All-In-One Web Analysis Tool with Machine Learning

Chew Wei Jie, Jevan Chow Jun Kiat, Tan Kian Yun, Yong Javen
Infocomm Technology Cluster
Singapore Institute of Technology
Singapore 567739
2101186@sit.singaporetech.edu.sg, 2101144@sit.singaporetech.edu.sg, 2103012@sit.singaporetech.edu.sg,
2100992@sit.singaporetech.edu.sg

***Abstract*—Web security is a critical aspect in the modern-day world, and as the world continues to adapt to online communications, detecting fraud is a crucial component of ensuring the safety of Internet usage. This document presents FraudFence, a machine learning tool designed to detect phishing websites by analysing their features. FraudFence includes various website reconnaissance features, such as Whois lookup, port scanning, DNS lookup, server location checking, web header checking, SSL information, trace route, directory busting, and web risk rating. The tool uses a random forest classifier to analyse website features and determine their legitimacy. The report includes a detailed analysis of the machine learning model used in FraudFence, including accuracy and feature importance. What was finalised is a tool that can gather web information and detect if a particular Urban Resource Locator (URL) is a legitimate website or not.**

I. Introduction

As technology advances exponentially in today's era, websites are slowly getting more prone to attacks. Defenses will also improve over time, and attackers will find new ways to execute malicious intents on a website. Fraudulent activities such as phishing or identity theft can cause significant financial losses and damage to a person or company. To mitigate fraud, multi-faceted approaches that involve identifying patterns of suspicious behavior, analyzing user data, and/or monitoring transactions for potential fraud indicators need to be implemented. Our project will be an all-in-one tool that can identify the nature of a particular website, whether it is safe or not, plus various features such as Whois checking, pinpointing the server location and various others for reconnaisance and intelligence gathering. It is aptly named FraudFence, to draw the line between fraud and integrity which our tool is designed to achieve.

II. Background Research

A. Literature Review

In paper [1], the proposed system detects phishing websites by conducting analysis of the webpage's URL. The system employs Alexa database and Phistank.com as its datasets. Phistank.com contains URLs that were already detected as phishing websites. The proposed system then carries out an API call to collect these blacklisted URL.

According to paper [2], some ways to create Phishing website detection solutions, include list-based approaches, heuristic strategies, and machine learning-based methods. The list-based approach is simply using a constantly updated whitelist or blacklist approach to detect whether a website is legitimate. Heuristic strategies identify a phishing website by comparing website features against legitimate ones, to see if it's just trying to imitate real websites or if it's legitimate. The machine learning method uses a machine learning model and uses many data sets of features to start learning what is legitimate or not so that it can eventually accurately predict whether a new site is a phishing website.

In paper [3], the authors mentioned that many existing anti-phishing tools suffer from generating false negatives or false positives results, this is a loophole which increases the success rate of phishing attempts and allows scammers to avoid detection by these tools. Therefore, it is essential that our tool addresses the issue of outputting false negatives and false positives results.

In paper [4], the authors mentioned how unsecured communication of sensitive information such as personal and credit card details could easily be sniffed by third parties. They also mentioned the performance of Security Sockets Layer (SSL)/Transport Layer Security (TLS) servers have decreased due to increased user requests or Denial of Services (DoS) attacks. Therefore, our tool will ensure that SSL information of a website is being safely acquired without crashing the website itself.

In paper [5], the authors went in-depth on the information which Whois offers and showed that 90% of such registrants fail to provide valid contact numbers, with the uncertainty of whether privacy and proxy services are used as a way of hiding contact details a possible factor. Therefore, the Whois function in our tool will help us determine whether the results of a URL look legitimate or suspicious.

B. Existing Tools and Solutions

As part of our research, our group took inspiration from various existing tools in which their functionality is somewhat related to the tool that we have mentioned in the Introduction section, and a brief description of some of the tools are listed below.

Phishai [6] is a machine learning model that takes screenshots of a website and compares them to a database created and maintained via computer vision, to see if it has similarity to known malicious websites, then it predicts whether the website is a phishing site so that it can defend against zero-day phishing websites.

Google Safe Browsing [7] is a service that lets a client's application to check URLs against Google's constantly updated lists of unsafe web resources. It then informs users of suspicious websites and warns them against accessing the websites. Google Safe Browsing also allows users to report any possible fraudulent websites. By allowing users' input, Google can better ensure that the blacklist stays updated.

Phishing-URL-Detection [8] is a tool that allows the input of a website URL, then uses machine learning models to identify characteristics of a phishing websites, thereafter, attempt to predict any possible phishing activities on the website. The various machine models enable the detection precision to fall between the range of 0.92 to 0.99.

BlockSec [9] is an extension that compares a site's SSL certificate against a blacklist on the EOS blockchain. The utilization of blockchain technology enables more flexibility and allows the blacklisting process to be decentralized. This makes it relatively less labor intensive and efficient to manage the database in the long run.

III. Proposed Solution

- *Machine Learning*

FraudFence utilises a machine learning model to analyse and identify phishing websites. Specifically, we implemented the Random Forest Classifier Algorithm in this project. The Random Forest Classifier Algorithm is an ensemble learning technique that incorporate multiple decision trees to improve the robustness and accuracy our our classification model. Fig. 1 shows a visual representation of the Random Forest Classifier Tree.

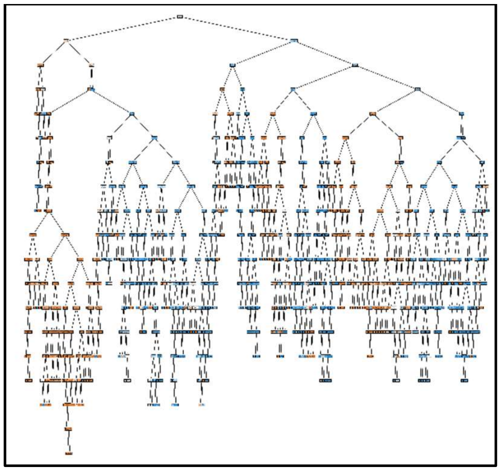


Fig. 1: Random Forest Classifier Tree

The individual decision tree in the forest is constructed by randomly choosing a subset of features from the dataset, thereafter the data is recursively partitioned into smaller subsets. Finally, each tree will provide a result, voting for the classification with the highest likelihood. The results will then be aggregated and a final classification will be determined.

The Random Forest Classifier Algorithm was chosen as it is known to reduce overfitting, a common phenomenon in machine learning where the model starts memorising noise instead of learning new underlying patterns. It is important for our tool to perform effectively with new data in order to detect zero-day phishing sites. The Random Forest Classifier Algorithm incorporates randomness in the data sampling and feature selection. This allows each decision tree to be unique, sensuring our tool to be less sensitive to outliers or noise in the dataset.

Our machine learning is trained using a dataset of over 11,000 known legitimate and phishing websites [10]. Result of –1 indicates that the feature is a phishing website feature, a result of 0 indicates that the feature is suspicious, and a result of 1 indicates that the feature is a legitimate website feature. It consists of the following features:

a) IP address in url: Having IP address in url, even in hexadecimal, will be classified as -1

b) Length of url: Length of over 75 are classified as -1

c) Shortened service used in url: Shortened url are classified as -1

d) '@' symbol in url: Having '@' symbol are classied as –1

e) '//' symbol in url: Having '//' symbol othen than the scheme are classied as –1

f) '-' symbol in url: Having '-' symbol are classied as –1

g) Sub-domains in url: Having more than 2 subdomains are classified as –1

h) SSL Certificate Authority (CA): Trusted CA and the age of certificate is more than or equal to 365 days are classified as 1

i) Domain registration duration: Domain registration duration of more than 365 days are classified as 1

j) Favicon: Having favicon that is loaded from external domains are classified as –1

k) Open ports: Ports 1 to 500 will be scan. Having uncommon ports opened will be classified as –1

l) Scheme of url: Url starting with 'https://' are classified as 1

m) Tags: Url with less than 22% of tags that are loaded from external domains are classified as 1

n) Redirects: Websites with more than 3 redirects are classified as –1

o) Right-click: Websites with right-click disabled are classified as –1

p) Pop-up windows: Websites with pop-up windows are classified as –1

q) Iframe: Websites with iframe tag frameBorder to create an invisible frame are classified as –1

Fig. 2 shows the feature importance graph. It is shown that SSL CA is the most important feature followed by the number of sub-domains detected in the URL.

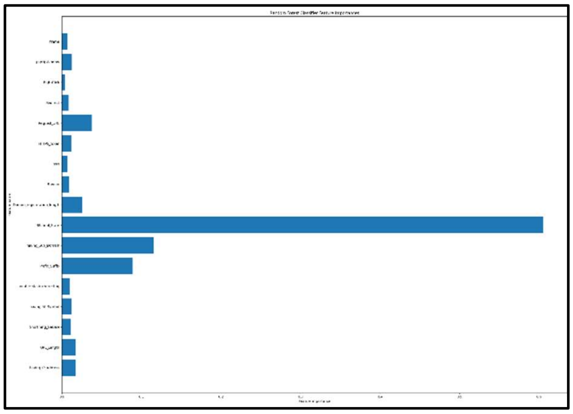


Fig. 2: Feature importance graph

Introspectively, our model can be improved by adding more features and evaluate their impact on overfitting.

Our machine learning model achieved an approximate accuracy score of 91%. Fig. 3 shows the learning curve.

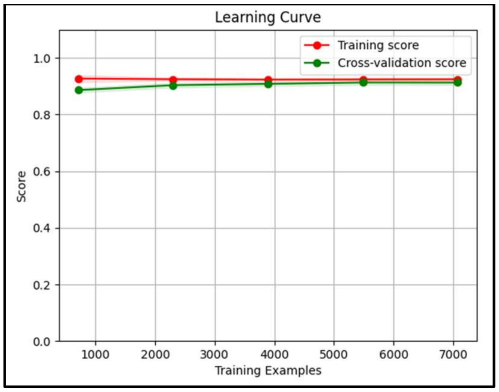


Fig. 3: Learning curve

Fig. 4 shows the confusion matrix. In this specific confusion matrix, the model predicted 837 phishing websites correctly (true positives), and 100 legitimate websites as phishing websites (false positives). It correctly classified 1191 legitimate websites as legitimate (true negatives) but misclassifed 94 legitimate websites as phishing websites (false negatives).

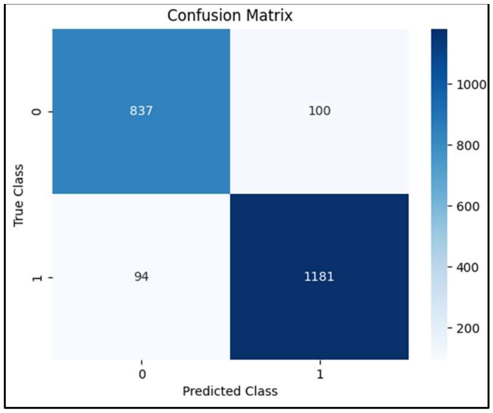


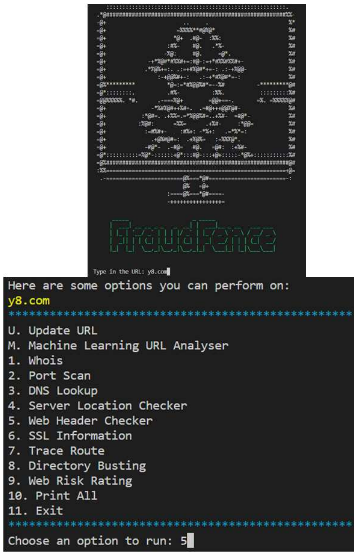
Fig. 4: Confusion matrix

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Overall, the accuracy of our model is approximately 0.91. The model is deem to be performaing relatively well. However, more dataset will improve the accuracy of our model, especially in reducing false positives and flase negatives.

- *Command Line Interface (CLI)*

Our program will first prompt the user to key in a URL, such as ‘y8.com’ or ‘singaporetech.edu.sg’. Adding the scheme to the URL (http/https) is optional. Upon entering the URL, a machine learning algorithm will be conducted on the URL, deeming whether the website is safe or suspicious. The user will be presented with a myriad of options, numbered 1 to 11, as shown in Fig. 5 and 6.



```
Type in the URL: y8.com
```

```
Here are some options you can perform on:
y8.com
**************************************************
U. Update URL
M. Machine Learning URL Analyser
1. Whois
2. Port Scan
3. DNS Lookup
4. Server Location Checker
5. Web Header Checker
6. SSL Information
7. Trace Route
8. Directory Busting
9. Web Risk Rating
10. Print All
11. Exit
**************************************************
Choose an option to run: 5
```

Figs. 5 and 6: CLI start menu and main menu

Upon entering their desired option, our tool will perform the respective function and will display the results in the command line interface (CLI). Users can input the option ‘u’ to change the URL they want to test. A brief description of each of the functions provided by our tool is explained below.

A. *Whois (Option 1)*

In this option, the url link that was entered by the user will be passed to the whois_check() module. The program will then use the request function to grab information from a database that stores information on the URL. Some of the information available are:

- Registered domain name
- Ip addresses
- Domain name owner’s information

B. *Port Scan (Option 2)*

Port Scanning allows user to discover open ports on a particular target that they have inputted. Our tool will prompt the user to enter either one of three options:

- Default: Our tool will read through a text file containing the 1000 most popular port numbers and scan the listed ports
- Full: Our tool will scan through all 65,535 port numbers
- Range: The user can input a range of ports that they wish to be scanned in the syntax A-B, e.g. 1-80 or 200-500. The ports within the specified range will then be scanned

Once the scan is completed, our tool will display the result of the indicated ports based on whether it is open at the top, followed by all the ports that are closed.

C. *DNS Lookup (Option 3)*

DNS Lookup involves querying a DNS server for the IP address associated with the domain name in the URL.

Additionally, by performing a DNS lookup on a particular URL, we can identify additional information such as:

- Live Hosts
- Subdomains
- Mail servers
- Name servers

Our tool uses the `dns.resolver` module to retrieve the following various types of DNS records for the URL: A, AAAA, ANY, CAA, CNAME, MX, NS, PTR, SOA, SRV and TXT.

D. Server Location Checker (Option 4)

Upon running this option, our tool will convert the URL to an IP address of the server which will then passed to a IP geolocation service to grab the latitude and longtitude of the server. With this information, our tool will then open a browser and pinpoint the location of the server of the website on a map. In the CLI, a display message indicating that the location of the IP address of the website was opened in a browser will be shown.

E. Web Header Checker (Option 5)

By performing the Web Header check, users are able to examine the HTTP headers of the website. HTTP headers contain critical information on the contents that are sent, the encoding used and the server software that's being used. This will give the user an understanding of how the website is handling requests.

When the user selects this option, a HTTP HEAD request is sent to the URL using the requests.head() function, and the response headers are retrieved via another function, and stored in a dictionary. Upon the completion of the retrieval, the results are presented in the CLI to the users.

F. SSL Information (Option 6)

When an SSL check is being performed, an SSL certificate will be returned. This certificate is a digital certificate that is used to establish the identity of the website and it allows secure communication between the user and website. The certificate includes information such as:

- Domain name
- Information of the organization that owns the domain
- Issuer
- Expiration Date

When this option is selected, our tool establishes a secure connection with the website before retrieving the SSL certificate. Our tool will then display the SSL certificate information on the CLI.

G. Trace Route (Option 7)

Traceroute is a network diagnostic tool that can track the pathway taken by a packet from the source to its destination. By performing a traceroute, users can identify the routes taken between two hosts which can be helpful in troubleshooting a network connectivity issue.

Our tool uses the `scapy` library to send an ICMP packet to the target URL one by one, increasing the time-to-live (TTL) by 1 for each iteration. When the return value is 0, the function breaks out of the loop. This indicates that the packet has reached its destination. The results stored in a dictionary will then be displayed to the user on the CLI.

H. Directory Busting (Option 8)

Directory Busting is a technique used to discover any hidden or non-linked directories and files on a web server. By performing this scan, users are able to identify vulnerabilities on the web and take any necessary precautions or measures to protect against them.

When the user selects this option, our tool will begin finding any hidden directories of the website. It does this by extracting a wordlist of common directory names from a text file, appending them to the URLs and then making requests to them. The CLI then starts displaying messages based on the returning response code, including success, failure or whether the directory is blocked from access. At the end of the scan, our tool will display any successful scans to the user on the CLI.

I. Web Risk Rating (Option 9)

Web Risk Rating is a way to determine the safety of a particular website. The ratings is determined by the Web of Trust (WOT) database [11]. The rating is derived from various factors such as phishing scams, presence of malware or any suspicious content found on the webpage. By conducting a check on the URL, users can protect themselves from potential threats and reconsider their decision to access a URL when the rating returns 'Suspicious'.

Based on the URL provided by the user, our tool will check the integrity of the given URL with the aid of Web of Trust (WOT) API. It will send a GET request to the API followed by retrieving a response containing information about the URL. The information retrieved will then be displayed to the user.

J. Print All (Option 10)

In this option, our tool will run all the functions mentioned from Options 1 to 9, plus the machine learning results. All the returned values are stored in a dictionary. These values are then formatted and sent to the HTML template. A HTML report that consists of all the results is then generated in the folder for the user to view.

- *Browser Extension*

We also designed a browser extension that analyses the website that a user is visiting. Upon loading a website in a browser, if a user clicks on the 'Extensions' icon on the top right of the browser, they will be able to see our extension option in the window. Clicking on it will reveal the results of the website with the safety rating and WOT category.

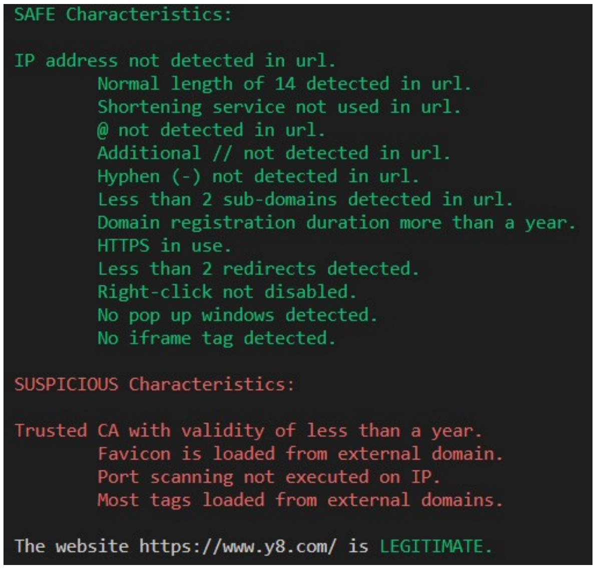
- *Graphical User Interface (GUI)*

We also designed a graphical user interface (GUI) for our tool using the PySimpleGUI library. The idea of a GUI version of this tool came into mind as not all users would prefer using a CLI to run our tool, and we would like to cater our tool to non-programmers who may not be familiar with a CLI environment as well. The GUI is still in the development stages and does not include every function that is available in the CLI. As of this report, five functions are available: Whois, DNS Lookup, Server Location Checker, Web Header Checker and SSL Information.

IV. Results and Analysis

- *Machine Learning*

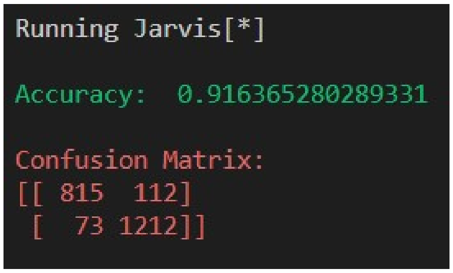
Fig. 7 shows the machine learning classification for https://*y8.com*. The website is classified as legitimate.



```
SAFE Characteristics:

IP address not detected in url.
        Normal length of 14 detected in url.
        Shortening service not used in url.
        @ not detected in url.
        Additional // not detected in url.
        Hyphen (-) not detected in url.
        Less than 2 sub-domains detected in url.
        Domain registration duration more than a year.
        HTTPS in use.
        Less than 2 redirects detected.
        Right-click not disabled.
        No pop up windows detected.
        No iframe tag detected.

SUSPICIOUS Characteristics:

Trusted CA with validity of less than a year.
        Favicon is loaded from external domain.
        Port scanning not executed on IP.
        Most tags loaded from external domains.

The website https://www.y8.com/ is LEGITIMATE.
```

Fig. 7: ML classification for https://y8.com

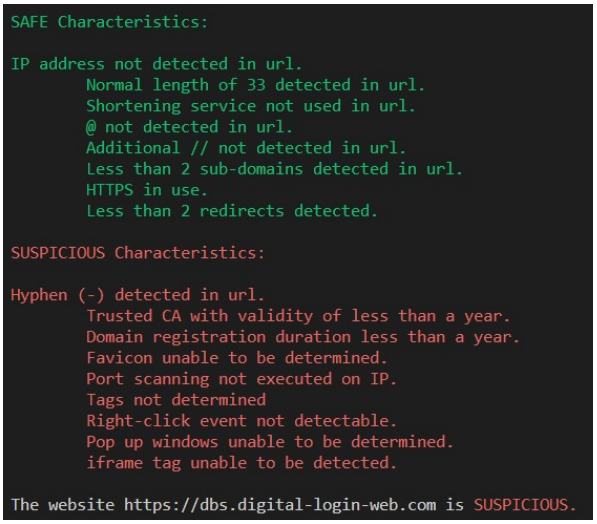
Fig. 8 shows the machine learning accuracy and confusion matrix for *https://y8.com*. The accuracy is determined to be approximately 0.916. There are 815 true positives, 112 false positives, 73 false negatives, and 1212 true negatives.



```
Running Jarvis[*]

Accuracy:  0.916365280289331

Confusion Matrix:
[[ 815  112]
 [  73 1212]]
```

Fig. 8: ML accuracy and confusion matrix for https://y8.com

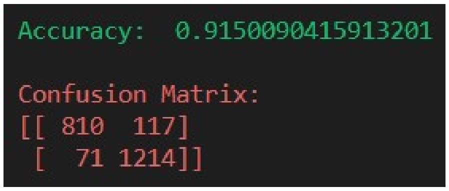
Fig. 9 shows the machine learning for *https://dbs.digital-login-web.com*, a website classified as malicious by OpenPhish database [12].



```
SAFE Characteristics:

IP address not detected in url.
        Normal length of 33 detected in url.
        Shortening service not used in url.
        @ not detected in url.
        Additional // not detected in url.
        Less than 2 sub-domains detected in url.
        HTTPS in use.
        Less than 2 redirects detected.

SUSPICIOUS Characteristics:

Hyphen (-) detected in url.
        Trusted CA with validity of less than a year.
        Domain registration duration less than a year.
        Favicon unable to be determined.
        Port scanning not executed on IP.
        Tags not determined
        Right-click event not detectable.
        Pop up windows unable to be determined.
        iframe tag unable to be detected.

The website https://dbs.digital-login-web.com is SUSPICIOUS.
```

Fig. 9: ML classification for https://dbs.digital-login-web.com

Fig. 10 shows the machine learning accuracy and confusion matrix for *https://dbs.digital-login-web.com.* The accuracy is determined to be approximately 0.915. There are 810 true positives, 117 false positives, 71 false negatives, and 1214 true negatives.



```
Accuracy:  0.9150090415913201

Confusion Matrix:
[[ 810  117]
 [  71 1214]]
```

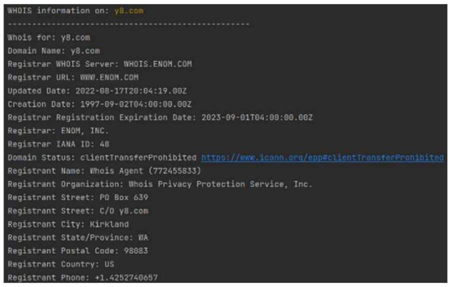
Fig. 10: ML accuracy and confusion matrix for https://dbs.digital-login-web.com

- *Command Line Interface (CLI)*

For most of the harmless functions, we will be testing our tool using the classic gaming websites *y8.com* and *miniclip.com*, whereas for the functions with more malicious intent, such as Directory Busting, we will be using one of our personal URLs that was created during our course of study and is unique only to us.

A. *Whois (Option 1)*

Upon inputting option 1 in the main menu, our tool extracted the information from the database and displayed them row by row in our output. Fig. 11 shows a snippet of the Whois information for *y8.com*, and such information which can be seen are domain name, creation date and registrar/registrant information.

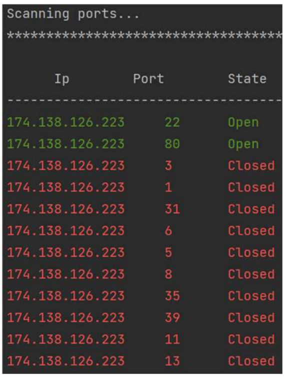


```
WHOIS information on: y8.com
--------------------------------------------------
Whois for: y8.com
Domain Name: y8.com
Registrar WHOIS Server: WHOIS.ENOM.COM
Registrar URL: WWW.ENOM.COM
Updated Date: 2022-08-17T20:04:19.00Z
Creation Date: 1997-09-02T04:00:00.00Z
Registrar Registration Expiration Date: 2023-09-01T04:00:00.00Z
Registrar: ENOM, INC.
Registrar IANA ID: 48
Domain Status: clientTransferProhibited https://www.icann.org/epp#clientTransferProhibited
Registrant Name: Whois Agent (772455833)
Registrant Organization: Whois Privacy Protection Service, Inc.
Registrant Street: PO Box 639
Registrant Street: C/O y8.com
Registrant City: Kirkland
Registrant State/Province: WA
Registrant Postal Code: 98083
Registrant Country: US
Registrant Phone: +1.4252740657
```

Fig. 11: Snippet of the Whois information

B. Port Scan (Option 2)

Using the same URL as option 1 and inputting a range of 1-100, we ran the port scan with our tool. Fig. 12 shows a snippet of the output.



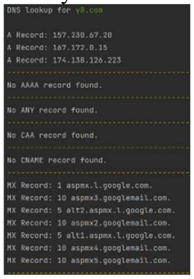
```
Scanning ports...
************************************

      Ip          Port       State
------------------------------------
174.138.126.223     22       Open
174.138.126.223     80       Open
174.138.126.223     3        Closed
174.138.126.223     1        Closed
174.138.126.223     31       Closed
174.138.126.223     6        Closed
174.138.126.223     5        Closed
174.138.126.223     8        Closed
174.138.126.223     35       Closed
174.138.126.223     39       Closed
174.138.126.223     11       Closed
174.138.126.223     13       Closed
```

Fig. 12: Snippet of the output for y8.com

From Figure x, we can see that only ports 22 (SSH) and 80 (HTTP) are open, while the other 98 ports are closed.

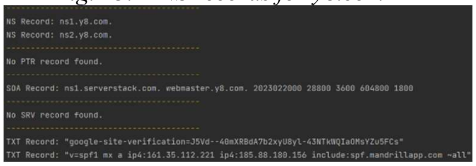
C. DNS Lookup (Option 3)

Running this function on y8.com will reveal its DNS records as seen in Fig. 13 and 14. In this example, y8.com does not have AAAA, ANY, CAA, CNAME, PTR and SRV records, while the other records that could be extracted by our function are displayed row by row in our CLI.



```
DNS lookup for y8.com

A Record: 157.230.67.20
A Record: 167.172.0.15
A Record: 174.138.126.223
----------------------------------------
No AAAA record found.
----------------------------------------
No ANY record found.
----------------------------------------
No CAA record found.
----------------------------------------
No CNAME record found.
----------------------------------------
MX Record: 1 aspmx.l.google.com.
MX Record: 10 aspmx3.googlemail.com.
MX Record: 5 alt2.aspmx.l.google.com.
MX Record: 10 aspmx2.googlemail.com.
MX Record: 5 alt1.aspmx.l.google.com.
MX Record: 10 aspmx4.googlemail.com.
MX Record: 10 aspmx5.googlemail.com.
----------------------------------------
```

Fig. 13: DNS records for y8.com



```
----------------------------------------
NS Record: ns1.y8.com.
NS Record: ns2.y8.com.
----------------------------------------
No PTR record found.
----------------------------------------
SOA Record: ns1.serverstack.com. webmaster.y8.com. 2023022000 28800 3600 604800 1800
----------------------------------------
No SRV record found.
----------------------------------------
TXT Record: "google-site-verification=J5Vd--40mXRBdA7b2xyU8yl-43NTkWQIaOMsYZu5FCs"
TXT Record: "v=spf1 mx a ip4:161.35.112.221 ip4:185.88.180.156 include:spf.mandrillapp.com ~all"
```

Fig. 14: DNS records for y8.com (cont.)

D. Server Location Checker (Option 4)

Two websites were used for comparison, *y8.com* and *miniclip.com* shown in Fig. 15 and 16. Both websites successfully opened the map on Google Chrome, with the server pinpointed to North Bergen, New Jersey and Singapore respectively.

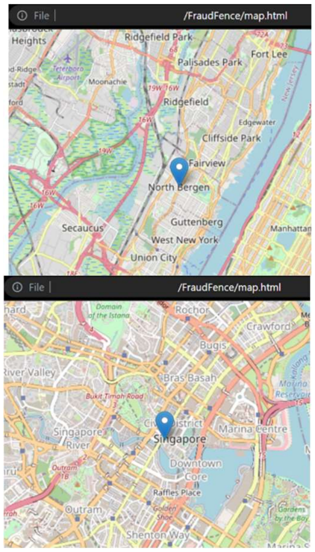
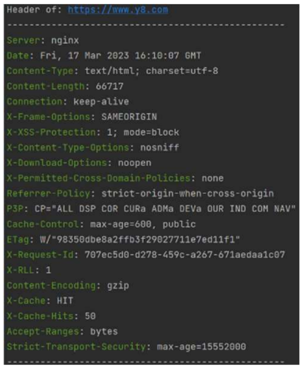


Fig. 15 and 16: Server locations for y8.com and miniclip.com

E. Web Header Checker (Option 5)

When the user selects this option, all extractable HTTP headers will be displayed row by row in our output as seen in Fig. 17.

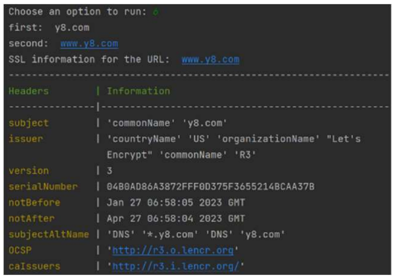


```
Header of: https://www.y8.com
--------------------------------------------------
Server: nginx
Date: Fri, 17 Mar 2023 16:10:07 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 66717
Connection: keep-alive
X-Frame-Options: SAMEORIGIN
X-XSS-Protection: 1; mode=block
X-Content-Type-Options: nosniff
X-Download-Options: noopen
X-Permitted-Cross-Domain-Policies: none
Referrer-Policy: strict-origin-when-cross-origin
P3P: CP="ALL DSP COR CURa ADMa DEVa OUR IND COM NAV"
Cache-Control: max-age=600, public
ETag: W/"98350dbe8a2ffb3f29027711e7ed11f1"
X-Request-Id: 707ec5d0-d278-459c-a267-671aedaa1c07
X-RLL: 1
Content-Encoding: gzip
X-Cache: HIT
X-Cache-Hits: 50
Accept-Ranges: bytes
Strict-Transport-Security: max-age=15552000
--------------------------------------------------
```

Fig. 17: HTTP headers for y8.com

F. SSL Information (Option 6)

Running this function on y8.com will reveal the SSL certificate information for the website, including the serial number, issue date(notBefore) and expiry date(notAfter), as seen in Fig. 18.

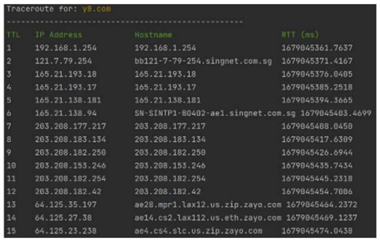


```
Choose an option to run: 6
first:  y8.com
second:  www.y8.com
SSL information for the URL:  www.y8.com
-----------------------------------------------------------------
Headers        | Information
---------------|-------------------------------------------------
subject        | 'commonName' 'y8.com'
issuer         | 'countryName' 'US' 'organizationName' "Let's
                 Encrypt" 'commonName' 'R3'
version        | 3
serialNumber   | 04B0AD86A3872FFF0D375F3655214BCAA37B
notBefore      | Jan 27 06:58:05 2023 GMT
notAfter       | Apr 27 06:58:04 2023 GMT
subjectAltName | 'DNS' '*.y8.com' 'DNS' 'y8.com'
OCSP           | 'http://r3.o.lencr.org'
caIssuers      | 'http://r3.i.lencr.org/'
```

Fig. 18: SSL Information for y8.com

G. Trace Route (Option 7)

The time taken to show all results will vary depending on the URL. As an example, this function was performed on *y8.com*, and it required 15 hops (TTL) to reach the target, with a total elapsed time of around a minute, as shown in Fig. 19.

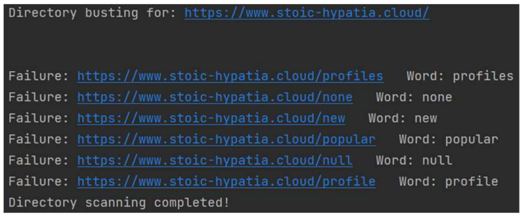


```
Traceroute for: y8.com
-----------------------------------------------------
TTL  IP Address       Hostname                        RTT (ms)
1    192.168.1.254    192.168.1.254                   1679045361.7637
2    121.7.79.254     bb121-7-79-254.singnet.com.sg   1679045371.4167
3    165.21.193.18    165.21.193.18                   1679045376.0405
4    165.21.193.17    165.21.193.17                   1679045385.2518
5    165.21.138.181   165.21.138.181                  1679045394.3665
6    165.21.138.94    SN-SINTP1-BO402-ae1.singnet.com.sg 1679045403.4699
7    203.208.177.217  203.208.177.217                 1679045408.0450
8    203.208.183.134  203.208.183.134                 1679045417.6309
9    203.208.182.250  203.208.182.250                 1679045426.6944
10   203.208.153.246  203.208.153.246                 1679045435.7434
11   203.208.182.254  203.208.182.254                 1679045445.2318
12   203.208.182.42   203.208.182.42                  1679045454.7086
13   64.125.35.197    ae28.mpr1.lax12.us.zip.zayo.com 1679045464.2372
14   64.125.27.38     ae14.cs2.lax112.us.eth.zayo.com 1679045469.1237
15   64.125.23.238    ae4.cs4.slc.us.zip.zayo.com     1679045474.0438
```

Fig. 19: Trace Route for y8.com

H. Directory Busting (Option 8)

As directory busting is considered illegal, we only performed this function on our personal websites. Fig. 20 shows the output for the URL *stoic-hypatia.cloud*, which is a URL given to us for testing in our module, Web Security. Being a rather plain website, there are no hidden files with common directory names found within this URL.

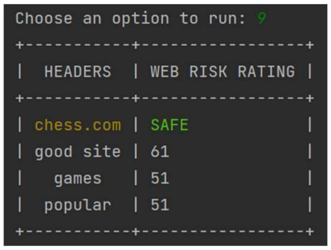


```
Directory busting for: https://www.stoic-hypatia.cloud/


Failure: https://www.stoic-hypatia.cloud/profiles   Word: profiles
Failure: https://www.stoic-hypatia.cloud/none   Word: none
Failure: https://www.stoic-hypatia.cloud/new   Word: new
Failure: https://www.stoic-hypatia.cloud/popular   Word: popular
Failure: https://www.stoic-hypatia.cloud/null   Word: null
Failure: https://www.stoic-hypatia.cloud/profile   Word: profile
Directory scanning completed!
```

Fig. 20: Directory busting for stoic.hypatia.cloud

I. Web Risk Rating (Option 9)

As a test, we performed this function on *chess.com*. The results were as follows.



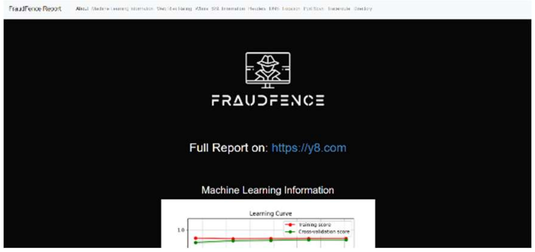
```
Choose an option to run: 9
+-----------+-----------------+
|  HEADERS  | WEB RISK RATING |
+-----------+-----------------+
| chess.com | SAFE            |
| good site | 61              |
|   games   | 51              |
|  popular  | 51              |
+-----------+-----------------+
```

Fig. 21: Results for Web Risk on chess.com

Being a legitimate website, which is popular among chess players, our tool is deemed to have successfully identified this website as Safe, with additional information describing this website, such as it being classified under 'games'.

J. Print All (Option 10)

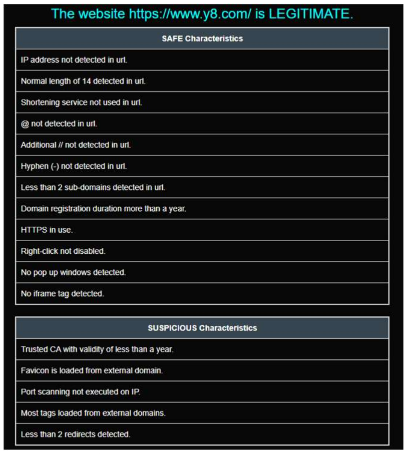
To see all the results in one page, we ran this function on *y8.com*. After running the function for about 45 seconds, a file named 'report.html' was generated with all the information gathered by the various functions provided by our tool, including the machine learning. A navigation bar is included at the top for convenience so that users need not scroll down too much to view their desired results. Shown below are several screenshots of the HTML file.



Full Report on: https://y8.com

Machine Learning Information

Fig. 22: Top of the HTML report with navigation bar

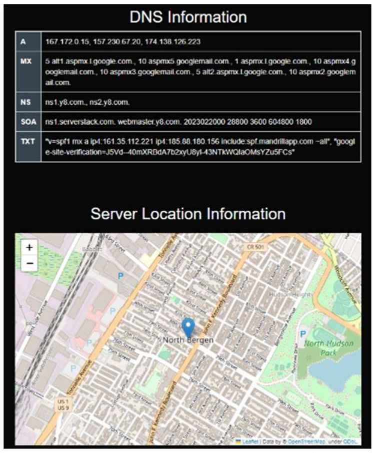


The website https://www.y8.com/ is LEGITIMATE.

SAFE Characteristics
IP address not detected in url.
Normal length of 14 detected in url.
Shortening service not used in url.
@ not detected in url.
Additional // not detected in url.
Hyphen (-) not detected in url.
Less than 2 sub-domains detected in url.
Domain registration duration more than a year.
HTTPS in use.
Right-click not disabled.
No pop up windows detected.
No iframe tag detected.

SUSPICIOUS Characteristics
Trusted CA with validity of less than a year.
Favicon is loaded from external domain.
Port scanning not executed on IP.
Most tags loaded from external domains.
Less than 2 redirects detected.

Fig. 23: Machine learning results for y8.com

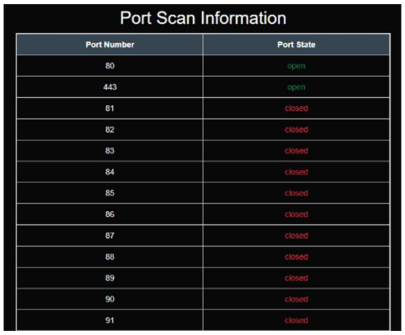


DNS Information

A	167.172.0.15, 157.230.67.20, 174.138.126.223
MX	5 alt1.aspmx.l.google.com., 10 aspmx5.googlemail.com., 1 aspmx.l.google.com., 10 aspmx4.googlemail.com., 10 aspmx3.googlemail.com., 5 alt2.aspmx.l.google.com., 10 aspmx2.googlemail.com.
NS	ns1.y8.com., ns2.y8.com.
SOA	ns1.serverstack.com. webmaster.y8.com. 2023022000 28800 3600 604800 1800
TXT	"v=spf1 mx a ip4:161.35.112.221 ip4:185.88.180.156 include:spf.mandrillapp.com ~all", "google-site-verification=J5Vd--40mXRBdA7b2xyU8yl-43NTkWQlaOMsYZu5FCs"

Server Location Information

Fig. 24: DNS and Server Location information



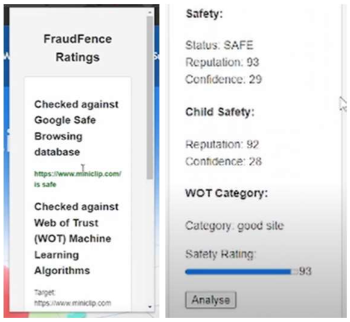
Port Scan Information

Port Number	Port State
80	open
443	open
81	closed
82	closed
83	closed
84	closed
85	closed
86	closed
87	closed
88	closed
89	closed
90	closed
91	closed

Fig. 25: Part of port scan information. A range of 80-445 was used

- *Browser extension*

To test the functionality of our browser extension, we used the URL *miniclip.com*. Clicking on the extension button revealed that miniclip.com is safe with a safety rating of 93. It also revealed that it has a high child safety reputation of 92 as it is a gaming website popular among children.



FraudFence Ratings

Checked against Google Safe Browsing database

https://www.miniclip.com/ is safe

Checked against Web of Trust (WOT) Machine Learning Algorithms

Target: https://www.miniclip.com

Safety:

Status: SAFE
Reputation: 93
Confidence: 29

Child Safety:

Reputation: 92
Confidence: 28

WOT Category:

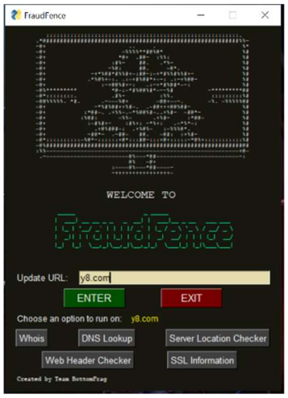
Category: good site

Safety Rating: 93

Fig. 26 and 27: Browser extension results for miniclip.com

- *Graphical User Interface (GUI)*

The GUI starts off similar to the CLI program, whereby the user would have to enter a URL in the input field, which will trigger the five options to appear in the GUI window, allowing the user to select their desired option, as seen in Fig. 28. As an example, Fig. 29 shows a snippet of the SSL information of *y8.com*.



WELCOME TO

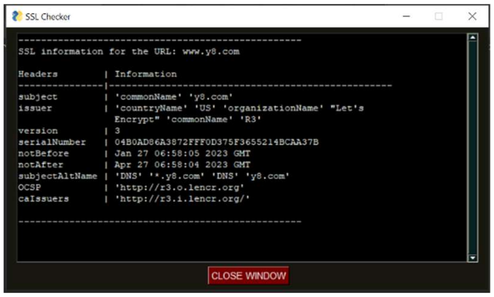
FraudFence

Update URL: y8.com

Choose an option to run on: y8.com

Created by Team BottomFrag

Fig. 28: Main menu of the GUI



SSL Checker

```
---------------------------------------------------
SSL information for the URL: www.y8.com

Headers         | Information
----------------|-----------------------------------------------
subject         | 'commonName' 'y8.com'
issuer          | 'countryName' 'US' 'organizationName' "Let's
                  Encrypt" 'commonName' 'R3'
version         | 3
serialNumber    | 04B0AD86A3872FFF0D375F3655214BCAA37B
notBefore       | Jan 27 06:58:05 2023 GMT
notAfter        | Apr 27 06:58:04 2023 GMT
subjectAltName  | 'DNS' '*.y8.com' 'DNS' 'y8.com'
OCSP            | 'http://r3.o.lencr.org'
caIssuers       | 'http://r3.i.lencr.org/'

---------------------------------------------------
```

Fig. 29: SSL information of y8.com

Clicking the red 'CLOSE WINDOW' button in Figure x will close the SSL and return to the main menu window, where the user can update the URL, choose a different option, or exit the program.

V. Conclusion

In conclusion, the FraudFence tool using machine learning proves to be a promising solution to the growing issue of phishing websites. The random forest classifier model was able to achieve an accuracy score of 0.91 and effectively distinguish between legitimate and malicious websites. Our tool also offers various features and a machine learning URL analyzer to conduct website reconnaissance. From the high accuracy of our results displayed in the HTML report and the respective functions, our team has concluded that our tool is indeed effective in analyzing any website. The all-in-one nature of our tool further justifies its effectiveness thanks to the convenience of easily switching between functions by just a few keyboard inputs. We do feel that the tool can still be improved by including more features to reduce overfitting and conducting further testing with larger datasets. Overall, our tool offers a promising solution to detect and prevent phishing attacks, potentially saving businesses and individuals from financial losses and other security risks. It is also an efficient tool for cyber professionals to conduct reconnaissance and intelligence gathering on suspicious websites.

Acknowledgment

ICT2206 Group 'BottomFrag' wishes to acknowledge Professors Goh Wei Han and Spiridon Bakiras for their unwavering help and guidance throughout the lectures and labs which enhanced our knowledge about Web Security and allowed us to understand how and why this module is vital for our future knowledge in our course of study.

References

[1] "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," *ResearchGate*, 2020, doi: 10.1109\/ICCCNT49239.2020.9225561.

[2] L. Tang and Q. H. Mahmoud, "A Survey of Machine Learning-Based Solutions for Phishing Website Detection," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 672–694, Aug. 2021, doi: 10.3390/make3030034.

[3] "Itrustpage: a user-assisted anti-phishing tool: ACM SIGOPS Operating Systems Review: Vol 42, No 4," *ACM SIGOPS Operating Systems Review*, 2023. https://dl.acm.org/doi/abs/10.1145/1357010.1352620?casa_token=_uQQvQuNrewAAAAA:wuwsLgmGrYv1-

Lr4NHztYk0KS3rOAFI0Fab0C_440la3Xy5Jh3mLr9Q0T_F0RXvwwoVkobGNmfve4Q. [Accessed 04-Feb-2023].

[4] A. Maheswaran, R. Kanchana, "Pasic: A Novel Approach for Page-Wise Web Application Security," [Online]. Available: https://www.researchgate.net/publication/235944405_Web_Application_Security_Using_SSL_Certificates. [Accessed: 16-Mar-2023].

[5] R. Clayton, T. Mansfield, "A Study of Whois Privacy and Proxy Service Abuse," [Online]. Available: https://www.cl.cam.ac.uk/~rnc1/whoisstudy.pdf. (Accessed: 16-Mar-2023).

[6] phishai, "phishai/phish-ai-api: Official python API for Phish.AI public and private API to detect zero-day phishing websites," *GitHub*, Apr. 26, 2018. https://github.com/phishai/phish-ai-api [Accessed: 17-Mar-2023].

[7] *Google safe browsing | google developers*. [Online]. Available: https://developers.google.com/safe-browsing. [Accessed: 17-Mar-2023].

[8] V. Bichave, Phishing URL Detection [Online]. Available: https://github.com/VaibhavBichave/Phishing-URL-Detection. [Accessed: 18-Mar-2023].

[9] Google, "Netcraft Extension", *Chrome Web Store* [Online]. Available: https://chrome.google.com/webstore/detail/netcraft-extension/bmejphbfclcpmpohkggcjeibfilpamia?hl=en. [Accessed: 18-Mar-2023].

[10] akashkr, "Phishing URL EDA and modelling 🕸," Kaggle.com, Jun. 30, 2020 [Online]. Available: https://www.kaggle.com/code/akashkr/phishing-url-eda-and-modelling/input. [Accessed 19-Mar-2023]

[11] "WOT Web Risk and Safe browsing," Rapidapi.com, 2023 [Online]. Available: https://rapidapi.com/mywot-mywot-default/api/wot-web-risk-and-safe-browsing [Accessed 19-Mar-2023].

[12] "OpenPhish - Phishing Intelligence," Openphish.com, 2023 [Online]. Available: https://openphish.com/ [Accessed 19-Mar-2023].