MAK HILL Publications ACE Journal of Computer Science and Engineering

Research Article
doi: 10.59218/makacejcse.2023.1.11

Phishing Websites Spotting with Help of Using Machine Learning Tools

Prasanth Baskaran, B.J. Ronald Isaac, R. Manju and G. Chandru

Department of CSE-Anna University, GKM College of Engineering and Technology, Chennai, Tamil Nadu

OPEN ACCESS

Key Words
Phishing websites, machine learning techniques, deep neural network (DNN), phish tank, web application

Corresponding Author
Prasanth Baskaran
Department of CSE-ANNa University, GKM College of Engineering and Technology, Chennai, Tamil Nadu

Received: 25 January 2023
Accepted: 5 February 2023
Published: 19 February 2023

Citation: Prasanth Baskaran, B.J. Ronald Isaac, R. Manju and G. Chandru, 2023. Phishing Websites Spotting with Help of Using Machine Learning Tools. ACE J. Comp. Sci. Eng., 2: 1-11, doi: 10.59218/makacejcse.2023.1.11

Copy Right: MAK HILL Publications

ABSTRACT
Phishing assaults cost internet users billions of dollars every year and are a constantly growing hazard in the cyberspace. It is illegal to gather sensitive information from consumers through a number of social engineering techniques. Email, instant messaging, pop-up messages, web pages and other forms of communication can all be used to identify phishing tactics. This study offers a model that can determine whether a URL link is genuine or fraudulent. The data set used for the classification was sourced from the University of New Brunswick dataset bank, which has a collection of benign, spam, phishing, malware and defacement URLs, as well as from an open source service called "Phish Tank," which contains phishing URLs in multiple formats such as CSV, JSON, etc. Phishing URLs are identified using a combination of deep neural network methods and more than six machine learning models. The goal of this study is to create a web application software that can identify phishing URLs from a database of more than 5,000 URLs that have been randomly selected, divided into 80,000 training samples and 20,000 testing samples and then divided again into equal portions of phishing and legitimate URLs. To distinguish between legal and phishing URLs, the URL dataset is trained and tested using feature selections like address bar-based features, domain-based features, HTML and JavaScript-based features. Finally, the study provided a model for classifying URLs as phishing or legitimate. This would be extremely useful in assisting individuals and businesses in identifying phishing attacks by authenticating any link provided to them to prove its validity.

INTRODUCTION

The Internet, particularly social media, has become an important part of our lives for gathering and disseminating information. According to Pamela, the Internet is a network of computers that contain valuable data; thus, many security mechanisms are in place to protect that data; however, there is a weak link: the human. Security mechanisms have a much more difficult time protecting a user's data and devices when they freely give away their data or access to their computer Abdelhamid[1].

Shaikh *et al*.[2] defines social engineering (a type of attack used to steal user data such as login credentials and credit card numbers) as one of the most common social engineering attacks. When an attacker tricks a victim into opening an email, instant message, or text message that appears to be from a trusted source, the attack occurs. When the recipient clicks the link, they are duped into thinking they have received a gift and unknowingly click a malicious link, which results in the installation of malware, the freezing of the system as part of a ransom ware attack, or the disclosure of sensitive information.

Computer security threats have grown significantly in recent years, owing to the rapid adoption of technological advancements, while also increasing the vulnerability to human exploitation. Users should understand how phishers operate and be aware of techniques to help protect themselves from being phished.

As a result, this is a rapidly evolving threat to individuals as well as large and small businesses. Criminals now have access to industrial-strength services on the dark web, resulting in an increase in the number of these phishing links and emails, as well as an increase in 'quality,' making them more difficult to detect.

LITERATURE REVIEW

According to Gupta *et al*.[3], Internet world stats ("Internet world stats usage and population statistics", 2014), the total number of Internet users worldwide in 2014 is 2.97 billion; that is, more than 38% of the world population uses the Internet. Hackers exploit insecure Internet systems to trick unsuspecting users into falling for phishing scams. On the Internet, phishing emails are used to defraud both individuals and financial institutions. (n.d., "RSA Anti-Fraud Command Centre") According to their website, the Anti-Phishing Working Group (APWG) is an international consortium dedicated to promoting research, education and law enforcement in order to eliminate online fraud and cybercrime.

Total phishing attacks increased by 160% in 2012 compared to 2011, indicating a record year for phishing volumes. In 2013, approximately 450,000 phishing attacks were detected, resulting in financial losses of more than $5.9 billion ("RSA Anti-Fraud Command Centre", n.d.). In 2013, total attacks increased by 1% over 2012. The total number of phishing attacks detected in the first quarter of 2014 was 125,215, representing a 10.7 percent increase over the fourth quarter of 2013. To fool users, more than 55% of phishing websites include the target site's name in some form and 99.4% of phishing websites use port 80 (“Anti-Phishing Working Group (APWG) Phishing activity trends report first quarter”)[4].

According to an APWG report published in the first quarter of 2014, the second-highest number of phishing attacks ever recorded occurred between January and March 2014 ("Anti-Phishing Working Group (APWG) Phishing activity trends report first quarter", 2014), with payment services being the most targeted industry. 123,972 unique phishing attacks were observed in the second half of 2014. Total financial losses in 2011 were 1.2 billion dollars and they increased to 5.9 billion dollars in 2013[4].

MATERIALS AND METHODS

An extensive review was conducted on related topics and existing documented materials such as journals, e-books and websites containing related information gathered, which was examined and reviewed to retrieve essential data to help better understand and improve the system.

The methodology used to achieve the previously stated goals is described below. The dataset is made up of phishing and legitimate URLs obtained from open-source platforms. To avoid data imbalance, the dataset was pre-processed, which means it was cleaned up of any anomalies such as missing data. Following that, expository data analysis was performed on the dataset in order to explore and summarise it. Once the dataset had been cleaned of all anomalies, website content-based features were extracted from it in order to obtain accurate features for training and testing the model. To best decide the classification models to solve the problem of detecting phishing websites, an extensive review of existing works of literature and machine learning models on detecting phishing websites was performed.

As a result, a series of machine learning classification models, including Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto encoder Neural Network and Random Forest, were deployed on the dataset to differentiate between phishing and legitimate URLs. Out of all the deployed models, the best model with the highest training accuracy was chosen and integrated into a web application. As a result, a user can enter a URL link into the web application to determine whether it is phishing or legitimate.

PROPOSED SYSTEM

This chapter describes the various processes, methods and procedures used by the researcher to achieve the stated goals and objectives, as well as the conceptual framework in which the research was carried out.

Any research work's methodology refers to the research approach taken by the researcher to address the stated problem. Because the efficiency and maintainability of any application are solely determined by how designs are created. This chapter contains detailed descriptions of the methods used to provide solutions to the research work's stated objectives.

System analysis, according to Merriam-Webster (11th edition), is "the process of studying a procedure or business to identify its goals and purposes and create systems and procedures that will efficiently achieve them." It is also the act, process, or profession of studying an activity (such as a procedure, a business, or a physiological function) typically through mathematical means in order to define its goals or purposes and discover operations and procedures for most efficiently accomplishing them. System analysis is used in every field where something is created. Prior to planning and development, you must thoroughly understand the old systems and use that knowledge to determine how well your new system can function.

Machine learning models and deep neural networks are used in the proposed phishing detection system. The machine learning models and a web application are the two main components of the system. Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network and Random Forest are among the models used.

These models are chosen based on the comparative performance of various machine learning algorithms. Each of these models is trained and tested on a content-based website feature extracted from both phishing and legitimate datasets.

As a result, the model with the highest accuracy is chosen and integrated into a web application that allows users to predict whether a URL link is phishing or legitimate.

MODEL DEVELOPMENT

The model development method starts with several models, tests them and then adds them to an iterative process until a model that meets the requirements is created. Figure 1 depicts the steps involved in the development of supervised and unsupervised machine learning models.

The following are the stages to machine learning model development for phishing detection systems.

Data collection: Data for the datasets on which the models are trained is obtained from various open-source platforms. The dataset set includes phishing and legitimate URL datasets.

The phishing URLs were gathered from Phish Tank, an open-source service. This service provides a collection of phishing URLs in various formats such as CSV, JSON and others that is updated hourly. This dataset can be accessed via the phishtank.com website. Over 5000 random phishing URLs are collected from this dataset to train the ML models.

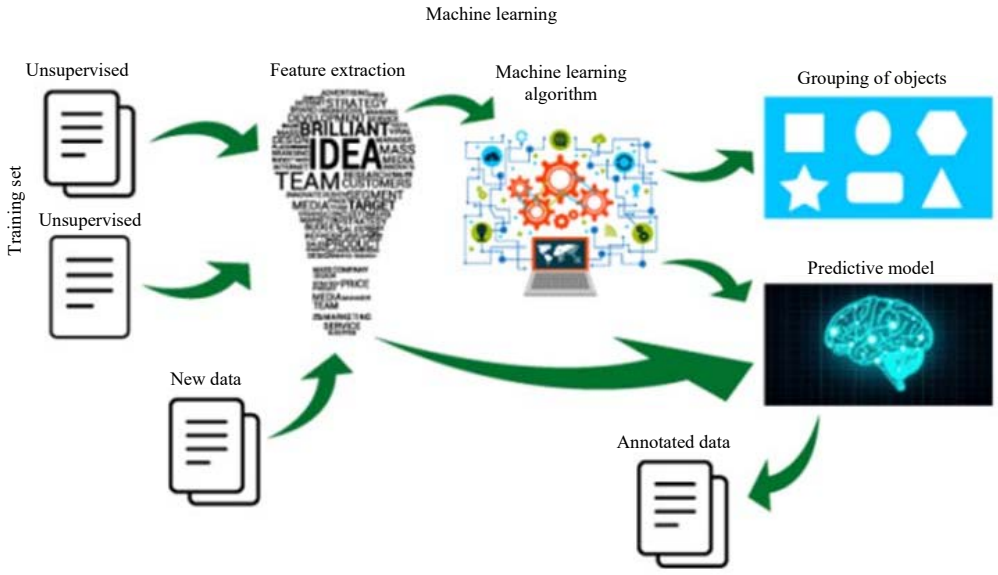


Fig. 1: Machine learning development process

The set of legitimate URLs was obtained from the University of New Brunswick's open datasets. This dataset can be found on the university's website. This dataset contains URLs that are benign, spam, phishing, malware, or defacement. The benign URL dataset is being considered for this project out of all of these types. Over 5000 random legitimate URLs are collected from this dataset to train the ML models.

Pre-processing: Following data collection, the first and most important step is data pre-processing. The raw dataset for phishing detection was prepared by removing redundant and irregular data and then encoded into a useful and efficient format suitable for the machine learning model using the One-Hot Encoding technique.

Exploratory data analysis: After a series of data cleaning steps, the dataset was subjected to exploratory data analysis (EDA). The data visualisation method was used to analyse, explore and summarise the dataset. These visualisations include heat maps, histograms, box plots, scatter plots and pair plots to uncover patterns and insights within data.

Feature extraction: The goal of feature extraction is to reduce the number of features in a dataset by generating new ones from existing ones. Thus, website content-based features such as the Address bar-based feature, which consists of 9 features, the Domain-based feature, which consists of 4 features and the HTML and JavaScript-based feature, which consists of 4 features, were extracted from phishing and legitimate datasets. As a result, a total of 17 features were extracted for phishing detection.

Model training: Model Training entails feeding data to Machine Learning algorithms to assist in identifying and learning good attributes of the dataset.

This study problem is the result of supervised learning and belongs to the classification problem. The phishing detection algorithms include supervised machine learning models and a deep neural network that was used to train the dataset. Decision Tree, Random Forest, Support Vector Machines, XGBooster, Multilayer Perceptron and Auto-encoder Neural Network are among the algorithms used. The dataset was used to train all of these models. As a result, the dataset is divided into two parts: Training and testing. The training model contains 80% of the dataset, allowing machine learning models to learn more about the data and distinguish between phishing and spam.

Model testing: Model testing is the process of evaluating the performance of a fully trained model on a testing set.

As a result, after 80% of the data has been trained, 20% of the dataset is used to evaluate the trained dataset to see how well the models perform[5].

Model assessment: Model evaluation entails estimating model generalisation accuracy and deciding whether or not the model performs better.

Thus, the Scikit-learn (sklearn matrices) module was used to implement several score and utility functions to measure classification performance in order to properly evaluate the phishing detection models.

System modelling: Figure 2 depicts the architecture of the proposed phishing detection system, in which a user enters a URL link, which is then passed through various trained machine learning and deep neural network models until the best model with the highest accuracy is chosen. As a result, the chosen model is deployed as an API (Application Programming Interface) and integrated into a web application. As a result, a user interacts with the web application, which is available on various display devices such as computers, tablets and mobile devices[6].

RESULTS AND DISCUSSIONS

Data collection: The dataset used for classification was obtained from a variety of sources, as detailed in the methodology.

The dataset used to classify the dataset into phishing and legitimate URLs was obtained from open source websites, with examples shown in Fig. 3 and 4.

Feature extraction on the datasets: The features extraction used on the dataset are categorized into:

- Address bar based features
- Domain-based features
- Html and java-script based features

Data analysis and visualization: Figure 5 depicts a distribution plot of how legitimate and phishing datasets are distributed based on the features chosen and how they are related to one another.

Figure 6 depicts a correlation heat-map of the dataset. The plot depicts the relationship between various variables in the dataset.

Figure 7 depict the feature importance in the model for the Decision tree classifier and the Random forest classifier, respectively.

Phishing detection model: According to the methodology, the proposed system makes use of machine learning models and deep neural networks. Decision Tree, Support Vector Machine, XG

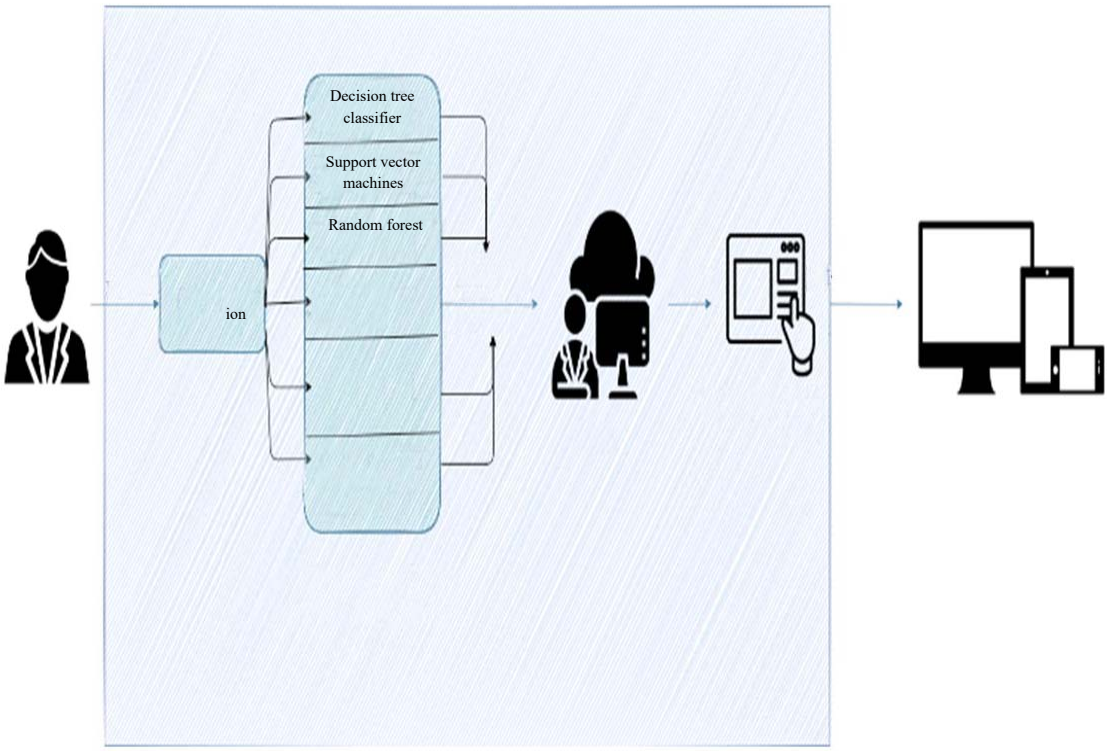
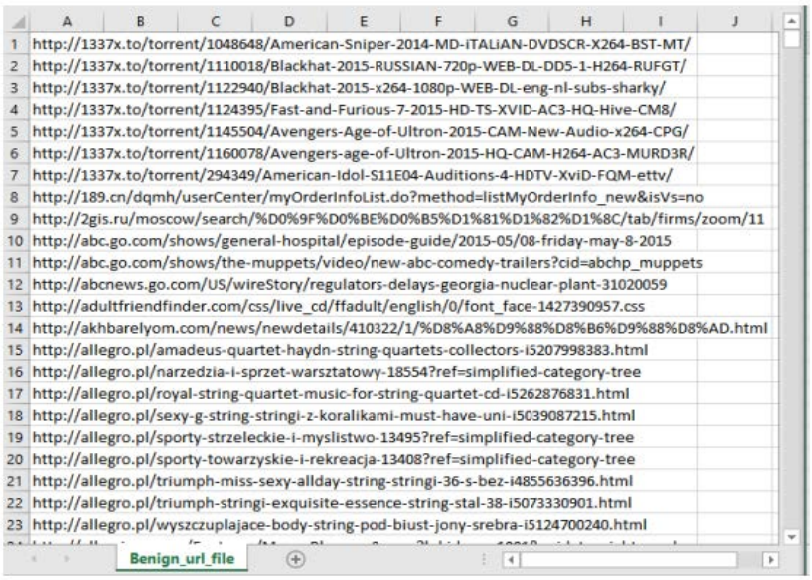


Fig. 2: Architecture diagram



	A
1	http://1337x.to/torrent/1048648/American-Sniper-2014-MD-iTALiAN-DVDSCR-X264-BST-MT/
2	http://1337x.to/torrent/1110018/Blackhat-2015-RUSSIAN-720p-WEB-DL-DD5-1-H264-RUFGT/
3	http://1337x.to/torrent/1122940/Blackhat-2015-x264-1080p-WEB-DL-eng-nl-subs-sharky/
4	http://1337x.to/torrent/1124395/Fast-and-Furious-7-2015-HD-TS-XVID-AC3-HQ-Hive-CM8/
5	http://1337x.to/torrent/1145504/Avengers-Age-of-Ultron-2015-CAM-New-Audio-x264-CPG/
6	http://1337x.to/torrent/1160078/Avengers-age-of-Ultron-2015-HQ-CAM-H264-AC3-MURD3R/
7	http://1337x.to/torrent/294349/American-Idol-S11E04-Auditions-4-HDTV-XviD-FQM-ettv/
8	http://189.cn/dqmh/userCenter/myOrderInfoList.do?method=listMyOrderInfo_new&isVs=no
9	http://2gis.ru/moscow/search/%D0%9F%D0%BE%D0%B5%D1%81%D1%82%D1%8C/tab/firms/zoom/11
10	http://abc.go.com/shows/general-hospital/episode-guide/2015-05/08-friday-may-8-2015
11	http://abc.go.com/shows/the-muppets/video/new-abc-comedy-trailers?cid=abchp_muppets
12	http://abcnews.go.com/US/wireStory/regulators-delays-georgia-nuclear-plant-31020059
13	http://adultfriendfinder.com/css/live_cd/ffadult/english/0/font_face-1427390957.css
14	http://akhbarelyom.com/news/newdetails/410322/1/%D8%A8%D9%88%D8%B6%D9%88%D8%AD.html
15	http://allegro.pl/amadeus-quartet-haydn-string-quartets-collectors-i5207998383.html
16	http://allegro.pl/narzedzia-i-sprzet-warsztatowy-18554?ref=simplified-category-tree
17	http://allegro.pl/royal-string-quartet-music-for-string-quartet-cd-i5262876831.html
18	http://allegro.pl/sexy-g-string-stringi-z-koralikami-must-have-uni-i5039087215.html
19	http://allegro.pl/sporty-strzeleckie-i-myslistwo-13495?ref=simplified-category-tree
20	http://allegro.pl/sporty-towarzyskie-i-rekreacja-13408?ref=simplified-category-tree
21	http://allegro.pl/triumph-miss-sexy-allday-string-stringi-36-s-bez-i4855636396.html
22	http://allegro.pl/triumph-stringi-exquisite-essence-string-stal-38-i5073330901.html
23	http://allegro.pl/wyszczuplajace-body-string-pod-biust-jony-srebra-i5124700240.html

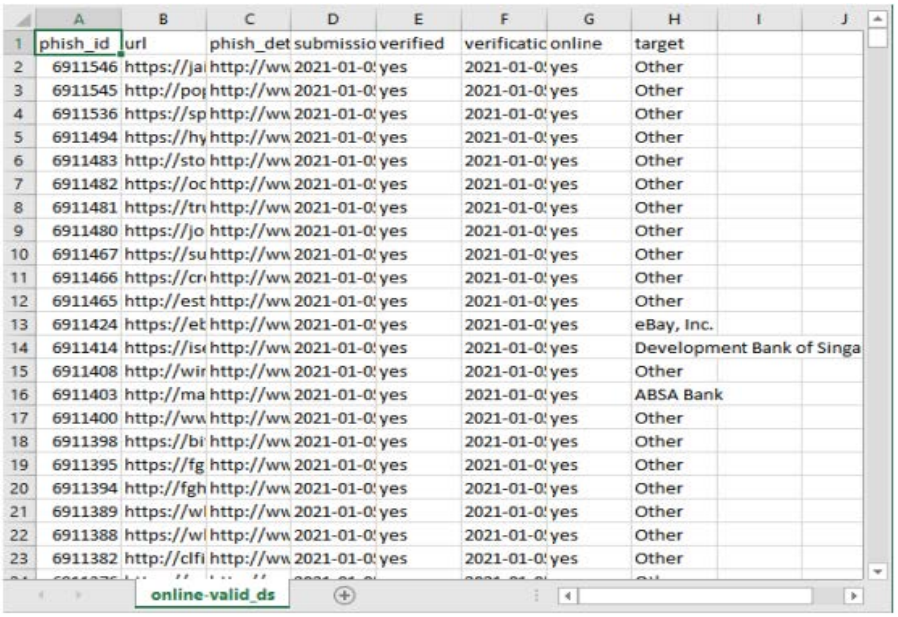
Benign_url_file

Fig. 3: Dataset of phishing URLs

Source: The dataset is collected from an open-source service called Phish-Tank. This dataset consists of 5,000 random phishing URLs which are collected to train the ML models

Booster, Multilayer Perceptions, Auto Encoder Neural Network and Random Forest are among the models used.

The models determine whether a website URL is legitimate or phishing. The models assist in providing a two-class prediction (legitimate (0) and phishing (1)).



	A	B	C	D	E	F	G	H	I	J
1	phish_id	url	phish_det	submissio	verified	verificatic	online	target		
2	6911546	https://ja	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
3	6911545	http://po	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
4	6911536	https://sp	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
5	6911494	https://hy	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
6	6911483	http://sto	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
7	6911482	https://oc	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
8	6911481	https://tr	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
9	6911480	https://jo	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
10	6911467	https://su	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
11	6911466	https://cr	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
12	6911465	http://est	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
13	6911424	https://eb	http://ww	2021-01-0	yes	2021-01-0	yes	eBay, Inc.		
14	6911414	https://is	http://ww	2021-01-0	yes	2021-01-0	yes	Development Bank of Singa		
15	6911408	http://wir	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
16	6911403	http://ma	http://ww	2021-01-0	yes	2021-01-0	yes	ABSA Bank		
17	6911400	http://ww	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
18	6911398	https://bi	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
19	6911395	https://fg	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
20	6911394	http://fgh	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
21	6911389	https://wl	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
22	6911388	https://wl	http://ww	2021-01-0	yes	2021-01-0	yes	Other		
23	6911382	http://clfi	http://ww	2021-01-0	yes	2021-01-0	yes	Other		

online-valid_ds

Fig. 4: Dataset of legitimate URLs

Source: The Dataset were obtained from the open datasets of the University of New Brunswick, The dataset

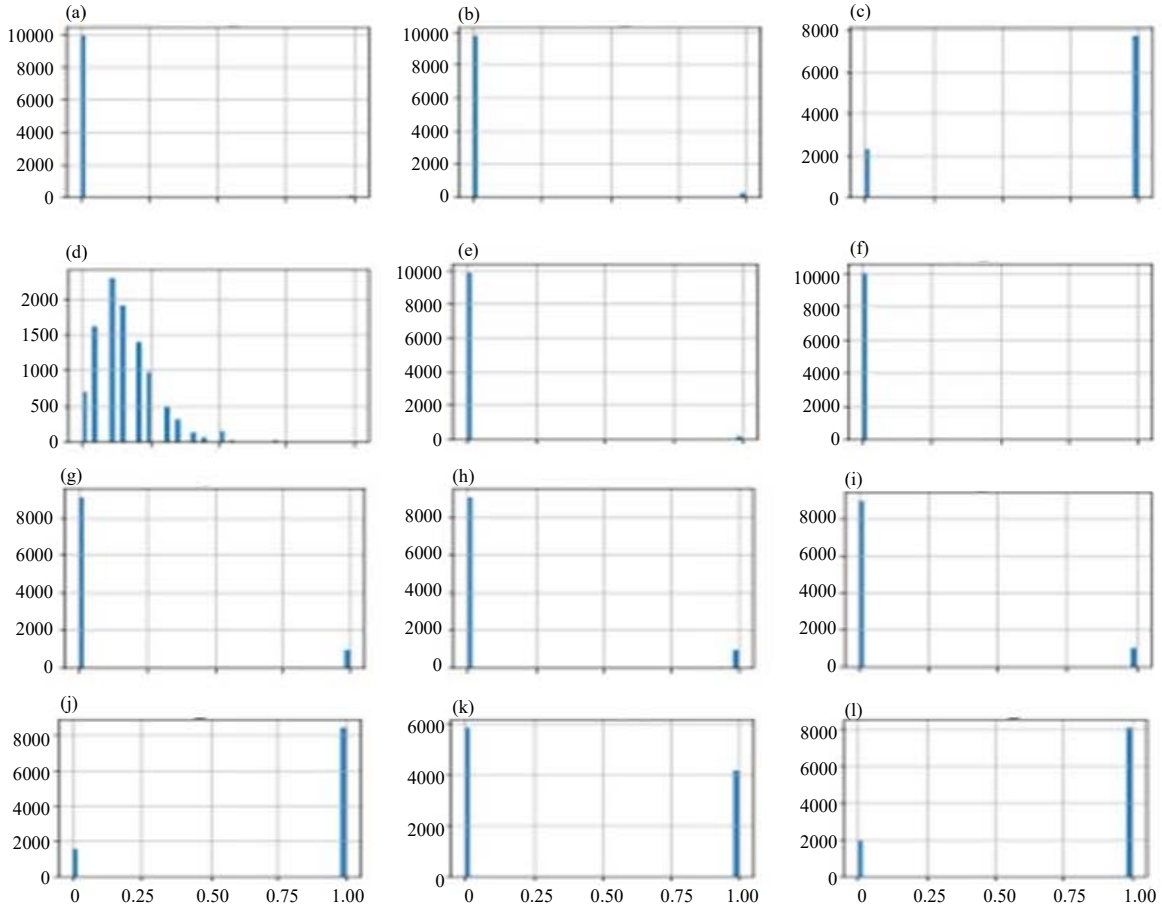


Fig. 5(a-q): Continue

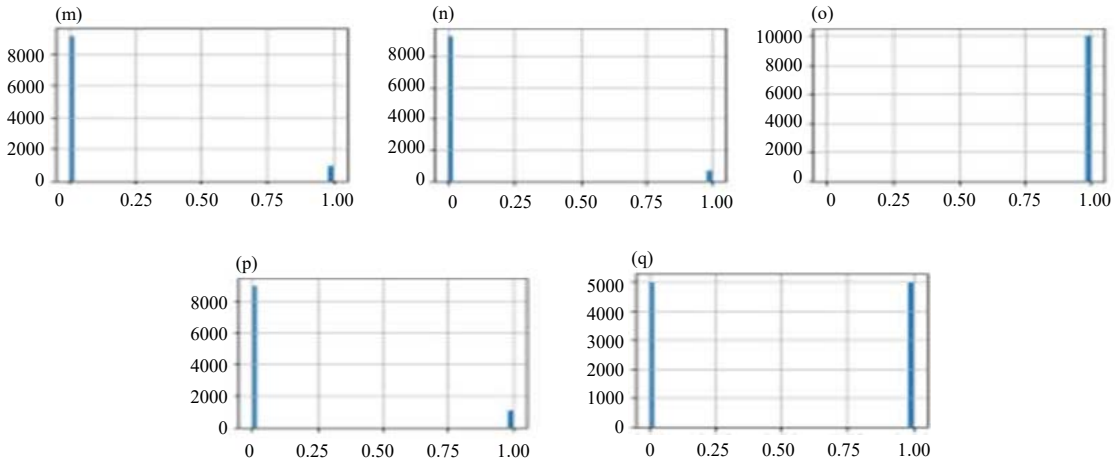


Fig. 5(a-q): Distribution plot of dataset base on the features selected, (a) Have_IP, (b) Have_At, (c) URL_length, (d) URL_depth, (e) Redirection, (f) http_domain, (g) Tiny_URL, (h) Prefix_suffix, (I) DNS_record, (j) Web_traffic, (k) Domain_age, (l) Domain_end, (m) iFrame, (m) Mouse_over, (o) Right_click, (p) Web_forwards and (q) Label

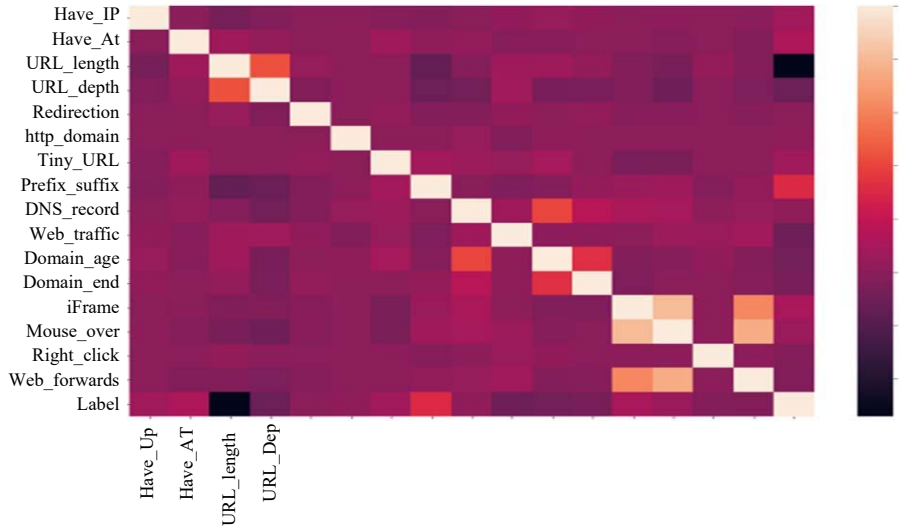


Fig. 6: Correlation heat map of the dataset

Over 6 machine learning models and deep neural network algorithms were used in the model development process to detect phishing URLs using Jupyter notebook IDE with packages such as pandas, Beautiful Soup, who-is, urllib and others.

The models are shown in Fig. 8 and their accuracy was tested using sklearn matrices with an accuracy score. The XG Booster model achieved the highest performance score of 86.6%, followed by the Multilayer Perceptions model at 86.5%, the Decision Tree model at 81.4%, the Random Forest model at 81.8%, the Support Vector Machine model at 80.4% and the Auto Encoder Neural Network model at 16.1%.

General working of the system: "Phish-BusterV2," a one-page phishing detection web application, has been developed to run on any browser. HTML, CSS, PHP and JavaScript were used in the development of the application.

The following pages are available in the phishing detection web application:

- **Home page:** The home page includes a session in which a user can enter a URL and predict whether it is phishing or legitimate.

Figure 9 a and b shows how it predicts the state of the URL based on the feature selection. The goal of this

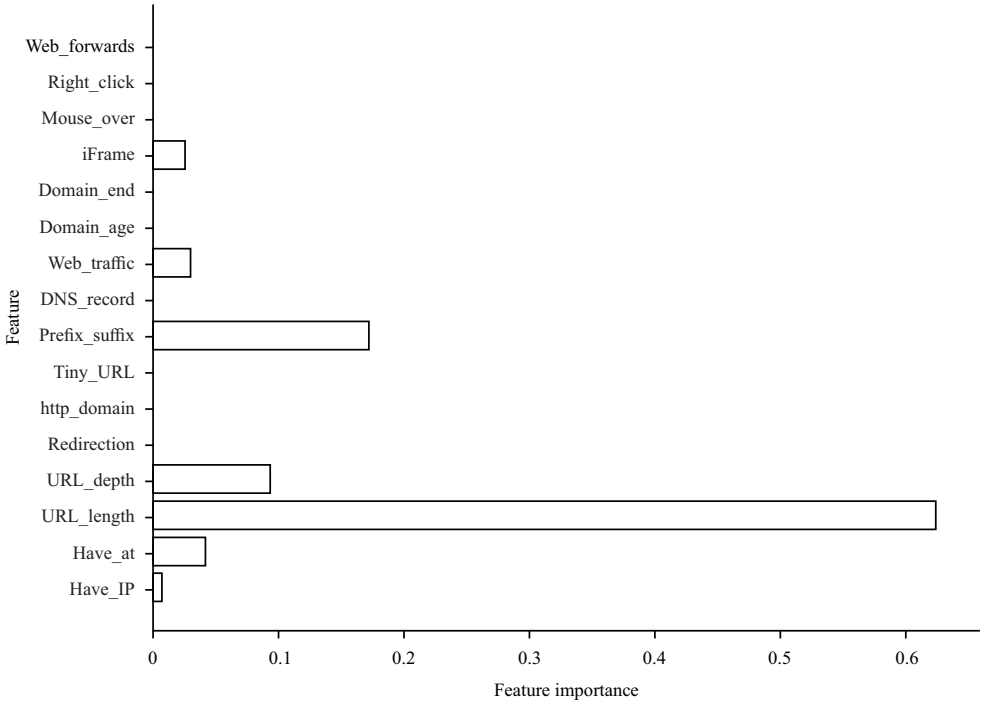


Fig. 7: Feature importance for decision tree classifier

```
#Sorting the dataframe on accuracy
results.sort_values(by=['Test Accuracy', 'Train Accuracy'], ascending=False)
```

49]:

	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.866	0.864
2	Multilayer Perceptrons	0.865	0.864
0	Decision Tree	0.814	0.812
1	Random Forest	0.818	0.811
5	SVM	0.804	0.794
4	AutoEncoder	0.161	0.177

Fig. 8: Accuracy performance of models

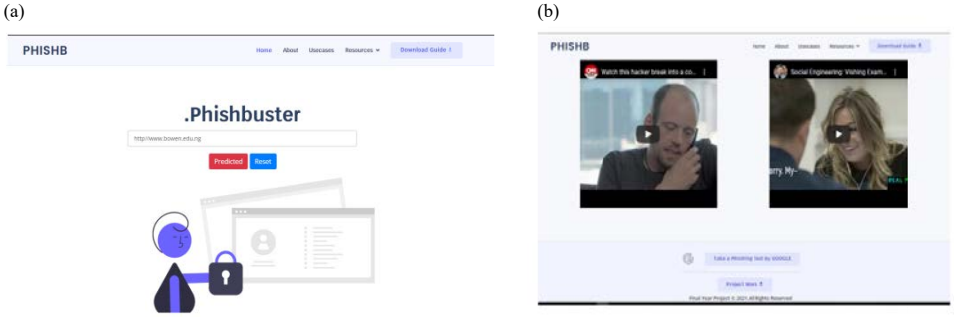


Fig. 9(a-b): (a) The home page and (b) The home page footer

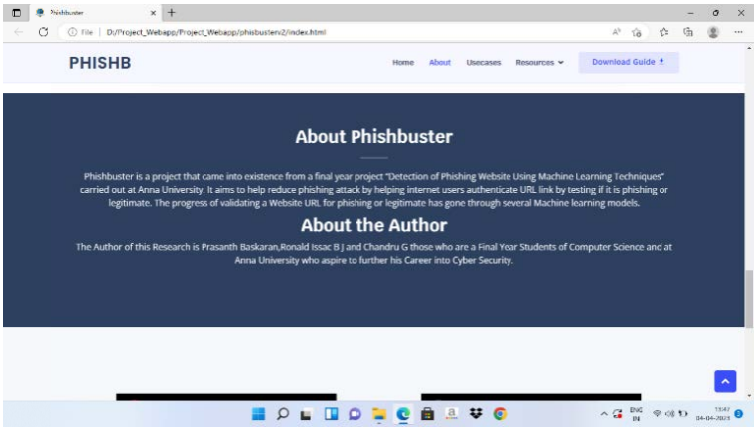


Fig. 10: The about page

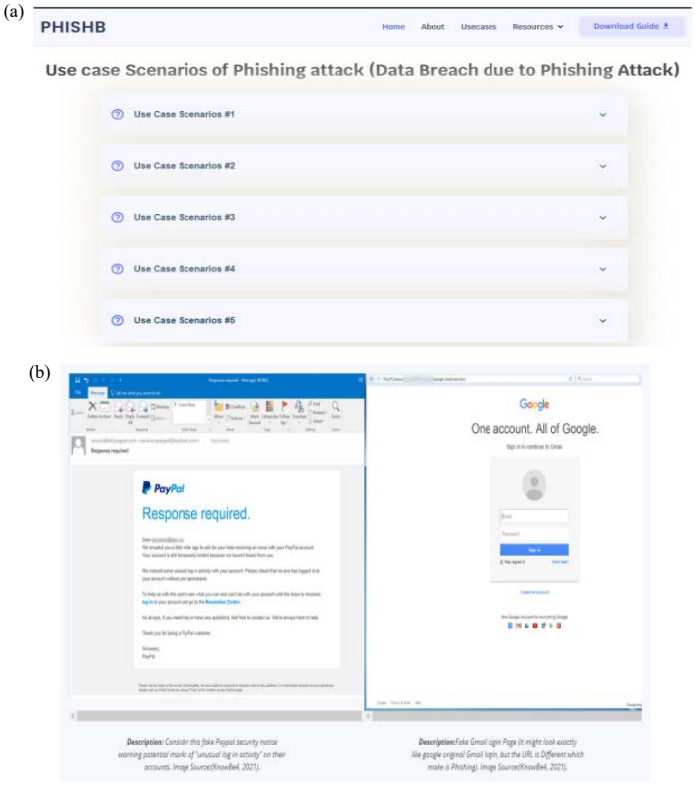


Fig. 11(a-b): Use case page

page is to assist users in validating a URL link as well as to provide various resources on phishing attacks. A Google phishing test can also be taken to help the user understand how to detect phishing messages and URLs. Users can also download a book containing information and other resources on phishing.

- **About page:** Figure 10 depicts the about page, which contains information about the application as well as the author of the project

The use case page: The use case page contains various case scenarios of phishing attacks that occurred in previous years to various companies. It also includes images of phishing attack messages sent by phishers, as illustrated in Fig. 11.

- **Resource page:** The resource page it contains various phishing resources, such as the definition, types and techniques of a phishing attack, as well as reference links to the source from which the

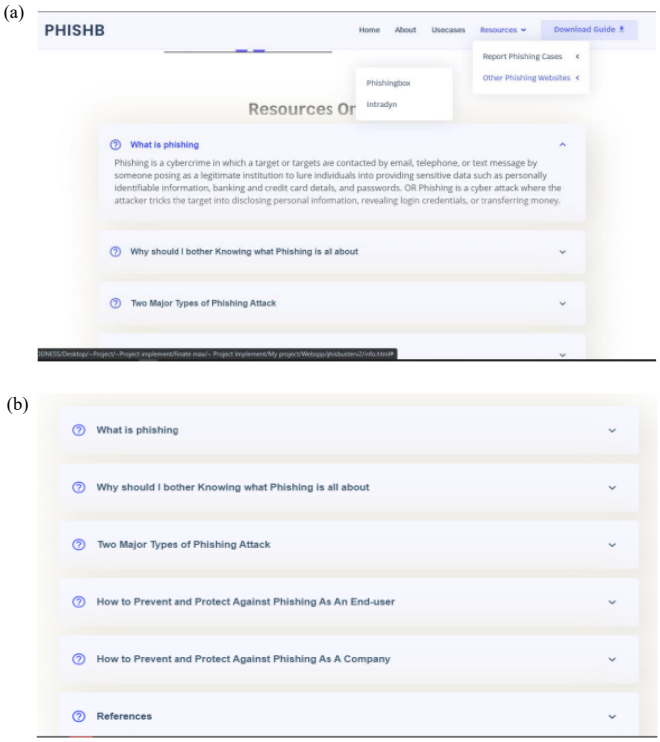


Fig. 12(a-b): Resource page

content was retrieved. It also contains 2 sub-session links: The first session reports a phishing case and the second session contains a phishing website

First session: The Report Phishing Case link directs users to Google Support and Google Safe Browsing to report any phishing attack, be it a URL or a phishing email.

Second session: Phishing website link directs users to 2 websites with useful information and a phishing test. Fig. 12 depicts two of these sites: Phishing box and intradyn.

CONCLUSION

Phishing attacks are a rapidly growing cyber threat that costs internet users billions of dollars each year. It entails employing a variety of socialengineering techniques to obtain sensitive information from users. As a result, Phishing techniques can be detected through a variety of modes of communication, such as email, instant messaging, pop-up messages and web pages. This project was able to categorise and recognise how phishers carry out phishing attacks, as well as the various ways in which researchers have assisted in the detection of phishing. As a result, the proposed system for this project used various feature selection, machine learning and deep neural networks to identify patterns, including Decision Tree, Support Vector Machine, XG Booster, Multilayer Perceptions, Auto Encoder Neural Network and Random Forest. The Model with the highest accuracy based on the feature extraction algorithm used to distinguish phishing URL links from legitimate URL links was integrated into a web application where users could enter website URL links to determine whether they were legitimate or phishing. Using machine learning models and deep neural network algorithms, the system developed determines whether a URL link is phishing or legitimate. The feature extraction and models used on the dataset aided in the unique identification of phishing URLs, as well as the performance accuracy of the models used. It is also surprisingly accurate at determining the legitimacy of a URL link.

RECOMMENDATION

Through this project, one can know a lot about phishing attacks and how to prevent them. This project can be taken further by creating a browser extension that can be installed on any web browser to detect phishing URL Links.

REFERENCES

1. Abdelhamid, N., F. Thabtah and H. Abdel-Jaber, 2017. Phishing detection: A recent intelligent machine learning comparison based on models content and features. EEE International Conference on Intelligence and Security Informatics (ISI), July 22-24, 2017, IEEE, China, pp: 72-77.
2. Shaikh, A.N., A.M. Shabut and M.A. Hossain, 2016. A literature review on phishing crime, prevention review and investigation of gaps. 10th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), December 15-17, 2016, IEEE, China, pp: 9-15.
3. Gupta, B.B., A. Tewari, A.K. Jain and D.P. Agrawal, 2016. Fighting against phishing attacks: State of the art and future challenges. Neural Comput. Applied, 28: 3629-3654.
4. Almomani, A., B.B. Gupta, S. Atawneh, A. Meulenberg and E. Almomani, 2013. A survey of phishing email filtering techniques. IEEE Commun. Surv. Tutorials, 15: 2070-2090.
5. Camp, W.G., 2001. Formulating and evaluating theoretical frameworks for career and technical education research. J. Vocational Educ. Res., 26: 4-25.
6. Kulkarni, A. and L.L. Brown, 2019. Phishing websites detection using machine learning. Int. J. Adv. Comput. Sci. Appl., Vol. 10. 10.14569/ijacsa.2019.0100702.