Phishing Website Detection

S. Pranav Ranjan[1], D. NithishKumar[2], S. R. Senthil Kumar[3], R. Reena[4]
[1, 2] Dept of Computer Science and Engineering
[3]Assistant Professor, Dept of Computer Science and Engineering,
[4]Associate Professor, Dept of Computer Science and Engineering
[1, 2, 3, 4] Prince Shri Venkateshwara Padmavathy Engineering College, Ponmar, Chennai

Abstract- *This proposed paper is on the detection of phishing websites that are used to steal sensitive information from users. These URLs are usually masked using emails with an urgent or congratulatory tone. This generally results in leaking of one's sensitive information for a scammer to use. The objective of this project is to use Machine Learning to identify patterns in these kinds of URLs and to prevent the user from getting scammed and enabling them to surf the internet safely.*

Keywords- Phishing, Machine Learning, Python, Gradient Boosting Classifier, Website, Browser

I. INTRODUCTION

Generally, websites are used to establish credibility and to build trust among its users but there are certain individuals who want to misuse the data that users enter while using the websites. This is called Phishing and it is commonly happening around the world to target innocent users of the internet.

The phishing website detector was designed keeping in mind the current disadvantages of phishing detectors and to help people to secure them from phishing threats. Phishing attacks are the simplest way to obtain sensitive information from innocent users. The aim of the scammers is to steal sensitive information like user credentials, credit card numbers etc. This data can be used for future advertisements or ITA (Identity Theft Attack). The frequently used attack method is to send emails to potential victims, which seem to be sent by banks, online organizations, or ISPs. Cybersecurity experts are currently looking for reliable and trustworthy detection techniques for detecting phishing websites. This paper describes machine learning technologies that detect phishing URLs by extracting and analyzing various characteristics of legitimate phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs using Gradient Boosting Classifier machine learning algorithm and to determine a safety score in terms of percentage (%) for each website.

II. EXISTING SYSTEM

So, to prevent this type of phishing attacks, a tool is needed which examines the URL entered by the user and checks if the website is malicious or not. In the paper [1], the authors mention the rise in the phishing websites and the method they used to detect and prevent it. For the detection of a safe or a phishing webpage, they implemented the Largest Common Substring (LCS) method. They prepared a database of phishing websites of their own and tested the LCS on the new website.

In [2]the authors first extracted the URLs and tested the Processed URLs with 3 machine learning algorithms Support Vector Machine (SVM), Naive Bayes (NB), Extreme Learning Machine (ELM). In that ELM has scored the highest 96.4%.

In this [3] described a new approach to detecting phishing websites using machine learning algorithms. First, build a web link network using nodes that represent web pages. Edges between nodes represent reference relationships that connect web pages through hyperlinks or similar textual content. SPWalk then applies network embedding techniques to map the nodes to a low-dimensional vector space. Use the node as a numerical characteristic to run the experiment and classify the website as legitimate or phishing. Achieve (95% or more).

In this paper we proposed a machine learning based algorithm to check if the URL is legitimate or not with also specifying accuracy score. To overcome the drawbacks from the above-mentioned researches this paper proposes a model and design to identify and detect the phishing URL that help the people to not getting scammed online. Our system can be used by both the public and also by cybersecurity enthusiasts.

III. PROPOSED SYSTEM

The proposed system uses Gradient Boosting Classifier algorithm to test the attributes extracted from the URLs to check for safety. Firstly, the user enters the URL in the search area then the URL is extracted into different attributes e.g., https, short URL, IP address, DNS, anchor URL etc. like this the URL is extracted in 30 different attributes. These 30 attributes act as a test data for the training model, then these test data is passed to the Gradient Boosting classifier along with the already tested datasets.

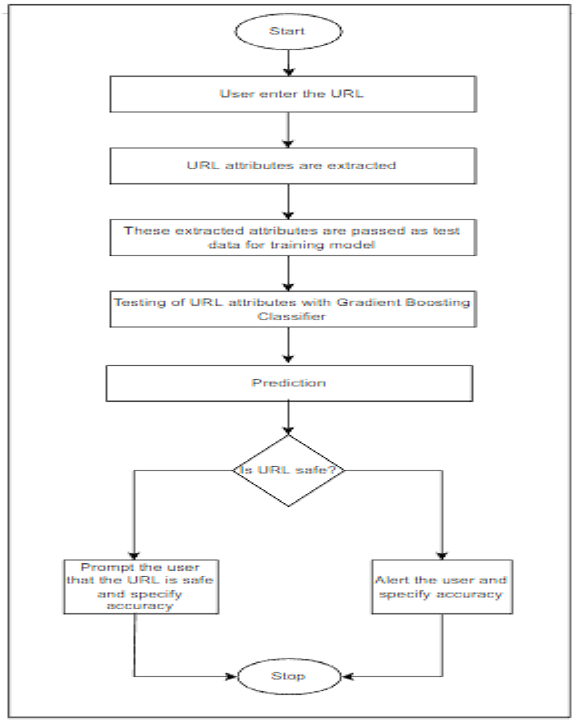

Start
User enter the URL
URL attributes are extracted
These extracted attributes are passed as test data for training model
Testing of URL attributes with Gradient Boosting Classifier
Prediction
Is URL safe?
Prompt the user that the URL is safe and specify accuracy
Alert the user and specify accuracy
Stop

Fig 1. Flowchart

The classifier model is already trained on phishing website datasets to test newly entered or newly generated URLs. The classifier will predict the entered URL as a malicious website or predict it as a safe website. If it is a phishing site, users will be warned that if they continue their credentials and other information is at risk of being hacked and if it's a secure site the user can go forward to access the page.

IV. ALGORITHM USED

Gradient Boosting Classifier: Gradient Boosting classifiers are a family of machine learning algorithms that combine many weak learning models to create powerful predictive models. Decision trees are typically used in this algorithm. Gradient boosting models are becoming more and more popular because they are effective in classifying complex data sets. The gradient boosting classifier depends on the loss function. User-defined loss functions are available and many standardized loss functions are supported by gradient boosting classifiers, but the loss functions must be differentiable.

The gradient boosting system has two other necessary parts. Weak learners and additional components. Gradient boosting systems use decision trees as weak learners. Regression trees are used by weak learners, and these regression trees output actual values. The output is an actual value, so you can sum the output of the regression tree when a new learner is added to the model to correct the prediction error. An additional component of the gradient boosting model comes from the fact that trees are added to the model over time, and when this happens, existing trees are not manipulated and their values remain fixed. A method similar to the gradient descent method is used to minimize the error between the specified parameters. This is done by getting the calculated loss and performing the gradient descent method to reduce that loss. Then change the parameters of the tree to reduce the residual loss. Then the output of the new tree is added to the output of the old tree used in the model. This process repeats until a certain number of trees are reached or the loss falls below a certain threshold.

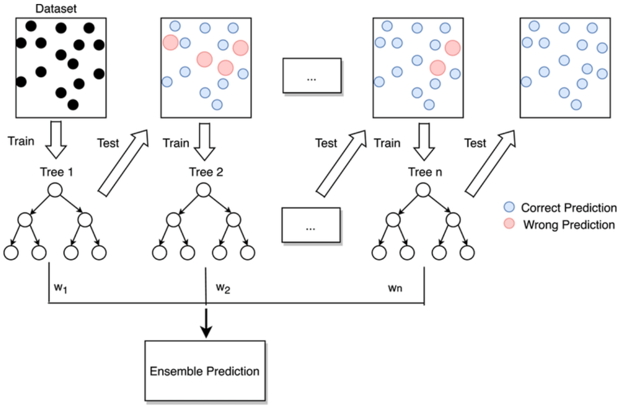
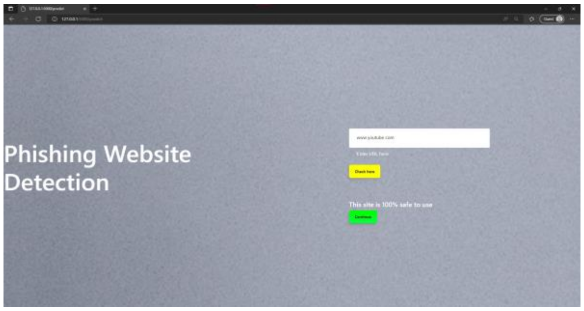


Fig 2: Gradient Boosting Classifier Flow diagram

V. IMPLEMENTATION AND TESTING

To evaluate a website extensive testing with both valid phishing sites and trusted sites was done. In this process the accuracy of the Gradient Boosting Classifier algorithm that was achieved was 97%.



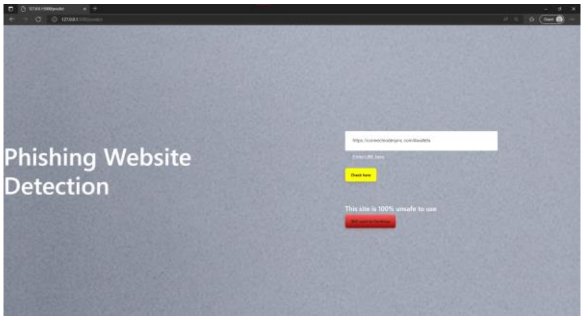
Phishing Website Detection

www.youtube.com

Enter URL here

This site is 100% safe to use

Fig 3: Output for Legitimate Website



Phishing Website Detection

Enter URL here

This site is 100% unsafe to use

Fig 4: Output for Phishing site

VI. CONCLUSION AND FUTURE WORK

Phishing attack is one of the common types of cyber-crime where the attackers can steal the user’s personal information by forgery the legitimate website with the masked one. The Proposed system uses Gradient Boosting Classifier Algorithm to detect the website is legitimate or not. The future work of the proposed system is to evaluate these machine learning classifiers with larger dataset, more accurate algorithms with higher accuracy

REFERENCES

[1] Wardman, B., Shukla, G., & Warner, G. (2009). Identifying vulnerable websites by analysis of common strings in phishing URLs. *2009 ECrime Researchers Summit.* https://doi.org/10.1109/ecrime.2009.5342610

[2] Sonmez, Y., Tuncer, T., Gokal, H., & Avci, E. (2018). Phishing web sites features classification based on Extreme Learning Machine. *2018 6th International Symposium on Digital Forensic and Security(ISDFS)*. https://doi.org/10.1109/isdfs.2018.8355342

[3] Liu, Xiuwen., & Fu, Jianming. (2020). SPWalk: Similar property oriented feature learning for phishing detection. *IEEE Access*, *8*, 87031–87045. https://doi.org/10.1109/access.2020.2992381.

[4] https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/

[5] https://www.researchgate.net/figure/Flow-diagram-of-gradient-boosting-machine-learning-method-The-ensemble-classifiers_fig1_351542039

[6] https://scikit-learn.org/stable/

[7] https://www.python.org/