IJIEMR Transactions, online available on 26th Apr 2022. Link

:http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 04

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 04/04

Title **PHISHING WEBSITES DETECTION USING RANDOM FOREST**

Volume 11, SPL ISSUE 04, Pages: 40-48

Paper Authors

Mr A. Chandra Mouli , M.N.V. Rohit, M. Teja, SK. Aakhil4 N. Sirisha

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

PHISHING WEBSITES DETECTION USING RANDOM FOREST

Mr A. Chandra Mouli [1], M.N.V. Rohit[2], M. Teja[3], SK. Aakhil[4] N. Sirisha[5]
[1]Associate professor, Department of CSE, PSCMRCET, Vijayawada, Andhra Pradesh
[2,3,4,5] Student, Department of CSE, PSCMRCET, Vijayawada, Andhra Pradesh
achandramouli@pscmr.ac.in[1] rohitmajety@gmail.com[2] maddinenimaniteja@gmail.com[3]
akhil071109@gmail.com[4] sirishaneelamb2@gmail.com[5]

Abstract

Emails that contain Uniform Resource Locators, also known as URLs, pose significant risks to businesses because they have the potential to compromise an organization's credentials in addition to the network security of the organisation through spear-phishing and general phishing operations directed at the employees of the organisation. An important academic topic with real-world ramifications is the identification and classification of URLs that link to harmful websites. This topic focuses on the identification and category of URLs that lead to hazardous websites. An organisation may protect itself by screening incoming emails and the websites that its employees are visiting by using an appropriate machine learning model to determine the maliciousness of URLs contained in emails and web pages. This screening can be done for both incoming and outgoing emails. Filtering like this may be used as a defence mechanism against cyberattacks. In this study, we compare the performance of popular deep learning framework models, such as Fast.ai and Keras-TensorFlow, with the performance of traditional machine learning algorithms, such as Random Forest, CART, and kNN, across CPU, GPU, and TPU architectures. We find that traditional machine learning algorithms perform better than deep learning framework models when it comes to performance. Random Forest, CART, and kNN are a few examples of the algorithms that fall under this category. We use the dataset ISCX-URL-2016, which is accessible to the general public and can be obtained here, to show the performance of the models across binary and multiclass classification tasks. When it came to the identification and categorization of dangerous URLs, we discovered that the Random Forest, Keras-TensorFlow, and Fast.ai models all performed in a manner that was comparable and had accuracies that were more than 96 percent. The Random Forest model, on the other hand, is the one that is recommended because of the time, performance, and complexity limitations that are involved. When we compared the results of using all of the features that were provided in the dataset to those of rating the features and using feature selection methods, we found that using the top five to ten features produced the best results. This was in comparison to using all of the features that were provided in the dataset.

Keywords: Malicious URLs, Phishing URLs, Deep Learning, Web Security, Machine Learning

I. INTRODUCTION

Phishing is a type of online fraud in which con artists pose as legitimate businesses on the web in order to obtain private information from innocent people by means of e-mail, text, online ads, or other types of interaction that actually occur on the internet. These con artists do this in order to obtain money or personal information from the victims. One of the most widespread kinds of fraudulent activity conducted online is known as phishing[1].

The majority of the time, this purpose is accomplished by including a link that gives the impression that it will direct you to the website so that you can fill out your information. This is the case the great majority of the time. However, despite the fact that the website in question gives off the impression of being quite authentic, it is in fact a fraud, and any information that someone enters into it will be sent directly to the unscrupulous persons who are behind the scam.

The term "phishing" is a play on the word "fishing" due to the fact that lawbreakers use a fake "lure" (an email, website, or advertisement that appears to be legitimate) in the desperate hope that consumers will "bite" and provide the information that the criminals have requested, such as credit card numbers, account numbers, passwords, usernames, or other valuable information. The term "phishing" is a play on the word "fishing." Hackers often make use of "lures" that are not what they seem to be (an email, website, or advertisement that appears to be legitimate)[1].

Phishing has evolved into a wide variety of extremely complicated activities since it was originally described in 1987, and the behaviour itself has become more ubiquitous over the course of that time period. Phishing was first defined in 1987. 1987 was the year that saw the first definition of what we now know as phishing.

This assault is always finding new ways to take advantage of vulnerabilities in the system as a result of the rapid pace at which information technology is advancing[1].

II. TYPES OF ATTACKS

Here is a rundown of eleven of the most typical types of phishing [1]:

Standard Email Phishing – Arguably the kind of phishing that is the most well-known to the general public, this sort of attack includes the attempt to collect sensitive information by way of an email that seems as if it was received from a legitimate company. This is not an attack that is directed at a specific target; rather, it is one that is designed to be carried out on a vast scale.

- Malware Phishing – Utilizing the same strategies as email phishing, this form of attack encourages targets to click on a link or download an attachment so that malware can be installed on the device.
- Email Phishing – This form of the attack sends an email with the intention of obtaining confidential information from a target. Phishing through email is an attack technique

in which the victim is encouraged to provide their email address on a website. Phishing attacks of this kind are currently the ones that are seen the most often.

- Spear Phishing – While the majority of phishing efforts include spreading their net over a wide area, spear phishing is an attack that is highly targeted, well studied, and is often directed at corporate executives, public figures, and other lucrative targets.
- Sending potentially malicious short URLs to people who use smartphones is the practise of smishing, which is also known as SMS-enabled phishing. It's common practise to mask these URLs as account reminders, award announcements, or political messaging of some kind.
- Search Engine Phishing is a kind of online attack in which hackers establish fraudulent websites with the purpose of gaining personal information as well as direct payments from victims. This type of attack is known as "search engine phishing." It's possible that these websites will show up in the results of organic search queries or as paid advertisements for terms that are commonly searched.
- Voice phishing, also known as vishing, is a type of social engineering in which a scammer calls a plaintiff and displays as a delegate of a respected firm, such as a financial institution or a government body, in an effort to trick them into divulging personal information, such as the details of their banking or credit card accounts.
- Pharming is a kind of phishing that is more technologically sophisticated than traditional phishing and takes advantage of the internet's domain name system. DNS poisoning is another name for this issue (DNS). Pharming is the practise of redirecting legitimate web traffic to a sham website without the user's knowledge or agreement, often with the goal of collecting private information.
- A dishonest participant gains access to a user's email account, adjusts an existing email by replacing a valid link, attachment, or another element with a malware one, and then sends the altered email to the victim's address book in order to promote a virus through the system.
- Attacking with a Person Placed in the Middle of You An eavesdropper will watch the conversation that is taking place between two parties who are ignorant of the intrusion when an assault is carried out utilising a man in the middle technique. These sorts of attacks are often carried out by putting up fake public WiFi networks in public areas like coffee shops, shopping malls, and other locations of a similar kind.

The man in the middle is able to phish for information or spread malware to devices once the connection has been established.

- The term "Business Email Compromise" (BEC) refers to a scam in which a fraudulent email is sent out, making it look as if it came from someone working at or linked with the firm of the target and asking for immediate action, such as sending money or purchasing gift cards. BEC is an abbreviation for "Business Email Compromise." It is estimated that this tactic was responsible for around half of the financial losses sustained by corporations in 2019 as a direct result of cybercrime.
- Malvertising is a kind of phishing that publishes adverts that, on the surface, seem to be safe but, in reality, incorporate malicious code. This is accomplished via the use of digital advertising technologies, which are widely used nowadays.

II. LITERATURE SURVEY

Phishing is a method of obtaining people's private information, as stated by the authors of the research [2,] who define it as the practise of creating fake websites and emails to fool individuals into giving their information. Phishing is a method of obtaining people's private information, as stated by the authors of the research. According to the people who carried out the investigation, phishing is a technique that is used to get the personal information of other people. [There must be other citations for this] A key problem is the fact that individuals are unable to carry out their operations via the internet as a result of phishing. This prevents people from using the internet. The identification of websites that are used for phishing is a task that falls squarely on the shoulders of the worldwide community of internet users. This is as a result of the fact that the existence of phishing websites has the ability to have a substantial effect on the results of financial transactions that are carried out online. Researchers have recently shown a greater interest in the RF method of intelligent machine learning as a result of the lightning-fast speed at which it classifies data as well as the high degree of accuracy it maintains throughout the process. This is due to the fact that the RF method was developed in the 1990s. This research endeavour focused on the problem of phishing websites, which resulted in the development of a machine learning model with the objective of locating connections between the features and extracting those correlations from straightforward and useful rules. The results of this study turned out to be quite important when it came to the construction of the model. In order to create a classifier model that is intelligently and independently capable of identifying websites that are used for phishing, the researchers who worked on this study[2] made use of datasets that were available to the general public. This allowed the researchers to identify websites that are used for phishing. We were successful in accomplishing this goal thanks to the utilisation of the data. In terms of classification accuracy, the area under the curve (AUC), and the F-measure, the RF classifier that was created has a performance that, all things considered, is quite great. The findings of the classifications were analysed, and this was one of the conclusions that was reached. Our study's findings also revealed that RF is a classifier that is more dependable than the others, in addition to having greater levels of accuracy and speed than the others. It takes a very short amount of time to run Random Forest, and in contrast to the other classifiers, it is able to recognise websites that are being employed in phishing schemes. Random Forest is the only game that has this functionality.

III. METHODOLOGY:

In this study, we compare the performance of popular deep learning framework models, such as Fast.ai and Keras-TensorFlow, with the performance of traditional machine learning algorithms, such as Random Forest, CART, and kNN, across CPU, GPU, and TPU architectures. We find that traditional machine learning algorithms perform better than deep learning framework models when it comes to performance. Random Forest, CART, and kNN are a few examples of the algorithms that fall under this category. We use the dataset ISCX-URL-2016, which is accessible to the general public and can be obtained here, to show the performance of the models across binary and multiclass classification tasks. When it came to the identification and categorization of dangerous URLs, we discovered that the Random Forest, Keras-TensorFlow, and Fast.ai models all performed in a manner that was comparable and had accuracies that were more than 96 percent. The Random Forest model, on the other hand, is the one that is recommended because of the time, performance, and complexity limitations that are involved. When we compared the results of using all of the features that were provided in the dataset to those of rating the features and using feature selection methods, we found that using the top five to ten features produced the best results. This was in comparison to using all of the features that were provided in the dataset.

A. Data set:

The use of the internet for unlawful activities has long ago established itself as a main arena for activities of this kind to take place online. Within this specific industry, URLs are utilised as the most common form of transit. In order to address these issues, members of the security community have focused their efforts on developing procedures that, in the majority of cases, require blacklisting malicious URLs.

Even supposing that this tactic is successful in shielding clients from potentially hazardous websites, it is still just addressing a small part of the overall issue at hand. The freshly produced malicious URLs that sprouted up in a significant number all across the world wide web acquired an early lead in this race. This was the situation in the vast majority of cases. In addition to this, reputable websites that have a high ranking on Alexa and have won the trust of their consumers are more likely to transmit fake

URLs that are known as defacement URLs. These fake URLs may be identified by their use of the term "defacement."

We have investigated a lightweight approach to the identification and classification of harmful URLs according to the type of attack they launch, and we have demonstrated that lexical analysis is an effective and efficient method for the proactive detection of these URLs. In addition, we have investigated a lightweight approach to the identification and classification of harmful URLs according to the type of attack they launch. We also research the influence that obfuscation techniques have on malicious URLs, with the intention of finding which kind of obfuscation techniques are more successful against specific types of malicious URLs.

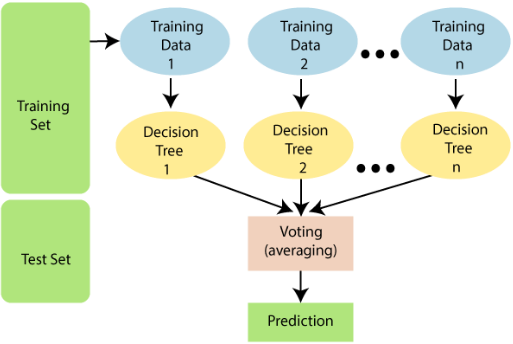
B. Random Forest classifier

The umbrella term of supervised learning, which is a more complete categorization, is where one of the most well-known machine learning algorithms, Random Forest, may be found. This category is far more extensive. Both classification and regression are types of machine learning tasks that are examples of the kinds of activities that can possibly benefit from their application. It is predicated on the idea of ensemble learning, which refers to the practise of including a number of distinct classifiers in order to solve a difficult issue and enhance the operational capabilities of the model.

The expression "The term "Random Forest"[3] refers to a method of categorization that, as its name suggests, "contains a number of decision trees on diverse subsets of the provided dataset and takes the average to increase the predicted accuracy of that dataset." Random Forest is also the name of a classification method.

" The Random Forest approach "takes the average in order to improve the accuracy of the forecasting for that dataset "according to the official account of what it is. The random forest model does not rely on a single decision tree; rather, it reviews the forecast from each tree in the forest and calculates the final output based on which tree's prediction won the majority of votes. This is done in order to ensure that the model is as accurate as possible. Within the context of the random forest model, the single decision tree does not play any role at all.

When there are a greater number of trees in a forest, it is not only feasible to achieve a higher degree of accuracy, but it also makes it possible to avoid the issue of overfitting, which may occur when there are insufficient numbers of trees[3].



The fact that training with random forest takes significantly less time compared to training with other algorithms was a significant consideration that contributed to our choice to use it.

Despite the high degree of precision with which it anticipates output, it can nevertheless handle a large dataset in a timely manner without sacrificing efficiency.

Additionally, it is able to retain its accuracy even in situations in which a sizeable portion of the data is lacking.

IV. IMPLEMENTATION:

If the data are not already in the required format, you will need to either convert them or make sure that they are already in a format that can be accessed easily.

Please describe all of the irregularities that may be seen without much effort, as well as any data points that are missing that may be required in order to gather the data that you want.

Build a model that can be used for machine learning.

Establish the template for the bare minimum of the quality that you want to achieve.

It is necessary to train the machine learning model that is data driven.

It is recommended that test data be utilised in order to offer an insight into the model.

Now that we have both sets of data, all that remains for us to do is do a comparison between the performance metrics of the test data and the data that the model predicted.

In the case that it does not live up to your expectations, you have the choice of either keeping your data up to date, improving your model in accordance with the new information, or turning to a different approach to data modelling.

At this stage, you are going to be tasked with doing an analysis on the data that you have gathered in order to provide an accurate report.

The first thing you need to do is import all of the required libraries.

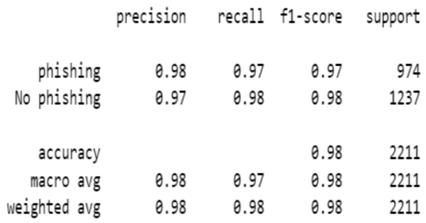
As a second step, you will need to import and print the dataset.

Adjusting the coefficients of the random forest regressor to fit the dataset is the third step.

The fourth step is to generate an up-to-date prediction of the findings.

Imagining how things will turn out is the fifth phase in the process.

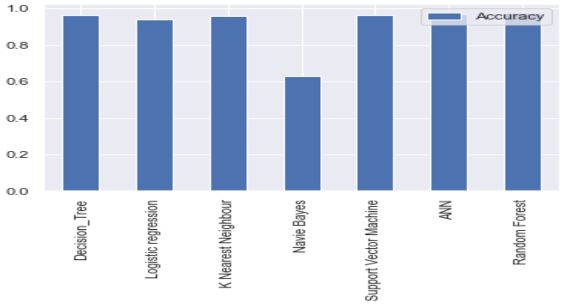
V. RESULTS:



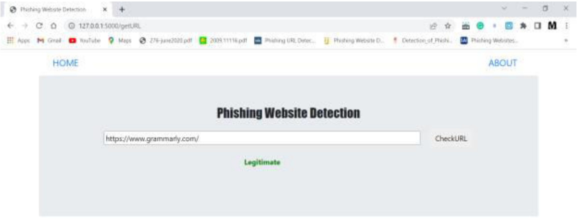
```
              precision    recall  f1-score   support

    phishing       0.98      0.97      0.97       974
 No phishing       0.97      0.98      0.98      1237

    accuracy                           0.98      2211
   macro avg       0.98      0.97      0.98      2211
weighted avg       0.98      0.98      0.98      2211
```



Random Forest has obtained high accuracy as compared to other algorithms.

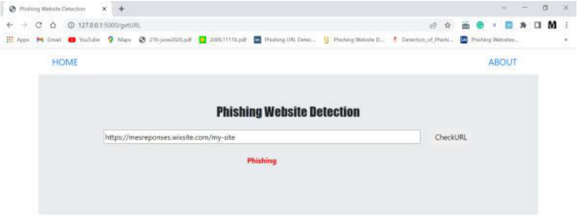


Phishing Website Detection

https://www.grammarly.com/

Legitimate

Legitimate



Phishing Website Detection

https://mesreponses.wixsite.com/my-site

Phishing

Phishing

VI. CONCLUSION

A phishing assault detection system supported by machine learning (ML) was steered. Phishing website Detection could be a system that uses machine learning approaches to observe phishing websites. To observe phishing, the investigation employs a spread of ways. The milliliter systems were fed customary datasets of phishing assaults from kaggle.com. to look at and select datasets for classification and detection, 2 common machine learning techniques, specifically call tree and random forest, are used. The elements of the datasets were known and classified as mistreatment principal part analysis (PCA). the website classified mistreatment as DT, and therefore the categorization was done as mistreatment RF. Finally, a confusion matrix was created to match the 2 algorithms performance. RF had less variance and will handle the matter of over-fitting. The random forest tree had a 97% accuracy rate. employing a convolution neural network, we are going to forecast phishing assaults from a logged dataset of attacks within the future (CNN).

REFERENCES

[1]"Email Phishing, Vishing & Other Types of", *Webroot.com*, 2022. [Online]. Available: https://www.webroot.com/in/en/resources/tips-articles/what-is-phishing. [Accessed: 26-May- 2022].

[2]A. Subasi, E. Molah, F. Almkallawi and T. J. Chaudhery, "Intelligent phishing website detection using random forest classifier," 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2017, pp. 1-5, doi: 10.1109/ICECTA.2017.8252051.

[3]www.javatpoint.com. 2022. *Machine Learning Random Forest Algorithm - Javatpoint*. [online] Available at: <https://www.javatpoint.com/machine-learning-random-forest-algorithm> [Accessed 26 May 2022].

[4]GeeksforGeeks. 2022. *Random Forest Regression in Python - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/random-forest-regression-in-python/> [Accessed 26 May 2022].