Domain Detector - An Efficient Approach Of Machine Learning For Detecting Malicious Websites

Mrs. T. Sai Priyanka[1], **Dr Naadem Divya**[2], A. Sruthi[3], S. Laxmi Prasanna[4], B. Sahithi[5], P. Jyothsna[6]

[1]Assistant Professor, Department of CSE (AI&ML), Vignan's Institute of Management and Technology for Women, HYD, India.

[2]Associate Professor, Dept of CSE-DS, Sreenidhi institute of science and technology

[(3,4,5,6)] B.Tech 4th year Student, CSE(AI&ML), Vignan's Institute of Management and Technology for Women, Hyderabad, India.

[1]saipriyankathotapalli9@gmail.com, [2]divya.n@sreenidhi.edu.in, [3]amudalasruthi2505@gmail.com, [4]prasannasirimalla20@gmail.com [5]sahithibuddha247@gmail.com, [6]pendurthijyothsna@gmail.com.

ABSTRACT:

Phishers employ social engineering and mimic sites to trick users and organizations into divulging personal details such as account IDs, usernames, and passwords. Phishing URL detection, hence, in the face of this is of paramount significance. Machine learning and deep learning algorithms have been created to identify phishing URLs automatically. We use a Gradient Boosting Classifier which has been trained on a wide range of features and an extremely large corpus of data in our process. This enables the system to learn in real-time, reacting to new threats by incorporating recently detected phishing techniques, actual domain changes, and notes by experienced analysts. Our system analyzes the content of sites for harmful patterns and adds reputation-based features like domain age to aid in detection. With such sophisticated means, our system is highly resistant to phishing attacks preventing loss of funds and safeguarding confidential information.

Keywords: Phishing, URL detection, Machine Learning, Gradient Boosting Classifier, Cyber Security.

I. INTRODUCTION:

Phishing is a deceptive cybercrime technique that exploits social engineering and technological methods to steal personal and financial data. As digital platforms become an essential part of everyday life, the rise of mobile and wireless technologies has also increased cybersecurity risks. There are several methods of attack that can be employed, such as hacking, malware, or fraud; however, phishing stands out as one of the most damaging techniques. Phishing often results in considerable financial losses and harm to reputation. Cyber criminals utilize strategies that closely resemble those of legitimate organizations, tricking users into revealing sensitive information through counterfeit emails and websites. Their methods go beyond just deception; they also

incorporate technical tactics, such as malicious software aimed at capturing and stealing user credentials. As we discussed, to avoid detection, attackers utilize strategies like URL manipulation and fast-flux networks, making it harder for traditional security systems to recognize these threats. Who said recognizing phishing attacks is easy? Traditional blacklist methods fail to catch new threats, whereas heuristic approaches frequently generate high false positives. To enhance security, researchers are exploring machine learning techniques like Logistic Regression, K-

Nearest Neighbors, Support Vector Machine, Naive Bayes Classifier, Decision Tree, Random Forest, Gradient Boosting Classifier etc., for more accurate phishing detection. As phishing tactics continue to evolve, developing effective countermeasures and raising awareness are crucial in mitigating risks. Phishing has become the most serious problem, harming individuals, corporations, and even entire countries. The availability of multiple services such as online banking, entertainment, education, software downloading, and social networking has accelerated the Web's evolution in recent years. Consequently, a huge volume of data is continuously downloaded and transmitted across the Internet. Cybercriminals use spoofed emails that masquerade as messages from trustworthy companies and organizations as part of social engineering tactics, leading consumers to fraudulent websites that trick them into providing sensitive financial informaation like usernames and passwords. Moreover, technical tactics often involve deploying malicious software on devices to capture credentials directly, while systems are regularly utilized to intercept users' online account usernames and passwords.

II. LITERATURE SURVEY:

Title	Author(s)	Year	Methodologies Used	Dataset Taken	Achievements	Limitations
[1] Phishing URL Detection with Gradient Boosting Classifier	Narayana Rao Appini, V. Bhuvana Kumar, N. Yedukondalu	2025	Gradient Boosting Classifier, Random Forest, SVM, CatBoost	Public phishing URL dataset	Shown GBC to be the best model among a few ML algorithms for phishing URL detection	Real-world applicability not addressed, scalability, and privacy concerns

[2] Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection	Mahdieh Zabihimay van, Derek Doran	2019	Fuzzy Rough Set Feature Selection, Random Forest, Gradient Boosting	Universal Phishing Dataset (custom aggregate d)	Applied Fuzzy Rough Set theory to minimize dimensionality and increase model interpretability	Sparse information on deployment suitability; less emphasis on deep learning or real-time detection
[3] Automatic Detection of Phishing Pages with LightGBM	Ömer Kasim	2021	Event-based request analysis, Deep-Hybrid Feature Extraction, LightGBM	Custom-harvested phishing and legitimate webpages dataset	Request flow analysis and deep hybrid features successfully combined for enhanced phishing detection	Model and data preprocessing complexity could pose a challenge for real-time application or resource-limited use
[4] Phishing Detection: Performanc e of Classical Machine Learning Models	J.O. Ajayi , A.O. Adetun mbi	2023	Voting Ensemble (includes Gradient Boosting), Decision Tree, Logistic Regression	Phishing websites dataset from UCI or similar platform	Demonstrated ensemble models perform better than traditional ones in terms of robustness and generalization	Does not provide deep understanding of feature selection or sophisticated modeling approaches

III. METHODOLOGIES:

The proposed phishing detection system utilizes a methodical approach encompassing the collection of preprocessing, feature extraction, model training, evaluation, and real-time deployment. These methodologies ensure the Gradient Boosting Classifier (GBC) optimally detects the phishing URLs based on their different distinguishing characteristics.

1. Data Collection

Here, a complete dataset of URLs was compiled which contained the phishing and non-phishing websites obtained from: Phish Tank and Open Phish (verified phishing URLs). Alexa Top Sites and non-phishing domains (verified non-phishing URLs). Further online databases for ensuring the diversity of the forms of URLs. A dataset of phishing URLs defined by the user constructed by manipulating the authentic ones in order to replicate common phishing methods. The data was balanced to prevent bias so that the model could classify both classes accurately.

2. Data Preprocessing

The URLs collected were subjected to preprocessing to ensure data quality by removing duplicates, null values, and irrelevant information. The preprocessing process involved several key steps. First, URL parsing was performed to break down each URL into its meaningful components such as the domain name, subdomains, and query parameters. Next, tokenization was applied to separate different parts of the URL—such as the protocol, domain, path, and parameters—allowing for the extraction of valuable features. Finally, the URLs underwent removal of unnecessary elements, including the elimination of unwanted characters and the normalization of text, ensuring consistency and cleanliness in the dataset for subsequent analysis.

3. Feature Extraction:

Phishing detection is dependent on this feature engineering to extract meaningful patterns from URLs. All the features extracted in this research were classified into two categories:

A. **Lexical Features (URL Structure Analysis):**

Lexical features, particularly those derived from URL structure analysis, play a crucial role in detecting phishing websites. Phishing URLs are often longer than legitimate ones, as they attempt to mimic legitimate domains using pseudo sub-domains. The presence of special characters such as '@', '-', or an unusually high number of periods ('.') can serve as strong indicators of malicious intent. Additionally, phishing URLs typically include an unusual number of digits and special characters, forming characteristic patterns that help in their identification. Alongside lexical features, domain-based features such as domain reputation also contribute significantly to phishing detection. Newly registered domains are frequently associated with phishing, as attackers often use fresh domains to avoid blacklists. WHOIS data can further enhance detection by validating the domain registrant's name and examining the domain’s expiration date, both of which can help identify suspicious or fraudulent websites.

B. **Keyword-Based Features (Content-Based Indicators):**

Keyword-based features, also known as content-based indicators, are essential for identifying phishing attempts through the analysis of specific terms within URLs. The presence of keywords such as "login," "verify," "secure," or "banking" often signals a phishing attempt, as these terms are commonly used to lure users into entering sensitive information. Additionally, keyword-based analysis helps detect homograph attacks, where deceptive domain names closely resemble legitimate ones—for example, and using "g00gle.com" in place of "google.com" to mislead users. To facilitate machine learning model training, all these extracted features are converted into numeric representations, enabling effective pattern recognition and classification.

4. Model Selection and Training:

The Gradient Boosting Classifier (GBC) was employed for its robust capability to learn complex patterns incrementally and reduce prediction errors through iterative refinement. The model training process involved several key stages. First, data splitting was carried out by dividing the dataset into 80% training data and 20% test data. To ensure uniformity and improve model performance, feature scaling and encoding were applied to convert numerical and categorical values into a consistent format. During model training, GBC utilized weak learners—specifically decision trees—where each new tree aimed to correct the errors made by the previous ones. To further improve the model's

performance and prevent overfitting, hyperparameter tuning was conducted, adjusting parameters such as the learning rate, number of estimators, and maximum depth of trees.

5. Model Evaluation:

The model was evaluated using standard classification metrics to assess its effectiveness in detecting phishing URLs. Accuracy was measured to determine the overall correctness of the model's predictions. Precision and recall were used to ensure that the model accurately identified phishing sites while minimizing false positives and false negatives. The F1-score, which combines precision and recall, provided a balanced measure of the model's classification performance. Additionally, the ROC curve and AUC score were employed to evaluate the model's ability to distinguish between phishing and legitimate URLs, offering insight into its confidence and robustness in classification tasks.

6. Real-Time Detection and Deployment

To evaluate the practical usability of the model, it was integrated into a real-time phishing detection system that allows users to input URLs for classification. A simple and intuitive frontend interface was developed where users can enter the URL to be checked. The backend processing handles feature extraction, loads the pre-trained Gradient Boosting Classifier model, and classifies the input URL as either phishing or legitimate. The system was optimized to provide fast prediction times, ensuring that users receive immediate feedback without noticeable delays, making it efficient and user-friendly for real-time deployment.

IV. ALGORITHM:

Gradient Boosting Classifier Algorithm for Domain Detector

Step 1:

Initialize the model with a constant value

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

Step 2: For m = 1,……., M

Step 3:

Complete the (pseudo-) residuals of the predictions and the true values\

$$r_{im} = -\left[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

Step 4:

$$(x_i, r_{im})_{i=1}^{n}$$

Fit a model (Weak Learner) $h_m(x)$ to the residuals, i.e., use the training data set

Step 5:

Find the optimized Solution for the Loss Function

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

Step 6:

Update the Model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Step 7:

Output the final model

$$F_M(x)$$

V. RESULTS:

Our GBC-based phishing detection system performed a remarkable 97.4% accuracy, making it extremely reliable at differentiating between phishing sites and legitimate sites. During testing, the model correctly identified most phishing attempts without causing unnecessary false alarms. In comparison with other machine learning models such as Random Forest, Decision Trees, and SVM, GBC performed the best, with its capability to identify even advanced phishing patterns. The system processes URLs in milliseconds, which is perfect for real-time scanning. Even when executed on newly designed phishing sites, it performed quite well, bearing witness to its adaptability. Even though the model is extremely efficient, some borderline cases—phishing sites that closely resemble legitimate sites with a high level of similarity—were difficult but only slightly so. With ongoing advancements in features, for example, the utilization of feature extractors for improved performance and ongoing learning for improved performance in new environments, this system is well on its way to being an even more powerful ally in the realm of cybersecurity, protecting users from web threats.

Fig 1: Legitimate Website **Fig 2: Phishing Website**

VI. COMPARATIVE ANALYSIS

S.NO	MODEL	VALUES
1.	ning URL Detection with Gradient Boosting Classifier [1]	Accuracy: 92.82%, Precision: 92.63%, Recall: 93.04%, F1-score: 92.82%
2.	Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection [2]	F1-score: 93%
3.	Automatic Detection of Phishing Pages with Light Gradient Boosted Machine Model [3]	Accuracy: 92%-93%
4.	Our Domain Detector	Accuracy: 97.4%, Precision: 98.9%, Recall: 98.8%, F1-score: 97.4%

VII. CONCLUSION:

The solution in the paper presented in the work provides a machine learning approach to phishing detection by employing the Gradient Boosting Classifier (GBC) in the process of attempting to label URLs as legitimate or phishing. The model helps detect phishing attacks by screening attributes within URLs and is cost-effective with regards to high accuracy and minimal false positives compared to the blacklist approach. By learning from a labeled data set, feature extraction, and preprocessing, the system demonstrates that it can identify complex patterns on phishing websites. The outcome shows that algorithmic approaches based on machine learning, such as ensemble approaches like GBC, improve phishing detection accuracy by a great margin through the identification of minute differences between malicious and secure URLs. Furthermore, the real-time system detection feature of the system makes the system a valuable asset in cybersecurity since it

helps users to steer clear of scam sites prior to their use. Unlike static detection tools, the model can potentially adapt to counter new phishing strategies in a bid to provide better protection against increasingly prevalent cyber-attacks.

VIII. FUTURE SCOPE:

Further work can be conducted to improve the model by utilizing ensemble models to achieve a higher accuracy score. Ensemble methods is a ML method that takes numerous base models and uses them to create an optimal predictive model. More extensive future work would be integrating several classifiers, which have been trained on various aspects of the same training set, to one classifier that can give a stronger prediction than any of the individual classifiers alone. The project can further encompass other types of phishing such as smishing, vishing. to make the system complete. Even looking ahead even further, the approach must be tested on how it could deal with collection growth.

IX. REFERENCES:

[1] Narayana Rao Appini, V. Bhuvana Kumar, "Phishing URL Detection with Gradient Boosting Classifier", Communications on Applied Nonlinear Analysis, Vol. 32 No. 3 (2025)

[2] Mahdieh Zabihimayvan, Derek Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection", arXiv (2019),

[3] Ömer Kasim, "Automatic Detection of Phishing Pages with Event-Based Request Processing, Deep-Hybrid Feature Extraction and Light Gradient Boosted Machine Model", Telecommunication Systems, Springer (2021)

[4] J.O. Ajayi , A.O. Adetunmbi, "Phishing Detection: Performance Evaluation of Both Ensemble and Classical Machine Learning Models", International Journal of Information Security, Privacy and Digital Forensics.

[5] Pawan Prakash, Manish Kumar, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks" 2022 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC)

[6] Shelby R. Curtis, Prashanth Rajivan, "Phishing attempts among the dark triad: Patterns of attack and vulnerability" October 2018 Sciencedirect

[7] K.N.S.B.V. Manjushal, Dr. D. Jaya Kumari2, "Detecting Phishing Links Analysis Using Machine Learning" 2024, IJFMR

[8] A. Alswailem, B. Alabdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications& Information Security (ICCAIS), 2019, pp. 1-6.

[9] J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 2020, pp. 43-46.

[10] M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Websites Using Machine Learning," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), 2020, pp. 111-114.

[11] A. Razaque, M. B. H. Frej, D. Sabyrov, A. Shaikhyn, F. Amsaad and A. Oun, "Detection of Phishing Websites using Machine Learning," 2020 IEEE Cloud Summit, 2020, pp. 103-107.

[12] Thomas Nagunwa, "Comparative Analysis of Nature-Inspired Metaheuristic Techniques for Optimizing Phishing Website Detection", MDPI, 2024

[13] D Shanthi , Smart Healthcare for Pregnant Women in Rural Areas, Medical Imaging and Health Informatics, Wiley Publishers,ch-17, pg.no:317-334, 2022

[14] Shanthi, R. K. Mohanty and G. Narsimha, "Application of machine learning reliability data sets", Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS), pp. 1472-1474, 2018.

[15] D Shanthi, N Swapna, Ajmeera Kiran and A Anoosha, "Ensemble Approach Of GPACOTPSOAnd SNN For Predicting Software Reliability", International Journal Of Engineering Systems Modelling And Simulation, 2022.

[16] Shanthi, "Ensemble Approach of ACOT and PSO for Predicting Software Reliability", 2021 Sixth International Conference on Image Information Processing (ICIIP), pp. 202-207, 2021.

[17] D Shanthi, CH Sankeerthana and R Usha Rani, "Spiking Neural Networks for Predicting Software Reliability", ICICNIS 2020, January 2021, [online] Available: https://ssrn.com/abstract=3769088.

[18] Shanthi, D. (2023). Smart Water Bottle with Smart Technology. In Handbook of Artificial Intelligence (pp. 204-219). Bentham Science Publishers.

[19] Shanthi, P. Kuncha, M. S. M. Dhar, A. Jamshed, H. Pallathadka and A. L. K. J E, "The Blue Brain Technology using Machine Learning," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1370-1375, doi: 10.1109/ICCES51350.2021.9489075.

[20] Shanthi, D., Aryan, S. R., Harshitha, K., & Malgireddy, S. (2023, December). Smart Helmet. In International Conference on Advances in Computational Intelligence (pp. 1-17). Cham: Springer Nature Switzerland.

[21] Babu, Mr. Suryavamshi Sandeep, S.V. Suryanarayana, M. Sruthi, P. Bhagya Lakshmi, T. Sravanthi, and M. Spandana. 2025. "Enhancing Sentiment Analysis With Emotion And Sarcasm Detection: A Transformer-Based Approach". Metallurgical and Materials Engineering, May, 794-803. https://metall-mater-eng.com/index.php/home/article/view/1634.

[22] Narmada, J., Dr.A.C.Priya Ranjani, K. Sruthi, P. Harshitha, D. Suchitha, and D.Veera Reddy. 2025. "Ai-Powered Chacha Chaudhary Mascot For Ganga Conservation Awareness". Metallurgical and Materials Engineering, May, 761-66. https://metall-mater-eng.com/index.php/home/article/view/1631.

[23] Geetha, Mrs. D., Mrs.G. Haritha, B. Pavani, Ch. Srivalli, P. Chervitha, and Syed. Ishrath. 2025. "Eco Earn: E-Waste Facility Locator". Metallurgical and Materials Engineering, May, 767-73. https://metall-mater-eng.com/index.php/home/article/view/1632.

[24] P. Shilpasri PS, C.Mounika C, Akella P, N.Shreya N, Nandini M, Yadav PK. Rescuenet: An Integrated Emergency Coordination And Alert System. J Neonatal Surg [Internet]. 2025May13 [cited 2025May17];14(23S):286-91. Available from: https://www.jneonatalsurg.com/index.php/jns/article/view/5738

[25] D. Shanthi DS, G. Ashok GA, Vennela B, Reddy KH, P. Deekshitha PD, Nandini UBSB. Web-Based Video Analysis and Visualization of Magnetic Resonance Imaging Reports for Enhanced Patient Understanding. J Neonatal Surg [Internet]. 2025May13 [cited 2025May17];14(23S):280-5. Available from: https://www.jneonatalsurg.com/index.php/jns/article/view/5733

[26] Srilatha, Mrs. A., R. Usha Rani, Reethu Yadav, Ruchitha Reddy, Laxmi Sathwika, and N. Bhargav Krishna. 2025. "Learn Rights: A Gamified Ai-Powered Platform For Legal Literacy And Children's Rights Awareness In India". Metallurgical and Materials Engineering, May, 592-98. https://metall-mater-eng.com/index.php/home/article/view/1611.

[27] Shanthi, Dr. D., G. Ashok, Chitrika Biswal, Sangem Udharika, Sri Varshini, and Gopireddi Sindhu. 2025. "Ai-Driven Adaptive It Training: A Personalized Learning Framework For Enhanced Knowledge Retention And Engagement". Metallurgical and Materials Engineering, May, 136-45. https://metall-mater-eng.com/index.php/home/article/view/1567.