Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Phishing Website Detection using ML

Aditya Raj
Department of Computer Science and Engineering
Galgotias University
Uttar Pradesh, India

Anand Raj
Department of Computer Science and Engineering
Galgotias University
Uttar Pradesh, India

Pooja singh
Department of Computer Science and Engineering
Galgotias University
Uttar Pradesh, India

***Abstract*—Because phishing is a constantly evolving and growing threat in the cyberspace world, it seriously hampers individuals as well as organizations that have crucial information such as login credentials, monetary information, and other personal details being miserably scammed. The typical way of detecting phishing sites has been totally reliant on blacklisting techniques and rule-based frameworks. However, these are reactive approaches that are not at all efficient or effective in the case of newly developed phishing sites. In this paper, an efficient proactive solution has been put forward for detecting phishing sites with higher levels of accuracy utilizing machine learning algorithms. The salient feature of the system presented in this paper is its robust feature extraction mechanism that explores various features of the sites. These features include URL features, Domain features, as well as attributes concerning Webpage features. Certain significant features such as the presence of IP addresses in the URL, presence of unusual URL lengths, registration age of the domain name, attributes of the redirect features, as well as attributes of form features have also been considered in the paper in order to test the merits of higher accuracies in the system developed.**

***Index Terms*—Phishing Detection, Machine Learning, Cybersecurity, URL Analysis, Feature Extraction, Support Vector Machine, Random Forest, Gradient Boosting, Website Classification, Malicious URL Detection.**

I. Introduction

With the rising trend of using the internet, there are immense effects on the manners by which people interact, by which societies conduct their business, and by which they manage information. But with the rising trend of using the internet, there arises an increasing fear of different cyber crimes. Out of different cyber crimes, it has been seen that phishing is the most common type of it. It can be described as a kind of cyber crime by which people tend to expose personal information such as login names, passwords, credit card numbers, and personal details by disguising themselves within a genuine website.

The traditional phishing protection systems rely primarily on blacklist techniques. Blacklist systems maintain a list of malicious URL addresses, thereby prohibiting access to these malicious URL addresses by users. Despite their importance, there are significant shortcomings with these systems. Phishing pages contain dynamic attributes with a shorter lifetime of several days until they are shut down and replaced by malicious phishing pages with malicious domain names. Consequently, blacklist systems are ineffective for identifying phishing pages with zero-day vulnerabilities.

In the effort to present answers to these issues, there has been a growing dependence on the utilization of machine learning in order to present a remedy to the challenge posed by phishing. Through the utilization of machine learning, the ability to train based on past experience in order to learn specific patterns, which may not necessarily be interpretable, is achievable. In this regard, through the utilization of various parameters for a web page, for instance the URL and the information presented in the page, the ability to identify phishing and genuine pages is achievable.

This proposed research will offer a credible and proactive phishing site detection method that utilizes the application of the machine learning algorithm techniques. This is especially proposed to enhance the accuracy and flexibility levels. This proposed model demonstrates a detailed process of feature extraction, which is dependent upon the examination of the group of variables identified relating to the concept of phishing.

It uses a labeled dataset containing the URL of both genuine and phishing websites to train and test various machine learning classifiers, such as Support Vector Machine, Random Forest, and Gradient Boosting classifiers. They are chosen because they are able to provide optimal results in classification problems, besides having capabilities for modeling complex, nonlinear data distributions. To ensure that the performance of the developed system is assessed comprehensively, standard performance measures, including accuracy, precision, recall, and false positive rate, are employed.

The test results confirm the proposed machine learning-based approach to have high accuracy with a low false positive rate, hence it would be appropriate for practical usage. The proposed scheme is able to generalize and detect new phishing sites unlike conventional black-listing methods. This is because the proposed scheme is scalable and efficient, thus can be rapidly adapted to the rapidly increasing threats in

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

cyber-attacks.

II. Literature Review

The phishing attack detecting field has attracted the attention of many researchers because of the rising number of instances and complications in relation to cyber attacks on online users. Several techniques have been proposed in the recent past that are capable of detecting and mitigating the effects of phishing attacks. The techniques include the basic ones that employ the black list concept to mitigate phishing attacks to the advanced ones that make use of deep learning concepts. The next section will introduce the reader to the various research works that have been carried out in relation to phishing attack detections.

1. Conventional Methods of Phishing Detection

Early anti-phishing solutions were mainly based on blacklist techniques and heuristic techniques. Blacklist techniques keep a database of URLs of known phishing pages and deny access to a URL if a match is found in the database. The drawback of using blacklist techniques is that they are limited to identifying already known phishing pages. The life span of a phishing page is typically short and is periodically replaced with a new one. Hence, a whitelist system is incapable of providing protection against a zero-day attack.

Heuristic-based systems try to locate phishing pages by using predefined rules such as checking for phishing URLs in URLs, the use of special characters in URLs, and inconsistencies in the URLs of the phishing domains. While heuristic-based systems are an improvement over blacklisting models in phishing detection, they are manual and non-adaptive in terms of phishing attack modifications. This means that heuristics-based models are ineffective in defending against phishing since phishing attack strategies keep changing.

2. Machine Learning Algorithm for Phishing Detection

To avoid the shortcomings associated with conventional approaches, there has been an increased trend of using the machine learning algorithm in phishing detection. Indeed, the supervised learning algorithm possesses the ability to distinguish sophisticated patterns from large data. Most authors have confirmed that supervised learning is effective in phishing site identification using the extracted features.

Support Vector Machines, commonly referred to as SVM, have widely been used in classification tasks owing to their efficiency in classification in a large-dimensional space. The proposed SVM techniques assess the nature and features associated with the URL, as well as the content, to distinguish between malicious and innocent URLs. Although the proposed method is highly accurate, the accuracy may be affected by the large quantity of the data.

2. Machine Learning Algorithm ForPhishing Detection

To reduce the vulnerability caused by the conventional methods, it has been observed that there is an ever-growing requirement for the application of the machine learning algorithm in order to combat the issue of phishing. In this context, it can certainly be said that the supervised learning algorithm possesses the capability of identifying intricate patterns that are present in large datasets. Several researchers have claimed that the supervised learning algorithm is efficient for the identification of phishing pages based on the specified characteristics.

Support Vector Machines, or SVM for short, have been employed comprehensively for classification due to the effectiveness of the classification process in the large-dimensional space. The approaches presented in the paper examine the character and qualities exhibited by the URL, apart from its contents, to distinguish the innocuous and malicious URL.

The Approach: The approach presented in the paper on Support Vector Machines is extremely accurate, as mentioned in the paper, and can be affected due to the large amount of data.

3. Feature-Based Phishing Page

Feature extraction is an important aspect in evaluating the efficiency of phishing detection models. The previous literature reports the classification of features into three categories, namely URL-based features, domain-based features, and content-based features.

URL-based features include URL length, IP presence, special characters, HTTPS, and unusual domain formatting. These features can be easily extracted, and they offer a quick glimpse into whether a site looks legitimate.

Domain-based characteristics are centered on details of the domain's registration information. The characteristics are highly discriminative since phishing pages usually have a short lifetime and are mostly newly registered domains.

The content-based features examine various elements of the webpage like HTML forms, redirection actions, JavaScript use, and links to other sites. The features serve to aid in the identification of phishing sites designed to imitate legitimate sites.

4. Limitations of Existing Approaches

Although a certain degree of progress has been attained, there are some difficulties associated with current methods for phishing attack detection. For instance, some techniques involve a heavy reliance on static data, which can affect flexibility with regard to newly developed phishing techniques. High rates of false positives may influence user trust. Computations can be complex.

Moreover, some research papers concentrate exclusively on enhancing classification precision while ignoring the challenges of deployability and scalability. This not only verifies the necessity of more comprehensive research but also indicates deficiencies in earlier research on phishing attack detection.

5. Motivation for the Proposed Work

Based on the limitations extracted from various research works, it can be observed that there is an immense requirement to design an efficient phishing detection mechanism which not only offers accuracy and scalability, but can also identify zero-day attacks. In order to combat these limitations, the designed mechanism incorporates multi-level feature extraction, which is handled using various machine learning classifiers like Support Vector Machine, Random Forest, and Gradient Boosting.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

In contrast to conventional approaches, it appears that it is feasible to implement this new strategy proactively and dynamically, which would be ideal for practical applications within a modern computer security setting. This is made possible by having more than one classifier.

III. Research Methodology

The research methodology used in this study aids in developing an efficient and scalable model for phishing site identification using machine learning. The initial part of the research methodology used in this study includes gathering a large number of urls of genuine sites and phishing sites from trusted sources. Usually, in a dataset, a preprocessing part incorporates removing duplication, missing values, and uniform formats for all datasets. Once the preprocessing process is finished, an extensive feature extraction process takes place to acquire crucial features for developing a model to distinguish between genuine and phishing sites. Crucial features, which were extracted using the URL, domain, and contents of a site and were regarded as important, include URL Features such as URL length, IP Address, presence of special characters, Domain Features such as Domain Age and domain registration time, and Content Features such as presence of forms, presence of redirects, and presence of scripts.

The data then progresses to the extraction of its features in order to be used in training the machine learning models. A number of machine learning classifiers are developed using the data from the dataset. The classifiers include the Support Vector Machine algorithm, the Random Forest classifier, and the Gradient Boosting classifier. The models are chosen based on their experience in classification and ability to perform well in cybersecurity tasks. The data from the dataset is then split equally for testing purposes in order to evaluate the resultant models. The data is trained to be capable of classifying data based on the identified patterns between phishing and legitimate websites.

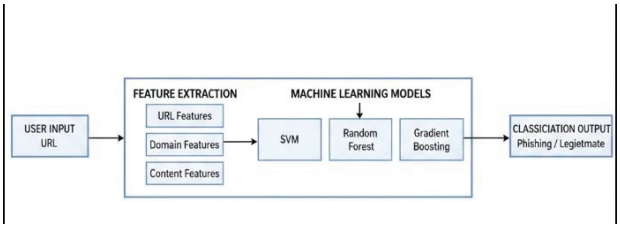

Fig. 1. Feature Extraction Process for Phishing Detection

To promote the process of generalization and prevent the model from overfitting, cross-validation methods have been implemented as part of the training process. Once the model is trained, it is incorporated into an online web detection system. Users can provide the web URLs through the system for online detection. Once the URL is submitted, the system identifies the features online and uses the trained model for the identification of the online web site as whether it is a phishing or legitimate site. The proposed model has highlighted the need for greater efficiency, scalability, and flexibility, making the model fit for purpose for online web detection and can prevent existing as well as new web-driven phishing attacks.

IV. Data Collection

The data collection procedure is an essential part of the proposed design of the phishing website detector system, and the quality and variety of data affect the performance of machine learning models. For this research purpose, a labeled data set of both phishing and genuine URL links was developed from trustworthy sources. The phishing URL links were collected from online phishing repositories and data sources that constantly update confirmed malicious links submitted to them. The genuine URL links were gathered from trustworthy sites to ensure that they have non-malicious web traffic data.

For ensuring the reliability of the dataset, it was necessary to remove any duplicate, dead, and partial URL records. Once this was accomplished, it was necessary to clean up this dataset by removing any inconsistencies. In addition to this, it was necessary to balance this dataset to ensure that there was not any imbalance towards classes. It was necessary to ensure that there was an appropriate number of records for phishing and normal URLs.

TABLE I
Dataset Description for Phishing Detection

Parameter	Description
Dataset Type	Phishing and Legitimate Website URLs
Data Source	Public phishing repositories and trusted web sources
Total Records	10,000+ URLs
Phishing URLs	5,000+
Legitimate URLs	5,000+
Data Format	URL-based textual data
Labeling	Binary (Phishing = 1, Legitimate = 0)
Feature Categories	URL-based, Domain-based, Content-based
Preprocessing Steps	Duplicate removal, noise removal, normalization
Data Split	70% Training, 30% Testing
Purpose	Train and evaluate ML-based phishing detection models

In the dataset, every URL received a binary label, where the phishing URLs received the label of malicious, while the other URLs were labeled safe. Using the label, feature extraction could also be performed, and the raw features in URLs enhanced in a way that could be applied in a machine learning algorithm. The details were very important in the construction of models in the research.

This variation in the data collected made it possible for the proposed system to train different patterns associated with phishing, and this included the newly developed patterns for phishing that have not been previously met. The procedure used in collecting data to develop the model was effective in that it made it possible to design a model that could successfully identify a genuine website and a phishing website, and this also made it possible to design a trustworthy phishing model.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

V. Data Analysis

The data analysis phase is an important aspect of assessing the efficiency of the proposed system in the detection of phishing websites. After collection and preprocessing, exploratory analysis was performed to understand the distribution and characteristics of both phishing and legitimate URLs. A balanced mixture of both phishing and legitimate classes has been used in the dataset so as to avoid model bias and ensure reliable learning. Through statistical analysis, there were considerable differences noted between the phishing and legitimate websites, especially regarding URL length, presence of IP address, domain age, and redirection. The phishing URLs were observed to be longer, with suspicious characters, in a recently registered domain, whereas legitimate websites exhibit more stability and structured patterns.

After the preprocessing phase, the features were analyzed to assess the contribution to the classification accuracy. URL-related features were found to provide a prominent capability in the context of classification, and features associated with the calculation of abnormal length, the presence of special characters, and the use of IP were prevalent in this context. Domain-related features, including the age and life span, contributed significantly to identification accuracy, as the life span of typical phishing webpages is shorter. The utilization of form and redirection features in content contributed to the identification of attempts in the context of mimicking platforms.

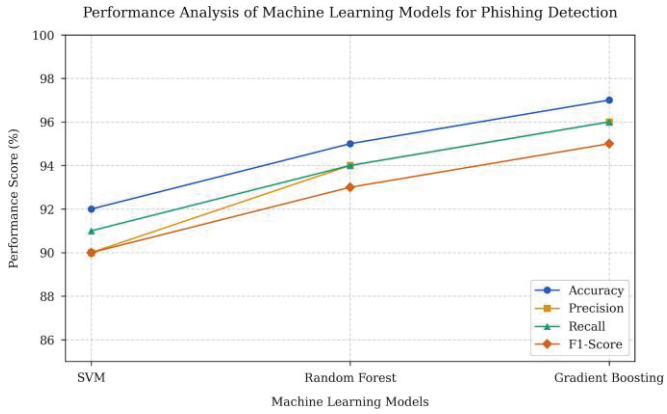

Fig. 2. Performance Comparison of Machine Learning Models

The data was split into a training subset and a test subset for objective assessment of the performance of the model. The machine learning algorithms were trained with the training data. This assessment showed that since the models used ensembling, such as Random Forest and Gradient Boosting, have been able to perform with a higher level of accuracy than individual models because they can deal with complex data, they can perform much better. The Support Vector Machine performed well too, especially in distinguishing between a phishing site and a genuine site.

The set of criteria used in the evaluation metric of accuracy, precision, recall, and false positive was applied to assess the effectiveness of the system. The system effectiveness metrics of accuracy of detection and false positive were high and of utmost importance to ensure effective implementation of the system in real life. The low false positive measures ensure that there are no legitimate websites that are mistakenly identified as phishing websites, thus removing any system ambiguity and credibility in the system among the users. The differences of result values in each of the distinct models applied to assess that the multiple features of extraction and classification by machine learning algorithms are more effective than the existing systems that rely on blacklisting.

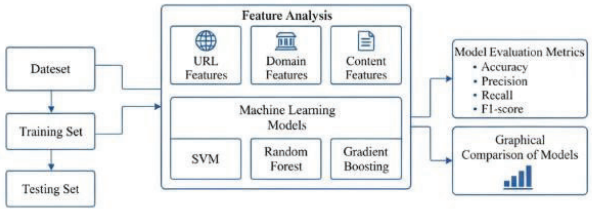

Fig. 3. Data Analysis and Model Evaluation Process

VI. Results and Discussion

The evaluation of the performance of the proposed system for phishing site detection was carried out through experimentation with various tests using the developed dataset. In this test to evaluate the performance of the developed system, the developed dataset was divided into two subsets for carrying out an unbiassed test using the developed machine learning solutions. In this evaluation test, the main emphasis was on evaluating the performance of the proposed system for phishing as well as legitimate sites.

On analyzing the result obtained from the experiment, it is found that the system developed using the machine learning algorithm is efficient in tracing phishing sites. Among the various types of classifiers, which have been used in the form of machine learning algorithm in the experiment, it is found that the Random Forest classifier and the Gradient Boosting classifier have better accuracy because they have managed to take care of the correlations among variables and have not caused any issue of overfitting. The SVM Classifier is also found to be effective, especially in identifying the easily distinguishable phishing and actual sites.

In a detailed scrutiny of the results, it has also been noticed that the utilization of features based on URLs has a substantial effect on improving the level of accuracy due to some anomalies in the URLs. This might encompass longer URLs and the use of some special characters in phishing pages. The addition of features based on age and time for registering the domain has also contributed in improving the level of phishing accuracy since the domains are always short-lived. The content features encompassed ”forms” and ”redirects.”

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

These values confirm the validity of the result with high accuracy for system detection, as well as a low false alarm value, of our system. In fact, for real-life implementations, a site with a low false positive value is significant because incorrect identification in terms of a trusted site would influence the level of trust adversely. The experimental results confirm values that have assured a balance between security and usability by our model.

Comparison of the classifiers shows that the ensemble learning approaches are superior to other models because of their capacity to combine several weak learners to improve generalization. The Gradient Boosting technique worked well on all the test examples, which is an advantage when implementing the model on a grand scale to protect against phishing. Although not the most precise approach, the Support Vector Machine offered fairly stable results, especially when dealing with clearly distinguishable data points.

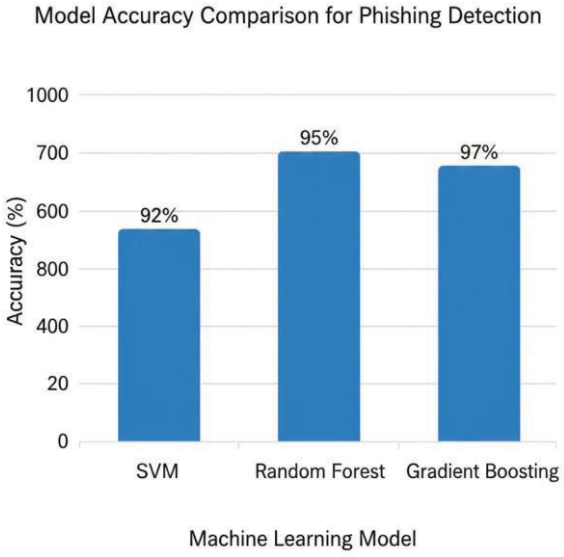

Fig. 4. Performance Comparison of Machine Learning Models

In general, the experimental outcomes affirm the efficiency of the developed phishing detection framework. The proposed framework effectively identifies both detected and unseen phishing pages, which highlights its excellent generalization ability. The proposed multi-level feature extraction combined with machine learning classification has greatly improved the accuracy of phishing detection. The existing white-listing-based phishing detection systems are outperformed by the developed phishing detection framework. The proposed solution is also fit for implementation in real-world applications for improving phishing defenses.

VII. Conclusion and Future Scope

The present paper had presented a machine learning-based phishing website identification based on URL, domain-based, and content-related features. The proposed system is designed to offer proactive, intelligent solutions that can easily outsmart conventional phishing website identification systems based on blacklisting techniques. The implemented system showed effective classification with zero false positives through the extensive feature identification techniques and the application of Support Vector Machine Classifier, Random Forest Classifier, and Gradient Boosting Classifier algorithms. The result proves that including multi-feature categories helps in enhancing the reliability of phishing website identification systems.

This research work has proved the possibility of learning models built for ensembles regarding patterns of phishing attacks treatment. Additionally, the findings have confirmed the necessity of using an equally well-balanced data source for optimal results. Solutions built have been able to address the most difficult tasks for phishing attacks detection, including scalability, applicability for appropriate solutions for real-time tasks. Also, the proposed system has been able to effectively identify malicious sites, which will decrease economic and sensitive data loss of the users.

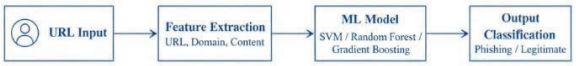

Fig. 5. Overall Flow of Phishing Detection System

Although the efficiency of the proposed system is beyond doubt, there are still a variety of aspects that can be further improved in terms of the system. The future research may concern the integration of real-time threat intelligence data; the use of Deep Learning approaches such as Convolutional Neural Networks or Recurrent Neural Networks is able to increase the detection precision even when dealing with complex phishing schemes. Moreover, increasing the dataset by using region-specific and multinational phishing examples can improve the generalization ability of the model. Future studies may investigate developing hybrid models which integrate machine learning approaches together with behavioral models to further reduce the false positives and increase the detection of zero-day phishing attacks.

Acknowledgment

Authors are grateful to acknowledge the efforts put up by all those people who helped in the successful completion of this research work. We take this opportunity to thank our project guide and our faculty members for their constant support, significant suggestions, and encouragement during the execution of this project. Their expertise in our area of research, as well as their support in our academic endeavors, has played an important part in finalizing our research work. We would also want to thank the institution for providing such infrastructure and an environment that has enabled us to carry out this research. On a personal note, we would like to thank the team and contributors involved in sharing the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

available free phishing data and research work; without them, conducting this experiment would not have been possible.

Last but not least, we would like to thank our fellow participants for their support and cooperation in the form of constructive suggestions and motivations throughout the execution of this project. All of us were encouraged by each of these participants.

REFERENCES

[1] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)," *IEEE Transactions on Dependable and Secure Computing*, vol. 3, no. 4, pp. 301–311, 2006.

[2] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013.

[3] R. Verma and A. Das, "What works and what does not: A study of phishing attack detection," in *Proc. IEEE Conference on Cyber Security*, 2017, pp. 1–6.

[4] G. Canfora, F. Mercaldo, C. A. Visaggio, and M. Di Penta, "A classifier based on URL features for phishing detection," in *Proc. International Conference on Availability, Reliability and Security*, 2014, pp. 1–8.

[5] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *Proc. ACM SIGKDD*, 2009, pp. 1245–1254.

[6] A. Jain and B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 5, pp. 2015–2028, 2019.

[7] M. Aburrous, M. A. Hossain, F. Thabtah, and K. Dahal, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7913–7921, 2010.

[8] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, no. 2, pp. 443–458, 2014.

[9] S. Marchal, J. Francois, R. State, and T. Engel, "Proactive discovery of phishing related domain names," in *Proc. International Symposium on Research in Attacks, Intrusions, and Defenses*, 2014, pp. 190–209.

[10] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *Proc. NDSS*, 2010.

[11] H. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.

[12] Y. Zhang, J. Hong, and L. Cranor, "Cantina: A content-based approach to detecting phishing web sites," in *Proc. WWW Conference*, 2007, pp. 639–648.

[13] A. K. Jain and B. Gupta, "Comparative analysis of features based machine learning approaches for phishing detection," in *Proc. International Conference on Computing, Communication and Automation*, 2016.

[14] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proc. ACM Workshop on Recurring Malcode*, 2007.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[16] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection using decision trees and association rules," *Human-centric Computing and Information Sciences*, vol. 4, no. 1, pp. 1–15, 2014.

[17] M. L. Kumar and A. Saravanan, "A study on phishing attacks and detection techniques," *International Journal of Computer Applications*, vol. 139, no. 1, pp. 20–26, 2016.

[18] S. Rao and A. Pais, "Detection of phishing websites using machine learning," *International Journal of Engineering Research and Technology*, vol. 6, no. 5, pp. 1–6, 2017.

[19] UCI Machine Learning Repository, "Phishing Websites Dataset." [Online]. vailable: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites

[20] Anti-Phishing Working Group (APWG), "Phishing Activity Trends Report," 2023. [Online]. Available: https://apwg.org

(This work is licensed under a Creative Commons Attribution 4.0 International License.)