade, Omkar Khose, Sugam Mahajan, Amisha Patil, Sheetal Gawande

Dept. of Information Technology, Pillai College of Engineering, New Panvel, India

Dept. of Information Technology, Pillai College of Engineering, New Panvel, India

Dept. of Information Technology, Pillai College of Engineering, New Panvel, India

Dept. of Information Technology, Pillai College of Engineering, New Panvel, India

Dept. of Information Technology, Pillai College of Engineering, New Panvel, India

ABSTRACT: Due to the rapid rise in the internet usage the threat of cybercrime has been increased drastically. Among this phishing is one of the most commonly used tools to conduct cyber-attacks. It is a way to obtain sensitive information from the users, and thereafter the information is exploited. The phisher's aim is to acquire sensitive information like name, password, credit card, debit card numbers, other bank account details and much more. The proposed system will detect the phishing URLs by extracting and analysing various features of legitimate and phishing URLs. The proposed system will use machine learning algorithms to do so. It will compare the accuracy of these algorithms and will select the best machine learning algorithm by comparing accuracy rate, true positive, true negative, false positive and false negative of each algorithm.

I. INTRODUCTION

Today, the Internet has become an important partof human life, and it is hard to imagine life without the Internet. The I nternet has brought convenience and various opportunities in people's lives, and it has also opened a void for cybercrim e, a new form of crime.Cybercrime is criminal activity against or using computers, computer networks or network equi pment.Most cybercrimes are committed by cybercriminals or hackers who want to make money. There are many types of cyber crime such as phishing, fraud, ransomware, malware, IoT hacking etc. As mentioned earlier, phishing is a very popular cybercrime today.Today, many people around the world havefallen victim to phishing due to lack of awarenes s and increased use of the internet.

In URL phishing, a user is tricked into clicking on a maliciouslink that appears legitimate.
Once the user clicks on the link and takes an action, all key information about the user that could have anyvalueisavaila ble.Users' financial information is one of the main targets of attackers. Once they have
this information, the user's entire financialsituation can be exploited.

The reason phishing hasbecomesopopularcompared to other cybercrimes is that it is so easyto trick users. It is a type of social engineering usedbycriminals to steal data, infect computer networks,computer software, cell phonesoranyotherel ectronic device.Hackers create fake websites that looklikelegitimatewebsites.However,thisfakewebsite is slightly differ ent.People thought theywere on the right site and accidentally clickedthrough and engaged.The biggest differencebetwe en the proposed system and the existingsystem is that we have created a Google Chromeextension that will notify the u ser when there is aphishing website, very user-friendly.

II. LITERATURE SURVEY

Purbay M and Kumar D [1] researched various ML methods to detect various URL components using machine learning and deep learning technologies. They addressed various supervised learning methods for the identification of phishing URLs based on WHOIS properties, lexicon, traffic rank, page rank information and page importance properties. They studied how the volume of different training data can create a huge impact in the accuracy of classifiers. The research includes Support Vector Machine (SVM), K-NN, random forest classification (RFC) and Artificial Neural Network (ANN) techniques for the classification.

| DOI: 10.15680/IJIRCCE.2023.1104108 |

Gandotra E and Gupta D [2] have presented the output with and without the functionality selection a study of ml algorithms is carried out in their study. 30 experiments were carried put on a phishing dataset including 4898 phished and 6157 benign web pages. Multiple ML algorithms were used to get a better outcome. A method for selecting functions is subsequently employed to increase model performance. In their work Random forests algorithm achieved the highest accuracy. The outcome of the experiment shown us that using machine learning algorithms with selection approach can boost the performance, effectiveness and reliability of the classification models for detecting phishing sites.

Hung le et al [3] proposed URL Net, a CNN-based deep-neural URL detection network. They stated that Bag of Words (Bow) usage by current methods such as features have suffered some essential limitations, such as the lack of automatic feature extraction, not able to detect sequential concepts in a URL string and the failure of unseen features in real—time URLs. They have developed a method i.e., CNNs and Word CNNs for charactering and configuring the network. On top of that, they suggested advanced techniques that were highly effective in handling terms that are uncommon, a common problem which exist in malicious URL detection tasks. This method can allow URL Net to identify embedding and use sub word information from invisible words during testing phase.

Kumar J .et al [4] researched that how effectively phishing URLs can be classified in the set of URLs which consists of benign URLs. They discussed characteristics engineering, randomisation and the extraction of characteristics with the help of statistical analysis and host based analysis. For the comparative study, various classifiers and they consistent results across. A convenient approach was argued by the authors which can remove functionality from URLs with the help of simple standard words. A better result was obtained with the help of additional features. The dataset which was used in the study have some older URLs. Thus, there is a high chance of lack of performance.

K. Chiew [5] have contributed a new feature selection framework, named as the Hybrid Ensemble Feature Selection (HEFS). They used a cumulative distribution function gradient called feature vector, which challenges the randomness of words/characters in a URL.

Y. Haung [6] proposed a capsule-based neural network for phishing URL detection. They implemented two capsule layers namely classification capsule layer and primary capsule layer. Classification capsule layer used averaged outputs, dynamic routing algorithm and squashing function from all the branches. On the other hand, in the primary layer, they extracted accurate features from shallow features which was generated by the former convolution layer

III. SYSTEM ARCHITECTURE

A. Proposed System Architecture

To circumvent the drawbacks of conventional phishing detection techniques, we suggest using machine learning. As a result of the abundance of data on phishing attack patterns, the problem of phishing detection is a prime choice for deploying a machine learning solution. The main concept is to create a model that can be used to categorise a given web page as either a phishing page or a valid web page in real time using a machine learning algorithm on a data collection of existing phishing websites. In order to stop phishing efforts, our objective is to create learnt patterns in a software solution that can be quickly delivered to end users. To do this, we chose to build a machine learning algorithm from the ground up usingJavaScript and use it to create a Chrome extension.

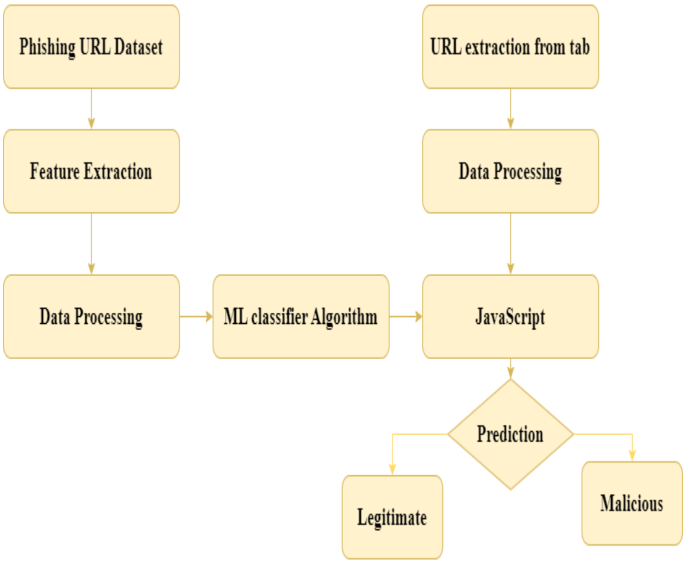


Fig 1 : Proposed System Architecture

B. Dataset

The data was obtained from the Kaggle. It has 11,055 Links, 6157 of which are phishing attempts, and 4898 of which are genuine. 16 features are present in each occurrence. Either a value of 1 (not phishing) or a value of -1 (phished URL) can be found in the result. The values 1, -1, or 0 are present in each column, which denotes a feature. '1' indicates a fully phished URL, '0' indicates a partially phished URL, and '-1' indicates a benign URL.
The dataset includes following features: IP address, Long and Short URL, Alphanumeric URL, having forms submitting to a blank page.

C. Machine Learning Algorithm

- **Support vector machine**

Support vector machine is one of the most powerful algorithms in machine learning technology. Here, a data item is put as a point in n-dimensional space after that a separate line for classification of two classes is constructed, the separating line known as hyper plane. SVM looks for the nearest points called as support vectors and once it finds those nearest points it draws a line which connects them. Support vector machine then creates a separating line which bisects and perpendicular to the connecting line. It is important that the margin is maximum which can classify the data correctly. The margin is a distance between hyper plane and support vectors. However, in real life scenario it is almost impossible to separate complicated and non-liner data, to resolve this particular problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

- **Neural networks algorithms**

The biggest advantage of using Neural networks algorithms is that it can recognize patterns without being explicitly programmed. It is a set of algorithms designed to help machines recognize patterns. They consist of a group of interconnected nodes. Neural networks process information with the help of mathematical or computational models. These neural networks are usually non-linear in nature, which helps them to model complex relationships between data input and output so they can efficiently find patterns in a dataset. Other benefits of using neural network are as follows: Storing of information is done on the entire network, meaning that the neural network can keep functioning even if some information is lost from one part of the neural network. They are cost as well as time effective once they are trained with a quality data set, as they take a shorter time to analyse data and present results. Training with high quality dataset also ensures that they are less prone to errors. They provide high accuracy.

- **Random Forest Algorithm**

Random forest algorithm is another one of the most powerful algorithms in machine learning technology as it based on a concept of decision tree algorithm. RF algorithms creates forest with high number of decision trees. As a result, is gives high detection accuracy. The trees are created with the help of bootstrap method. In this method features and samples of dataset are selected randomly with replacement to construct a single tree. It will choose best splitter among randomly selected features for the classification and similar to decision tree algorithm; RF algorithm also uses Gini index and information gain methods by which it can find the best splitter. The process will get going on until random forest creates n number of trees. All the trees in the forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally, the target with highly voted prediction will be considered as the final predication.

D. Chrome Extension

The JavaScript-based browser plugin uses a model that was learned in Python. The content.js file collects the URL's features and uses the ML algorithms to determine whether it is a phished URL or not. Background.js connects the content.js to the frontend, and manifest.json includes the extension's meta data. The Chrome extension's front end appears as an alert pop-up with a 'OK' button that indicates if the URL is phished or not.
Functions implemented in the content.js are:

IP Address():to check IP in url
Long URL():Its checks for the length of an URL
Tiny URL(): It again checks for the length of an URL if URL is less than 20
Alphanumeric URL():It checks for @ character in an URL
Redirecting URL():It checks whether the URL is a redirecting URL or not
Hyphen URL():It checks whether the URL contains hyphen or not
Multidomain URL():It checks whether the URL belongs to a multi domain website or not
Illegal HTTPs URL():It checks current website's URL contains an "https" string after the "//" characters.
StatusBarTampered():It checks whether the status bar has been tampered with on the current webpage.
Iframe Present():It checks if there are any iframes present in the current web page.
Img from different domain():It checks whether the images on the current webpage are loaded from different domains or not.
Anchor from different domain ():It checks whether the hyperlinks (anchor tags) on the current webpage are linking to different domains or not.
Form Action Invalid():It checks whether the action attribute of all the <form> elements on the current webpage are valid or not.

IV. RESULT AND ANALYSIS

Machine learning methods have been imported using the Scikit-learn tool. The data set is split into training and testing sets in the following ratios: 70:30. Each classifier is trained using a training set, and the performance of the classifiers is assessed using a testing set. The accuracy score, false negative rate, and false positive rate of classifiers have all been calculated in order to assess their performance.

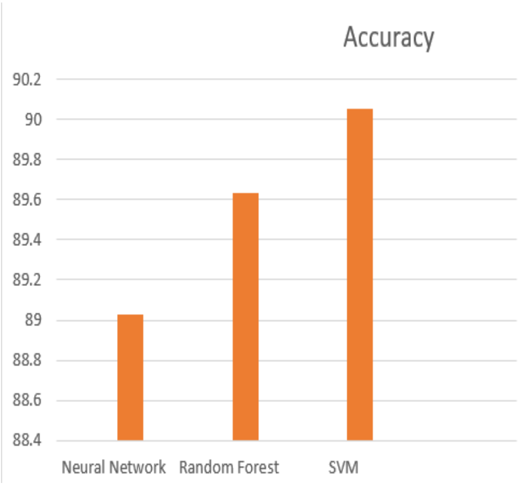
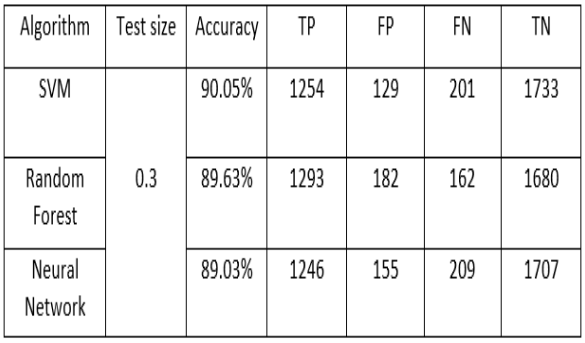


Fig 2: Accuracy of Algorithum

TABLE 1



Algorithm	Test size	Accuracy	TP	FP	FN	TN
SVM	0.3	90.05%	1254	129	201	1733
Random Forest		89.63%	1293	182	162	1680
Neural Network		89.03%	1246	155	209	1707

V. CONCLUSION

This paper uses machine learning technologies to detection of phishing websites. Using the Neural Network, Random Forest and SVM algorithms. The accuracy of algorithms are 89.03%,89.63% and 90.05% respectively in which SVM shows the highest accuracy with low false positive rate.

In the future, hybrid technology will be utilised to more accurately identify phishing websites, using both the blacklist method and the SVM algorithm of machine learning technology.

REFERENCES

[1] Purbay, M. and Kumar, D., 2021. Split behavior of supervised machine learning algorithms for phishing URL detection. In *Advances in VLSI, Communication, and Signal Processing: Select Proceedings of VCAS 2019* (pp. 497-505). Springer Singapore.

[2] Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", *Algorithms for Intelligent Systems*, Springer, Singapore, 2021, 10.1007/978-981-15-8711-5_12.

[3]Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", *Conference'17*, Washington, DC, USA, arXiv:1802.03162, July 2017.

[4] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," *2020 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.

[5] Chiew, K.L., Chang, E.H. and Tiong, W.K., 2015. Utilisation of website logo for phishing detection. *Computers & Security*, *54*, pp.16-26.

[6] Huang, Y., Qin, J. and Wen, W., 2019, October. Phishing URL detection via capsule-based neural network. In *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)* (pp. 22-26). IEEE.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details