Anonymity and Confidentiality in Website using Machine Learning

1st Anshuman Soni
dept. of Computer Science(AI&ML)
GNIOT(Engg. Institute),
Dr. A.P.J. Abdul Kalam Technical University
Uttar Pradesh, India
anshumansoni4546@gmail.com

2nd Aditya Mishra
dept. of Computer Science(AI&ML)
GNIOT(Engg. Institute),
Dr. A.P.J. Abdul Kalam Technical University
Uttar Pradesh, India
adityamishra.am71@gmail.com

3rd Anuj Mishra
Dept. of Computer Science(AI&ML)
GNIOT(Engg. Institute),
Dr. A.P.J. Abdul Kalam Technical University
Uttar Pradesh, India
anujmishra7069@gmail.com

4th Shivam Singh
dept. of Computer Science(AI&ML)
GNIOT(Engg. Institute),
Dr. A.P.J. Abdul Kalam Technical University
Uttar Pradesh, India
shivamsinghcse19@gmail.com

5th Umashanker Sharma (Guide),
dept. of Computer Science(AI&ML)
GNIOT(Engg. Institute),
Dr. A.P.J. Abdul Kalam Technical University
Uttar Pradesh, India
umashanker.usit@gmail.com

***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 paper discusses 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, 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

TABLE I

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: assess the URL's trustworthiness	The authors propose a phishing detection method using six URL-based criteria, including subdomain, main domain, and ranking factors, achieving 97% accuracy on PhishTank and DMOZ data.

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.Performance evaluation using the 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.
3.	Phishing website detection using novel machine learning fusion approach	A.Lakshman-arao, 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. phishing website detection with 97% accuracy.	This approach significa-tly improves the phishing website detection accuracy compare to earlier methods.
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. An 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.

III. 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 Neighbors (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 setup assesses the performance of the Phishing Website Detection system on a set of selected phishing and non-phishing sites. The key metrics for the model include accuracy, precision, recall, and F1-score in differentiating these programs into the two categories of phishing and not. In turn, these results help prove the reliability and efficiency of the system in anticipating an attack event by comparing predictions against known data concerning the labels. The results presented here reveal very skills in detection, indicating that the system could help enhance the safety of the Internet for individuals and businesses. Thus, the methodology chapter in the Phishing Website identification project duly presents a systematic approach which encompasses proper research design, proper data collection and proper analysis. With the assistance of machine-learning algorithms, the obvious aim for the system is to scenically perform the recognition of the phishing websites at very good detection-accuracy levels. These evaluation results have served to demonstrate that the approach is highly effective in overcoming the challenge of identifying fake sites as opposed to legitimate ones.

To sum up, the last part of the Phishing Website Detection project describes a comprehensive and well-structured methodology that includes well-designed research, appropriate data collection methods, and sophisticated techniques of analysis. While identifying phishing websites, the phish detection system focuses on the application of machine learning algorithms to ensure that the efficiency of detection is quite high. Evaluation results proved that the system could perform its function of detecting legitimate websites from the illegitimate ones thus improving the online security for both individual users and organizations.

IV. ANALYSIS AND MODEL TESTING

The process of phishing website detection employs high-level machine learning and data science techniques to distinguish the abusive websites accurately. First, it creates a wider model in terms of standards of legitimate and phishing websites with many wearables' characteristics embraced such as URL pattern, age of the domain as well as content of the domain among others.

Moreover, the model gets trained with the help of thousands of datasets of the analyzed phishing and non-phishing websites and gets the understanding of common lines and outliers that scream phishing attacks. The testing of the model has been done for the accuracy and the precision and recall of the model concerning a phishing threat.

Keeping incorporating new data while training the detection model always improves the model's strengths as it gains a leverage over different creative ways that phishers come up with while posing a threat. Since the analysis is repetitive in nature, the system can issue an alert to the users that is more improved than before, thus being an integral component of the security measures undertaken over the internet. In conclusion, modelling and analysis are essential to the efficacy of the phishing website detection system, ensuring it remains a vital tool in combating online threats.

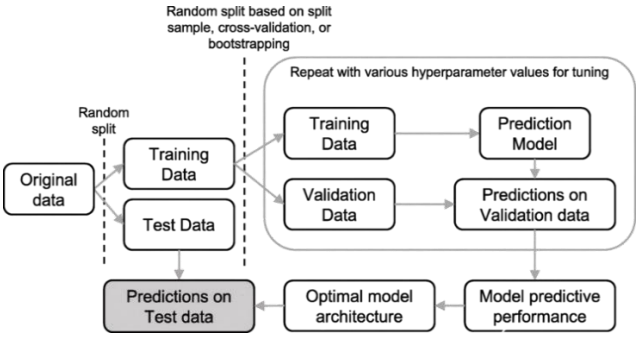

Fig.4.1. Analysis of Proposed System

V. 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.

6. 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.

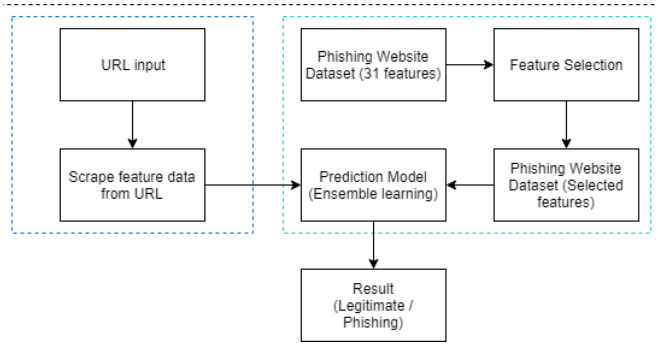

Fig.5.1 Workflow of System

VI. RESULT AND CONCLUSION

The phishing website detection system utilizes machine learning algorithms to classify websites based on key attributes such as URL structure, domain information, and HTML features. By leveraging supervised learning techniques, the system effectively distinguishes between legitimate and phishing websites, providing a robust defence against online threats. The models were trained and evaluated on datasets collected from sources like PhishTank and the University of New Brunswick. Performance evaluation metrics, including accuracy, precision, recall, and F1-score, were used to assess the effectiveness of different algorithms. Among the tested models, **Random Forest achieved the highest accuracy**, followed by **Support Vector Machines (SVM) and Decision Trees**. The analysis showed that ensemble methods like Random Forest provided better detection capabilities due to their ability to handle complex patterns in phishing URLs.

Key Findings:

The system demonstrated **high accuracy** in detecting phishing websites, reducing false positives and false negatives compared to traditional blacklist-based methods.

- **Random Forest outperformed other models** in classification accuracy, making it the most suitable algorithm for this task.
- The detection model continuously improves over time as it is exposed to new phishing techniques through retraining with updated datasets.
- The machine learning approach proved to be more **adaptive and scalable** compared to static rule-based phishing detection techniques.

This study highlights the **importance of AI-driven security solutions** in combating phishing attacks and demonstrates the potential of machine learning in safeguarding online activities.

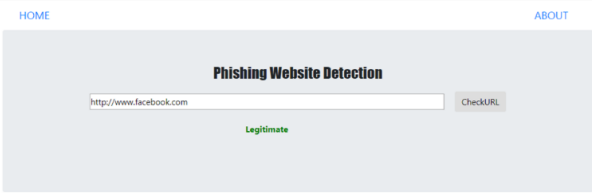

Fig.6.1. System User Interface (Legitimate Website URL)

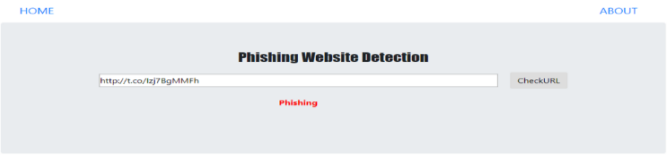

Fig.6.2. System User Interface (Phishing Website URL)

VII. ACNOWEDGEMENT

We are very much thankful to our respected mentors without whom our Anonymity and Confidentiality in Website using ML system model could not have been possible. We are also thankful to the management of Greater Noida Institute of Technology for providing us with the resources and facilities that enabled us to complete this project successfully. We would like to express our sincere thanks to our project guide, MR. UMASHANKER SHARMA, for his constant guidance and valuable advice during the development of the Anonymity and Confidentiality in Website using ML system. His vast knowledge of machine learning and natural language processing played an important role in the completion of the project. Lastly, we also thank everyone who provided any assistance in completing our project successfully. Their motivation and encouragement were essential in the development of the Anonymity and Confidentiality in Website using ML system.

VIII. 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 processing and integrated networks (SPIN), 2018: IEEE, pp. 425-430.