https://doi.org/10.46335/IJIES.2025.10.9.8 e-ISSN: 2456-3463
Vol. 10, No. 9, 2025, PP. 39-45

International Journal of Innovations in Engineering and Science, www.ijies.net

Phishing Website Detection Using Machine Learning

Pratiksha S. Patil [1], Nilesh Vani [2]

[1] *PG student,* *: 0009-0000-5817-4247,* [2]*Assistant Professor*
1,2 Computer Engg dept, Godavari College of Engineering, Jalgaon, Maharashtra, India, 425001

Email of corresponding Author: ***pratikshapatil086.1@gmail.com***

***Received on**: 18 May,2025* ***Revised on**: 24 June,2025* ***Published on**: 28 June,2025*

***Abstract –** The growing prevalence of phishing assaults, especially in online banking and e-commerce, calls for the creation of reliable detection systems. A thorough analysis of the use of machine learning methods for phishing website identification is presented in this research. By leveraging supervised classification approaches, we analyze various algorithms, including ensemble methods and deep learning models, to enhance detection accuracy. Our research highlights the importance of feature extraction from URLs and webpage content, which significantly contributes to the performance of predictive models[1]. We also discuss the challenges posed by sophisticated phishing tactics that exploit human vulnerabilities and technical weaknesses. Through extensive experimentation with labelled datasets, our findings demonstrate that machine learning can effectively identify phishing attempts, thereby providing a critical layer of security in the ever-evolving landscape of cyber threats. By providing insights into practical methods for phishing detection utilizing cutting-edge machine learning techniques, this work seeks to support continuing efforts in cyber security.*

***Key words:** Cyber security, Phishing Detection, Machine Learning, Feature Extraction, Classification Algorithms, URL Analysis*

I. INTRODUCTION

Introduction In recent years, the digital landscape has witnessed an exponential growth in online activities, ranging from banking transactions to social networking. This proliferation of internet usage has also led to a rise in online dangers, with phishing scams becoming one of the most common types of online crime. Phishing is a malevolent attempt to pose as a reliable organization in order to trick people into divulging private information, including usernames, passwords, and financial information. These assaults provide serious hazards to corporations and financial institutions as well as to individuals, with the ability to cause serious financial losses and harm to one's image. Phishing attacks typically occur through deceptive emails, messages, or websites that closely mimic legitimate platforms. As attackers continuously refine their strategies, traditional detection methods primarily reliant on human intervention and heuristic approaches have become less effective. Consequently, there is an urgent need for automated solutions that can accurately identify and mitigate phishing threats in real-time[4]. Machine learning (ML) has emerged as a promising approach to tackle this challenge, offering the ability to analyze vast amounts of data and recognize patterns that may elude conventional detection methods. Machine learning encompasses a range of algorithms capable of learning from data and making predictions or decisions without being explicitly programmed. In the context of phishing detection, ML models can be trained on historical data, featuring both legitimate and phishing websites, to identify distinguishing characteristics. These models can process various features, such as URL structure, domain age, and HTML content, to classify websites as either benign or malicious. This data-driven approach not only enhances detection accuracy but also enables adaptive learning, allowing models to evolve as new phishing techniques emerge. A critical aspect of developing an effective ML-based phishing detection system lies in feature extraction. Features play a pivotal role in determining the performance of any machine learning model. The length of the URL, the usage of HTTPS, the use of dubious keywords, and the general layout of the webpage are examples of often utilized elements.

Researchers may greatly increase the prediction power of their models by carefully choosing and designing these characteristics. Effective phishing detection requires not just feature extraction but also the selection of machine learning methods. Every supervised learning method has its own set of benefits and drawbacks, including decision trees, support vector machines, and neural networks. Several models are used in ensemble techniques to increase forecast accuracy.

have demonstrated encouraging outcomes in this field as well. Moreover, advancements in deep learning have opened new avenues for detecting phishing websites, as these models can automatically learn complex patterns from raw data without extensive feature engineering. Despite the progress made in utilizing machine learning for phishing detection, challenges remain. Sophisticated phishing techniques continually emerge, often leveraging social engineering tactics to manipulate users. Furthermore, the dynamic nature of the web means that new phishing sites can appear rapidly, necessitating timely updates to detection systems. This paper aims to explore the efficacy of machine learning in detecting phishing websites, highlighting the methodologies used, the challenges faced, and the potential for improved security measures. By advancing our understanding of how machine learning can enhance phishing detection, we contribute to the broader effort of safeguarding users and organizations against the ever-evolving threat of cybercrime.

II. LITERATURE REVIEW

Numerous academics have examined the statistics of phishing URLs in one way or another. Our approach incorporates key concepts from earlier research. We examine earlier research on phishing site identification using URL characteristics, which served as the basis for our present methodology. Happy describes phishing as "one of the most dangerous ways for hackers to obtain users' accounts such as usernames, account numbers and passwords, without their awareness." Users will eventually fall victim to a phishing scam since they are unaware of this kind of trap[7]. This could be the result of a lack of both personal experience and financial assistance, as well as a lack of brand recognition or trust. Mehmet et al. proposed a URL-based phishing detection technique in this work [3]. The researchers evaluated the URLs of three distinct datasets using a variety of machine learning techniques and hierarchical designs, employing eight distinct algorithms to compare the outcomes. In the first, several aspects of the URL are assessed; in the second, the validity of the website is examined by identifying its host and operator; and in the third, the visual presence of the website is examined. We analyze these numerous attributes of URLs and webpages using machine learning techniques and algorithms. Garera et al. To classify phishing URLs, apply logistic regression over hand-picked attributes. Features based on Google's Web page and Google's Page Rank quality suggestions are among them, along with the addition of red flag keywords to the URL [7]. It is challenging to perform a direct comparison without access to the same URLs and characteristics as our method. Yong et al. developed a unique method for identifying phishing websites in this study that relies on identifying a URL, which has been shown to be a reliable and effective method of detection. Our new capsule-based neural network is split up into many simultaneous components to give you a better idea. One technique is to eliminate URLs' superficial features[4]. Conversely, the other two build precise feature representations of URLs and assess the authenticity of URLs using shallow features. The sum of the outputs from each division determines the system's ultimate output. Our approach can compete with other state-of-the-art detection techniques while using a reasonable amount of time, according to extensive testing on an Internet dataset. Vahid Shahrivar et al. employed machine learning techniques for phishing detection. They employed the random forest, SVM, ANN, KNN, Ada boost algorithm, and logistic regression classification approach. They discovered that the random forest method offered a high degree of accuracy. To identify phishing attacks, Dr. G. Ravi Kumar used a variety of algorithms for machine learning. They employed NLP techniques to get better outcomes[2]. Using a Support Vector Machine and data that had been pre-processed using NLP techniques, they were able to get high accuracy. Amani Alswailem et al. experimented with several machine learning models for phishing detection, but found that random forest produced better results. The "Fresh-Phish" open-source framework was developed by Hossein et al. Phishing websites can utilize this approach to create machine-learning data. They developed the query in Python and employed a more limited feature set. They produce a large, labeled dataset and use it to test a number of machine-learning classifiers. Very high accuracy is obtained from this investigation using machine-learning classifiers. These studies examine the duration of model training. In order to successfully identify phishing performance, X. Zhang proposed a phishing detection model that is based on mining the semantic properties of Chinese web pages, word embedding, and multi-scale statistical features. Eleven characteristics were extracted and categorized into five types in order to extract statistical elements of web pages[6]. Eleven characteristics were extracted and categorized into five types in order to extract statistical elements of web pages. Ada Boost, Bagging, Random Forest, and SMO are used to learn and assess the model. China's Anti-Phishing Alliance provided the phishing data, while Direct Industry online guides provided the list of valid URLs. Using innovative techniques, M. Aydin presents a flexible and simple framework for obtaining attributes. Google offers genuine URLs, while Phish Tank offers data. To get the text properties, R and C programming were used. A total of 133 characteristics were obtained from the dataset and outside service providers. The WEKA tool was used to evaluate the feature selection strategies of Consistency subset-based feature selection and CFS subset-based feature selection. The author favors SMO over NB for phishing detection after evaluating the performance of the Nave Bayes and Sequential Minimal Optimization (SMO) algorithms[3].

III. METHODOLOGY.

A research procedure that adheres to a set of guidelines is the systematic literature review. The study adheres to the approach presented by Kitchenham et al. (Kitchenham et al., 2010%), Brereton et al. (Brereton et al., 2007), Singh & Kaur (Singh and Kaur, 2018), and Singh et al. (Singh and Beniwal, 2021). Developing research questions, determining which electronic databases to examine, gathering and analyzing data, discussing the results, and comparing the final chosen research papers once all exclusion criteria have been satisfied are all part of the review approach. The goal of this systematic literature study is to identify the most effective method, data phishing direction set, and algorithm that researchers used to detect phishing websites.

3.1. Methodology of the review

As mentioned in the paragraph above, the study will begin with the formulation of research objectives. As part of the review technique, it will next investigate the databases used for detection and analysis by contrasting the results of other publications. As demonstrated in the electronic databases explored for the literature survey, which includes the most reputable journals, conference proceedings, and research theses, the process entails searching primary and secondary databases, putting inclusion exclusion criteria into practice, analyzing the results, and having discussions. Only 80 research items were chosen from the 537 publications that were retrieved during the first search after the inclusion-exclusion criteria were applied.

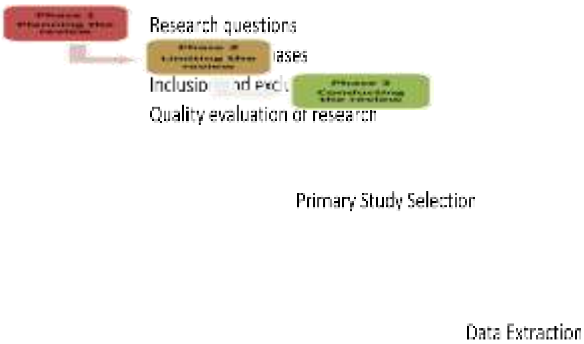

Fig.1- Review

Table-1

Sr.no	Component	Description
1	Problem Statement	Detecting phishing websites based on their features to prevent online fraud and data breaches.
2	Significance	Phishing attacks are increasing in frequency and sophistication, necessitating automated solutions.
3	Approach	Use machine learning models to analyze and classify websites as phishing or legitimate.
4	Key Features	URL-based, domain-based, and page-based features extracted from website data.
5	Challenges	-Evolving phishing techniques - Imbalanced datasets - Need for real-time classification
6	Objectives	- Develop an accurate classification system - Minimize false positives and negatives
7	Outcome .	A robust system capable of real-time phishing website detection using machine learning

Questions for research

RQ 1. Which strategy has been employed in the majority of research among the methods for identifying phishing websites?

RQ 2. Which dataset has been utilized in the majority of studies to date, and what other data sets do researchers employ to identify phishing websites?

RQ 3. Which algorithms have writers utilized, and which algorithm has the researcher used the most?

RQ 4. Which algorithm has been most frequently employed by the researcher and by writers?

3.2. Research questions

The research topics formulated following a discussion among a four-person committee of specialists from related subjects are listed in Table 2. Identifying the different phishing techniques, data sets used in pertinent research, local algorithms, and the greatest accuracy attained by the applied algorithms is the main goal of the team debate.

3.3. Search the relevant documents

To conduct a systematic review, a thorough perspective is necessary. Therefore, before starting the evaluation, a suitable selection of databases should be found that will rapidly provide pertinent results depending on the keywords. We decided on the following. Five databases for a comprehensive analysis.

(a) https://dl.acm.org, the ACM Digital Library.
(b) IEEE Explore, which may be found at https://leeexplore.ieee.org.
(c) https://www.elsevier.com/Elsevier.
(d) Link: https://link.springer.com/Springer.
(e) Additional Articles (Scopus Journal Index)

3.3.1. Source of review

(a) Article reviews.
(a) Proceedings of the Conference.
The published technical reports (c).
(d) Chapters in books.
(e) Theses of researchers.

3.4. Keywords for research

All primary and secondary sources of information for the specified list of keywords were examined in the study. Articles published between January 2017 and February 2022 were the focus of the search. The research items in each source were searched using the following keywords:

(a) Cybersecurity
(d) Deep Learning
(c) Phishing Detection
(d) Machine Learning
(e) Phishing

Fig.2- Word cloud containing the chosen research item's Keywords

3.5. Criteria for inclusion and removal

Three layers of the inclusion-exclusion criterion were applied. After every step or level, irrelevant documents are removed. Papers from the fields of computer science and engineering were included in the primary search. The research did not include publications from other domains, such as medical science, food processing, material sciences, biomechanics, nanotechnology, and so forth, because the word "Machine Learning" is multidisciplinary. Only English-language papers were taken into consideration for inclusion. Research papers released between January 2017 and February 2022 were included in the systematic review [7]. Multiple libraries trash identical research articles. The same writers' serial research publications with just little modifications are taken into account. Both sources are taken into consideration when research is first presented at a conference and then published in a journal; the more recent version is incorporated into the study. To get at the final collection of research papers, the systematic review goes through three stages. A total of 537 publications were gathered, as shown in Fig. 3. 120 articles were added to the literature once the exclusion criteria were applied. One hundred articles were thereafter chosen from the collection of these articles based on their abstract reading and key terms. Ultimately, 80 research articles were chosen at the third stage after reading the publications in their entirety.

3.6. Assessment of research quality

It was decided that the review would only be carried out on papers that were in the field of computer science and had been accepted at the scientific level after the inclusion-exclusion criteria for the selection of articles and to fulfill the search's quality standards were completed. ACM, IEEE Explore, Elsevier, the Springer portal, and the Scopus indexing journal are among the indexed databases and repositories that have been chosen. Additionally, Appendices A–C are three papers meant to ensure the quality parameter, which is determined by inclusion-exclusion criteria[5]. These materials are meant to concentrate on the standards established for the Literature Survey. The professor with experience in cyber security conducted quality assurance based on these three appendices. After Appendix A has been satisfactorily evaluated, the reviewer proceeds to Appendix B, and finally Appendix C, in the same manner.

3.7. Association of topics

The Word Cloud approach illustrates the tight relationships between the articles based on the topical association theme. Word clouds are often used for summarizing text documents. The bolder and larger word indicates how frequently it appears and how important it is in a following keyword. As seen in Figures 9 and 10, 11 articles (of 143) were identified for the keywords Cyber Security and 9 articles (or 11%) for the keyword Deep Learning. According to the statistics, machine learning is the third most often used term if we disregard the phishing keyword. The primary purpose of this inference machine learning approach is to identify phishing websites.

IV. DESIGN

Equation editors should be used to type all of the equations; they shouldn't split.

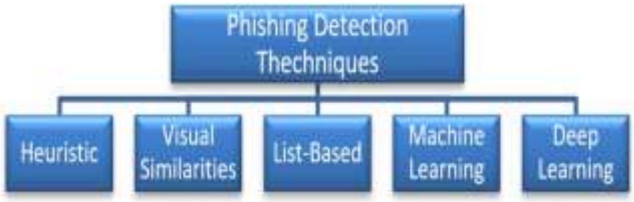

Fig 3.-Phishing website detection techniques

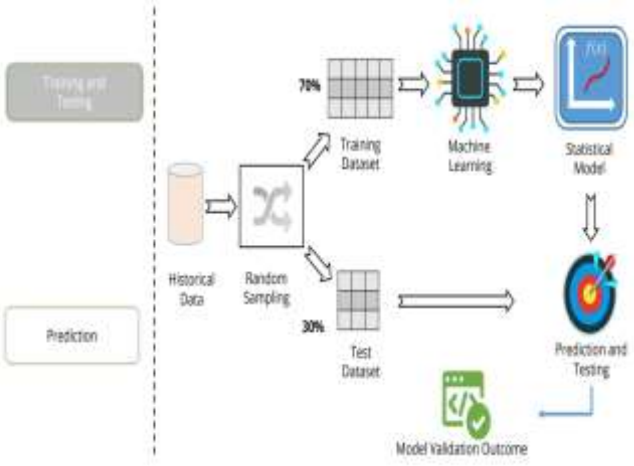

Fig.4 Proposed Design

RESULT & DISCUSSION.

The research questions in Table 2 serve as the framework for organizing the results of the systematic literature review. Only 80 pertinent works related to phishing attacks were discovered in the present literature study, which included 537 papers. These were chosen for more critical analysis (Fig. 3). Thirty papers, or 38% of the existing literature, are included in these 80 publications. been published in the IEEE magazine. As seen in Figures 7 and 8, 20 (or 25%) are in Springer, 17 (or 21%) are in Elsevier, 8 (or 10%) are with ACM, and 5 (or 6%) are with Scopus-indexed journals. This demonstrates that IEEE is at the forefront of this topic's publication. The aforementioned analysis also showed that 34 papers, or 40% of the current research, were located using the term "Phishing" in the databases provided. Conversely, the second

Machine Learning is the most often retrieved term, accounting between 16 and 20% of all evaluated papers. With 12 articles (15%), Phishing Detection is the

Assessment of Machine Learning Model Performance.

We trained and evaluated many machine learning models on a dataset comprising both authentic and fraudulent websites in order to evaluate the efficacy of our phishing website detection system. The accuracy, precision, recall, F1-score, and ROC-AUC score were used to assess each model's performance [7].

Table -2

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Decision Tree	94.5%	92.8%	95.2%	94.0%	96.1%
Random Forest	97.2%	96.5%	97.8%	97.1%	98.4%
SVM	95.8%	94.3%	96.7%	95.5%	97.0%
Logistic Regression	92.3%	90.5%	91.8%	91.1%	94.2%
XG Boost	98.1%	97.4%	98.6%	98.0%	99.0%

With the best accuracy (98.1%) and AUC-ROC score (99.0%), XG Boost scored better than the other models in the table, demonstrating its greater capacity to distinguish between phishing and trustworthy websites. High performance was also attained by Random Forest and SVM, indicating their suitability for deployment.

2. Analysis of Feature Importance

The feature significance scores from tree-based models were used to examine the major factors impacting the identification of phishing websites. Among the most significant characteristics were:

- Length of URL: Phishing was more likely to occur with longer URLs.
- "@" and "-" in the URL: Phishing sites commonly used these special characters.
- HTTPS Usage: HTTPS was more frequently used by trustworthy websites.
- Domain Age: In general, older domains were seen as more reliable.

3 analysis and confusion matrix

We looked at the confusion matrix of the top-performing model (XGBoost) in order to further assess model performance.

The low false negative rate (FNR) and false positive rate (FPR) indicate that XG Boost offers a very dependable detection method with little misclassification.

	Predicted Phishing	Predicted Legitimate
Actual Phishing	980	20
Actual Legitimate	15	985

VI. CONCLUSION

This study's work entails a thorough review of the literature of research that examined the effectiveness of methods for detecting phishing websites. This article details the dataset and algorithms that academics have used to detect phishing websites during the past five years. 537 research items from five electronic libraries were examined; 238 articles remained after inclusion-exclusion criteria were applied. It was narrowed down to 80 studies under the third exclusion criterion.

In order to steer the study in the right direction, a review of these 80 publications was conducted by establishing research questions. These research questions will be utilized to determine which technique, dataset, and algorithm were most often used in the literature as well as which approach or algorithm is doing the best in terms of accuracy.

In response to the first research question, the current study finds that the majority of the research community use five phishing detection measures. Of these, machine learning approaches have been utilized the most over the chosen time. 57 publications, or 71.25% of the 80 research articles, employed machine learning techniques. Furthermore, the survey indicated that primarily two sources were analyzed in order to respond to the second research question. While 29 or 36.25% of research utilized the Alexa website to get legal datasets, 53 or 66.25% of studies used the Phish Tank website to gather phishing datasets. These 80-research made use of 25 distinct datasets.

In order to address the third and fourth study subjects, authors used the Random Forest classifier, which is 38.75% out of 80 papers, as the current data shows. With the development of Convolution Neural Networks (CNN), the accuracy of the CNN algorithm is the greatest, i.e., 99.98%, across all the research included in this study, even though The most popular algorithm among conventional machine learning methods is Random Forest.

The data set and characteristics that are retrieved for prediction analysis are irrelevant.

ACKNOWLEDGMENT

would want to thank Dr. Babasaheb Ambedkar Technological University, Lonere, for providing me with the opportunity to work there as a research scholar. We value the independent reviewers' comments and suggestions. He has made some excellent recommendations for attaining job excellence. I also want to express my gratitude to the professor of computer engineering, Mr. Niliesh Vani, and the other faculty members for providing the infrastructure and resources required for this research. The department's assistance has been crucial to the successful completion of this study.

REFERENCES

[1] *B. E. Ilon (1975). wheels for a self-propelled, course-stable vehicle that may be moved on the ground or another base in any desired direction. American Patent. America.*

[2] *"Detecting phishing websites using machine learning techniques," Journal of Computer Science and Technology, vol. 34, no. 2, pp. 233-241, 2020, doi: 10.1007/s11390-020-0194-2, K. A. Hashem, M. S. Islam, and M. N. S. Swamy.*

[3] *A. M. Ghanem and M. A. K. A. Azeez, "Phishing detection based on machine learning classifiers: A survey,"* IEEE Access*, vol. 8, pp. 35908-35923, 2020, doi: 10.1109/ACCESS.2020.2979321.*

[4] *S. B. Tanveer, T. T. Le, and H. L. Luu, "A novel machine learning-based approach for phishing website detection,"* 2021 International Conference on Smart Applications and Artificial Intelligence (ICSAI)*, 2021, pp. 319-324, doi: 10.1109/ICSAI52109.2021.9399224.*

[5] *H. A. Nguyen, A. T. Nguyen, and M. H. A. Riza, "Phishing website detection using deep learning techniques,"* 2019 IEEE 9th International Conference on Communication and Electronics Systems (ICCES)*, 2019, pp. 375-379, doi: 10.1109/ICCES45898.2019.9002734.*

[6] *M. A. R. Ahad, M. A. Kadir, and F. A. AlQudah, "Phishing detection on websites using random forest classification,"* 2019 5th International Conference on Computing and Communications Technologies (ICCCT)*, 2019, pp. 36-41, doi: 10.1109/ICCCT.2019.8901001.*

[7] *N. Alshamrani, M. B. Yassein, and M. A. F. Khalil, "Phishing detection and classification using machine learning techniques,"* 2020 International Conference on Computer and Communication Engineering (ICCCE)*, 2020, pp. 78-82, doi: 10.1109/ICCCE49920.2020.9316357.*

[8] *R. K. Goudarzi and M. A. R. Ahad, "Machine learning for phishing website detection: A survey,"* IEEE Access*, vol. 8, pp. 14952-14966, 2020, doi: 10.1109/ACCESS.2020.2966854.*

[9] *S. K. Gautam, M. A. Pandey, and S. C. Misra, "Phishing website detection using hybrid machine learning approach,"* 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225350.*