Phishing Website DetectionModel using Machine Learning

Prajwal T S
Computer Science and Engineering
RV College of Engineering,
Bangalore, India
prajwalts.cs20@rvce.edu.in

Prasanna Suresh Naik
Computer Science and Engineering
RV College of Engineering,
Bangalore, India
prasannasn.cs20@rvce.edu.in

***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 work offers a model that can determine whether a URL link is genuine or fraudulent. The goal of this project 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 separated again into equal portions of phishing and legal 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 & JavaScript-based features. The study offered a model for dividing real URLs from phishing URLs. By authenticating every link that is sent to them to confirm its legitimacy, this would be highly helpful in assisting people and businesses in recognizing phishing attacks.**

***Keywords*—Decision Tree, Random Forest, Dataset, accuracy, diagnosis, efficiency.**

I. INTRODUCTION

Numerous benefits and conveniences have been made possible by the internet's explosive growth, but it has also given rise to cybersecurity risks, with phishing being one of the most widespread and harmful types of assaults. Phishing attacks involve criminals posing as legitimate organizations in an effort to trick users into disclosing sensitive information, such login passwords, financial information, or personal information. Traditional techniques for detecting phishing websites frequently rely on signature-based techniques, which find it difficult to keep up with the continuously changing strategies employed by hackers. Machine learning has recently gained attention from researchers and cybersecurity specialists as a potential defense against this constantly changing threat.

Phishing is a growing problem for people, companies, and governments alike since it preys on the human psychology that is the weakest link in cybersecurity. In order to trick victims into providing sensitive information, phishing attempts use social engineering techniques that take advantage of human inclinations like trust and haste. Development of reliable and adaptable detection techniques is essential given the prevalence and sophistication of phishing assaults. Machine learning has enormous potential for enhancing the effectiveness of phishing website detection due to its capacity to learn trends and spot abnormalities. Machine learning models can detect tiny trends that are suggestive of phishing websites by analyzing enormous volumes of data, offering a pro-active defense against these harmful acts.

Maintaining vast databases of well-known phishing websites and utilizing signature-based methods to match against incoming URLs are the foundations of traditional phishing detection techniques. This method has a number of drawbacks, though. By using obfuscation techniques or making slight changes to URLs, new phishing websites can quickly avoid detection, making signature-based tactics useless. A dynamic and scalable alternative is provided by machine learning algorithms, which can learn from real-time data and adjust to new phishing tactics. Machine learning models may create complex classifiers that discriminate between legitimate and phishing websites by utilizing variables like website content, structure, and user behavior. This improves the overall security posture of people and businesses.

II. LITERATURE SURVEY

There have been many journal articles about the Phishing website detection using machine learning. Some of the following journals and articles that we considered and reviewed are listed below.

Dourlens et al. (2017) proposed a machine learning-based method for phishing website identification that was centered on feature extraction and classifier creation in their foundational work. To obtain more accuracy, they suggested a hybrid technique that integrated lexical and visual information. Lexical characteristics examined URLs and content, whilst visual features looked at the structure and design of websites. Their research showed how integrating different feature sets might increase detection accuracy and decrease false positives. It also brought attention to the necessity of frequent upgrades to keep the classifier current with changing phishing techniques.

Liu and colleagues (2018) investigated the use of deep learning algorithms for phishing website identification. They examined website screenshots using convolutional neural networks (CNNs) to find visual phishing assault patterns. The benefit of CNNs is that they can automatically identify pertinent features from unprocessed data, eliminating the need for manual feature engineering. The study outperformed conventional machine learning techniques in terms of accuracy and resilience, yielding outstanding results. It also brought up issues with interpretability because deep learning models frequently act as "black boxes," making it difficult to comprehend the thought process that went into them.

Alazab et al. (2019): Decision trees, support vector machines (SVMs), and random forests were all included in a thorough review of several machine learning methods for phishing website identification. Their research emphasized the importance of choosing the right characteristics and striking a balance between the detection accuracy and false positive rates. To ensure the generalizability of the models, they also underlined the significance of datasets containing a wide variety of phishing scenarios. The researchers also talked about the possibility of combining different classifiers using ensemble approaches to enhance detection performance.

The literature presented here demonstrates the developments and current work in machine learning-based phishing website identification. Researches have investigated a range of strategies, including deep learning methods, hybrid methodologies, and real-time detection for mobile platforms. Even though machine learning has showed a lot of promise in thwarting phishing assaults, issues like dataset quality, interpretability, and striking a balance between accuracy and false positives continue to be an issue. In order to develop more reliable and trustworthy phishing detection systems that can shield consumers from emerging cyber dangers, future research should concentrate on tackling these problems.
.

III. OBJECTIVE OF STUDY

The major goal of this study is to carry out a thorough assessment into how machine learning techniques are used to identify phishing websites. The study seeks to accomplish the following particular goals

Evaluation of Machine Learning's Effectiveness: The study compares machine learning algorithms to more established signature-based techniques in an effort to determine which is more effective in spotting phishing websites. The research aims to ascertain whether these methods offer a more robust and adaptive approach to combat the dynamic nature of phishing assaults by comparing the performance of several machine learning models.

Determine Key qualities for Phishing Detection: Another goal is to determine the most pertinent and instructive qualities that aid in the successful identification of phishing websites. The study will investigate how many variables, including lexical, visual, and behavioral ones, may be used to train efficient machine learning models.

Address Interpretability and Explain ability: The work tries to address the issue of interpretability and explain ability given the "black box" character of some machine learning algorithms. The goal of the research is to increase confidence and trust in the detection system by investigating ways to make the models' decision-making process more transparent.

Analyze Real-Time and Mobile Platform Detection: The research will be focused on determining if machine learning-based phishing detection can be implemented in real-time situations and on mobile platforms. In particular for people accessing the internet via mobile devices, this purpose is to investigate resource-saving methods and lightweight models suitable for quick and efficient identification.

IV. IMPLEMENTATION

A. Extracting features

It is possible to extract features from URLs, and the resulting binary values can be used to determine if a website is a phishing website or not.

The features that we can extract to identify fraudulent URLs are listed below.

1. If an IP address appears in the URL, the feature is set to 1; otherwise, it is set to 0. Most trustworthy websites don't include an IP address in the URL of a webpage that can be downloaded. The use of an IP address in the URL suggests that the attacker is attempting to gather sensitive data.

2. If a "@" symbol appears in the URL, the feature is set to 1; otherwise, it is set to 0.When hackers add a special "@" symbol to a URL, the browser ignores everything before the "@" symbol and frequently ignores the true address that follows.

3. Prefix or suffix to domain separated by dash (-): If the domain name is separated by the dash (-) symbol, the feature is set to 1; otherwise, it is set to 0. Rarely does the '-' symbol appear in valid URLs. Phishers append the hyphen (-) to the domain name to give users the impression that they are visiting a trustworthy website. For instance, the real website address is http://www.flipkart.com, but phishers can create a false version of it to deceive unwary customers.

4. Host name length: The average benign URL is 25 characters long. If the URL is longer, the feature is set to 1, otherwise it is set to 0.

5. The presence of an HTTPS token in the URL causes the feature to be set to 1 otherwise. To trick users, phishers may tack on the "HTTPS" token to the domain portion of a URL.

6. URL rerouting: If "//" appears in the URL path, the feature is set to 1 otherwise. The user will be directed to another website if the URL path contains the character "//".

B Decision-tree

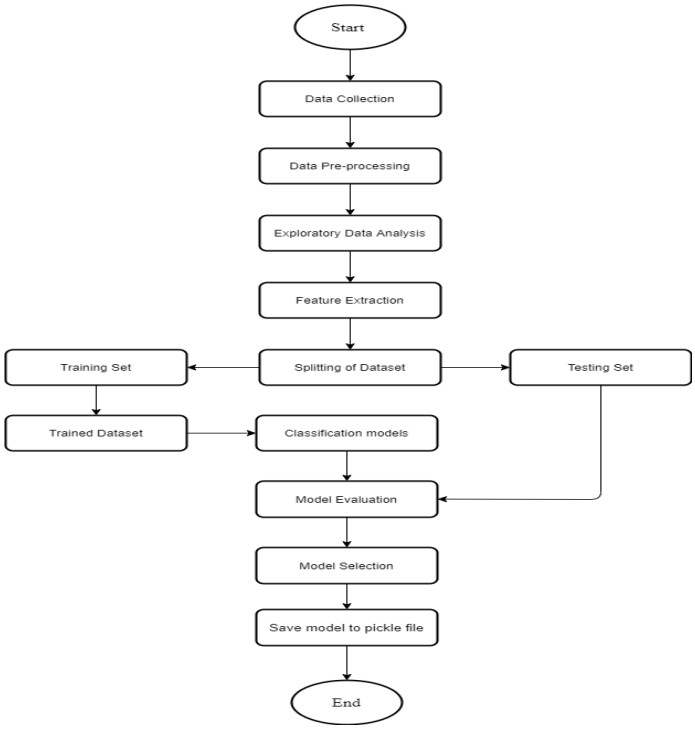
A supervised learning technique called a decision tree can be used to solve classification and regression problems. however it works best when dealing with classification issues. It is a tree-structured classifier, in which the internal nodes stand in for the dataset's features, the branches for the decision-making processes, and the leaf nodes for the results. There are two nodes in the decision tree: the decision node and the leaf node. While Leaf nodes are the results of those decisions and do not contain any additional branches, Decision nodes used to make numerous decisions and had various branches.

C. Random Forest Algorithm

Random Forest is a machine-learning algorithm that may be used for Classification and Regression issues; it is a part of the supervised learning methodology. It is based on the idea of associative learning, which is the process of integrating numerous distinct classifiers to solve a complex problem and increase the effectiveness of the model. This concept is used in the Phishing Website Detection using Machine Learning System Implementation. As the name implies, Random Forest is a classifier that uses numerous decision trees on different subsets of the provided dataset and averages the results to improve the dataset's predicted accuracy. Instead, then relying on a single decision tree, the random forest uses predictions from all of the trees to determine the final result based on the majority of those predictions.

V. PROPOSED ARCHITECTURE

Obtain a thorough dataset with instances of both authentic and phishing websites that have been labeled. The machine learning model will be trained and evaluated using this dataset.
The dataset should include numerous website elements, such as content, SSL certificate details, and URL attributes.
Data preprocessing: To handle missing values, outliers, and normalize the features, clean and preprocess the dataset. This phase is essential for guaranteeing model performance and data quality.
Feature Choice: Determine pertinent characteristics that can help users identify between phishing and trustworthy websites. To choose the most important features, use methods like feature importance ranking and correlation analysis.
Model selection: Select the best machine learning technique for the job. Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines, and Deep Learning models are often used methods for phishing website identification.
Split the pre-processed dataset into training and testing sets for the model. Train the chosen machine learning model using the training set on the labelled samples.
To improve the performance of the model, adjust its hyperparameters using methods like cross-validation.
Model assessment: Assess the trained model's accuracy, precision, recall, F1-score, and ROC-AUC using the testing set. These metrics will allow us to gauge how well the model can identify phishing websites.
Addressing Imbalance: To address imbalance and prevent biased results, utilize methods like oversampling, undersampling, or class weighting if the dataset contains more legitimate websites than phishing ones.
Model fine-tuning: Using the evaluation results as a guide, adjust the model to increase accuracy and decrease false positives and false negatives.
Application in real-time: Use the real-time model that has been trained and refined. For real-time phishing threat mitigation, incorporate the model into online browsers, network appliances, or security software.
Updates and Continuous Monitoring: Continue to keep an eye on how the model performs in practical situations. Update the model frequently with fresh data to retain its efficacy and adjust to new phishing techniques.



VI. PROPOSED METHODOLOGY

A methodology for Alzheimer's disease prediction using a deep learning model typically involves the following steps:

Requirements gathering and analysis: Gathering and evaluating requirements involves gathering a significant dataset of clinical and neuroimaging data from patients with Alzheimer's disease and healthy controls, preprocessing the data, establishing and training a deep learning model using CNN, and evaluating the model's performance using metrics like accuracy and AUC. The ultimate objective is to create a predictive model capable of early, accurate Alzheimer's disease diagnosis.

System design: involves creating the user interface, system architecture, and other elements.

Development: Using programming languages and software development tools, the system is developed based on the design.

Testing: Testing the system to make sure it meets the specifications and runs properly.

Implementation: Building up the software such that medical practitioners can utilize it as a web application or a mobile app to help with the early detection of Alzheimer's disease.

Updating and maintaining the system regularly is necessary to repair bugs and introduce new features that will increase accuracy.

To adapt to the changing requirements of the users such as practitioners and doctors, the methodology for predicting Alzheimer's disease using a deep learning model should be flexible, adaptable, and scalable.

VIII.RESULTS AND CONCLUSION

In order to get a well desired output we need algorithms that constantly learn and adapt to new examples and features of phishing URL's. And thus we use online learning algorithms. This new system can be designed to make use of maximum accuracy. Using different approaches altogether will improve the precision of the system, providing an efficient protection system. The drawback of this system is detecting of some minor false positive and false negative results. These disadvantages can be abolished by introducing much enhanced feature to feed to the machine learning algorithm that would result in much higher accuracy. We need algorithms that continuously learn from new phishing URL instances and characteristics in order to produce the appropriate results. Thus, we make use of online learning algorithms. The design of this new system may take full use of accuracy. The precision of the system will be increased by combining several methods, creating a reliable defense. The system's weakness is the detection of a few tiny false positive and false negative outcomes. By supplying considerably improved features to input to the machine learning algorithm, which would produce far higher accuracy, these drawbacks can be eliminated.

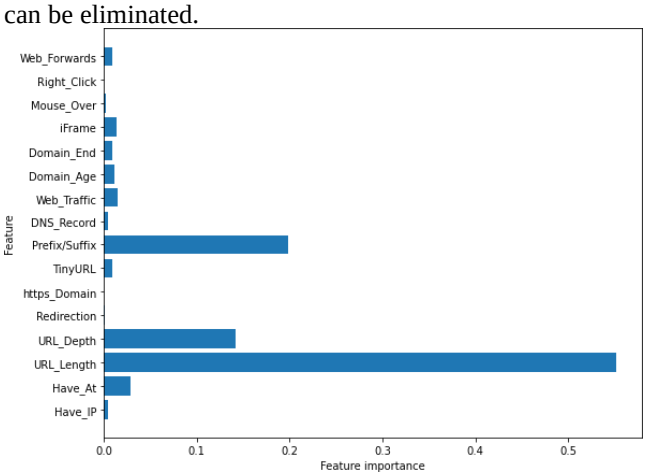


Fig 2. Feature importance for Random Forest classifier

When referring to a Random Forest classifier, the term "feature importance" describes the percentage that each feature in the dataset adds to the model's ability to predict the future. For a number of reasons, including determining the most important variables, choosing pertinent features for model building, and learning about the underlying connections between features and the target variable, understanding feature importance is essential. Based on its ensemble of decision trees, Random Forest offers a simple method to determine feature relevance.

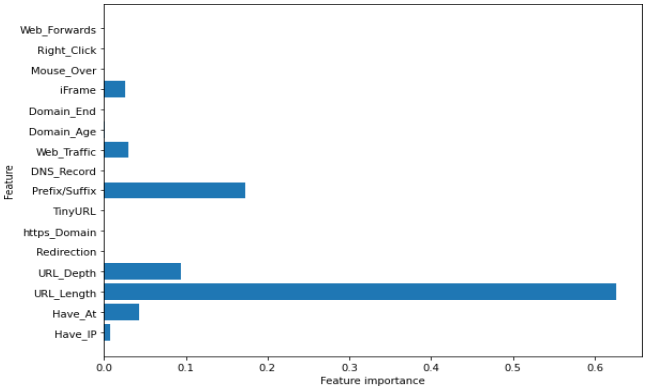


Fig 3. Feature importance for Decision Tree classifier

IX. FUTURE ENHANCEMENTS

Future development should concentrate on adding the project directly to the Chrome extension so that if a user clicks on a specific URL and if that URL is a phishing site, a pop-up warning message appears. by including this data. Moreover, clinical data is frequently accessible and simple to add to the dataset, making it a useful and deployable improvement.

X. REFERENCES

1.Abdelhamid, N., Thabtah F., & Abdel-Jaber, H. Phishing detection: A recent intelligent machine learning comparison based on models' content and features," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2017, pp. 72-77, DOI: 10.1109/ISI.2017.8004877.

2.Anjum N. S., Antesar M. S., & Hossain M.A. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Chengdu, China, 2016, pp. 9-15, DOI: 10.1109/SKIMA.2016.7916190.

3.Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques, Proceedings of IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2070–2090.

4.Ashritha, J. R., Chaithra, K., Mangala, K., & Deekshitha, S. (2019). A Review Paper on Detection of Phishing Websites using Machine Learning.Proceedings of International Journal of Engineering Research & Technology (IJERT), 7, 2. Retrieved from www.ijert.org.

5.Anti-Phishing Working Group (APWG) Phishing activity trends report the first quarter. (2014) Retrieved from http://docs.apwg.org/reports/apwg trends report q1 2014.pdf

6. APWG report. (2014). Retrieved from http://apwg.org/download/document/245/APWG Global Phishing Report 2H 2014.

pdf 111 Ayush, P. (2019). Workflow of a Machine Learning project.

Retrieved fromhttps://towardsdatascience.com/workflow-of-a-machine-learning-projectec1dba419b94

7.Camp W. (2001). Formulating and evaluating theoretical frameworks for career and technical education research. Journal of Vocational Education Research, 26(1), 4- 25.

8.DeepAI (n.d.). About clinical psychology. Retrieved from https://deepai.org/machine-learning-glossary-and-terms/feature-extraction

9. Engine K., & Christopher K. (2005). Protecting Users Against Phishing Attacks. Proceedings of the Oxford University Press on behalf of The British Computer Society, OxfordUniversity,0,2005,Retrievedfrom: https://sites.cs.ucsb.edu/~chris/research/doc/cj06_phish.pdf

10.Gandhi, V. (2017). A Theoretical Study on Different ways to identify the Phishing URL and Its Prevention Approaches: presented at International Conference on Cyber Criminology, Digital Forensics and Information Security at DRBCCC Hindu College,Chennai.Retrievedfrom https://www.researchgate.net/publication/319006943_A_Theoretical_Study_on_ Different_ways_to_Identify_the_Phishing_URL_and_Its_Prevention_Approache s