www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.03
TPI 2019; 8(1): 764-770

www.thepharmajournal.com
Received: 02-10-2018
Accepted: 06-11-2018

Anshu Sharma
Assistant Professor, Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Correspondence
Anshu Sharma
Assistant Professor, Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Phishing detection system using machine learning: Safeguarding online security with advanced analytical models

Anshu Sharma

DOI: https://doi.org/10.22271/tpi.2019.v8.i1m.25411

Abstract
Since many of our regular activities-including financial transactions, work-related activities, and other daily activities-now take place online, we are more vulnerable to the dangers posed by cybercrime. Phishing attacks based on URLs are among the most frequent threats that internet user's encounter. Rather than taking advantage of software bugs, the attacker in this kind of attack targets human weakness. The malicious software preys on people and institutions, deceiving them into opening secure websites, and either obtains private data or infects our computers with malware. Various machine learning algorithms are being used to identify phishing URLs, In order to distinguish between genuine and phishing URLs, researchers constantly strive to enhance the precision and overall efficiency of existing techniques. Our objective in this study is to validate various system learning approaches, as well as datasets and URL functionalities utilized to train these machine learning models. We examine and discuss the effectiveness of various machine learning algorithms as well as techniques for raising their accuracy metrics. Developing a survey resource to educate researchers about recent advancements in the field is the aim. They will be able to contribute to phishing detection models that produce outcomes that are more precise as a result.

Keywords: Phishing, machine, Safeguarding, models

Introduction
The evolution of communication technologies and the advent of digitalization–has enabled life to become faster and more accessible, particularly when the COVID-19 pandemic lockdown occurred and everything had to be ordered online, including basic necessities, i. e. instead of needing to complete transactions and shopping in person [1]. All you have to do is turn on your smart device, look up the website you want, and finish your everyday chores. For instance, a drugstore, retail establishment, educational website, or bookshop. E-services are also becoming more popular, which raises the possibility that hackers will obtain or misuse a user's personal data, including name, phone number, identity, and credit card number [2]. As a result, users now have to deal with a variety of online threats and cyberattacks on a daily basis. One of them, phishing, occurs primarily through emails. Phishing is a type of cybercrime in which victims' personal identity information or financial account credentials are obtained via social engineering and technical deception [3].
Phishing is a form of cybercrime in which perpetrators fabricate counterfeit versions of reliable websites in an attempt to trick victims into disclosing sensitive data, including passwords, banking or credit card information, and usernames. Customers may receive these phishing URLs via text, instant messaging, or email.
The Phishing detection system is designed to deal with cyberattacks caused due to fake web links that steal confidential information using malware viruses, etc. This first of all dataset of phishing and legitimate emails will be collected and processed further for analysis. A particular machine learning algorithm is used either a decision tree or a neural network to analyse the dataset. Machine learning (ML) is an artificial intelligence subfield that focuses on using parts to make predictions. It is similar to (and often overlaps with) computational measurements.

Machine learning is tightly related to scientific advancement, which establishes the field's theories, methods, and application domains. However, data mining is a subset of data mining that focuses more on the initial collection of information; machine learning is sometimes added to it. This process is known as unsupervised learning. Additionally, ML can be used in an unsupervised manner to identify significant anomalies by first learning and establishing patterns for a variety of entities [4].

Machine learning (ML) techniques are finding more and more applications in cybersecurity. One of the best solutions to counter zero-day attacks is machine learning, which starts with the classification of IP traffic and separates malicious traffic for intrusion detection. By using ML techniques and measurable traffic characteristics, new exploration is being conducted. The year 1987 saw the introduction of the word "phishing." Phishing is a type of cybercrime where personal information is stolen, including identity and private data. It is a form of extortion in which the attacker gains total access to the personal information of others. Many ways to deal with the issue of phishing attacks have been put forth as their frequency has grown [5]. There exist multiple approaches to constructing a framework that ensures a phishing attack resolution. It is also possible to use various Other methods for detecting phishing attacks. Among them are image-based detection techniques, machine learning-based detection, canteen-based detection, fuzzy rule-based detection, black lists, heuristic detection, and white list-based detection. Numerous other studies cover various approaches and strategies for identifying various phishing attack types. Many people struggle to identify phishing websites because they appear to be legitimate websites. Certain browsers have built-in ant phishing techniques [6].

Background

Phishing Detection: Sending users malicious links that appear genuine in an effort to trick them into clicking on them is known as a URL-based phishing attack. Phishing detection involves analysing an incoming URL for a number of traits to identify whether or not it is phishing, after which it is classified correctly. To discern between a genuine URL and a phishing attempt, differing machine learning algorithms are trained on disparate datasets of URL attributes [7].

Phishing detection Approaches: Two lists-the whitelist and the blacklist-are used in the list-based approach to distinguish between authentic and phishing URLs. Only if the URL is whitelisted is access to the website allowed. Blacklisting is employed. The structure of a phishing URL is examined in the heuristic-based method [8]. A phishing URL pattern is established. URLs are categorized based on this pattern. Websites are accurately classified by comparing the visual similarity of their pages, and the visual similarity of the URL is a major factor in this process. To determine whether websites are fraudulent or not, a server-side.

Analysis is conducted. Given that fraudulent websites frequently have designs that closely mimic authentic ones, this data is then compared to the original website using image processing techniques. Techniques for image processing are more successful in spotting small differences that users might miss [9].

The content of pages is analysed by the content-based approach. This method extracts features from the page content as well as from external services like search engines and DNS servers. These features may include words like brand names that attackers incorporate into the URL to create the illusion of a genuine website. The presence of these words at different positions in the URL is assigned weights to determine their significance. After selecting the most likely terms, Yahoo Search is asked to return the domain name that appears the most frequently among the top 30 results [10]. To determine whether or not a website is phishing, the domain name owners are compared. In order to verify the legitimacy of web pages, a logo image was used in conjunction with a comparison between legitimate and fraudulent websites [11].

Uncertain variables can be processed using a fuzzy rule-based method, and human experts can be added to classify the variables and their relationships. This method uses a predefined set of rules and metrics to categorize web pages according to the level of phishing they contain. Fuzzy logic systems have fewer features, according to the experimental findings reported in the paper. The efficacy of a classifier will decline if irrelevant features have an impact on a fuzzy logic algorithm, and vice versa.

In the machine learning-based approach, a given URL is classified as either legitimate or a phishing attempt by building machine learning models that make use of supervised learning algorithms. To find out each model's performance, several algorithms are tested after being trained on a dataset [12]. The Performance of the model is directly impacted by any inconsistencies in the training data. Nonetheless, this approach offers effective phishing detection techniques and is a significant area of research. On the subject of machine learning-based phishing detection, many papers have been published.

Machine learning Algorithm

Decision Tree Algorithm

The machine learning algorithm that is most frequently used. The decision tree approach is simple to use and comprehend. Among the characteristics that can be utilized for classification, the ideal splitter is selected at the start of the decision tree's procedure, also known as the tree's root. The tree is constructed by the algorithm up until a leaf node is encountered [13]. The decision tree generates a training model that forecasts the target value or class in a tree representation. Every leaf node in the tree belongs to the class label, and every internal node to the attribute. Nodes in decision tree algorithms are computed using the information gain method and the Gini index [14].

Random Forest Algorithm

One of the most potent algorithms in machine learning technology is the random forest algorithm, which was developed from the decision tree algorithm. It operates by organizing multiple decision trees into a forest. The accuracy of detection rises as the number of trees increases. One way to create trees is with the bootstrap method. To create a single tree, the bootstrap method randomly selects features and dataset samples, then replaces them with new ones [15].

The random forest algorithm chooses the best splitter among a set of randomly chosen features, much like the decision tree algorithm does. The Random Forest algorithm selects the optimal splitter for classification by utilizing information gain techniques and the Gini index. This process will keep going until n number of trees are produced by the random forest. The algorithm then counts the votes for each predicted target after each tree in the forest makes a prediction regarding the

target value. The final prediction made by the random forest algorithm is based on which predicted target received the most votes.

Neural Network

Neural networks are one type of machine learning model that takes its cues from the structure and functions of the human brain. It is composed of a vast number of layer-organized neurons, which are basic processing units. The neurons in each layer are connected to neurons in the previous and subsequent layers, forming a network. After processing information from neighbouring neurons, each neuron quickly calculates and transmits the result to neurons higher up in the layer. The input layer, hidden layer(s), and output layer are the three primary layer types that make up a neural network's structure. Following the data's receipt by the input layer, it is processed by one or more hidden layers using various computations to extract relevant features. The final output, which may be a probability distribution, a prediction, or a classification label, is then generated by the output layer. The computations performed by the neurons are based on weights and biases. The weights determine the strength of the connection between neurons, while the biases determine the threshold at which a neuron is activated. In order to reduce the discrepancy between the expected and actual outputs, the weights and biases of the neurons are changed during training. This
usually carried out with the aid of an optimization algorithm, like stochastic gradient descent. The capacity of neural networks to recognize intricate patterns in data is one of its main advantages. To do this and generate predictions that are more accurate, weights and biases are adjusted during training. Backpropagation is the process of propagating the error from the output layer back through the network, changing the weights and biases in the hidden layers. The capacity of neural networks to produce precise predictions on Their use in a range of applications, including natural language processing, picture classification, and speech recognition, has grown in popularity recently. For instance, by teaching a neural network to identify features like ear shape, fur color, and skin texture, one can train the network to distinguish between images of cats and dogs.

Literature Survey

a perceptive piece that summarizes the current state of knowledge, including important findings and theoretical and methodological dedications to a specific subject, is called a literature survey. Email's introduction has made communication easier, which has resulted in an increase in unsolicited bulk emails—particularly phishing attacks—and unsolicited messages overall. The issue of phishing attacks has led to the development of numerous anti-phishing strategies. The way messages are represented is one of the primary classification factors. This paper focuses on determining whether important emails are spam or if they are important emails. Making decisions about which features to use and how to use them during the categorization process is essential. It is well known that many researchers have used artificial intelligence to build intelligent systems, and many of them have applied deep learning to cybersecurity systems as well. Using an optimal feature selection technique and also using a neural network to detect phishing websites (OFS-NN), a very effective model of phishing website detection is presented. An index known as the feature validity value (FVV) has been created in this proposed model to evaluate the impact of each feature on the identification of such websites.
An algorithm is now developed to identify the best features from the phishing websites based on this newly generated index. The neural network's over-fitting issue can be mostly resolved with the help of this chosen technique. The neural network is then trained using these ideal features to create an optimal classifier that recognizes phishing URLs.
While feature knowledge will heavily influence the model's accuracy, feature engineering is a key component in the search for solutions in order to identify phishing websites.The time required to gather these features is the limitation, even though the features drawn from all of these different dimensions make sense. Web Crawler based Phishing Attack Detector (WC-PAD), a three-phase detection method, has been proposed to accurately detect phishing attempts. The authors have proposed a multidimensional phishing detection feature approach that uses deep learning (MFPD) to address this flaw by emphasizing a fast detection method. In this case, the input features are the URL, traffic, and web content. This completes the classification after these features are taken into account. Deep learning-based Phishing Net is a technique for quickly detecting phishing URLs.
A detection system was created to adapt to both phishing websites and the ever-changing environment. In this approach, different types of distinctive features are taken into accountfrom the source code of URLs and webpages, negating the need for third parties to be involved. One method that has been proposed to ascertain whether a webpage is phishing or authentic is parse tree validation. This is a clever way to find these websites: using Google's API, intercept each hyperlink on a current page, then create a parse tree using all of the intercepted hyperlinks. The Depth-First Search (DFS) algorithm is used in this method to start parsing from the root node and check whether any child nodes share the root node's value.
For the detection of phishing websites by URL method, the Random Forest classifier was used as a solution. The program integrates phishing website detection, validation, and collection. This online tool checks and identifies phishing websites while continuously monitoring the PhishTank blacklist. This framework, called "Fresh-Phish," uses phishing websites to generate machine learning data. Thirty distinct website features are queried using Python to generate a very large dataset, which is then used to compare the various ML classifiers to see which has the highest accuracy. This model looks at the model's training time and accuracy.
In a recent study, Qabajeh et al. (2018) conducted a comparison of conventional and automated techniques for detecting phishing attempts. Traditional methods for combating phishing involve various strategies such as increasing awareness, educating users, providing periodic training or workshops, and implementing legal measures. On the other hand, automated anti-phishing techniques employ list-based and machine learning-based approaches. The research paper investigates the similarities, advantages, and drawbacks of these approaches from both the user's perspective and performance standpoint.
The study came to the conclusion that rule induction and machine learning work well to thwart phishing attacks. The review of only 67 research items and the absence of Deep Learning techniques for phishing website detection are the study's two main limitations.
Kunju and associates. (2019) used a survey approach to detect

phishing attacks and put forth a number of fixes and techniques for doing so. The study found that a large number of suggested remedies were ineffective at stopping phishing attempts. There are just 14 studies from 2007 to 2019 in the literature reviewed for this work. Only machine learning methods for phishing website detection are covered in this study.

Athulya and Praveen (Athulya and Praveen, 2020) talked about different kinds of phishing attacks, the newest methods that phishers are using, and counter-methods. The article also seeks to highlight and increase awareness of phishing attacks techniques used to identify phishing attempts. This study suggests that educating users about the various forms of phishing attacks is an effective way to prevent phishing attacks. To identify phishing attacks, The right security software, such as anti-phishing browser extensions, can be chosen by users. The literature review for this work includes nine research items. The study excludes the use of deep learning techniques to detect phishing websites.

Arshad and associates. (Arshad and others, 2021) offered a range of phishing and anti-phishing techniques during their inquiry. The SLR's assessment indicates that spear phishing, email spoofing, phone phishing, and email manipulation are the most often used phishing techniques.

This study found that machine learning techniques produced the highest accuracy. There are only 20 studies in the research.

Catal and associates. Catal et al. 2022) conducted a thorough review of the literature and addressed nine research issues. Finding, evaluating, and synthesizing Deep Learning methods for phishing detection is the primary goal of the project. 42 out of 43 studies, according to this study, used supervised machine learning algorithms. DNN was the most popular algorithm, and it and the hybrid DL algorithm produced the best results.

Methodology

Data collection

Assembling a dataset of phony and authentic emails. The process of obtaining information for a particular goal or research project from a variety of sources, including websites, databases, and sensors, is referred to as data collection. To train and test a machine learning model for phishing detection through machine learning, data collection entails gathering a representative dataset of phishing and legitimate emails. The data can be obtained from various sources, such as publicly available datasets, online repositories, or through manual collection by researchers. The collected data should be diverse and balanced in terms of the number of legitimate and phishing emails, and should represent different types of phishing attacks and techniques. The quality and quantity of the collected data is crucial for the success of the machine learning model and its ability to accurately detect phishing emails in real- world scenarios.

Data pre-processing

Cleaning and pre-processing the dataset by turning the text that remains into numerical features and eliminating information that isn't relevant, like email headers or signatures.

Data processing refers to the transformation of raw data into a more meaningful form that can be used for analysis or decision-making. In the context of phishing detection using machine learning, data processing involves various steps such as cleaning, pre- processing, feature engineering, and selection.

Cleaning refers to the removal of any noise or irrelevant data from the dataset, such as duplicates, irrelevant features, or incomplete records. Pre-processing is putting the data in a format that is appropriate for machine learning algorithms. Examples of this include scaling the data, handling missing values, and turning categorical variables into numerical values.

Feature engineering involves selecting relevant features or variables from the dataset that can help the machine learning algorithm to differentiate between legitimate and phishing emails. This may include features such as the sender's email address, subject line, content, or metadata. Feature selection involves reducing the number of features to only those that are most important or informative for the model [16].

In order to enhance the quality of the dataset and the predictive accuracy of the model, data processing is a crucial stage in the creation of a machine learning-based phishing detection system.

Feature extraction

Identifying and selecting relevant features that can discriminate between legitimate and phishing emails. The process of choosing a subset of pertinent and educational features is known as feature selection. (also known as variables or attributes) from a larger set of features in a dataset. In machine learning, the choice of features used to train a model is critical to the model's performance. Irrelevant or redundant features can lead to overfitting or poor generalization, while informative features can improve the model's accuracy and interpretability. Feature selection techniques aim to identify and retain only the most relevant and informative features for a given task, while discarding the rest.

This enhances the model's efficiency, lowers its complexity, and raises the caliber of the predictions the model makes. The following feature category is taken from the URL data: 1. Address Bar-based Features: Nine features are extracted in this category. 2. Four features are extracted in this category based on domains. 3. Four features are extracted from this category, which is based on HTML and Javascript.

Training

Utilizing machine learning techniques to train a classification model on the chosen features, such as logistic regression, decision trees, or neural networks. The process of teaching a model to make accurate predictions through exposure to labeled examples-data with known inputs and outputs-is known as training in machine learning. Accurately mapping inputs to outputs is made possible by the model's ability to be trained with a set of weights or parameters. The model iteratively modifies its parameters during training in response to variations in the actual outputs in the training data and its predicted outputs. We refer to this procedure as parameter estimation or optimization.

The model learns from its mistakes and updates its parameters to reduce prediction errors. The data is divided into 80-20 groups prior to the ML model training process. e. two thousand testing samples and eight thousand training samples. The input URL for this data set is categorized as either legitimate (0) or phishing (1), posing a categorization issue. To train the dataset for this project, supervised machine learning models (classification) such as Decision Tree,

Random Forest, Multilayer Perceptrons, XGBoost, Autoencoder Neural Network, and Support Vector Machines were considered.

The quality of the training data, the choice of model architecture, and the training algorithm used are all critical factors that can affect the performance of the trained model. The trained model can then be used to make predictions on new, unseen data.

Evaluation

Testing the trained model on a separate validation dataset and assessing its effectiveness with the use of metrics like F1-score, recall, accuracy, and precision. An assessment of a phishing detection system's effectiveness entails determining how well it detects and categorizes phishing attacks. A a collection of assessment metrics, such as area under the ROC curve (AUC), recall, accuracy, precision, and F1 score are commonly used to achieve this.

To evaluate a phishing detection system, Usually, a dataset comprising authentic and well-known phishing emails or URLs is used. The system is trained on a subset of the dataset, and the remaining data is used for testing. The Based on the dataset, it is clear that this is a supervised machine-learning task. Regression and classification are the two primary categories of supervised machine learning problems.

system's performance is then evaluated by comparing its predictions against the known classifications in the test dataset.

The evaluation of a phishing detection system is important because it allows researchers and practitioners to determine the effectiveness of the system and identify any areas for improvement.

Deployment

Integrating the trained model into a real-time phishing detection system that can classify incoming emails as either legitimate or phishing. Deployment of a phishing detection system using machine learning involves making the system available for use in a production environment. This usually entails making sure the system is scalable, dependable, and secure in addition to integrating it with the hardware and software infrastructure already in place.

The deployment process may involve additional testing and validation to ensure that the system works as intended in real-world scenarios. The system may also need to be continuously monitored and updated to maintain its effectiveness in detecting new and evolving phishing attacks.

Deployment also involves considerations such as user interfaces, documentation, and training for end-users and administrators. It is important to ensure that the system is easy to use and understand, and that users have the necessary knowledge and skills to effectively utilize the system.

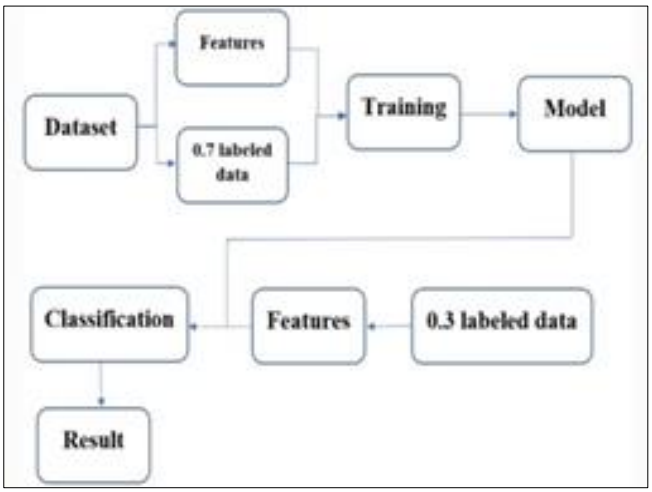

Fig 1: Flowchart of the phishing detection system.

Implementation and Result

Five data files are used for extraction of data and to analyse which url is phishing and which is legitimate.

1. Benign_list_big_final
2. online-valid
3. legitimate
4. urldata
5. phishing

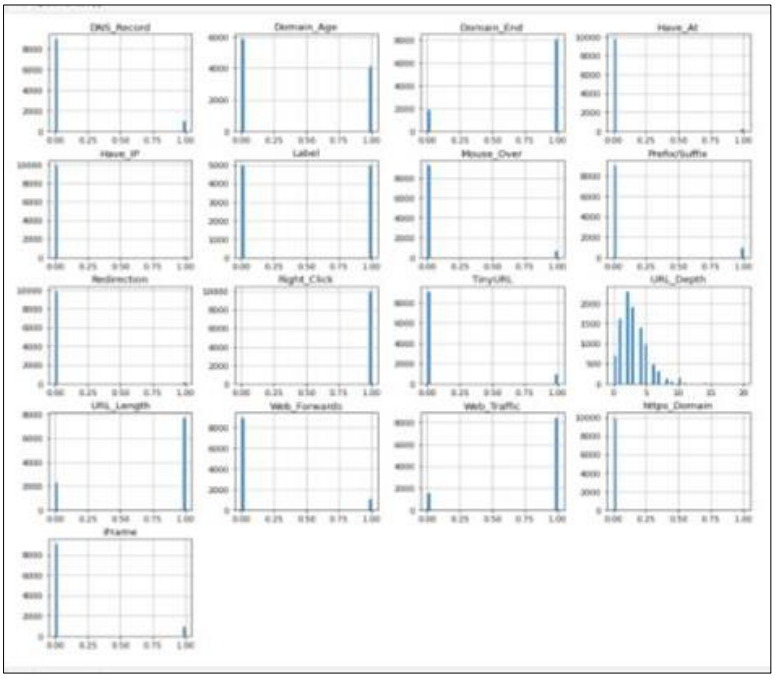

Fig 2: Plotting the data distribution

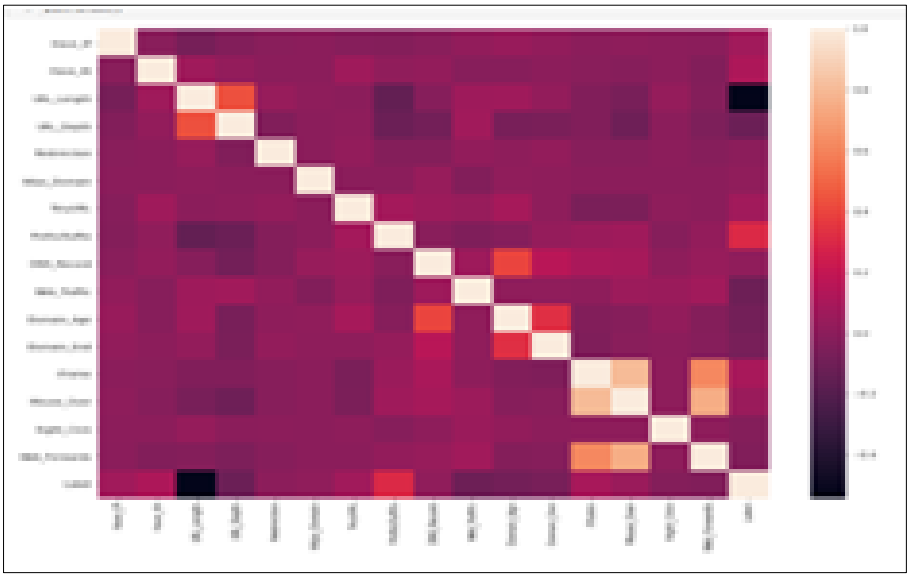

Fig 3: Correlation heatmap

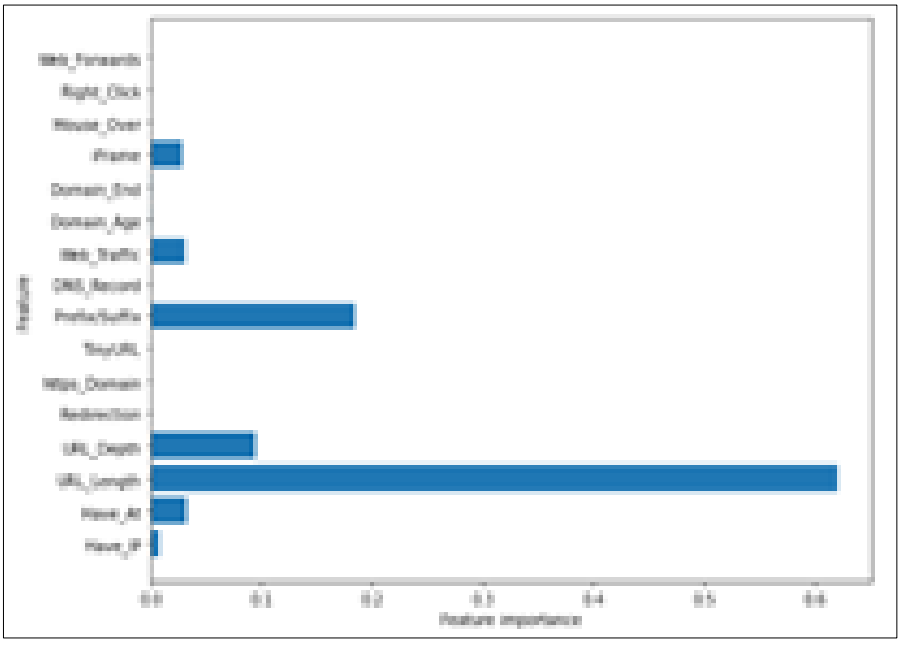

Fig 4: Checking the feature importance in the Model using decision tree classifier

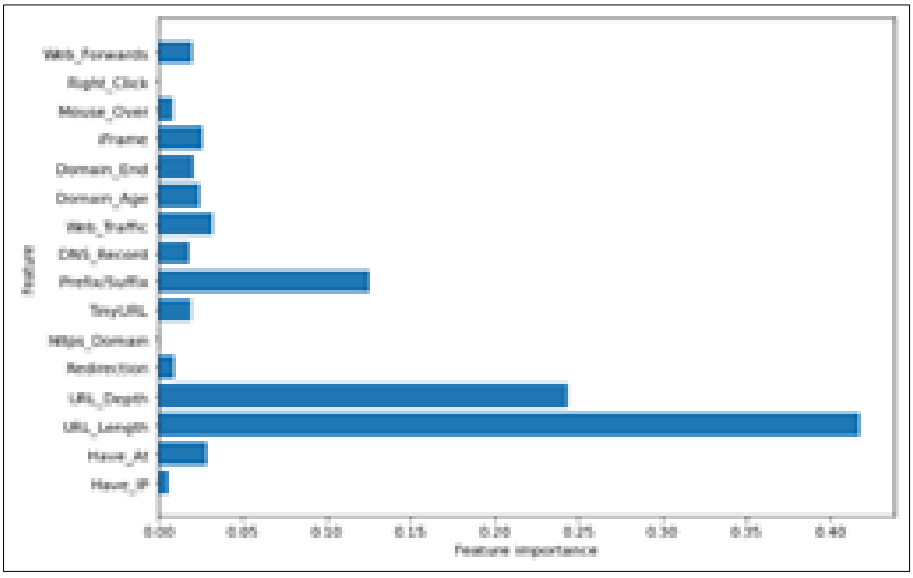

Fig 5: Checking the feature importance in the Model using Random Forest classifier.

Out[90]:

	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.866	0.864
2	Multilayer Perceptrons	0.858	0.863
1	Random Forest	0.814	0.834
0	Decision Tree	0.810	0.826
4	AutoEncoder	0.819	0.818
5	SVM	0.798	0.818

Fig 6: Comparison of models

Conclusion

Phishing emails have become a common problem in recent years. Phishing emails are cleverly constructed social engineering emails in which the target is tricked into sending sensitive information to the sender via email. It is more likely that young users will fall prey to phishing attacks because of their naiveté. Additionally, users with the trait of agreeable behaviour are more likely to be lured by phishing scams than those with the trait of disagreeable behaviour. Women are more likely than men to give fraudulent emails and websites access to their personal and financial information. Internet usage patterns can explain the causal relationship between gender and social engineering. Therefore, it's essential to detect that kind of email. There are several methods available for identifying phishing emails. There are some restrictions, though, like the poor accuracy. The content may be identical to that of a legitimate email, making detection difficult and low in detection rate. The purpose of this study was to classify emails as either phishing or non-phishing by capturing inherent attributes of the email text and other characteristics. To achieve better results, machine learning techniques were employed. The results of this research have improved the accuracy of detecting phishing emails. A comparison between three supervised datasets was conducted between these classifiers. Many machine learning algorithms demonstrate robust performance metrics and effective classification. The phishing detection procedure and methods discussed in the most recent research literature were examined in this article. New researchers can use this study as a reference to better understand the procedure and create more accurate phishing detection systems.

References

1. Sarker IH. Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective. SN Comput Sci. 2021;2:154.
2. Kim D, Kim Y-H, Shin D. Fast attack detection system using log analysis and attack tree generation. Cluster Comput. 2019;22(1):1827-1835.
3. Phishing Detection using Machine Learning-based URL Analysis: A Survey. Available from: https://www.ijert.org/research/phishing-detection-using-machine-learning-based-url-analysis-a-survey-IJERTCONV9IS13033.pdf.
4. Kaushik P, Yadav R. Reliability design protocol and blockchain locating technique for mobile agent. J Adv Sci Technol (JAST). 2017;14(1):136-141. https://doi.org/10.29070/JAST.
5. Kaushik P, Yadav R. Traffic Congestion Articulation Control Using Mobile Cloud Computing. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(1):1439-1442. https://doi.org/10.29070/JASRAE.
6. Kaushik P, Yadav R. Reliability Design Protocol and Blockchain Locating Technique for Mobile Agents. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE.
7. Kaushik P, Yadav R. Deployment of Location Management Protocol and Fault Tolerant Technique for Mobile Agents. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE.
8. Kaushik P, Yadav R. Mobile Image Vision and Image Processing Reliability Design for Fault-Free Tolerance in Traffic Jam. J Adv Scholarly Res Allied Educ (JASRAE). 2018;15(6):606-611. https://doi.org/10.29070/JASRAE.
9. Available from: https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100210.html.
10. An intelligent cyber security phishing detection system using deep learning techniques. Available from: https://link.springer.com/article/10.1007/s10586-022-03604-4.
11. Available from: https://www.academia.edu/43230944/A_REPORT_on_DETECTION_OF_PHISHING_WEBSITE_USING_MACHINE_LEARNING.
12. Louridas P, Ebert C. Machine learning. IEEE Software. 2016;33:110–115. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8504731/.
13. Ariyadasa S, Fernando S, Fernando S. Detecting phishing attacks using a combined model of LSTM and CNN. Int J Adv Appl Sci. 2020;7:56–67.
14. Saoji S. Phishing detection system using visual cryptography. 2015.
15. Patil S, Dhage S. A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework. In: 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS); c2019. p. 588–593.