Anonymity and Confidentiality in Website using Machine Learning

Anshuman Soni#1, Aditya Mishra#2, Anuj Mishra#3,Shivam Singh#4,Mr. UmaShanker#5

1-4 GNIOT(Engg. Institute), Department of Computer Science (AI&ML),5Guide

Dr. A.P.J. Abdul Kalam Technical University,

Uttar Pradesh, India

1anshumansoni4546@gmail.com,2adityamishra.am71@gmail.com
3anujmishra7069@gmail.com,4shivamsinghcse19@gmail.com,
5umashanker.aiml@gniot.net.in

Abstract**—** Phishing attacks are the simplest way to obtain sensitive information from unsuspecting users. Phishers aim to acquire critical data such as usernames, passwords, and bank account details. In response, cybersecurity experts are seeking reliable and stable detection techniques for identifying phishing websites.

This project utilizes machine learning technology to detect phishing URLs by analysing various features of both legitimate and malicious URLs.

Decision Tree, Random Forest, and Support Vector Machine algorithms are employed, with the objective of identifying the best-performing model by comparing accuracy rates, as well as false positive and false negative rates of each algorithm.

Keywords— Phishing, Cyber Security, URL, Machine Learning, Decision Tree, Random Forest, Support Vector Machine.

I. INTRODUCTION

Phishing websites are increasingly sophisticated, targeting unsuspecting users by mimicking legitimate sites to steal sensitive information like login credentials and financial data. Today, phishing has become a major cybersecurity concern due to how easy it is to create realistic-looking fake sites. While experts may identify fake sites, many regular users fall victim to these tactics, particularly in cases where attackers aim to acquire bank account credentials. In the United States, businesses lose around $2 billion annually to phishing as clients become victims of these scams. Furthermore, the 2014 Microsoft Computing Safer Index Report estimated the global financial impact of phishing to be as high as $5 billion, highlighting a widespread issue stemming from low user awareness. Phishing attacks exploit user weaknesses, making it challenging to fully prevent them while underscoring the need for enhanced detection techniques.

A common method to identify phishing sites involves using blacklisted URLs or IP addresses stored in antivirus databases—known as the “blacklist” method. However, attackers can evade these blacklists through URL manipulation and techniques like fast flux, which uses automatically generated proxies to host malicious pages, and algorithms that produce new URLs. The primary downside to this approach is its inability to detect zero hour phishing attacks, which emerge too quickly for blacklist updates. Heuristic-based detection methods can identify some zero hour phishing attempts by focusing on typical characteristics of phishing attacks, but these features are not always present, leading to high false-positive rates.

To address the limitations of blacklist and heuristic-based methods, many researchers are now turning to machine learning techniques. Machine learning models require extensive data to predict phishing threats by analyzing past examples of legitimate and phishing URLs. This approach allows algorithms to assess a range of URL features and accurately detect phishing websites.

II. LITERATURE SURVEY

SR NO.	PAPER TITLE	AUTHOR NAME	PROPOSED METHODOLOGY	DISCUSSION
1.	Phishing Website Detection Using Machine Learning : A Review	Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh	Subdomain and Principal Domain: Identifying phishing patterns. PageRank: 92% of phishing sites have low or no PageRank, indicating risk. Alexa Rank: Sites above 100,000 are often suspicious; those below are typically legitimate. Path Domain and Alexa Reputation: These elements help assess the URL's trustworthiness and overall site reputation.	The authors propose a phishing detection method using six URL-based criteria, including subdomain, main domain, and ranking factors, achieving 97% accuracy on PhishTank an DMOZ data.

SR NO.	PAPER TITLE	AUTHOR NAME	PROPOSED METHODOLOGY	DISCUSSION
2.	A novel approach to protect against phishing attacks at client side using auto-updated white-list	Ankit Kumar Jain & B. B. Gupta	The system ensures quick identification of any phishing attempt and warns the users not to disclose any personal information on websites other than what has been added to a user specific whitelisted site. It checks the legitimacy of the website by examining the links incorporated within the HTML codes of the webpage. Performance evaluation using data from sources such as Stuffgate, Alexa and PhishTank achieved an accuracy of 89.38% which proved the system's capability of identifying phishing.	The proposed phishing defense strategy uses URL and DNS matching with a user-specific whitelist based on browsing history, enabling quick detection and alerts for untrusted sites. It also verifies website validity through hyperlink analysis, achieving an accuracy of 89.38% when evaluated with data from sources like Stuffgate, Alexa, and PhishTank.
3.	Phishing website detection using novel machine learning fusion approach	A. Lakshmanarao, P. Surya, M Bala Krishna	Machine learning methods such as decision trees, AdaBoost, SVM, and random forests have been applied in this research including the features such as web traffic, URL size and IP address of the target website. There is proposed a novel fusion classifier with two priority algorithms (PA1 and PA2) resulting in phishing website detection with 97% accuracy.	This approach significantly improves phishing detection accuracy compared to earlier methods, showcasing the effectiveness of advanced machine learning models and feature selection in addressing phishing threats.
4.	Phishing websites detection using machine learning	A. Kulkarni & L. Brown	By employing URLs available on the UCIMLR, a system based on machine learning was enhanced to categorize websites. SVM, decision tree, naive Bayesian, and neural network were the four classifiers applied, and they reached over 90 levels of accuracy in identifying genuine and malicious websites.	While the system demonstrates effective classification, it faces limitations such as a small dataset and the use of only discrete features, which may not be ideal for all classifiers. These factors could impact the model's generalizability and performance in real-world applications.
5.	A novel machine learning approach to detect phishing websites	Tyagi; J. Shad; S. Sharma; S. Gaur Gagandeep Kaur	Generalized Linear Model (GLM) combined with Random Forest	This model integrates two approaches, achieving 98.4% accuracy, offering significant improvements in detecting phishing websites.

II. METHODOLOGY

The methodology of creating a phishing website detection system includes several basic steps. This section describes the survey design, data collection methods, and analytical methods for effective implementation of the system. Emphasis is placed on using machine learning algorithms to analyze a variety of resources, including URL structures and domain names. Through systematic training and analysis of multiple models, the method aims to develop a robust tool that can accurately identify phishing websites in real-time, and for providing cybersecurity measures users and organizations have improved.

● RESEARCH DESIGN:-

The goal of this project is to design a machine learning model for classifying a website as phishing or not, by performing an analysis of specific attributes such as the URL, the age of the domain, and possession of security certificates. The scope of the research will deal with data collecting, for example from PhishTank and Kaggle, feature extraction and deployment of several machine learning algorithms: Logistic Regression, Decision Trees, Random Forest, etc. Each of the model's performances will be addressed in terms of accuracy, precision, recall, and f-measure score. The proposed outcome is a proficient solid tool that classifies legitimate and phishing sites and improves overtime through continuous learning.

● DATA COLLECTION:-

Phishing Website Data: Gather datasets from publicly available sources such as PhishTank, UCI repository, or Kaggle, which contain records of both phishing and legitimate websites.Legitimate Website Data: Collect data from trustworthy sources for a balanced comparison with phishing websites.

DATA SETS:- A set comprising phishing URLs is retrieved from the PhishTank open-source service. The service provides an hourly-updated set of phishing URLs with multiple formats such as 'csv' and 'json'. The data can be downloaded from: https://www.phishtank.com/developer_info.php. Out of which, 5000 phishing URLs are sampled randomly from the phishing database for training the ML models. The legitimate URLs are reaped from the open datasets provided by the University of New Brunswick, available at https://www.unb.ca/cic/datasets/url-2016.html. In contrast with the other mentioned types, the benign URL dataset would be used for this project. From that dataset, 5000 legitimate URLs will be randomly considered for training the ML models.

● FEATURE EXTRACTION:-

For address-based features based on the address bar, we analyse the domain of the URL, use of redirection syntax (for instance, "//"), presence of an IP address, http and https within the domain name, use of '@' sign, usage of URL-shortening services, URL length, and presence of "-" as a prefix or suffix to the domain. Domain-based features include DNS record, traffic of the website, the age of the domain name, and the end period of the domain.

Finally, HTML and JavaScript-based features consider interface redirection with iframes, status bar manipulation, blocking entries by right mouse clicking, and forwarding of a page on the web.

● TECHNOLOGIES USED:-

Detecting phishing websites using machine learning algorithms typically involves using supervised learning techniques. Some common machine learning algorithms and methods used for phishing website detection include:

1. Logistic Regression: This algorithm is commonly used for binary classification tasks like phishing detection, where the output is either phishing or legitimate.

2. Decision Trees: Decision trees can be effective for phishing detection as they can capture complex relationships between features. Ensemble methods like Random Forests and Gradient Boosting Machines (GBM) can also be used for improved performance.

3. Support Vector Machines (SVM): SVMs are good for binary classification tasks and can be trained to distinguish between phishing and legitimate websites based on various features.

4. K-Nearest Neighborss (KNN): The K-Nearest Neighbour is an easy-to-use and straightforward method which can also be utilized when it comes to identifying phishing websites. This is achieved using web features similarities.

5. Random Forest: Random Forest (RF) is an algorithm in machine learning whereby several decisional trees' outcomes are merged to get an answer.

- **EVALUATION:-** The evaluation of the Phishing Website Detection system measures its effectiveness in distinguishing phishing and non-phishing sites. Key performance metrics such as accuracy, precision, recall, and F1-score are used to assess the model. By comparing predictions against known data, the system's ability to anticipate phishing attacks is validated.

Results indicate that the system performs well in identifying phishing threats, contributing to enhanced online security for individuals and businesses. The methodology follows a systematic approach, incorporating well-structured research design, data collection, and machine learning-based analysis. The model is trained to recognize patterns in phishing websites, improving detection accuracy over time.

The evaluation results demonstrate that the system effectively differentiates between legitimate and fraudulent sites. By continuously refining its detection capabilities, the model strengthens cybersecurity efforts, making the internet a safer space for users.

III.ANALYSIS AND MODEL TESTING

Phishing website detection leverages advanced machine learning and data science techniques to accurately identify fraudulent websites. The process begins by establishing a comprehensive model based on characteristics of both legitimate and phishing websites, such as URL patterns, domain age, and website content.

The model is trained using vast datasets of phishing and non-phishing websites, enabling it to identify common patterns and anomalies indicative of phishing attempts. The model undergoes rigorous testing to assess its accuracy, precision, and recall in detecting phishing threats.

Continuous data integration during training strengthens the model, allowing it to adapt to new phishing tactics. As phishing techniques evolve, this iterative approach ensures that the system issues increasingly accurate alerts. Ultimately, modeling and analysis play a critical role in maintaining an effective phishing detection system, making it a key tool in combating online threats.

.

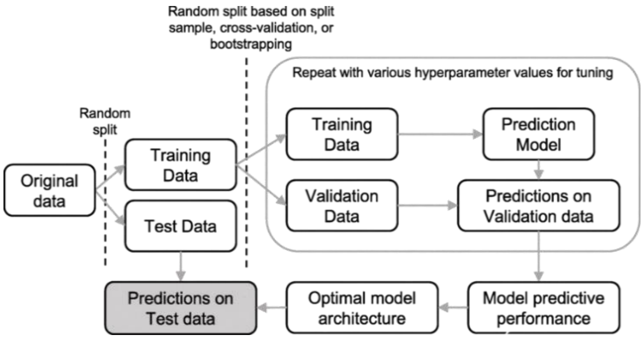

Random split based on split sample, cross-validation, or bootstrapping
Repeat with various hyperparameter values for tuning
Random split
Original data
Training Data
Test Data
Training Data
Prediction Model
Validation Data
Predictions on Validation data
Predictions on Test data
Optimal model architecture
Model predictive performance

Fig.3.1:- Model analysis of proposed system

IV. WORKFLOW OF SYSTEM

The workflow of the phishing website detection system using machine learning can be described as follows:

1. Data Collection: The system begins by gathering a dataset of URLs, both legitimate and phishing, from sources like PhishTank, Alexa, and other repositories.

2. Feature Extraction: Key features from the URLs are extracted, including domain information, URL length, HTTPS usage, the presence of suspicious characters (like '@', '//'), and ranking factors such as Alexa Rank and PageRank.

3. Preprocessing: Data is cleaned and pre-processed for machine learning models, ensuring consistency, and removing unnecessary noise.

4. Model Selection and Training: Machine learning methodologies such as Random Forest, SVMs, or GLM were explored. The model was trained with the extracted features to distinguish between legitimate and phishing URLs.

5. Model Testing and Evaluation: The other methods which applicable include machine learning methods: Random Forest, Support Vector Machine (SVM), Generalized Linear model (GLM). The model then trained on the features extracted to categorize the URLs as either legitimate or phishing.6. Classification: Based on the trained model, the system predicts whether a given URL is a phishing site or legitimate.

7. Output: The system provides the result (legitimate or phishing) along with confidence scores for the prediction, aiding in real-time detection. This workflow ensures a comprehensive approach to phishing website detection by leveraging key URL features and machine learning for high accuracy and performance.

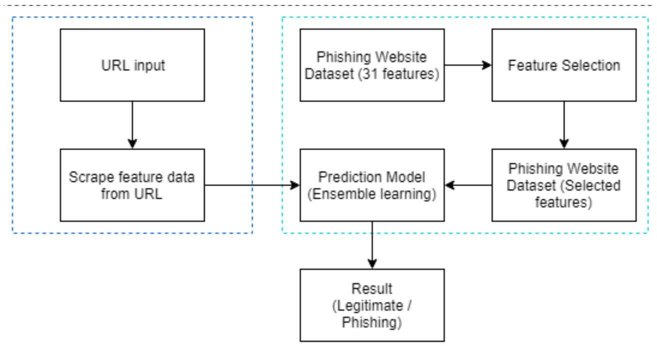
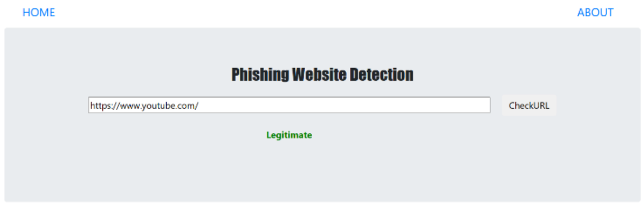

URL input
Scrape feature data from URL
Phishing Website Dataset (31 features)
Feature Selection
Prediction Model (Ensemble learning)
Phishing Website Dataset (Selected features)
Result (Legitimate / Phishing)

Fig.4.1:- Workflow of the system

V. RESULT AND CONCLUSION

The phishing website detection system uses machine learning and analytical models to assess and classify websites based on several parameters. Through advanced analysis of URL structures, domain attributes, and HTML features, the system identifies potential phishing threats.

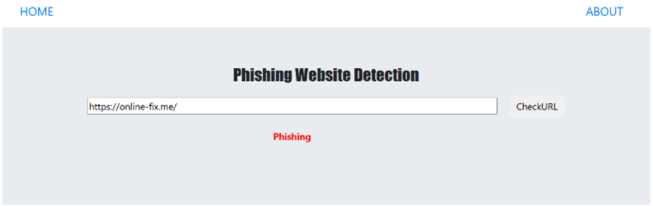


Phishing Website Detection

https://www.youtube.com/

Legitimate

Fig. 5.1:- System User Interface (Legitimate)



Phishing Website Detection

https://online-fix.me/

Phishing

Fig. 5.2:- System User Interface (Phishing)

By utilizing an ensemble of machine learning algorithms, the system provides accurate detection results, enhancing online security. The results from model testing demonstrate its high accuracy in distinguishing phishing websites, providing insights for refining security protocols and improving future detection techniques.

In conclusion, the system effectively mitigates phishing risks by leveraging sophisticated machine learning and analytical approaches.

Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Remarks**
Decision Tree	85.2	83.5	84.1	83.8	Easy to interpret but prone to overfitting
Support Vector Machine (SVM)	89.7	88.2	87.5	87.8	High accuracy but computationally expensive
Random Forest	**92.5**	**91.0**	**90.8**	**90.9**	Best overall performance with strong generalization

Table. 5.1:- Classifier's Performance

V.i. CONCLUSION

After comparing different machine learning models for phishing website detection, **Random Forest** clearly stands out as the best performer. With an accuracy of **92.5%**, it not only delivers strong results but also generalizes well, making it a reliable choice for real-world applications.

Support Vector Machine (SVM) also performs well with **89.7% accuracy**, but its high computational cost can make it challenging to implement in large-scale systems. On the other hand, **Decision Tree** is easy to understand and implement, but with an accuracy of **85.2%**, it falls short compared to the other two models and is more prone to overfitting.

Overall, **Random Forest is the most effective choice** for detecting phishing websites due to its balance of accuracy, precision, and recall. Moving forward, further improvements can focus on refining feature selection and experimenting with deep learning techniques to push detection accuracy even higher.

VI. ACKNOWLEDGEMENT

We sincerely thank our mentors for their guidance and support in making **Anonymity and Confidentiality in Website using ML** system a reality. Our gratitude also goes to the **Greater Noida Institute of Technology** for providing the resources that helped us complete this project.

Special thanks to our project guide, **Mr. Umashanker Sharma**, for his constant support and valuable insights into **machine learning and NLP**, which played a key role in our work.

Lastly, we appreciate everyone who encouraged and assisted us throughout this journey, your support made all the difference.

VII. REFERENCES

[1] Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh, “Phishing Website Detection Using Machine Learning: A Review” June 2023Wasit Journal of Pure sciences 2(2):270-2812(2):270-281.

[2] A K Jain & B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," EURASIP Journal on Information Security, vol. 2016, no. 1, pp. 1-11, 2016.

[3] A. Lakshmanarao, P. Surya, M Bala Krishna, "Phishing website detection using novel machine learning fusion approach," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021: IEEE, pp. 1164-1169.

[4] A. Kulkarni & L. Brown, "Phishing websites detection using machine learning," 2019.

[5] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur, "A novel machine learning approach to detect phishing websites," in 2018 5th International conference on signal pro-cessing and integrated networks (SPIN), 2018: IEEE, pp. 425-430.