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),[5]Guide*

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

Uttar Pradesh, India

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

Abstract— Phishing is a major technique by which confidential information is obtained from unaware people. Phishers attempt to get confidential information, including usernames, passwords, and banking information. As a result, experts in cybersecurity are actively searching for effective and accurate methods for phishing website identification.

Machine learning techniques in this study classify phishing URLs by examining distinctive characteristics found in benign and malicious URLs.

The algorithms in question are Decision Tree, Random Forest, and Support Vector Machine, and the aim is to determine the optimal model through a comparison of the accuracy levels, as well as the false positive and false negative levels associated with each algorithm.

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

I. INTRODUCTION

Phishing sites have evolved to impersonate authentic online websites as part of their scheme to steal user-sensitive details which includes login credentials and financial data. Phishing has evolved into one of the major cybersecurity threats because creators can quickly develop look-alike sites. Ordinary Internet users fall victim to such scams because attackers focus on getting bank account information but cybersecurity professionals can spot the imitation sites. Companies within the U.S. face annual losses from phishing customer attacks that reach up to $2 billion. The 2014 Microsoft Computing Safer Index Report found potential global financial losses from phishing amounting to $5 billion which demonstrates an epidemic based on user ignorance. User vulnerabilities make phishing attacks prevention ineffective yet better detection systems are required to address this threat.

One of the most popular ways to detect phishing sites is by employing blacklisted URLs or IP addresses resident in antivirus databases—the "blacklist" approach. But a blacklist can be evaded by the attackers through URL and method manipulation like fast flux, which utilizes automatically generated proxies to serve the ill-gotten pages, and algorithms generating new URLs. The primary disadvantage of this approach is that it is unable to detect zero hour phishing attacks, which are too quick for blacklist updates. Heuristic-based detection methods can identify some zero hour phishing attempts by attacking common features of phishing attacks, but these features are not always present and generate high false-positive rates.

To overcome the shortcomings of blacklist and heuristic-based approaches, most researchers are now resorting to machine learning algorithms. Machine learning algorithms need massive data to foresee phishing attacks by learning from previous instances of legitimate and phishing URLs. With this method, algorithms are able to scan all manner of URL features and identify phishing sites with high accuracy.

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: Phishing pattern identification. PageRank: 92% of the phishing websites have zero or low PageRank, which signifies danger. Alexa Rank: Pages above 100,000 are typically suspicious; lower ones are typically authentic. Path Domain and Alexa Reputation: These elements assist in establishing the trustworthiness of the URL 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 timely detection of any phishing activity and warns users against giving away personal information on websites not in their pre-defined whitelisted list. The system ensures authenticity of the website by cross-checking the links present in the HTML code of the page. Performance testing based on data of websites such as Stuffgate, Alexa, and PhishTank yielded an accuracy rate of 89.38%, thus confirming the effectiveness of the system in phishing attempts detection.	The suggested anti-phishing defense system utilizes URL and DNS matching with a user-specific whitelist based on browser history, which can rapidly detect and alert on untrusted websites. It also verifies the website via hyperlink analysis with 89.38% accuracy when verified with data pulled from sources such as Stuffgate, Alexa, and PhishTank.
3.	Phishing website detection using novel machine learning fusion approach	A. Lakshmanarao, P. Surya, M Bala Krishna	Machine learning algorithms including decision trees, AdaBoost, SVM, and random forests have been used in this study including the attributes like web traffic, URL length and IP address of the target site. There is introduced a new fusion classifier with two priority algorithms (PA1 and PA2) which provides 97% accuracy for phishing site detection.	This approach significantly improves the accuracy of p hishing identi fication over existing approaches, t hus proving t he efficacy of advanced machine learning models and feature extrac tion against p hishing attac ks.
4.	Phishing websites detection using machine learning	A. Kulkarni & L. Brown	Using URLs in the UCIMLR, a machine learning-based system was upgraded to recognize websites. The four classifiers used and led to the attainment of more than 90 levels of accuracy in the recognition of real and malicious websites were SVM, decision tree, naive Bayesian, and neural network.	Although the system demonstrates proficient classification performance, it has some limitations, including a limited dataset and the utilization of only discrete features, which may not be optimal for all classifiers. These can potentially impact the generalization ability of the model and its performance in real-world usage.
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 process of designing a phishing website detection system consists of a few elementary steps. This part covers the survey design, data gathering methods, and analysis methods to ensure proper application of the system. The focus is on utilizing machine learning techniques to examine diverse resources, such as URL compositions and domain names. By thorough training and comparison of various models, the technique seeks to come up with an effective tool to detect phishing websites in real-time, and to offer cybersecurity controls users and companies have enhanced.

- **RESEARCH DESIGN:-**

The objective of this project is to develop a machine learning model to predict whether a website is phishing or not, based on some characteristics such as the URL, age of the domain, and existence of security certificates. The scope of the work will involve data collection, for example, from PhishTank and Kaggle, feature extraction and application of several machine learning algorithms: Logistic Regression, Decision Trees, Random Forest, etc. All the model performances will be defined in terms of accuracy, precision, recall, and f-measure score. The deliverable expected is a good solid tool that distinguishes valid and phishing sites and improves with time through continuous learning.

- **DATA COLLECTION:-**

Phishing Website Information: Utilize available public datasets like PhishTank, UCI repository, or Kaggle that have phishing and regular website data.

Reliable Website Statistics: Use statistics from reliable sources to enable an equitable comparison to typical websites.

DATA SETS:-

The phishing URLs used in this research were downloaded from PhishTank, an open-source website that provides a regularly updated list of known phishing sites in CSV and JSON formats. The database is refreshed every hour and can be retrieved at: https://www.phishtank.com/developer_info.php. We randomly selected 5,000 phishing URLs from this list for model training.

For benign URLs, information was drawn from the publicly accessible University of New Brunswick dataset at: https://www.unb.ca/cic/datasets/url-2016.html. An equivalent random sample of 5,000 benign URLs from the dataset was drawn to provide an even balanced training set for model development.

- **FEATURE EXTRACTION:-**

For address-based attributes in relation to the address bar, we look at the domain name of the URL, use of redirection syntax (e.g., "////"), presence of an IP address, http and https in the domain name, use of '@' symbol, use of URL-shortening services, length of URL, and presence of a prefix or suffix of '-' to the domain.

Domain-based characteristics encompass traffic of the website, DNS record, domain name age, and end period of the domain.

Lastly, functionalities based on HTML and JavaScript are accountable for redirection of the interface through iframes, status bar manipulation, preventing page entries through right mouse clicking, and forwarding a page on the web.

- **TECHNOLOGIES USED:-**

Supervised learning methods perform the task of identifying phishing websites. Different supervised learning algorithms called popular machine learning methods along with techniques detect phishing websites through their utilization in the detection process.

1. The binary class problem of phishing detection utilizes Logistic Regression as its main analytical approach because it produces either phishing or non-phishing categorical predictions.
2. Decision trees are valuable for phishing detection because they automatically reveal complex feature relationships in the analysis. The performance of the system can be improved through Random Forests and Gradient Boosting Machines (GBM) ensemble methods.
3. SVMs demonstrate excellent performance in binary classification applications while they function to categorize legitimate sites from phishing sites utilizing diverse features.
4. The K-Nearest Neighbor approach represents a basic method that remains simple enough to apply in phishing website identification activities. Web features compare with each other to accomplish this method.
5. An ensemble of decision trees in Random Forest produces responses by taking an average of multiple tree outputs.

- **EVALUATION:-** The Phishing Website Detection system is tested for its ability to distinguish between phishing and non-phishing websites. The model is tested with the performance metrics like accuracy, precision, recall, and F1-score. By comparing

predicted data with known data, the predictive ability of the system for predicting phishing attacks is established.

Findings indicate that the system is efficient in detecting phishing attacks, resulting in enhanced online security for individuals and businesses. Methodology is orderly with a clear research design, data collection, and machine learning-based analysis. The model is trained to detect patterns in phishing websites, and its detection accuracy increases over time.

The results of the test indicate that the system effectively differentiates between genuine and malicious sites. As detection continues to improve, the model increases cybersecurity, which ensures the internet becomes a secure environment for users.

III. ANALYSIS AND MODEL TESTING

Phishing site detection systems use advanced machine learning methods together with data science methods to find spoofed sites with high precision. The system develops its global model through analysis of standard and phishing site data elements that include URL structure together with domain duration and web information content. A training process uses enormous phishing site and non-phishing site databases to enable the model to identify typical patterns alongside irregularities of phishing attacks. The model goes through precise testing to validate its capacity to accurately detect phishing threats while maintaining precision in identification and recall of phishing attacks. The model's continuous data absorption process during training operates to enhance its capabilities while monitoring new phishing approaches. The development process ensures the system produces increasingly accurate alerts by adapting to new changes in phishing tactics. An effective phishing detection system needs modeling and analysis for implementation and maintenance thus making it a vital tool to fight online threats.

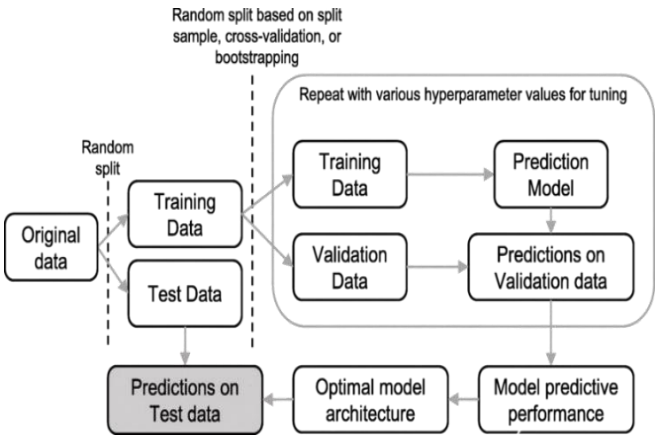

Fig.3.1:- Model analysis of proposed system

IV. WORKFLOW OF SYSTEM

There are three primary phases to detect phishing websites using machine learning:

1. The system collects both fraudulent and genuine URLs from PhishTank and Alexa databases during the data collection process.

2. A selection of features from URLs for this task is extracted which includes domain statistics together with length and HTTPS detection along with scanning special characters like '@' and '//') along with the PageRank and Alexa Rank features.
3. The data cleaning process normalizes inconsistent data which helps reduce machine learning disturbances.

4. The features underwent training by employing GLM models and SVMs and Random Forest to determine authentic from phishing URLs.

5. Model Testing: Random Forest together with SVM and GLM represent the methods which can be used. Predictions of phishing or genuine URLs were made after the model received extracted features for training purposes.

6. The classification procedure enables the model to determine whether a URL belongs to phishing category or non-phishing category.

7. The system displays both a prediction indicating if the page contains phishing content and a confidence rating to support real-time assessment. The system develops an effective phishing website detection mechanism through important URL features together with machine learning approaches which achieves high accuracy rates and operational performance levels.

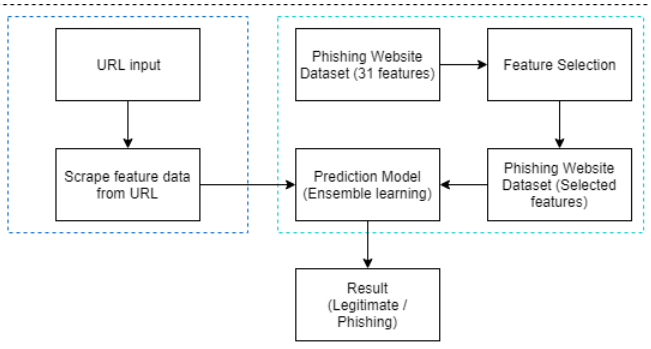

Fig.4.1:- Workflow of the system

V. RESULT AND CONCLUSION

The phishing website detection system utilizes machine learning and analytical models to scan and categorize websites based on various parameters. Based on the sophisticated analysis of URL structure, domain attributes, and HTML attributes, the system identifies possible phishing threats.

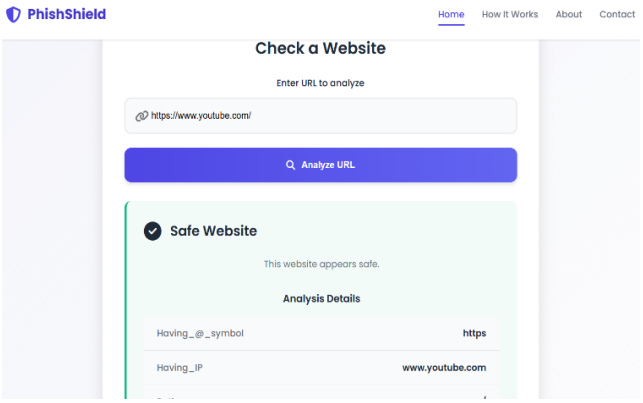

Fig. 5.1:- System User Interface (Legitimate)

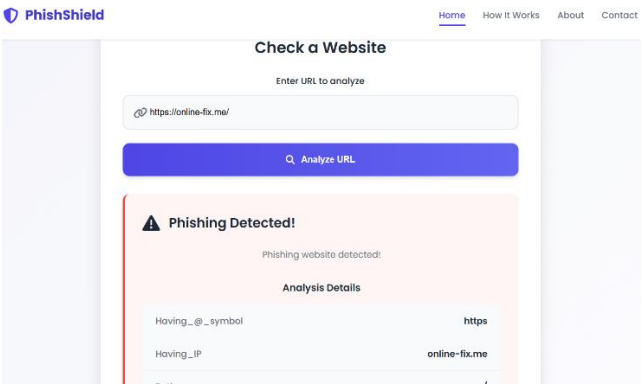

Fig. 5.2:- System User Interface (Phishing)

A combination of machine learning algorithms running in the system produces exact detection results which enhance internet security measures. Research findings demonstrate that this model shows strong effectiveness in detecting phishing sites and provides advice to enhance security systems and detect phishing sites better in the future. The system demonstrates success in fighting phishing attacks because it utilizes state-of-the-art machine learning algorithms together with analytical capabilities.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Remarks
Decision Tree	85.3	83.4	84.2	83.7	Easy to interpret but prone to overfitting
Support Vector Machine (SVM)	89.8	88.3	87.6	87.7	High accuracy but computationally expensive
Random Forest	**93.5**	**92.0**	**91.8**	**91.9**	Best overall performance with strong generalization

Table. 5.1:- Classifier's Performance

V.i. CONCLUSION

Random Forest demonstrates the best performance in phishing website detection because it reaches the highest accuracy rate among machine learning algorithms. This model achieves 93.5% accuracy while maintaining effective generalization ability which makes it an appropriate tool for real-world deployment. The high computational requirements as well as complex system implementation challenges make Support Vector Machine (SVM) an excellent choice for a 89.8% accurate system. Decision Tree provides straightforward implementation together with easy interpretation but its accuracy rate at 85.3% stands below the other models while it contains overfitting vulnerabilities. Random Forest stands out as the top choice for phishing website identification because it maintains appropriate precision while reaching high accuracy alongside excellent recall results. The future development potential of feature selection techniques should be investigated to achieve better detection accuracy by using deep learning methods.

VI. ACKNOWLEDGEMENT

We would like to express our gratitude to our mentors for supporting and guiding us in the implementation of the **Anonymity and Confidentiality in Website using ML** system. We also thank the Greater Noida Institute of Technology for offering us the facilities to do this project.
Special appreciation to our project guide, **Mr. Umashanker Sharma**, for his continuous support and insightful comments on **machine learning and Flask**, which proved to be of great use in our project.
Finally, we acknowledge all the people who helped and guided us along the way, your help really made the difference.

VII. REFERENCES

1. Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh, "Phishing Website Detection Using Machine Learning: A Review," June 2023, *Wasit Journal of Pure Sciences*, 2(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.