National College of Ireland

DETECTION OF WEB-BASED PHISHING URL USING MACHINE LEARNING, WHITELIST AND BLACKLIST APPROACH

MSc Research Project
MSc in Cybersecurity

Sandhya Rathore

Student-ID:- x20199805

School-of-Computing
National-College-of-Ireland

Supervisor- Liam McCabe

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sandhya Rathore

Student ID: X20199805

Programme: MSc in Cybersecurity **Year:** 2021

Module: Research in Computing

Supervisor: Liam McCabe

Submission Due Date:- 16/12/21……

Project Title:- DETECTION OF WEB-BASED PHISHING URL USING MACHINE LEARNING, WHITELIST AND BLACKLIST APPROACH

Word Count:- .....4241... **Page Count:-**.....20.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sandhya Rathore

Date: 15/12/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Can Support Vector Machine be employed to successfully implement a listing approach for improving accuracy and efficiency of detecting Phishing Websites?

Sandhya Rathore
x20199805

Abstract

Phishing cases have only been rising with each passing day, 36% of the data breaches are Phishing-based as per the report submitted by Verizon for the Year 2021. Individuals are deceived using various methods including emails, instant messages and advertisements on the website to click or browse to a malicious website. A lot of money transactions, around 13 percent of total E-commerce sales are occurring on daily basis involving Internet banking and Online-shopping, users are being directed to malicious websites which appear similar to the identical websites. Deceptive phishing can be done in various ways for obtaining the credentials of a user. Anti-phishing tools have been developed for detecting phishing attacks still the rate of phishing cases is rising every year.

Due to the increasing number of phishing websites which is 640% in the year 2020 revealed by Webroot, there is a chance of Cyber-attacks taking place. To prevent the cyberattacks taking place, experts have suggested to improve the existing anti-phishing methods. It is suggested to use multiple anti-phishing techniques at once to detect the phishing attacks. Recently, due to the covid break individuals and companies are facing a greater number of phishing attacks, they are being directed to fake websites and asked to enter their credentials. Cybercriminals are finding a greater number of opportunities to do the phishing attempts. Kernel-methods are popular which includes the Support Vector Machine algorithm (SVM). SVM can classify the given data into linear and non-linear. The solution suggested for this project is the combination listing-based methods and SVM to detect malicious URLs. [1]

Contents

1 Introduction

Hackers can attempt to do phishing by stealing sensitive data by deceiving a victim into opening an email, instant messaging, or text message by impersonating a reputable entity. The sensitive data can be credit card details, personal data of a user or data related to his/her company. The phishing attack can be done manually or there are tools available which makes the job easier or a hacker might start the attack with help of an automated tool and complete it manually.

Around the year 1995, the first phishing email is thought to have been sent. The "phishing" word was first used in the year 1994, where a group of people used AOL for manually grabbing the credit card details of the user who were unaware. In the year, the same group of people developed an automated software AOHell, which will retrieve the credit card details on its own. Teenagers have started using the same kind of attack methods leading to increased security risk. [2]

New methods have been developed by hackers such as email phishing, HTTPs phishing etc. to obtain sensitive data of people who are unaware of the phishing methods. Hackers are responsible for the third-party malicious applications which are presently being used, 99.9% of mobile malware are hosted by third party app stores according to the statistics shown by Purplesec. Some of the malicious applications have also been developed by ethical hackers to conduct penetration testing on the application. The malicious actors try to find how to use a given application for their benefits.

The Attacker sends an email to a victim, that emails seem to be a legitimate email. The victim unknowingly clicks on the malicious link and visits the phishing website. The Victim enters their details on the website, those details are used by the hacker to access the victim’s legitimate account.

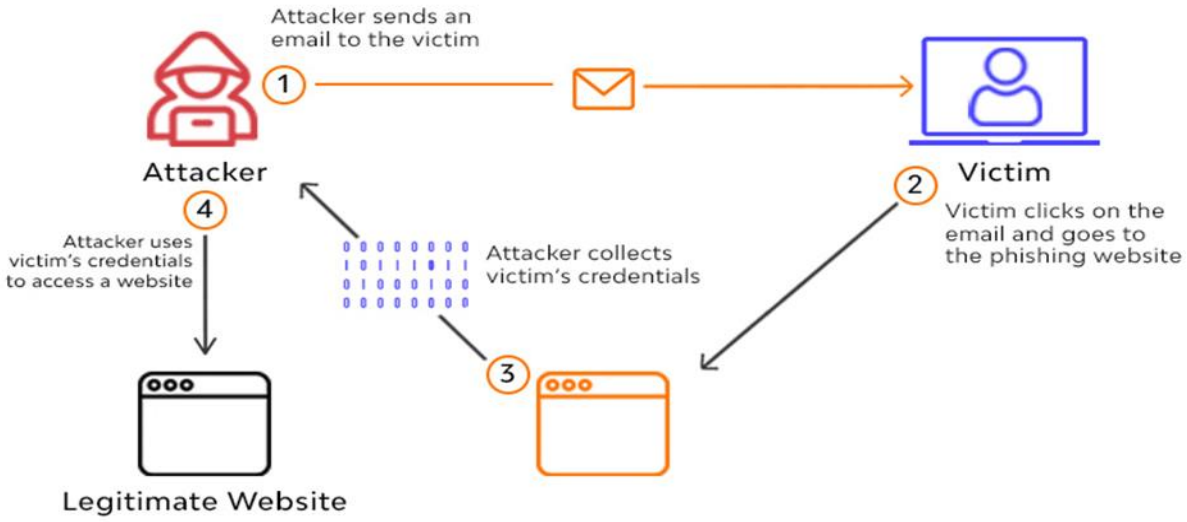


Figure 1: Phishing attack, attacker trying to steal credentials of a victim

1.1 Research Question

On the basis of research done, the solution is of using machine learning algorithm, due to its accuracy in detecting phishing. List based approach is also widely in use to detect malicious URL's. Therefore, the project will be using a combination of Support Vector Machine and List-based approach to detect phishing URLs.

The research question is as follows:

Can Support Vector Machine be employed to successfully implement a listing approach for improving accuracy and efficiency of detecting Phishing Websites?

2 Related Work

The number of phishing websites are only rising and they are affecting the most to the individuals who are not aware of Cybercrime methods, due to this reason phishing detection methods need to be improved. For improved phishing detection, it is suggested using multiple techniques at once which includes combining machine learning algorithms. Cyber-criminals

and hackers tend to attempt phishing attacks since the mid-1990s when hackers imitated as employees of AOL, used messaging to steal user's credentials and accessed their accounts, suggested by the Verizon company. This is the reason why a combination of a machine learning algorithm with a method such as blacklisting and whitelisting is required.

Below are some of the research studies evaluated for identifying the best anti-phishing technique.

2.1DIFFERENT MACHINE LEARNING APPROACHES

Users are the weakest point in the security chain, suggested in one of the articles by TechRepublic. Phishing mainly focuses on the weak points of humans. Phishing is a concern and it cannot be prevented by using only one technique. Hence, it is considered that using multiple techniques will be a stronger solution to detect phishing.

In a research paper "Anti-phishing system using LSTM and CNN", Mr. Yazhmozhi V.M illustrates how Convolution Neural Networks (CNN) can be applied to URLs for the detection of a malicious webpage. This technique does not involve feature engineering, In this CNN make use of its hidden layer for the extraction of information from a given URL. A combination of LSTM and CNN can be used for a large dataset, this combination is proven to be 96% accurate as compared to the other anti-phishing models.

In the research paper "A Novel Ensemble Machine Learning Method to Detect Phishing Attack" Mr Abdul Basit has identified that the combination of different machine learning algorithms (including Random Forest, K-Nearest etc) at once is proven to be more accurate as compared to the previous research done earlier. Going through the above two mentioned research papers, it was identified that SVM is a new approach for detecting phishing.

2.2 PHISING DETECTION ACCURACY

Artificial Intelligence (AI) strategies are crucial for today's phishing detection methods. Although, there are some disadvantages as well while using AI, including false alarms and sometimes they cannot find out the phishing methods which are in use by the attackers.

One of the research papers "Different types of Phishing Attacks and Detection Techniques: A review" by Dr. Sunil N Pawar identifies that combining multiple approaches is not reliable as the whole focus of the approach (which includes designing, assessing) is to determine the accuracy of the model. The evaluation factor does not match the standard of real-world applications. Although, the main aspect of a security device is that it should be efficient at the same time giving the right results.

A combination including URL and HTML features are used multiple methods can be ineffective as the focus is more on achieving phishing detection rate as compared to the evaluation which does not align with real world requirement. The main requirement for a security device is that it should be working well in overall terms including efficiency, accuracy, evaluation and it should be deployable.

As per the reports submitted through APWG and Phish tank, The phishing websites were expected to rise by the year 2020. Studies have been conducted for bringing up this issue, this includes studying the existing phishing webpages, how they are using a web page and the

methodologies involved. It was identified that Listing method can be combined with SVM as there are trained datasets to cross-check the working of the model.

2.3 SUPPORT VECTOR MACHINES

The main aim of support vector machine is to define a hyperplane that will categorize data points correctly. There are many kinds of hyper-planes from which one can choose to classify two different datasets. The fundamental part is to find the largest distance between the data points from two different classes. If the greatest margin is considered at first, it will give some flexibility for classifying newer data points. [3]

By determining the best hyperplane to distinguish the two groups, an existing dataset was used to train the classifier Support Vector machine, which further predicted if a given website was legitimate or malicious. [4] Using the property of Support Vector Machine, the URLs could be classified into legitimate and malicious for the anti-phishing model.

2.4 COMBINATION OF MULTIPLE TECHNIQUES AND IDENTIFYING RESEARCH GAPS

Day by day phishing cases have been rising, for detecting phishing different techniques have been carried out. There are many preventive methods including machine blacklisting, still there are phishing cases. [5]

From the research paper "A Novel Ensemble Machine Learning Method to detect Phishing Attack" it was identified that the combination of support vector machine and listing methods has not been used. Various machine learning algorithm combinations, Such as RF Classifier, SVM and ANN algorithm have been used already. The use of Artificial-intelligence; URL and HTML feature extraction with the machine learning algorithms.

3 Research Methodology

The Research methodology for this tool, will be using is TDSP (Team Data Science Process lifecycle). This lifecycle was created with data science projects in mind that are shipped as part of intelligent apps, for example- applications which identify and predict diseases, Personalizing Healthcare recommendations. For predictive analytics, these apps use machine learning or AI algorithms. This technique can also be beneficial for exploratory data science or improvised analytics projects.

The below figure is a visual representation of the Team Data Science Process lifecycle.[6]

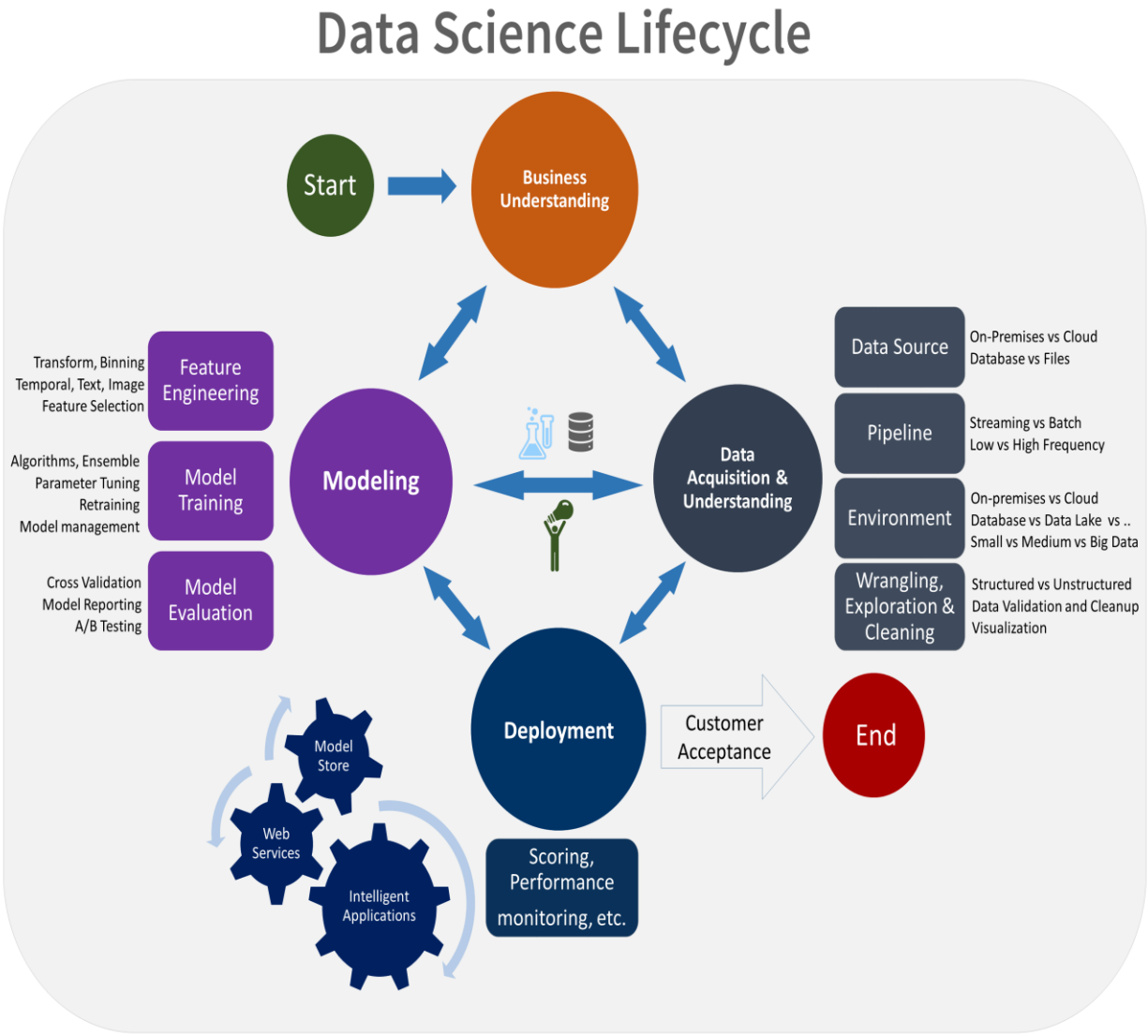


Figure 2 Various stages in Team Data Science Process Cycle

3.1 UNDERSTANDING BUSINESS REQUIREMENTS:

A combination of Support vector machine algorithm and Chrome-extension will be used to detect a malicious URL. The SVM model was trained with a dataset, using 16 URL and DOM-based features-extraction.

1. isIPInURL() – If an IP address is present in the URL
2. isLongURL() – Checks if the length of the URL is more than 75 characters
3. isTinyURL()- checks if the URL length is less than 20 character
4. isAlphaNumericURL() – checks for '@' in the URL
5. isRedirectingURL() – if '//' is present multiple times in an URL
6. isHypenURL()- If '-' is present with the URL name
7. isMultiDomainURL()- Domain name should have top-level domain, country-code and second level domain
8. isFaviconDomainUnidentical()- Checks if there are links loaded from other domains
9. isIllegalHttpsURL()- If multiple 'https' is their in the URL string
10. isImgFromDifferentDomain() – If images on the web-page are linked to other domain
11. isAnchorFromDifferentDomain() – It will detect if the links are from other domain
12. isScLnkFromDifferentDomain() – If scripts are linked to other domains
13. isFormActionInvalid() – It will detect blank forms submitted
14. isMailToAvailable() – Identifies anchor tag including mailto
15. isStatusBarTampered() – It checks if the status bar has been manipulated
16. isIframePresent() – It checks if iframes are present in the DOM

The tool evaluates the URLs by assessing it on the basis of 16 features mentioned above. After, predicting if the URL is malicious or safe, the dataset file is present in a structured manner (1, -1,0). Every website has a set of characteristics that indicate if it is real or not. Machine learning processes can use data as an input.[7] The above 16 features can be used for feature extraction and help the Support Vector Machine algorithm to classify the URLs as Malicious or Legitimate.

3.2 DATA ACQUIISTION AND UNDERSTANDING

The test file has been trained using the below datasets to learn the machine learning model.

UCI DATASET: It includes of databases, domains and data generators which are used to test the machine learning algorithms. David Aha and fellow PhD students at UC Irvine launched the archive as an ftp repository in 1987. [8]

PHISHTANK DATASET: PhishTank is a phishing prevention website. Phishtank was discovered in the year 2006, it was an outgrowth of OpenDNS by entrepreneur David Ulevitch. The platform provides group-based phishing detection system. The users can submit questionable urls and the other users can then vote if the given url is malicious or not. [9]

MALCRAWLER DATASET: The dataset has been compiled with the help of a web crawler named MalCrawler. The dataset includes JavaScript code, raw URL, the extracted parameters. It can be used for supervised as well unsupervised learning. [10]

3.3 MODELING

In this step, different types of machine learning algorithms were studied Four of these algorithms have been uncovered to improve accuracy in detecting Phishing Websites, as shown below.

1- SUPPORT VECTOR MACHINES: Support the Vector Machine is a supervised machine learning technique that can be used in data mining to accomplish classification tasks. Hyper planes are used to construct boundaries between the two or more classes. The usage of a hyper plane sets boundaries between the classes. The hyper plane is chosen in such a way that it separates the classes while also maximizing the margin (distance from the nearest point) among them. After the hyper plane is built, a boundary between the classes is established. Any data point can now be assigned to a class by determining which class it belongs to.[11]

2- RANDOM FOREST: Random Forest is a supervised machine learning technique that may be used to accomplish data mining tasks such as regression and classification. It is a classification technique that is based on an ensemble of techniques. It employs a number of classification trees (such as decision trees) before presenting the final result.

This approach works by randomly generating a large number of classification trees. These trees are generated by combining multiple samples from the same dataset and employing different types of features at different times. As a result, all of the trees are generated at random using multiple subsets of the same dataset, and the features are chosen at random for each tree. Random Forest assures that, unlike decision trees, it does not overfit the data by doing so. Each tree can be classified now by obtaining the results of each tree and then assigning it to the class with the most trees.[11]

3- ARTIFICIAL NEURAL NETWORKS: In the field of Information technology, artificial neural network comprises of hardware or software which are arranged as per the neurons present in brain. It is a type of deep learning technology which also fall under the category of artificial intelligence.[12]

4- LONG SHORT-TERM MEMORY (LSTM): It uses artificial neural network for its deep learning architecture. LSTM comprises of feedback connections which is not the case for feedforward neural networks. It can make use of datapoints (for example photos) as well as data streams (example-video) [13]

3.3.1 TRAINING AND TESTING

The structured dataset was divided into 70 % training and 30 % testing. The results are evaluated with the help of classifiers Neural Networks, Random Forest, Support Vector Machines to get the results in terms of Accuracy and Precision etc. The machine learning algorithm also assesses the efficiency of the train and test datasets separately using model predictions. The performance of all models is evaluated using the given parameters. Using these results, it is proved that SVM is the most accurate machine learning algorithm for the classification of phishing website.

3.4 DEPLOYMENT

The deployment of the anti-phishing tool depends on the project requirements. For this one, Support vector machine has been implemented in the Anti-phishing chrome extension which is the GUI to display the output (If the URL is legitimate or malicious). With the machine learning method, the dataset file will be predict at real time, for increasing the detection accuracy. This combination has been concluded by going through Research papers and their recommendations for a better anti-phishing system. There are datasets which provide a list of blacklisted URLs. For this tool, dataset will be used to identify the blacklisted and safe URLs.

4 Design Specification

This section includes about the applications required to run the project and also the working flow of the project.

4.1 PROJECT REQUIREMENT: The below software needs to be installed for the anti-phishing tool to run properly:

- Google Browser: In Developer Mode, this browser is used to add Chrome-Extensions. This Extension will serve as a user interface for the Anti-Phishing program.

- Python: Python3.8 is used to create this project. As a result, Python3.8 is highly suggested for proper application functioning.

- Visual Studio Code and Jupyter Notebook: This IDE is used to test the response by running the.py scripts locally.

4.2 SOLUTION ARCHITECT:

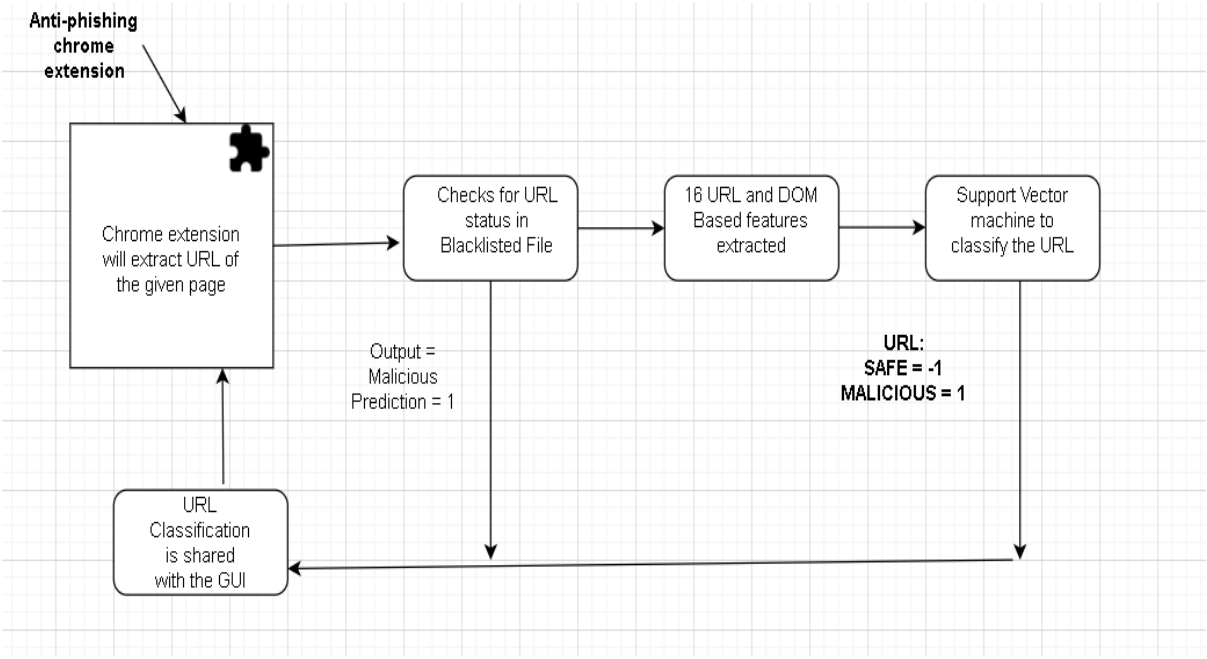


Figure 3: Block diagram for the proposed solution

Here, the chrome extension extracts the URL of a given page, it will cross-check if the URL is Malicious with the Blacklist file. If the URL is found in the Blacklist file. The result is displayed in the Chrome GUI as the URL being malicious (Prediction = 1). The process of feature extraction is carried out where 16 URL and DOM based features are extracted. The URL prediction depends on how effectively the process of feature extraction has been carried out. The data is then passed to the Support Vector Classifier which predicts whether the given URL is safe (-1) or malicious (1). This process repeats for every new URL.

4.3 URL FEATURE

In this Project, 16 features are extracted in total. Below table specifies phishing detection rules for each feature of extraction. If the prediction is -1 it is safe and prediction = 1, it is malicious.

FEATURES	RULE
1. isIPInURL()	If there IP address in Domain, result will be 1
2. isLongURL()	Url length > 54, then prediction = 1
3. isTinyURL()	Tiny url, prediction = 1
4. isAlphaNumericURL()	If an alphanumeric character is present, prediction = 1
5. isRedirectingURL()	prediction = 1
6. isHypenURL()	If Hyphen is present in the URL, prediction = 1
7. isMultiDomainURL()	prediction = 1
8. isFaviconDomainUnidentical()	Favicon from external domain, prediction = 1
9. isIllegalHttpsURL()	prediction = 1

10. isImgFromDifferentDomain()	If image from different domain, prediction = 1
11. isAnchorFromDifferentDomain()	<**ahref=>** greater than 31%, prediction = 1
12. isScLnkFromDifferentDomain()	totalCount-identicalCount)/totalCount)>=0.17, prediction = 1
13. isFormActionInvalid()	form[action*=""]').length<0, prediction = 1
14. isMailToAvailable()	('a[href^=mailto]').length>0), prediction = 1
15. isStatusBarTampered()	a[onmouseover*='window.status']").length>=0) || (document.querySelectorAll("a[onclick*='location.href']").length>=0), prediction = 1
16. isIframePresent()	Using Iframe = 1 else -1, prediction = 1

Table1: Feature extraction, list of 16 features used to classify the URL(Legitimate/malicious)

4.4 TRAIN AND TEST MODEL

For training and testing purpose, the dataset is acquired which consists of all malicious and legitimate URLs, which are obtained after feature extraction, the malicious URL are labelled as '1' and safe URL are labelled as '-1'. The train and test dataset are divided in the ratio **70:30 ratio.** The below diagram represents how each model was evaluated using training and testing files.

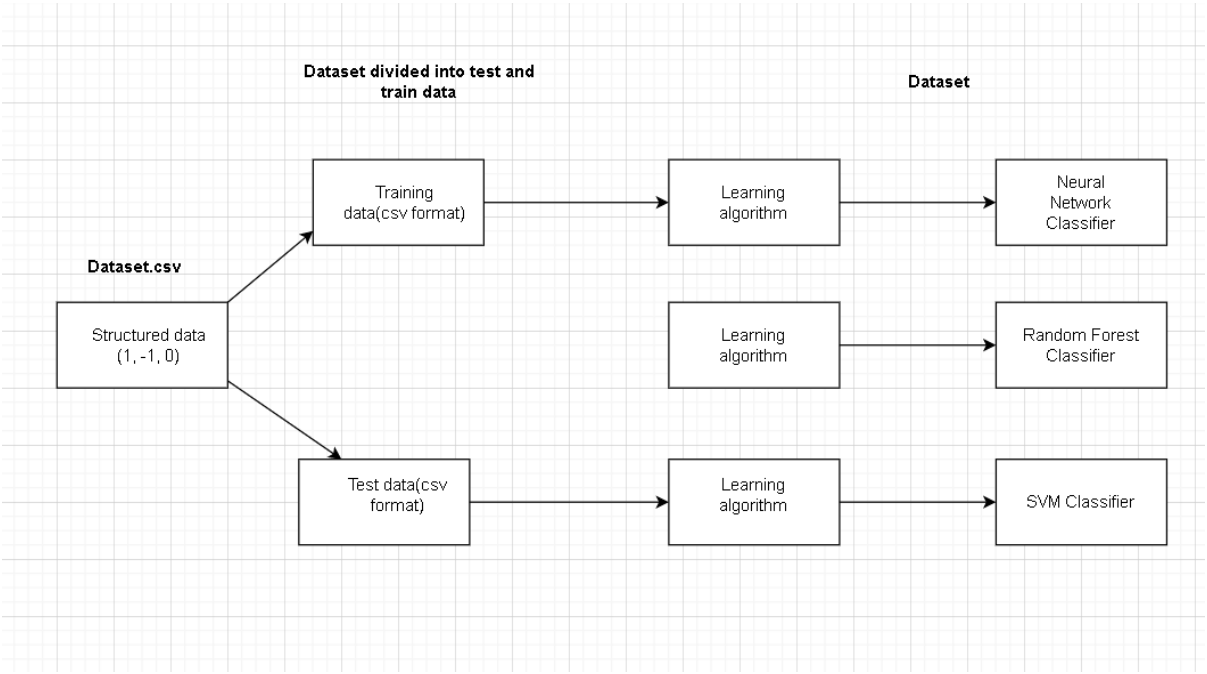


Figure 4: Train and test model

4.5 ALGORITHMS

For training and testing, Classifier algorithms was used which is included in the Python Scikit Learn library:

- The RF Algorithm utilizes RandomForestClassifier.
- SVM Algorithm makes use of SVC.
- For Neural Network Algorithm, MLPClassifier is used.

4.6 MODEL EVALUATION:

The model accuracy and F1-score are calculated using learnt metrics based on the model's predictions. It also generates a confusion matrix based on model prediction that includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Model Accuracy, Precision, Recall, and F1-score are all calculated using these characteristics to assess the accuracy of model predictions.

4.7 BLACKLIST

The URL entered in the GUI will be passed to the manifest.js which will forward it to the checker.js file. All 16 parameters of the URL will be checked and accordingly, it will be concluded if it is a safe (-1) or a malicious (1) URL. As per the result, the dataset.csv file will predict based on the 16 features and then the classifier will further work on the output. This is how blacklisting approach will be used with machine learning algorithm.

5 IMPLEMENTATION

This section of the report discusses about the implementation stage of the model. The Application working and process workflow is discussed in this section. The anti-phishing model uses Support vector machine which extracts 16 URL and DOM-based features, and utilize them to predict the result. The SVM model accuracy is 90.05%. This application also checks if the given URL is malicious or not by evaluating it on the basis of 16 URL and DOM based features.

5.1 IMPLEMENTATION OF THE APPLICATION

The diagram below depicts how the application accesses internal files in order to run the architecture.

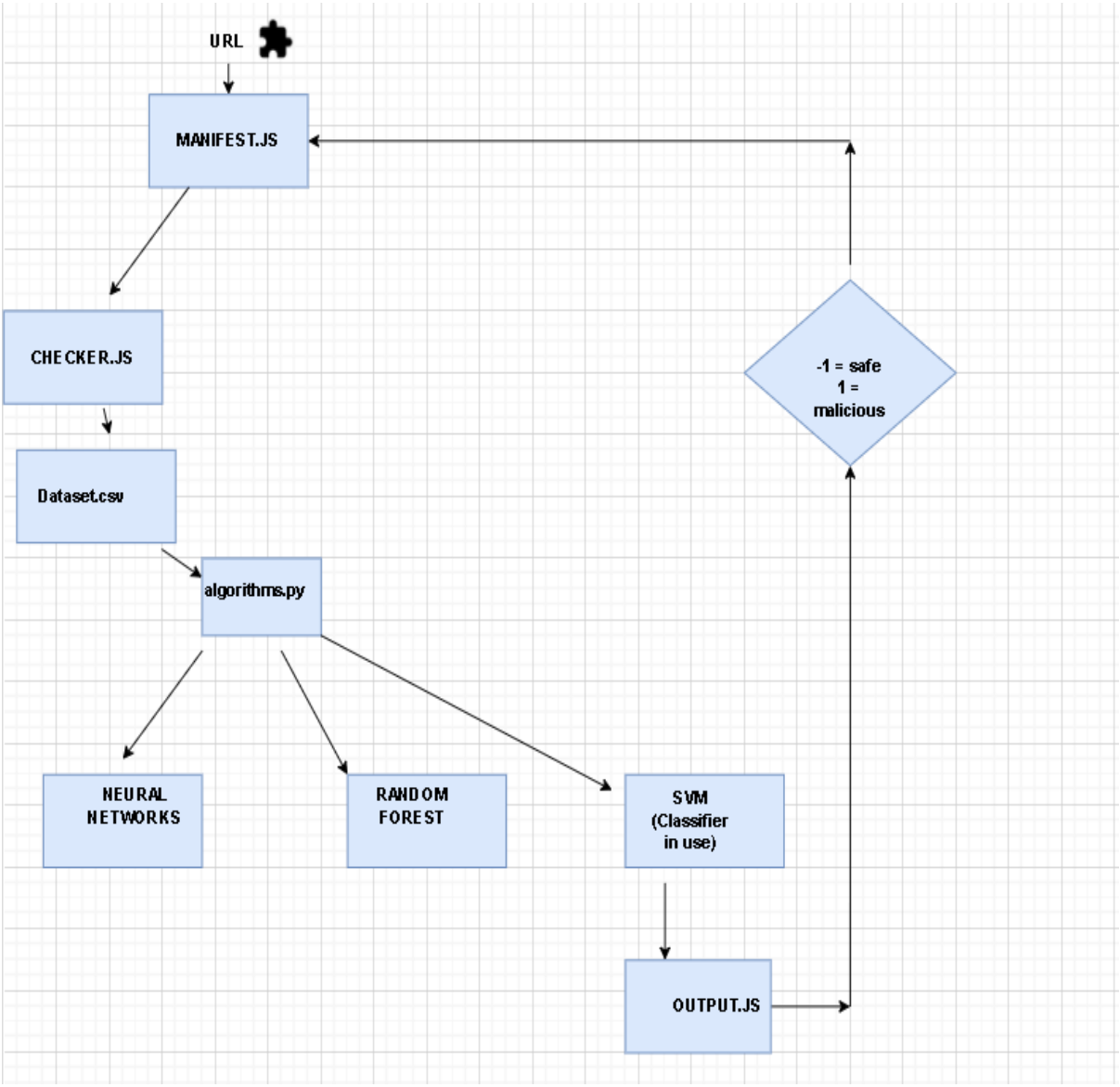


Figure 5: This figure represents the working flow of the model

1. Chrome GUI: When you visit any URL, the extension present in the Chrome will extract the links present in the URL and pass them to the Manifest.js file.

2. Manifest.js: This file provides chrome with information related to the extension, the name and files associated with the extension folder.

3. Checker.js: It runs in its own unprivileged JavaScript environment with full DOM access. For the portal under evaluation, the following functions construct a feature vector:

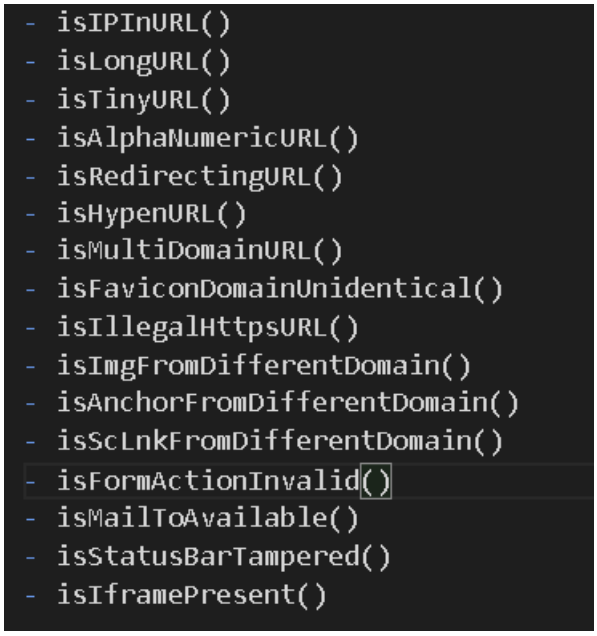


Figure 6: List of 16 URL and DOM based features, used for feature extraction

4. Dataset.csv: This will do the prediction every time, the result from Checker.js is passed to it in terms of (1, -1,0), where -1 = safe, 1 = malicious, 0 = suspicious

5. Algorithms.py: The trained 'SVM model' was utilized as a persistent model to classify webpages (weights generated in./ML Algorithm Evaluation/run algorithms.py).

6. Output.js: Access to external extensions or APIs are needed here, A mechanism of communication should be developed between checker.js and privileged areas of the extension, which is referred to as message passing. Message forwarding enables the extension's many components to work together.

NOTE: The extension verifies each URL call; in the event of URL redirection, it also verifies each intermittent URL hit.

6 EVALUATION

The Anti-phishing model is enhanced at this point, and based on the data examined, it looks to be qualitative. As a result, it's critical to assess the model and review the procedures taken to generate it in order to meet the project's goal.

6.1MODEL RESULTS

For Neural Networks, the below results were received:

```
accuracy = 89.03%
[[1246  209]
 [ 155 1707]]
(2, 2)
TP      FP      FN      TN      Sensitivity     Specificity
1246.0  155.0   209.0   1707.0
        0.86            0.92
1707.0  209.0   155.0   1246.0
        0.92            0.86
0.9036527263102171
runtime = 29.918338537216187 seconds
```

Figure 7: Results obtained for Neural Networks

The above figure shows the analysis of Neural Networks by using confusion matrix. The accuracy is 89.03% and the runtime obtained is 29.918 seconds.

For Random Forest Algorithm, the below results were received:

```
accuracy = 89.63%
[[1293  162]
 [ 182 1680]]
(2, 2)
TP      FP      FN      TN      Sensitivity     Specificity
1293.0  182.0   162.0   1680.0
        0.89            0.9
1680.0  162.0   182.0   1293.0
        0.9             0.89
0.9071274298056156
runtime = 2.544670343399048 seconds
```

Figure 8: Results obtained for Random Forest Algorithm

The above figure shows the analysis of Random Forest Algorithm by using confusion matrix. The accuracy is 89.63% and the runtime obtained is 2.544 seconds.

For the SVM classifier, the below results were received:

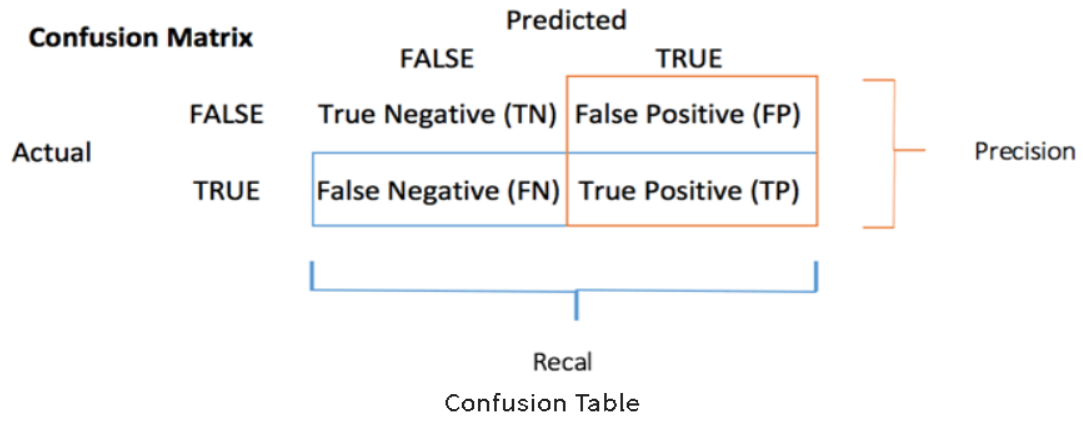
```
accuracy = 90.05%
[[1254  201]
 [ 129 1733]]
(2, 2)
TP      FP      FN      TN      Sensitivity     Specificity
1254.0  129.0   201.0   1733.0
        0.86            0.93
1733.0  201.0   129.0   1254.0
        0.93            0.86
0.9130663856691253
runtime = 6.011983394622803 seconds
```

Figure 9: Results obtained for Support Vector Machine Algorithm

The above figure shows the analysis of Neural Networks by using confusion matrix. The accuracy is 90.05% and the runtime obtained is 6.011 seconds.

6.2 EVALUATION USING THE CONFUSION MATRIX

The model is analysed using various factors including the Number of True Positives, True Negatives, False Positives and False Negatives. These four factors were used to determine the accuracy of the model. [14]



- TP: True Positive: Predicted values correctly predicted as actual positive
- FP: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive
- FN: False Negative: Positive values predicted as negative
- TN: True Negative: Predicted values correctly predicted as an actual negative

You can compute the **accuracy test** from the confusion matrix:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

-**True positive:** The phished URLs percentage are classified correctly. For the support Vector machine algorithm, output obtained 1254.
- True Negative: The legitimate URLs percentage are classified correctly. For the support Vector machine algorithm, output obtained 1733.
-False Positive: The phished URLs percentage are classified incorrectly. For the support Vector machine algorithm, output obtained 129.
-False Negative: The phished URLs percentage are classified incorrectly. For the support Vector machine algorithm, output obtained 201.

7 CONCLUSION

The major goal of this study was to increase the model's ability to detect phishing with greater accuracy. The study, on the other hand, offers a high level of accuracy and precision, as well as a low number of false positives. The accuracy of SVM classifier model is 90.05 %, which was obtained after running the machine learning model python file. The 16 features are utilized to predict the nature of the URL. Therefore, the 16 features are enough to predict the model accuracy of 90.05%. The SVM is implemented in an anti-phishing Chrome extension with a list-based method. This combination has improved the application detection and efficiency to achieve its target.

8 ACKNOWLEDGEMENT

I would like to express my gratitude to Mr Liam McCabe, who has supervised and guided me throughout this project. His contribution was critical to the effective completion of this project. I'm also grateful to National College of Ireland for providing the necessary resource.

References

[1] Rosenthal, M. (2020). Must-Know Phishing Statistics: Updated 2020. [online] Tessian. Available at: https://www.tessian.com/blog/phishing-statistics-2020/.

[2]Cofense. (n.d.). History of Phishing. [online] Available at: https://cofense.com/knowledge-center/history-of-phishing/.

[3] Gandhi, R. (2018). Support Vector Machine — Introduction to Machine Learning Algorithms. [online] Towards Data Science. Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[4]Anupam, S. and Kar, A.K. (2020). Phishing website detection using support vector machines and nature-inspired optimization algorithms. Telecommunication Systems, 76(1), pp.17–32.

[5]Tandale, K.D. and Pawar, S.N. (2020). Different Types of Phishing Attacks and Detection Techniques: A Review. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/9299624.

[6]marktab (n.d.). What is the Team Data Science Process? - Azure Architecture Center. [online] docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview.

[7]Vrbančič, G. (2020). Phishing Websites Dataset. data.mendeley.com, [online] 1. Available at: https://data.mendeley.com/datasets/72ptz43s9v/1.

[8]archive.ics.uci.edu. (n.d.). UCI Machine Learning Repository: About. [online] Available at: https://archive.ics.uci.edu/ml/about.html.

[9]Wikipedia. (2021). PhishTank. [online] Available at: https://en.wikipedia.org/wiki/PhishTank [Accessed 7 Dec. 2021].

[10]Anon, (n.d.). [online] Available at: https://www.acadpubl.eu/hub/2018-118-21/articles/21e/49.pdf.

[11]Artificial (2019). What is an Artificial Neural Network (ANN)? [online] SearchEnterpriseAI. Available at: https://searchenterpriseai.techtarget.com/definition/neural-network.

[12]Wikipedia Contributors (2018). Long short-term memory. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Long_short-term_memory.

[13]Johnson, D. (n.d.). Confusion Matrix in Machine Learning with EXAMPLE. [online] www.guru99.com. Available at: https://www.guru99.com/confusion-matrix-machine-learning-example.html [Accessed 7 Dec. 2021].