International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

Detection of Phishing Sites Using Machine Learning Techniques

Ayan Mahmood[1], Vishal Pandey[2], Rohit Raj[3], Gouri Shankar Mishra[4]

[1,2,3]UG, Computer Science & Engineering, Sharda University, Greater Noida, Uttar Pradesh, India.

[4]Associate Professor, Computer Science & Engineering, Sharda University, Greater Noida, Uttar Pradesh, India.

Email id: *2020401758.ayan@ug.sharda.ac.in[1], 2020490363.vishal@ug.sharda.ac.in[2], 2020561.rohit@ug.sharda.ac.in[3], gourishankar.mishra@sharda.ac.in[4]*

Orcid id: *0009-0005-8733-8868[1], 0009-0001-9865-1164[2], 0009-0003-3305-2114[3], 0000-0002-9103-3478[4]*

Abstract

Phishing is a very commonly occurring attack in which the attacker attempt to get the private information of the user like card details, their passwords their transaction details etc. using fake copy websites. Attackers uses the websites very similar to the original websites and not possible for common people to identify. Phishing is the biggest loop-hole in the cyber world. Phishing became a successful business for phishers. Other than fake websites phishers uses different methods to do this job using messaging, spoofed links to make money and to counter Phishing various method are proposed some anti-phishing techniques are blacklist, whitelist but due to exponential growth on new innovations and technologies these techniques falls a little below as new websites contains dynamically work in it due to that these techniques cannot outperform todays websites.

Keywords: *Domain, Phishing, Machine Learning, SVM, MLP, Heat Map.*

1. Introduction

Phishing represents a deceptive strategy that melds elements of social manipulation with technological trickery to illicitly gain access to individuals' personal information and financial assets. It remains a cornerstone and prevalent method employed by malicious attackers seeking to compromise the security of sensitive data [1]. In a typical phishing scenario, perpetrators craft misleading emails that pose as genuine communications from reputable companies or organizations. These deceptive messages are artfully designed to entice recipients into visiting counterfeit websites, where they are coerced into divulging confidential information, including email login details, passwords, financial data, and more. The initial step in these attacks often entails sending out spurious emails that closely mimic authentic correspondences or feign legitimacy as requests from well-established businesses [2]. If the assailant replicates the authentic website entirely, mirroring approximately 50% of the recent attacks we have investigated, identifying such instances becomes a straightforward challenge [13]. These emails frequently incorporate links that steer unsuspecting victims toward fake websites cleverly constructed to mirror their legitimate counterparts. Once ensnared on these counterfeit sites, victims are induced to input sensitive information, which the malicious attackers can then exploit for various nefarious purposes. Phishing attacks possess the capacity to inflict substantial economic losses on a global scale. They tend to concentrate their efforts on targeting payment companies and webmail services, as evidenced by recent research conducted by the Anti-Phishing Working Group (APWG) [15]. Phishing websites often closely mimic legitimate websites, making them appear almost identical. The structure of phishing websites tends to be relatively simple, allowing them to be created and deployed quickly when compared to the more time-consuming development process of legitimate websites. Phishing

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

sites usually lack the complexity of design seen in typical, authentic websites [3]. Current phishing detection technology primarily focuses on analyzing the features of individual webpages while overlooking the overall topology and structure. The rapid development of global networking and communication technology has led to a significant shift in users' daily activities, with electronic banking, social networking, and e-commerce moving to the digital realm. However, this migration to an undisclosed, unchecked, and cyberspace infrastructure has created a fertile ground for cyberattacks, which pose severe security vulnerabilities. These vulnerabilities affect not only inexperienced users but also entire networks. Despite the importance of experienced users, it is challenging to completely shield them from falling victim to phishing scams [4]. Cyberpunk often exploits the individual characteristics of skilled users in their attempts to trick them. To counter the menace of phishing, numerous studies have scrutinized various facets of malicious websites, encompassing domain characteristics such as website URLs and content, combinations of these elements, as well as examinations of source code and screenshots. Additionally, there is a pressing demand for effective anti-phishing tools capable of identifying malicious URLs within a company's network to protect its users. Machine Learning (ML) techniques have emerged as a promising solution for detecting malicious URLs on the internet. Conventional methods for URL detection predominantly rely on the maintenance of a blacklist containing known malicious websites, necessitating routine updates to scrutinize URLs for potential threats. However, cybercriminals have devised techniques, including Domain Generation Algorithms (DGA), to evade detection by crafting new, malicious URLs that elude identification using established methods. In response to these challenges, researchers have introduced Machine Learning-based approaches that leverage diverse features to identify counterfeit URLs. Machine learning (ML) is a versatile approach utilized in supervised learning for the creation of predictive models. Addressing the issue of phishing attacks involves transforming it into a classification problem. This entails using labeled historical data related to websites to train and assess the model's performance. When these models are incorporated into web browsers, they become capable of detecting phishing activities effectively. Nonetheless, the effectiveness of automated phishing detection models can be further enhanced when website features are integrated into the input dataset, encompassing a broad array of websites.

2. Experimental Methods or Methodology

Vaibhav Patil and colleagues in their study introduces a machine learning-based approach to differentiate between legitimate and malicious websites, utilizing a diverse data-set comprising URL attributes, domain information, and website content. Diverse machine learning models, comprising decision trees, support vector machines, and neural networks, are employed for classification. The study emphasizes the importance of feature selection and engineering to enhance performance and minimize false positives.Logistic Regression achieved an accuracy of 96.23%, accurately classifying 6447 accurate negatives and 2287 accurate positives, but showcased 325 wrong positives and 17 wrong negatives.The Decision Tree algorithm gained the same accuracy as Logistic Regression (96.23%), with 6393 true negatives and 2341 true positives, but had 326 wrong positives and 16 wrong negatives.Random Forest excelled both, achieving 96.58% accuracy, with the highest true positives (2374) and the lowest wrong positives (297) and wrong negatives (13). In summation, this research paper shows the importance of machine learning in identifying and thwarting phishing websites, with Random Forest providing the best performance by gaining high accuracy while minimizing false positives and negatives [8-10]. In this research paper by Arun Kulkarni and Leonard L. Brown, the main focus is on the detection of phishing attacks and the requirement for automated methods to identify such malicious websites.Their suggested model utilized a dataset consisting of 1,353 samples , encompassing nine characteristics, which

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

encompassed Backend Form Handler, Pop-up Windows,Transport Layer Security (TLS), Request Uniform Resource Locator (URL),Anchor URL, Online traffic, URL size, Domain Age, Presence of IP Address in the URL, and Category. The authors employed various machine learning techniques in their research, including Naïve Bayesian classifier, Decision Tree, Neural Network and Support Vector Machine. The results of the experiments conducted in this research showed that the Decision tree classifier gained the highest accuracy rate of 91.5%, followed by Support Vector Machine (SVM) at 86.69%, Naïve Bayes' Classifier at 86.14%, and Neural Network with the lowest accuracy of 84.87% [5]. In the article titled "Detecting Phishing Websites Using Machine Learning " the author introduces a method to tackle the problem of phishing. They propose an approach that utilizes the Random Forest technique to build a phishing detection system. This system employs machine learning. Is specifically designed as a browser extension making it easily integrable in practical scenarios. The author conducted research to analyze various features associated with phishing websites aiming to identify the most effective combination, for training the classifier. The investigation yielded results with an accuracy rate of 98.8% achieved by carefully selecting a set of 26 features. Overall this system shows promise in enhancing internet security through proactive detection of phishing websites [6]. The article titled "Detecting Phishing Websites Using Multidimensional Features and Deep Learning" introduces an approach called Multidimensional Feature Phishing Detection (MFPD) to combat phishing websites. Traditional methods of detecting phishing often rely on feature engineering and prior knowledge which can be time consuming and less efficient. MFPD on the hand utilizes deep learning to quickly and accurately detect phishing. In the step it examines character sequences in URLs and employs deep learning for classification eliminating the need for external assistance or prior knowledge of phishing indicators. In the step MFPD combines features from various dimensions such, as URL statistics webpage code webpage text and outcomes from deep learning classification. This integration simplifies the process of establishing a detection threshold. The authors assessed MFPD using a dataset comprising both phishing and legitimate URLs, achieving an impressive accuracy rate of 98.99% [7].

3. Methodology

3.1 Collection of Data

Gather a dataset of websites, labelled as phishing or legitimate with the attributes which could help more than being good or bad as the result may in good accuracy as well as good precision. Extract impactful features such as URL, domain age, IP address etc. As shown in Figure 1 Which describes the attributes quanity. Pre-process the data by handling missing values, encoding categorical features, and scaling numerical features. Now based on the dataset we can select the features based upon address bar, HTML & JavaScript, Domain [11]. And these features can be expanded further like address bar consist of URL, special symbols for sites, length of the domain, redirection ‘//’, http, URL shortening techniques as well as suffix and prefix.

Domain Features are:

- DNS Record
- Traffic on Website
- Domain Age
- Duration of Domain

Also one of the most important factor that can play a major role is age of any website or domain. As for **HTML & JavaScript:**

- I-Frame Redirection
- Disabling Right Click
- Status Bar Customization
- Forwarding Websites

And all the features are already converted into the forms of data that can be read by our machine.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

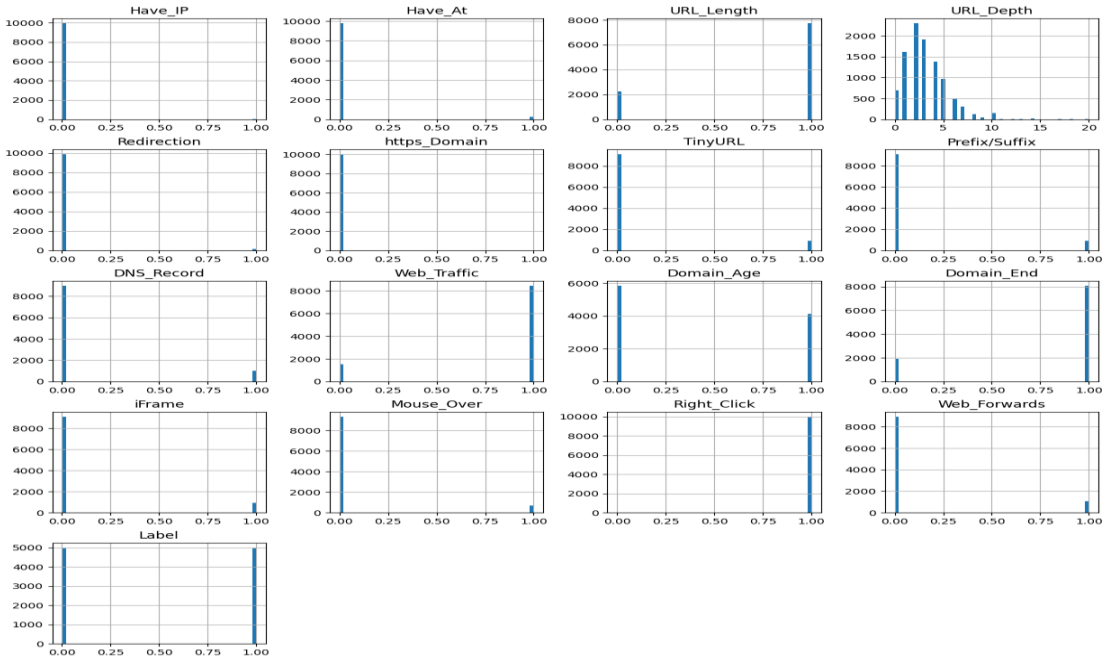

Figure 1 Classification of Attributes

3.2 Data Splitting

Split the dataset into three subsets: training, validation, and testing. Common splits are 80% for training, 20% testing [12].

3.3 Feature Selection

Use techniques like feature importance from Random Forest or XGBoost to selecting the most relevant and explorable features, reducing dimensionality and potentially improving model performance. Figure 2 (Describes the Relation between the variables or attributes)

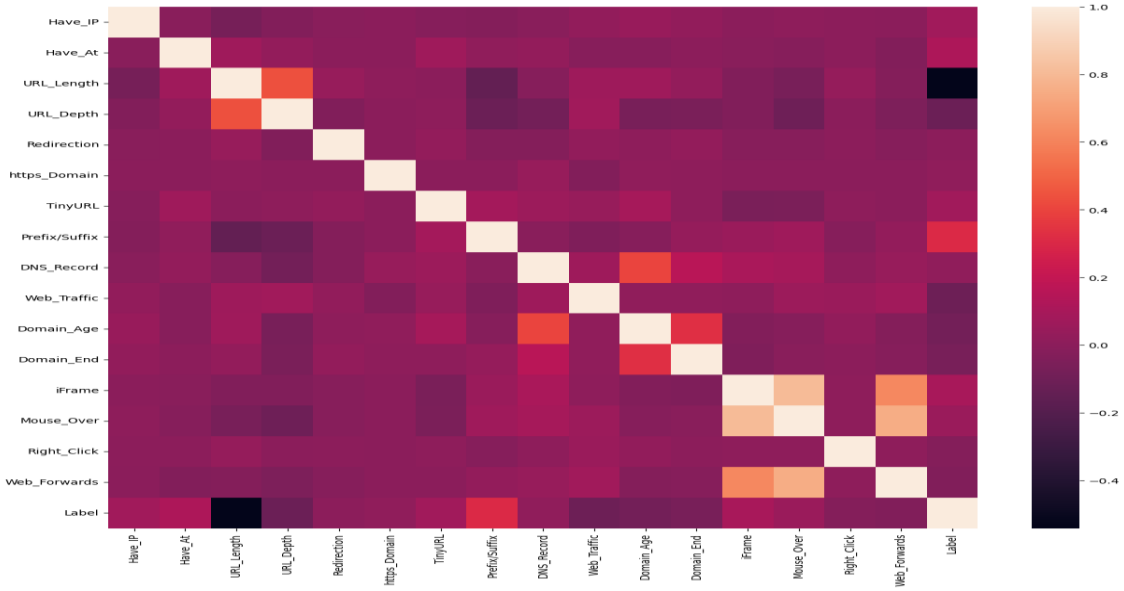

Figure 2 Heat map of Attributes

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

3.4 Model Selection

Choose the five models we have mentioned (SVM, MLP, Random Forest, XGBoost, Decision Trees) for building classifiers.

3.5 Model Training

Training each model according to the relevant most relevant features according to each models which may result in good accuracy.

3.6 Model Evaluation

Evaluate each model's performance on the test dataset using appropriate metrics. Common metrics include precision, accuracies on test data and training data. Assess the models' ability to correctly classify phishing websites while minimizing false positives.

3.7 Model Deployment

Once satisfied with the models' performance, deploy them in a production environment for real-time or batch processing of new website URLs. Implement a monitoring system to continuously evaluate model performance and update them as needed.

3.8 Education and Awareness

Educate users and organizations about the risks of phishing attacks and encourage best practices for safe online behavior. Remember that the effectiveness of each model may vary depending on the characteristics of the dataset and the specific features chosen. Regularly updating and re-evaluating the models is crucial for maintaining their accuracy in detecting evolving phishing threats.

4. Models

4.1 Support Vector Machines (SVM)

SVM stands for support vector machine which comes under supervised learning which can be used for problems like regression and classification. It is well suited for binary classification problems, where the goal is to separate data into classes. SVM work significantly as its find the most optimal hyperplane that separates it to dimensional space which is relevant as it may find easy to find-out the most useful features [14].

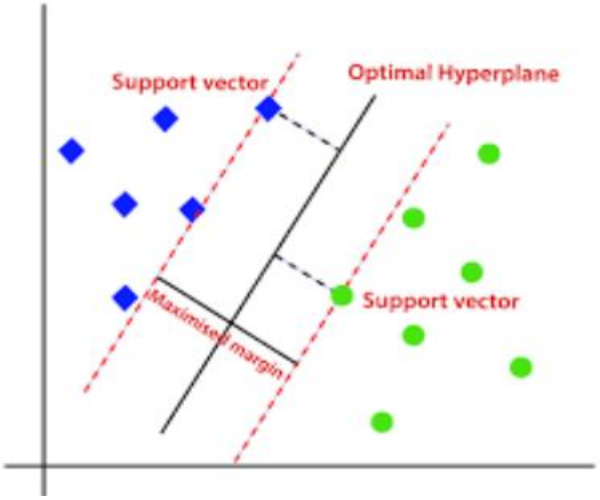

Figure 3 Optimal Hyperplane

It will start with the dataset where each data point is with an associated class. The primary goal of SVM model is to identify a decision boundary (hyperplane) shown in Figure 3,4 that separate data into classes. The boundary should maximize the margin i.e. the distance between the hyperplane and the nearest data point from the present class these classes are also known as support vectors. The margin is maximized by finding the hyperplane that has the largest minimum distance to the support vectors and the accuracy is 80.3 %. The equation for the hyperplane can be described as:

$$wTx+ b = 0$$

The distance between the point's xi and the decision boundary can be measured as:

$$di = wT.xi + b / ||w||$$

So to sum it up here are some key features:

- SVM can handle high-dimensional data efficiently, making it suitable for a variety of applications, including different classification applications.
- SVMs have regularization parameters that help prevent overfitting, ensuring that the model generalizes well to fresh data.

SVM can be used with different kernel functions (linear, polynomial etc.) to model complex, non-linear relationships between data points.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

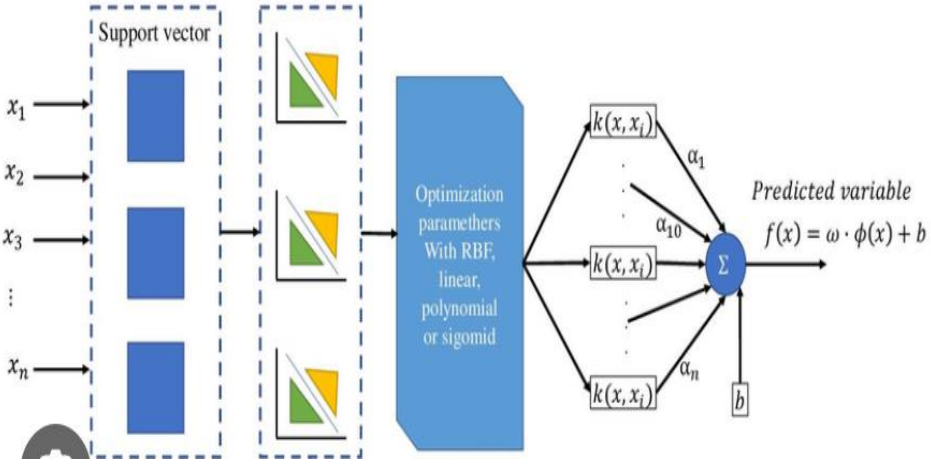

Figure 4 SVM Architecture

4.2 Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron is a type of Neural Network used in both machine learning as well as deep learning. It may be used for variety of use such as classification, regression, pattern recognition which makes it one of the finest algorithm available to reap the maximum fruits.

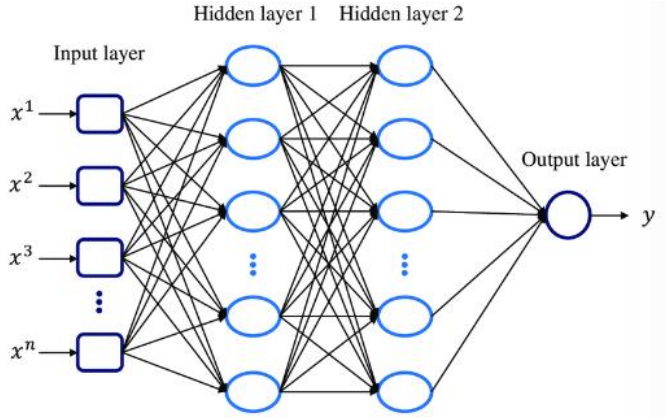

Figure 5 MLP Model

It starts with the interconnected artificial neurons also known as nodes which are inspired by neurons inspired by human brain. Each node consist of set of inputs, perform a weight sum of those inputs and adds a bias then passes and calculated the results with the activation function and produce output as shown in Figure 5.

Formula:

Weight = weight + learning rate * (expected - predicted) * x

Activation functions brings in non-linearity to the model. Some of the common Activation functions are Sigmoid, Tanh, ReLU. They help the network model complex relationships in the data whose accuracy comes out to be 85.8%. MLPs are versatile and can be used into a wide variety of tasks, including image classification, NLP, speech recognition, and financial forecasting.

- As we know deep learning is one of the finest algorithm that explores through the inherent patterns which a human may have difficult to find on.
- Its little bit expensive as the computational cost increases
- MLPs are a fundamental building block of deep learning, and their significance lies in their ability to learn complex patterns and representations from data.
- MLPs are the principle for Neural Network, which have achieved a significant amount of results in various fields like computer vision, NLP, and speech recognition.
- MLPs can impulsively learn hierarchical features from dataset, eliminating the work of hand-crafted features engineering in many cases.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

4.3 Decision Trees

A Decision Tree is a popular supervised machine learning algorithm used for both classification and regression tasks. It's a tree-like structure where each internal node represents a feature, each branch represents a decision or a rule, and each leaf node represents an outcome or a class label. Decision trees are particularly valued for their simplicity, interpretability, and versatility.

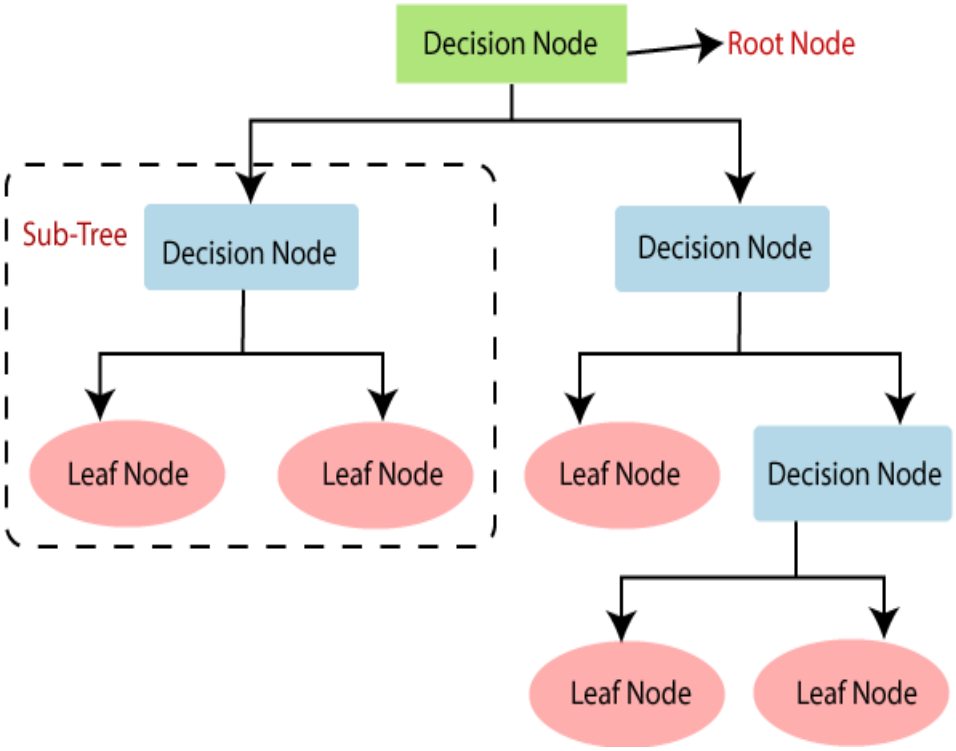

Figure 6 Architecture

It works like each internal node represents a decision rule based on the selected feature and threshold. For example, in a binary classification task, a decision node might split the data into "if feature A is greater than X, go left, otherwise go right." Shown in Figure 6. And Accuracy comes out to be 81.6%.

Decision trees provide a clear, interpretable representation of decision-making processes, which is valuable in fields like medicine and finance.

4.4 XG-Boost

XG-Boost, which stands for e-Xtreme Gradient Boosting, is a powerful machine learning algorithm that has proven to be highly effective in various applications, including phishing detection and the architecture has been described in Figure 7 means how does it works.

Formula:

$$\text{Similarity Score} = (\Sigma \text{ Residuals})2/\Sigma[P\ (1 - P)] + \lambda$$

P = Probability

λ= Regularization Parameter

XGBoost is an ensemble learning method, which means it combines the predictions of multiple weak learners (typically decision trees) to create a stronger, more accurate model. XGBoost has gain popularity in the field of phishing detection as it have the ability to handle imbalanced datasets, its robustness against overfitting, and its high predictive accuracy which is 86.7 %.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

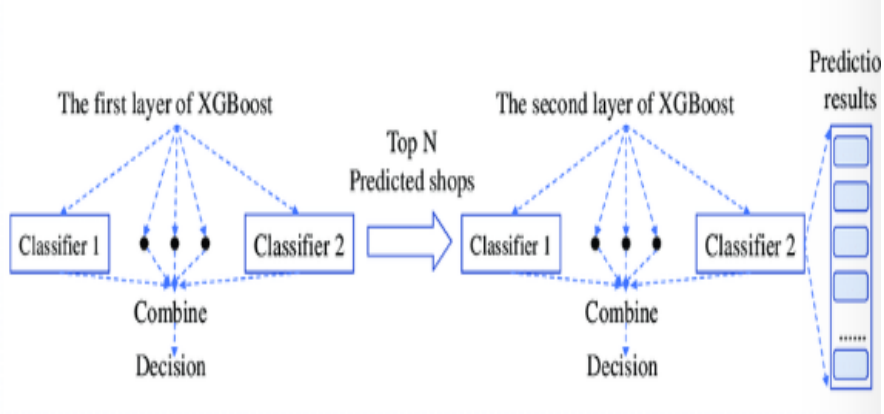

Figure 7 XG Boost Architecture

XG-Boost (Extreme Gradient Boosting) is an ensemble learning method based on decision trees. Its significance includes:

- XG-Boost is known for its exceptional predictive performance and has been used to win numerous machine learning competitions on platforms like Kaggle.
- It offers built-in regularization techniques to prevent overfitting, making it more robust.
- XG-Boost can naturally handle missing data, reducing the need for data pre-processing.

4.5 Random Forest

Random Forest is one of the finest machine learning algorithms that is commonly been using for various classification tasks, including the detection of phishing websites [12]. It is a band learning method that builds multiple decision trees and merge their predictions to improve accuracy and reduce overfitting. . It is a versatile and impactful tool for phishing detection due to its ensemble nature and ability to capture complex patterns between URL features and the likelihood of a website being a phishing site [10]. And accuracy comes out to be 82.1%.

- Random Forest is another ensemble learning method based on decision trees, and its significance lies in its robustness and versatility:
- By averaging the predictions of multiple decision trees, Random Forest helps in reduces overfitting and provides more stable results.
- It can estimate the importance of features, helping in feature selection and understanding the data.

Results

The results produce by the models shown in Table 1.

Table 1 Results Produced by Models

S.no	Model	Training Accuracy	Testing Accuracy
1.	Decision Tree	81.6%	80.3%
2.	Random Forest	82.1%	81.4%
3.	MLP	85.8%	84.7%
4.	SVM	80.3%	79.6%
5.	XGBoost	86.7%	85.3%

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 02 February 2024
Page No: 210 - 219
https://irjaeh.com

And the most appropriable and consistent results are produced by XGboost. Which have the best accuracy of 87% also it produces more accurate results other than previous modelws used in the previous times.

```
63/63 [==============================] - 0s 3ms/step
PREDICTED :
Mallicious
Mallicious
Non Mallicious
Mallicious
Mallicious
Non Mallicious
Non Mallicious
Non Mallicious
Non Mallicious
Mallicious


ACTUAL :
Mallicious
Mallicious
Non Mallicious
Mallicious
Mallicious
Non Mallicious
Non Mallicious
Non Mallicious
Mallicious
Mallicious
```

Figure 8 Predicted Results

Conclusion

A well-established principle for effective phishing website detection is the need for a strong balance between real-time performance, accuracy, and a low false positive rate. Our proposed Multi-Feature Phishing Detection (MFPD) approach aligns with this principle. It utilizes a dynamic category decision algorithm to swiftly identify suspicious URLs without relying on prior phishing knowledge, ensuring high-speed detection. Furthermore, the inclusion of multidimensional feature detection guarantees accurate identification. To validate the effectiveness of our MFPD approach, we conducted a series of experiments using a substantial dataset containing numerous phishing and legitimate URLs. The results demonstrate that MFPD is highly effective, exhibiting impressive accuracy, a minimal false positive rate, and swift detection capabilities. Our future developments will explore the integration of deep learning for enhanced feature extraction from webpage code and text. Additionally, we plan to create a browser plugin to seamlessly incorporate our approach into web browsers.

References

[1] Deshpande, Atharva, Omkar Pedamkar, Nachiket Chaudhary, and Swapna Borde. "Detection of phishing websites using Machine Learning." International Journal of Engineering Research & Technology (IJERT) 10, no. 05 (2021).

[2] Odeh, Ammar, Ismail Keshta, and Eman Abdelfattah."Machine learningtechniquesfor detection of website phishing: A review for promises and challenges." In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0813-0818. IEEE, 2021.

[3] Alkawaz, Mohammed Hazim, Stephanie Joanne Steven, and Asif Iqbal Hajamydeen. "Detecting phishing website using machinelearning." In 2020 16th IEEE International Colloquium onSignal Processing

& Its Applications (CSPA), pp. 111-114. IEEE, 2020.

[4] Ubing, Alyssa Anne, Syukrina Kamilia Binti Jasmi, Azween Abdullah, N. Z. Jhanjhi, and Mahadevan Supramaniam. "Phishing website detection: An improved accuracythrough feature selection and ensemble learning." International Journal of Advanced Computer Science and Applications 10, no. 1 (2019).

[5] Kulkarni, Arun D., and Leonard L. Brown III. "Phishing websites detection using machine learning." (2019)

[6] Alswailem, Amani, Bashayr Alabdullah, Norah Alrumayh, and Aram Alsedrani. "Detecting phishing websites using machinelearning." In 2019 2nd International Conference on Computer Applications &Information Security (ICCAIS), pp. 1-6. IEEE, 2019.

[7] Yang, Peng, Guangzhen Zhao, and Peng Zeng. "Phishing website detection based on multidimensional features driven by deep learning." IEEE access 7 (2019): 15196-15209.

[8] Patil, Vaibhav, Pritesh Thakkar, Chirag Shah, Tushar Bhat, and S. P. Godse."Detection and prevention of phishing websites using machine learning approach." In 2018 Fourth international conference on computing communication control and automation (ICCUBEA), pp. 1-5. Ieee, 2018.

[9] Islam, Mazharul, and Nihad Karim Chowdhury. "Phishing websites detection using machine learning based classification techniques." In International Conference on Advanced Informationand Communication Technology, Chittagong, Bangladesh 2016.

[10] A.Odeh, I. Keshta and E. Abdelfattah, "Machine LearningTechniquesfor Detection of Website Phishing: A Review for Promises and Challenges," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, 2021, pp. 0813-0818, doi: 10.1109/CCWC51732.2021.9375997. keywords: {Support vector machines;Deep learning;Whitelists;Phishing;Conferences;Training data;Tuning;ensemble learning;deep learning;cyber-attacks}

[11] Peng Yang, Guangzhen Zhao, Peng Zeng. "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning", IEEE Access, 2019

[12] Choon Lin Tan, Kang Leng Chiew, Kelvin S.C. Yong, San Nah Sze, Johari Abdullah, Yakub Sebastian. "A Graph-Theoretic Approach for the Detection of Phishing Webpages", Computers & Security, 2020

[13] S. Afroz and R. Greenstadt, "PhishZoo: Detecting Phishing Websites by Looking at Them," 2011 IEEE Fifth International Conference on Semantic Computing, Palo Alto, CA, USA, 2011, pp. 368-375, doi: 10.1109/ICSC.2011.52.

[14] V Shruthy, S Pragathishwari, M Nediga, Rajesh George Rajan, M Sreyaa. "Phishing Prediction on Website Updates with Novel Features Through Machine Learning", 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), 2023

[15] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.