International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

PHISHING DETECTION SYSTEM USING ENSEMBLED METHODS

Prof. Indira Joshi, Atharv Sawant , Sarth Malankar, Shashank Mane

Computer Engineering
New Horizon Institute of Technology and Management ,Thane

Abstract

This study proposes a robust content- grounded system exercising machine learning models for detecting phishing websites and distinguishing them from safe bones using Support Vector Machine, Gaussian Naive Bayes, Decision Tree, Random Forest, Adaboost, Nearest Neighbor, and Neural Network algorithms, the system aims to directly identify fraudulent websites. By assaying features similar as URL, HTML content, and runner structure, the models are trained to fete patterns reflective of phishing attempts. Through expansive trial and evaluation, this exploration demonstrates the effectiveness of employing different machine literacy ways in combating cyber pitfalls. The proposed approach offers a promising result for enhancing online security measures, furnishing druggies and associations with a dependable tool to identify and alleviate the pitfalls associated with vicious websites, therefore securing sensitive information and conserving digital trust.

Important keywords for research

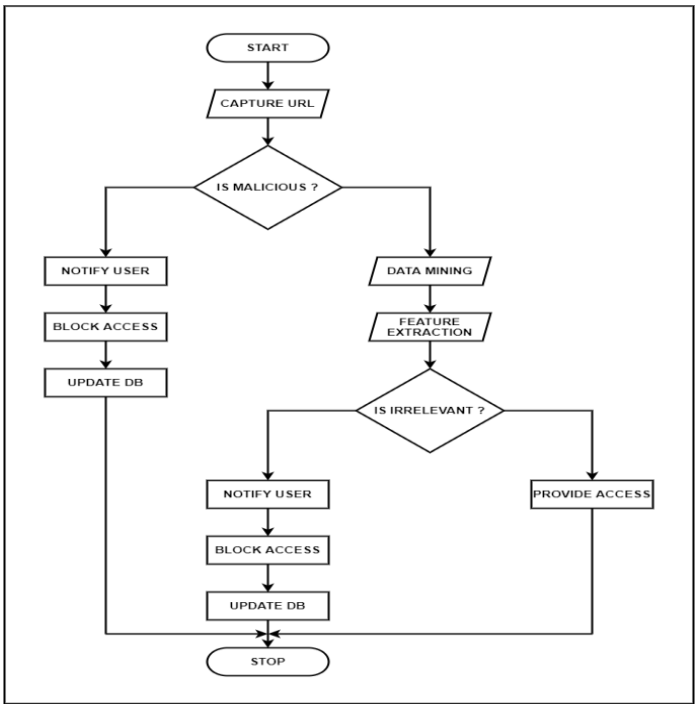
The study explored all primary and additional sources of information for the given list of keywords. The search was conducted for papers published from January 2017 to February 2022. The following keywords were used to search the research items in each source:

- Phishing
- Phishing Detection
- Deep Learning
- Cyber Security
- Machine Learning

INTRODUCTION

In an era where the internet serves as a ubiquitous platform for communication, commerce, and information exchange, the proliferation of phishing websites poses a significant threat to users' security and privacy. Phishing, a form of cybercrime, involves deceptive tactics to trick individuals into divulging sensitive information such as passwords, credit card numbers, or personal identification. To combat this menace, researchers and cybersecurity experts have turned to innovative approaches, among which content-based methods employing machine learning models have emerged as a promising avenue.
This paper explores the application of content-based methods leveraging various machine learning algorithms to detect phishing websites effectively. By analyzing the content and structural features of web pages, these methods aim to discern patterns indicative of phishing attempts, thereby enabling proactive identification and mitigation of potential threats.
The study investigates a range of machine learning models, including Support Vector Machine (SVM), Gaussian Naive Bayes, Decision Tree, Random Forest, Adaboost, Nearest Neighbor, and Neural Network. Each model offers distinct advantages in terms of its ability to classify websites based on their content features, such as HTML attributes, domain information, and textual content.
Support Vector Machine (SVM), known for its capability to handle high-dimensional data and nonlinear relationships, serves as a robust classifier in discerning subtle distinctions between phishing and legitimate websites. Gaussian Naive Bayes, on the other hand, relies on probabilistic principles to classify websites, making it suitable for rapid processing and scalability.
Decision Tree and Random Forest algorithms excel in capturing complex decision boundaries and handling large datasets, making them invaluable tools for detecting phishing websites amidst vast amounts of online content. Adaboost enhances classification accuracy by iteratively training weak learners on subsets of data, while Nearest Neighbor algorithms leverage similarity metrics to identify patterns resembling known phishing schemes.
Moreover, the study explores the application of Neural Networks, particularly deep learning architectures, in extracting intricate features from web content and making nuanced predictions regarding website authenticity.

By leveraging these diverse machine learning techniques, this research aims to enhance the efficacy of phishing website detection, thereby fortifying users' digital safety and fostering trust in online interactions. Through rigorous experimentation and evaluation, insights gained from this study hold the potential to inform the development of more resilient cybersecurity solutions in the ever-evolving landscape of cyber threats.



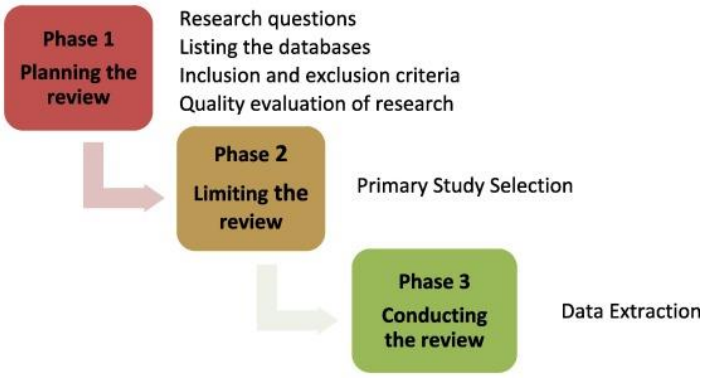
Background of related work

Many authors have explored the detection of phishing websites. However, only a few have conducted a systematic literature review on the topic, as described below.

Qabajeh et al. (Qabajeh et al., 2018) recently worked on conventional vs automated phishing detection techniques. The conventional anti-phishing methods include raising awareness, educating users, conducting periodic training or workshop, and using a legal perspective. The Computerized or automated anti-phishing approaches talks about list-based and Machine Learning Based techniques. More importantly, the paper compares these approaches' similarities, positive and negative elements from the user and performance perspectives. According to this study, Machine Learning and rule induction are suitable for combating phishing attacks. The limitations of this work are: the review is based on 67 research items, and the study does not include Deep Learning techniques for phishing website detection. Prior research in the field of phishing detection has explored various methodologies and techniques aimed at mitigating the risks associated with phishing attacks. Traditional approaches typically rely on rule-based heuristics, blacklists, or machine learning algorithms to identify phishing attempts. While these methods have shown some effectiveness, they often exhibit limitations in terms of accuracy, scalability, and adaptability to evolving threats. Ensemble methods, which involve combining multiple classifiers to improve performance, have gained traction in recent years as a promising avenue for enhancing phishing detection systems. The effectiveness of ensemble methods lies in their ability to leverage diverse base classifiers, each capturing different aspects of the data, and aggregating their predictions to make more informed decisions. Several studies have investigated the application of ensemble methods in phishing detection with promising results. Bagging, Boosting, Random Forest, and Gradient Boosting are among the popular ensemble techniques employed in these systems. Bagging, for instance, involves training multiple classifiers on bootstrap samples of the dataset and combining their predictions through averaging or voting. Boosting algorithms, such as AdaBoost, iteratively train weak learners by emphasizing misclassified instances, thereby constructing a strong classifier. Random Forest combines multiple decision trees trained on random subsets of features and instances, providing robustness against overfitting and enhancing generalization. Researchers have also explored feature engineering and selection strategies to extract informative signals from diverse sources of data, including URLs, email headers, textual content, and behavioral patterns. Feature extraction plays a crucial role in phishing detection systems, as it enables the ensemble classifier to discern subtle indicators of phishing while minimizing false positives. Moreover, advancements in model evaluation techniques, such as cross-validation, ROC analysis, and precision-recall curves, have facilitated the rigorous assessment of ensemble-based phishing detection systems. By benchmarking against baseline methods and evaluating performance metrics across diverse datasets, researchers have demonstrated the efficacy and scalability of ensemble approaches in real-world settings. While existing literature has made significant strides in leveraging ensemble methods for phishing detection, there remain opportunities for further exploration and improvement. Future research directions may include the integration of deep learning architectures, the incorporation of contextual information, and the development of ensemble models capable of detecting zero-day phishing attacks with high accuracy and efficiency. In summary, the body of related work underscores the potential of ensemble methods as a robust and adaptive framework for phishing detection, offering insights and methodologies that can inform the design and implementation of advanced phishing detection systems.

Methodology of the review

As discussed in the above para study will start by designing research questions and then explore the databases used for detection and analysis by comparing the findings of other literature as a part of the review methodology. The procedure includes searching primary and secondary databases, implementing inclusion–exclusion criteria, analyzing results, and discussions are all part of the process, as shown in Fig. 2. Only electronic databases are explored for the literature survey, which includes the most reputable journals, conference proceedings, and research thesis. During the initial search, 537 papers were found, and only 80 research items were selected after applying the inclusion–exclusion criteria (see Fig. 3).



1.Literature Search: The review begins with a comprehensive search of academic databases, journals, conference proceedings, and relevant literature repositories. Keywords such as "phishing detection," "ensemble methods," "machine learning," and variations thereof are used to identify relevant studies published in peer-reviewed journals and conference proceedings. Both recent publications and seminal works are considered to provide a comprehensive overview of the topic

2. Inclusion and Exclusion Criteria: A set of inclusion and exclusion criteria are established to ensure the relevance and quality of the selected literature. Inclusion criteria may include studies focusing on phishing detection using ensemble methods, empirical evaluations of ensemble-based approaches, and publications written in English. Exclusion criteria may encompass studies unrelated to phishing detection, non-peer-reviewed sources, and duplicates.

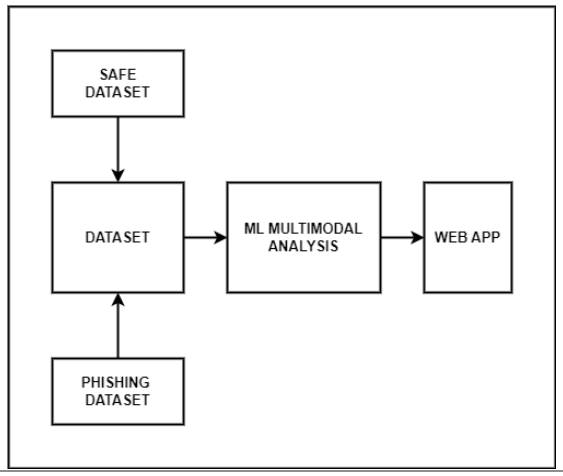
3.Screening and Selection: Each identified publication is screened based on the established criteria. Titles and abstracts are initially reviewed to assess their relevance to the topic of phishing detection using ensemble methods. Subsequently, full-text articles are obtained for further evaluation. The screening process is conducted independently by multiple reviewers to minimize bias, and discrepancies are resolved through consensus.

4.Data Extraction: Relevant information from the selected studies is extracted systematically. This includes details on the ensemble methods employed (e.g., Bagging, Boosting, Random Forest), feature extraction techniques, datasets used for evaluation, performance metrics, and key findings. Data extraction is performed using a predefined template or data extraction form to ensure consistency across studies.

5.Synthesis and Analysis: The extracted data are synthesized and analyzed to identify common themes, trends, strengths, and limitations across the selected studies. Comparative analysis may be conducted to evaluate the performance of different ensemble methods, feature sets, and evaluation metrics. Additionally, qualitative insights into the effectiveness, scalability, and practical implications of ensemble-based phishing detection systems are discussed.

6. Quality Assessment: The quality of the selected studies is assessed using established criteria, such as methodological rigor, clarity of presentation, and relevance to the research topic. Studies are rated based on their adherence to sound research principles and the validity of their findings. High-quality studies are given greater weight in the synthesis and analysis process.

7. Discussion and Conclusion: The review concludes with a comprehensive discussion of the findings, implications, and future research directions. Key insights derived from the synthesized literature are highlighted, and recommendations for researchers, practitioners, and policymakers are provided. The review may also identify gaps in the existing literature and propose avenues for further investigation to advance the field of phishing detection using ensemble methods.



Comparisons of the reviews

Author & Year	SLR or not	Analyzed articles	Aim	Main findings	Limitations
Qabajeh et al., 2018	No	67	This review paper compares traditional anti-phishing methods, which includes raising awareness, educating users, conducting periodic training or workshop, and using a legal perspective.	Machine Learning and rule induction are suitable to combat phishing due to their high detection rate and, more importantly, the easy-to-understand outcomes.	Sixty-seven studies were analyzed in work, and the research did not discuss Deep Learning techniques.
Zuraiq and Alkasassbeh, 2019	No	18	This study examines several phishing detection methods, including heuristic, content-based, and fuzzy rule-based methods.	The study indicated no perfect method for identifying phishing websites.	The work analyzed only 18 studies and did not include Machine Learning, List-based, and Deep Learning approaches.
Kunju et al., 2019	No	14	This paper provides an overview of many Machine Learning algorithms for identifying phishing websites, including kNN, Naive Bayes, Decision tree, SVM, Neural Network, and Random Forest.	According to this study, phishing website detection using a single approach is insufficient.	The literature in this work included only 14 studies discussing Machine Learning techniques.
Benavides et al., 2020	Yes	19	This systematic literature review aimed to evaluate various other scholars' proposals for identifying phishing attacks using Deep Learning algorithms.	In conclusion, there is still a significant gap in the area of Deep Learning algorithms for phishing attack detection.	This work includes 19 studies, and only research articles on phishing and Deep Learning are considered in this study.
Athulya and Praveen, 2020	No	9	The research addressed phishing attacks, phishers' most recent phishing	According to this research, the best method to mitigate	The literature in this work is based on nine research items, and

Author & Year	SLR or not	Analyzed articles	Aim	Main findings	Limitations
			tactics, and anti-phishing techniques. In addition, the article aims to raise awareness regarding phishing attacks and strategies used for phishing detection.	phishing attacks is to raise awareness for the users and select the best anti-phishing security software tool.	the study does not include Deep Learning techniques for phishing website detection.
Basit et al., 2020	No	21	For phishing detection, the study examines Artificial intelligence approaches such as Machine Learning, Hybrid Learning, Scenario-based, and Deep Learning.	The study proved that Machine Learning procedures give the best results.	The work analyzed

Conclusion

In conclusion, the utilization of content-based methods employing various machine learning models presents a promising approach for detecting phishing websites. The integration of Support Vector Machine, Gaussian Naive Bayes, Decision Tree, Random Forest, Adaboost, Nearest Neighbor, and Neural Network algorithms enables robust identification of both phishing and safe websites. These models leverage features extracted from website content to discern malicious intent, thereby enhancing cybersecurity measures. However, while these methods demonstrate effectiveness, ongoing research and development are imperative to address evolving phishing tactics and enhance detection accuracy further. Additionally, considerations for scalability, efficiency, and real-time detection capabilities are crucial for practical implementation in diverse online environments. By continuously refining and adapting these techniques, we can bolster defenses against phishing attacks, safeguarding users and organization.

Source of review

1. Review Articles.
2. Conference Proceedings.
3. Technical Reports that have been Published.
4. Books Chapters.
5. Researchers' theses.

Search the relevant documents

A comprehensive view is required to perform a systematic review. As a result, before beginning the review, an appropriate selection of databases should be identified to deliver relevant results based on the keywords quickly. We selected the following five databases for systematic review.

1. ACM Digital Library (https://dl.acm.org).
2. IEEE Explore (https://ieeexplore.ieee.org).
3. Elsevier (https://www.elsevier.com).
4. Springer (https://link.springer.com).
5. Other Articles (Indexed in Scopus Journals (https://scopus.com).