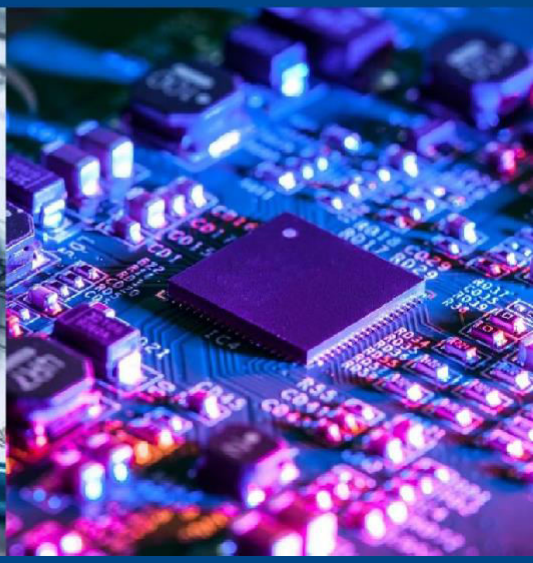
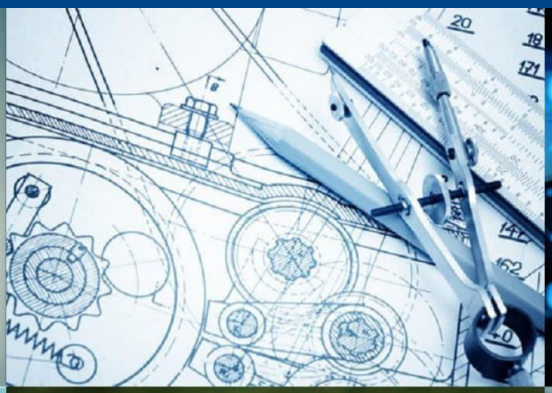
ISSN: 2582-7219

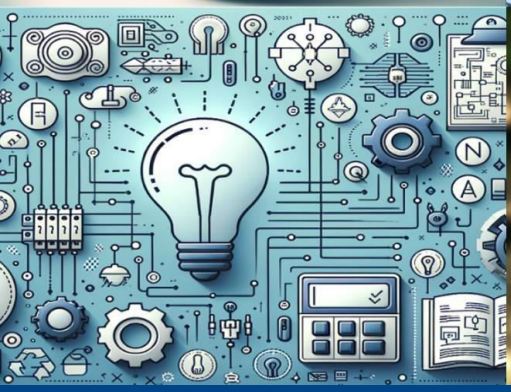




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 7, July 2025**

© 2025 IJMRSET | Volume 8, Issue 7, July 2025| DOI:10.15680/IJMRSET.2025.0807079

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Phishing URL Detection using Machine Learning

Lavanya M R, Dr. Udaya Kumar N L

M.Tech (CSE) Student, Shreedevi Institute of Engineering and Technology, Tumkur, Karnataka, India

Associate Professor, Dept. of CSE, Shreedevi Institute of Engineering and Technology, Tumkur, Karnataka, India

ABSTRACT: Cyber criminals' impersonation is one of the most common schemes, including them pretending to be someone else in order to steal personal information bearing on username and password, credit card numbers, etc. Phishing websites are of central importance to these attacks because they generate very similar-looking URLs to a genuine site. Since these methods are new to the world, the Blacklisting approach has been realized to be inadequate especially because now phishers are getting creative with the methods they use.

This paper aims at solving the emerging problem of detection of phishing URLs through the use of machine learning. Thus, through transforming URLs and by studying various structural and behavioral characteristics we are to create a strong model that will give minimum errors in the distinguishing of phishing and legitimate links. Our approach is to utilize a number of features such as URL length, HTTPS presence, domain age, etc. Moreover, we stress the feature extraction step, and look into the effectiveness of different ML techniques for classifying the phishing URLs. This work also shows the drawbacks of using the same system for phishing detection in different environment, showing that more dynamic approaches are necessary.

Based on the large data set and comparing the results of multiple algorithms, this work can help improve the performance of detecting phishing URLs and at the same time, propose more solutions for combating this type of cyber threat proactively.

KEYWORDS: Phishing URL Detection, Machine Learning, Feature Extraction, Cybersecurity, Classification Algorithms, Dynamic Detection, URL Structure

I. INTRODUCTION

Phishing itself constitutes one of the most common cyber threats and the attackers pretend to belong to a credible entity and will immediately request an individual's usernames, passwords, or even financial information. Phishing sites are at the core of these attack since they develop links that look almost authentic. While the advancement in techniques of phishing makes the use of check list ineffective because they can barely handle new development.

The research problem focuses on the modern threat of some websites, and more specifically, the identification of phishing URLs using machine learning approaches. In this paper, we seeks to classify phishing URLs from legitimate URL using structural and behavioral analysis of URLs in order to achieve a high level of accuracy. Our strategy involves the use of multiple characteristics such as URL length, existence of HTTPS and domain's age, among others. Furthermore, we also discuss the criteria for feature selection and compare the performances of the different algorithms that we implanted for the detection of phishing URLs. This research also discusses the shortcoming of extending the phishing detection mechanisms for the distinct environment, again proving that there is a need for models that can overcome the dynamics of the phishing threats.

By using a data-driven approach and comparing multiple algorithms, this research aims to enhance the accuracy and efficiency of phishing URL detection, providing a more proactive solution to a persistent cybersecurity threat.

Problem Statement:
Given the ever changing types of scams, spoting suspicious links has become almost impossible especially due to the incidence of more complex phishing attacks. Other early methodologies like blacklists will never work since they're unable to identify the newly produced or camouflaged URLs which are very dangerous to the users. Since the current phishing sites look very much like the original sites with slight differences, there is a strong requirement for a smart detection system that can identify phishing attempts from the structural and behavioral characteristics of URLs.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Objectives:

1. Create a classifier model that would differentiate between the phishing URLs and benign ones by considering URL characteristics.
2. Compare the output of using several algorithms in the task of diagnosing cases of phishing.
3. List and prioritize the features used for the recognition of the phishing attacks.
4. Algorithms should come up with a general solution that can counter current as well as emerging phishing methodologies.

II. LITERATURE REVIEW

Phishing attacks have become one of the most common cybersecurity threats, as they exploit users' trust by mimicking legitimate websites. In response, numerous studies have explored machine learning approaches to detect phishing URLs effectively. Over the past few years, researchers have developed various models, algorithms, and feature extraction techniques to address the limitations of traditional phishing detection methods.

Alam et al. [1] described various machine learning techniques and their types in case of variable detection for phishing URL. Their work also explained how supervised learning techniques like decision tree, random forest and support vector machines (SVM) can be used to categorize the certain URLs according to the features. They also stressed the specifics of feature extraction pointing out that phishing sites often contain IP addresses or have rather close domain names skewing the users. Nonetheless, Alam et al. admitted that one drawback of this approach is the difficulty of faithfully preserving the high efficiency of the detection needed given the ever-constant change in the way phishers work.

Unlike Gandotra and Gupta [2], spoofed website detection was made a primary object of enhancement. They proposed this improved feature selection technique in their study to help in minimising the false positive and false negative results. Specifically, the paper emphasized on the need to study subdomain, URL length and use of prefixes and suffixes within the URL as they are normally adjusted in phishing attacks. , some of the issues which are generally experienced with existing detection systems were solved by the enhanced feature selection process proven to increase the accuracy and generalization of the detection models.

Jalil et al. [3] provided a state of the art analysis of the classifiers employed in the detection of phishing URLs. Their work discussed various machine learning classifiers such as, logistic regression, decision tree classifiers, and ensemble classifiers. For random forests and gradient boosting, the authors learned that ensemble models always outperformed single models because of the feature of aggregating several weak classifiers. Jalil et al. also pointed out that issues of overfitting also affect the phishing detection models especially when the models are trained on outdated data sets. The obtained results highlighted that the fresh training data would be critical for the model and robots' stability in actual environment conditions.

In the light of this, Tanwar et al. [4] have proposed empirical study related to phishing detection deployability of the machine learning techniques. Many of these features they used in their study included whether the site used HTTPS connection, the length of the domain registration, and the URL length of the site. The authors discovered that legitimate sites usually take longer to register than phishing sites, as well as the fact that legitimate sites are likely to use HTTPS protocols while phishings sites are likely to use shorter registration periods or none at all or HTTPS. This research therefore shed light into the importance of URL metadata in the detection of phishing attacks and established that models that majors on Content Analysis Alone have limited capabilities.

A more specific study work has been made by Sánchez-Paniagua et al. [5] to analyse the detection of phishing URL in the framework of login pages. Real-world scenario based study used by them employed machine learning algorithms to analyze login URLs. As identified in the findings, the refinement increased the detection ratios using additional characteristics of URLs, including login form behavior and subdomain structures. Thus, analyzing the case of login URLs the work contributed to the further development of the discussion on the effectiveness of the targeted phishing detection models.

Most recently, Mahajan et al. [6] proposed a new approach of using correlation coefficients in the modeling of the methodology in identifying phishing URLs. As the authors of the discussed work pointed out, in their study they found that features of the URLs demonstrate different level of association, and it is within this context that their findings

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

indicate that these features could be enhance model. Thus, extending their works, they integrated the correlation analysis with the traditional machine learning to improve the accuracy of the phishing detection. Here, this approach was found to be useful in detecting otherwise, unnoticed patterns of phishing URLs that perhaps other conventional means of detection would fail to recognize.

Other works have also attempted to use deep learning methods for phishing URL detection. Zhao et al. [7] used deep neural networks to classify the URLs. From their work they were able to show that classifiers developed under the deep learning approach performed better than the regular classifiers under the machine learning approach, especially where use was made of large datasets. Another advantage of using deep learning models is the fact that the features needed for the model are often learned from raw data, there is little need for feature engineering. However, the study also pointed out that there is a problem of computation time and training duration when using deep neural networks which can affect their usability in real time systems.

Rathee et al. [8] used a hybrid approach where they propose a phishing detection model that incorporates multiple machine learning algorithms. The model that they used in their case study realizing in the real-time environment the function of the concrete result of the classifiers' cooperation, which revealed the excess of additives and the lack of the other and at the same time is more effective. By using the proposed ensemble model, the identification effectiveness and the rates of false alarms were increased making the solution suitable for practical use for monitoring phishing threats. Rathee et al. also stressed that feature diversities must be taken into consideration since the offered ensemble model used both content-based and metadata-based features to improve detection.

Finally, Wundsam et al. [9] investigated a feature fusion approach to fighting phishing. Pham et al. identified and evaluated five URL features, namely domain registration information, SSL certificate expiry date, and URL format, etc., for constructing their phishing detection model. In the similar way Wundsam et al shown that their model is capable of detecting more complex kind of features that are typically not noticed by the conventional approach. They found that the feature fusion technique was most efficient at differentiating between what is essentially a 'similar' domain structure of a well-known domain and a phishing site.

In the literature, the progress of phishing detection models has been categorized in the following simple taxonomy: startPosition => black list model => machine learning model => deep learning model. Although the authors point out a number of achievements towards this end, the studies also highlight the importance of refining datasets and detection models at regular intervals owing to the dynamism of phishing attackers. The proposed combination of URL features and the mixture of both types of models deserves further exploration in the development of efficient approaches to address phishing URL detection.

III. PROPOSED METHOD

This research focuses on the following goals:- To create a machine learning based efficient and accurate system for identifying phishing URLs. The proposed study is an attempt to overcome the limitations of the Black List methods by incorporating certain characteristics of the URLs such as the structure of domain name and the usage of www or https. The purpose of this work will be to compare the effectiveness of diverse machine learning algorithms for URL classification as phishing or legitimate and the effectiveness of the diverse URL attributes in boosting a model's accuracy.

Method: To achieve the objectives, the following methodology is proposed:

1. **Dataset Collection**: A comprehensive dataset consisting of both phishing and legitimate URLs is used. Key URL features, including the presence of IP addresses, the length of the URL, HTTPS usage, and domain registration length, are extracted for analysis.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

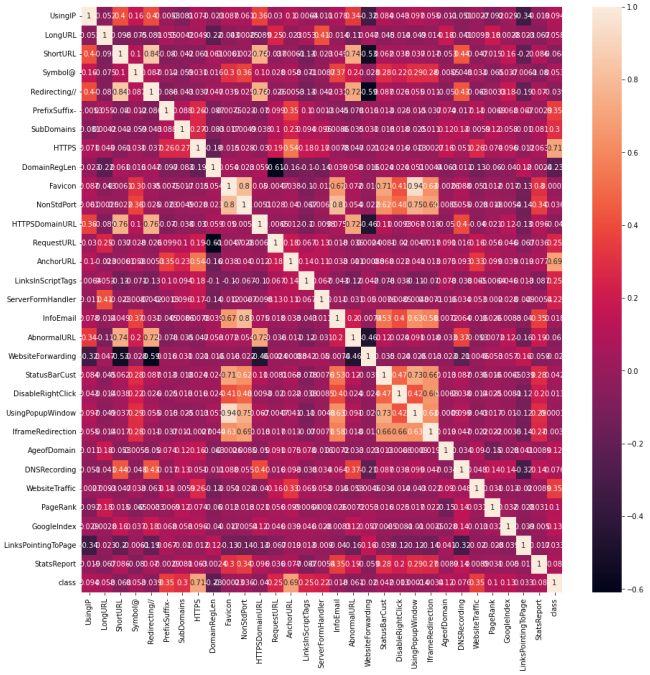


Fig 1: Correlation Matrix with the selected features from the dataset

Dataset Overview

The dataset used in this study consists of URLs labeled as phishing (-1) or legitimate (1). Each URL is represented by various features that describe its structure and behavior:

1. UsingIP: Indicates if the URL uses an IP address instead of a domain name. Phishing sites often use IP addresses to avoid being tracked.
2. LongURL: Checks if the URL is excessively long. Phishing URLs are often long to confuse users.
3. ShortURL: Determines if a URL shortening service is used, common in phishing URLs to hide their destination.
4. Symbol@: Detects the presence of the "@" symbol, which is used to deceive users by redirecting them to phishing websites.
5. Redirecting//: Indicates if "//" appears in the middle of the URL, often used to obfuscate the true destination of phishing URLs.
6. PrefixSuffix-: Checks if a hyphen is used in the domain name, which is frequently seen in phishing domains.
7. SubDomains: Counts the number of subdomains. Phishing websites often use multiple subdomains to imitate legitimate domains
8. HTTPS: Checks if the URL uses HTTPS. Some phishing websites use HTTPS to appear legitimate.
9. DomainRegLen: Measures the domain's registration length. Legitimate domains are typically registered for longer periods.
10. Favicon: Indicates if the favicon is loaded from the same domain. Phishing websites may load favicons from external sources.
11. NonStdPort: Detects the use of non-standard ports, which can bypass security protocols.
12. HTTPSDomainURL: Verifies whether the domain uses HTTPS. Legitimate websites usually enforce HTTPS.
13. RequestURL: Detects if webpage resources (e.g., images) are loaded from external domains, which is common in phishing sites.
14. AnchorURL: Checks whether webpage links are legitimate or lead to suspicious domains.
15. LinksInScriptTags: Counts external links hidden inside script tags, often used by phishing websites to hide malicious content.
16. ServerFormHandler: Checks if form actions send data to external servers, common in phishing sites to steal user input.
17. InfoEmail: Indicates if the website requests personal information via email, which is unusual for legitimate sites.
18. AbnormalURL: Detects deviations from expected URL structures, common in phishing URLs.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

19. WebsiteForwarding: Indicates if the URL redirects to multiple sites, often done to hide phishing attempts.
20. StatusBarCust: Detects whether the webpage customizes the browser's status bar, which is used by phishing sites to hide link destinations.
21. DisableRightClick: Indicates if right-click functionality is disabled, preventing users from inspecting webpage content.
22. UsingPopupWindow: Detects if pop-up windows are used, common in phishing websites to capture user input.
23. IframeRedirection: Detects hidden iframes used to redirect users without their knowledge.
24. AgeofDomain: Measures domain age. Older domains are generally more trustworthy, while phishing sites tend to use newer domains.
25. DNSRecording: Checks if the domain has a valid DNS record. Phishing websites may lack valid DNS records.
26. WebsiteTraffic: Estimates website traffic. Legitimate websites usually have higher traffic than phishing websites.
27. PageRank: Measures search engine ranking. Legitimate websites typically have higher PageRank values.
28. GoogleIndex: Indicates whether the website is indexed by Google, which phishing websites often are not.
29. LinksPointingToPage: Counts the number of external links pointing to the page. Legitimate websites typically have more external backlinks.
30. StatsReport: Indicates whether the website appears in phishing statistical reports, marking it as suspicious

2. Feature Engineering: Several key features are preprocessed and transformed to improve their suitability for phishing detection. The following operations are performed:

1. **Model Training**: Machine learning algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM) are employed. These models are trained on the dataset to classify URLs as phishing or legitimate based on the selected features.

2.1 Logistic Regression

The probability that a URL is phishing is modeled using logistic regression, represented by the following equation:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Where:

- $P(y = 1 \mid x)$ is the probability of the URL being phishing,
- $x_1, x_2, \ldots, x_n$ are the feature values of the URL,
- $\beta_0, \beta_1, \ldots, \beta_n$ are the model parameters learned during training.

2.2 Support Vector Machine (SVM)

To separate phishing URLs from legitimate ones, an SVM constructs a decision boundary. The decision function is given by

$$\omega . x + b = 0$$

Where:

- ω is the weight vector,
- x is the input feature vector,
- b is the bias term.

2.3 Gradient Boosting

The Gradient Boosting classifier minimizes the error iteratively by combining weak learners. The error minimization is expressed as:

$$L(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:

- $L(y, \hat{y})$is the loss function,
- y_i is the actual label,
- $\hat{y}_i$ is the predicted label

3. Random Forest

For Random Forest, each decision tree is built by splitting data based on criteria like

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Gini Index is computed as:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

Where p_i the probability of class i

- Information Gain:

$$IG(S,A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v)$$

4. Evaluation: The models are evaluated using metrics such as accuracy, precision, recall, and F1 score to determine their performance. The feature importance is also analyzed to identify which characteristics contribute most to the detection process

5. The proposed phishing URL detection system is structured to ensure a comprehensive and efficient classification process. As illustrated in Figure 1, the system begins with data collection from various sources, both phishing and legitimate URLs. Following this, preprocessing and feature extraction steps are applied to prepare the data for model training. After feature selection, the machine learning model is trained and evaluated based on several performance metrics. A feedback loop ensures continuous model improvement

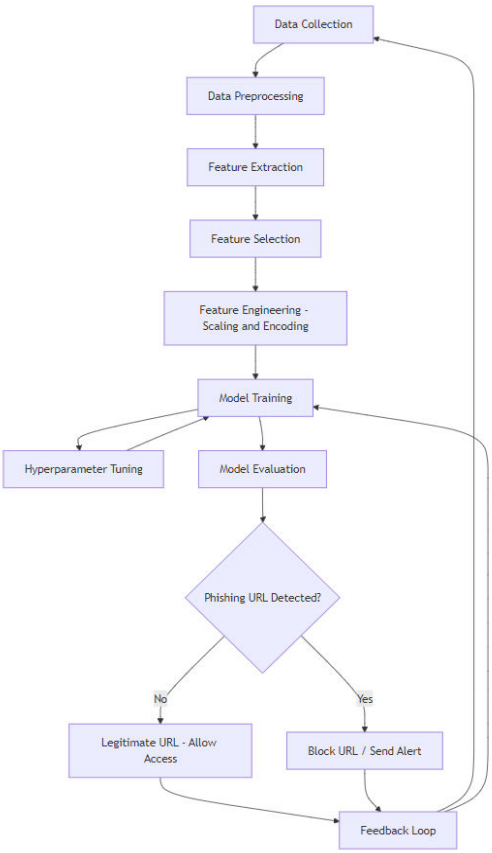


Fig 1: Flowchart of the Proposed Phishing URL Detection System

© 2025 IJMRSET | Volume 8, Issue 7, July 2025| DOI:10.15680/IJMRSET.2025.0807079

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. RESULTS AND DISCUSSION

In this assessment, we tested myriad of models in order to identify the approximate of the set of URLs in two categories: phishing or not. The effectiveness of the developed models was assessed by various evaluation parameters including accuracy, precision, recall, and F1-score. Bar chart depicting the results of each classifier in the light of evaluation metrics Performance of each classifier in terms of length of code, number of features used in training, accuracy, precision, recall and f1 score is shown in the following figures

ML Model	Accuracy	f1_score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
CatBoost Classifier	0.972	0.975	0.994	0.989
Multi-layer Perceptron	0.971	0.974	0.992	0.985
XGBoost Classifier	0.969	0.973	0.993	0.984
Random Forest	0.967	0.97	0.992	0.991
Support Vector Machine	0.964	0.968	0.98	0.965
Decision Tree	0.961	0.965	0.991	0.993
K-Nearest Neighbors	0.956	0.961	0.991	0.989
Logistic Regression	0.934	0.941	0.943	0.927
Naive Bayes Classifier	0.605	0.454	0.292	0.997

Among the models, the Gradient Boosting Classifier achieved the highest accuracy of 0.974 with a precision of 0.977 and recall of 0.994, followed closely by the CatBoost Classifier (accuracy: 0.F1 score: 971: Precision: 0,975, Recall: 0,994. The two ensembles showed very good average performance, though higher in terms of recall which means both approaches can accurately detect phishing URLs.

The last proposed model is the Multi-layer Perceptron (MLP) with accuracy 0.971 and the XGBoost with accuracy 0.969. Given the explicit and significant phishing threats that these models can achieve, they offer attractive convergence with conventional methods of detection.

On the other hand, traditional classifiers such as Logistic Regression and Naive Bayes were slower to learn and Naive Bayes' precision (0.454) and recall (0.292) are poor to make it as an efficient classifier for this task. New suggested classificator Logistic Regression had accuracy of 0.934, still all modern classifiers such as Gradient Boosting or especially CatBoost outperformed it

IV. CONCLUSION

This paper aims to compare several machine learning techniques that have been proposed in the literature for detecting phishing URLs while evaluating them using several Evaluation metrics namely: Accuracy, Precision, Recall, and F1-score. Among all the tested models the ensemble methods were indicated as the most efficient including the Gradient Boosting, CatBoost and XGBoost where the accuracy of the Gradient Boosting reached the maximum value making

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

97.4%. These models show a good precision and recall, which makes it optimal especially for applications for which false negatives are detrimental, such as phishing detection.

Logistic Regression and Naive Bayes, being traditional classifiers, proved to be comparatively worst, especially Naive Bayes showing very low precision and recalls and hence, not suitable for this task at all. The poor performance of these models can be however attributed to the fact that the models fail to handle the interaction between the features in data sets containing phishing URLs well. On the other hand, ensemble methods performed well because it has the strength of integrating numerous weak learners which could easily address intricate pattern platforms.

Still, non-dominating two top models were found to have some limitations such as ensemble methods involving higher learning steps and hence in terms of computational complexity were slightly higher. However, because of their enhanced overall performance, particularly in terms of identifying phishing URLs, these models are good candidates for practical cybersecurity uses, for example, as browser add-ons, e-mail filters or web activity scanners.

As for future work, features including real-time behavioral data or website addresses should be inserted; Also, the combination of neural networks and ensemble method should be developed to enhance the detection function. Furthermore, extending this scope to other types of cyberattacks, for example speakers or social engineering, could improve the performance of these models in practice.

REFERENCES

[1] Alam, M.N., et al., "Phishing Attacks Detection Using Machine Learning Approach," *IEEE ICSSIT Conference*, 2020.
[2] Sánchez-Paniagua, M., et al., "Phishing URL Detection: A Real-Case Scenario Through Login URLs," IEEE Access, vol. 10, pp. 42949–42960, 2022
[3] Gandotra, E., Gupta, D., "Improving Spoofed Website Detection Using Machine Learning," Cybernetics and Systems, 2021
[4] Tanwar, S., et al., "Detection of Phishing Websites Using Machine Learning Algorithms," Springer Cybernetics Systems, vol. 52, pp. 169–190, 2022.
[5] O. K. Sahingoz, E. BUBEr and E. Kugu, "DEPHIDES: Deep Learning Based Phishing Detection System," in IEEE Access, vol. 12, pp. 8052-8070, 2024, doi: 10.1109/ACCESS.2024.3352629
[6] V. Borate, A. Adsul, R. Dhakane, S. Gawade, S. Ghodake, and M. Jadhav, "A Comprehensive Review of Phishing Attack Detection Using Machine Learning Techniques," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 435–441, Oct. 2024, doi: 10.48175/ijarsct-19963.
[7] B. T. Geetha, P. Malathi, T. Thirumalaikumari, V. Janakiraman, H. A. Basha, and S. Devi, "Machine Learning Approaches for Proactive Phishing Attack Detection," pp. 757–762, Aug. 2024, doi: 10.1109/iccpct61902.2024.10672638.
[8] Y. Padmini and U. C. Sree, "Phishing Website Detection using Machine Learning," *Journal of Innovation and Technology*, vol. 2024, no. 1, Nov. 2024, doi: 10.61453/joit.v2024no30.
[9] M. S. Alzboon, M. S. Al-Batah, M. Alqaraleh, F. Alzboon, and L. Alzboon, "Phishing Website Detection Using Machine Learning," *Gamification and Augmented Reality.*, vol. 3, p. 81, Jan. 2025, doi: 10.56294/gr202581.
[10] Y. S. Tambe, "Phishing URL Detection Using Machine Learning," *Journal of advanced research in production and industrial engineering*, Sep. 2023, doi: 10.24321/2456.429x.202301.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com