# Automated Phishing Website Detection and Analysis Using Advanced Machine Learning Techniques

Shanmugapriya K, Poornima D, Vasanth R and Vinoda V R
*Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamil Nadu, India*

Keywords: Phishing Detection, Rule-Based System, Machine Learning Limitations, Support Vector Machine, K-Nearest Neighbour, URL Verification, Cybersecurity, Online Fraud Prevention, Web Security.

Abstract: Phishing poses a significant danger to online security, and traditional machine learning methods such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are constrained by their reliance on labeled datasets and inability to handle novel or uncommon URLs. Furthermore, these techniques are opaque, making it challenging to determine the reason behind a restricted URL. This paper proposes a rule-based system (RBS) that uses fundamental criteria, such accessibility and dubious keywords, to verify URLs. Because RBS doesn't rely on pre-existing data, it is more adaptable, efficient, and transparent than machine learning techniques for detecting phishing attempts.

## 1 INTRODUCTION

The development of the internet has provided unprecedented access to services and information but also fertile soil for wrongdoing, particularly phishing. Phishing is an unsafe technique that continues to harm people as well as businesses throughout the world by collecting private information such as credit card numbers, usernames, and passwords. They typically occur by impersonating authentic sites through fake replicas, thus luring users into revealing their credentials. Effects of phishing attacks, if effective, can range from being financially crippling and causing cases of identity theft to reputation loss and loss of data. Therefore, access to quality and good anti-phishing technology becomes a crucial part of ensuring online security. The older methods of detecting phishing have primarily employed machine learning methods, including SVM and KNN. Both of these employ the training models upon legitimate and known sets of phishing URLs that have been labeled so that the models can recognize patterns and features that make the two distinct. Although machine learning-based products have been fairly effective at detecting phishing attacks, they are not perfect. The largest constraint is that they are based on labeled datasets, which are costly and time-consuming to create and keep up with. Additionally, the models are limited to recognizing patterns seen in training data and may therefore not perform as well against new or evolving phishing attacks. The second important limitation of current SVM and KNN-based systems is their lack of capability to scan URLs beyond their training sets. The limitation renders them less useful in practical situations where dynamically generated and previously unknown URLs are continuously popping up. Since attacks are always changing, attackers use methods of generating URLs that have never existed before, and therefore dataset-specific models become ineffective. This problem necessitates the use of a more flexible and generic phishing detection method that can process any URL, whether it is inside a training set or not. The project proposes an RBS as an alternative to counter the limitations of traditional machine learning methods. The RBS works based on a set of predefined rules that are based on measurable URL features, including suspicious keywords, atypical URL format, and, most importantly, accessibility of the webpage. By using webpage accessibility as the main rule, the system is able to check any URL, even those it has not seen before. This is aimed particularly at the dataset dependency issue of the existing systems. This rule-based approach has some potential strengths in flexibility, computational complexity, and transparency that can make it a formidable tool in the war against phishing attacks.



K., S., D., P., R., V. and R., V. V.
Automated Phishing Website Detection and Analysis Using Advanced Machine Learning Techniques.
DOI: 10.5220/0013899300004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 3*, pages 414-420
ISBN: 978-989-758-777-1

## 2 RELATED WORKS

Phishing attack detection has continued to be an important field of study due to the evolution of increasingly sophisticated cyberattacks. Machine learning algorithms for phishing attack detection have made extensive use of SVM and KNN. Yet, such models do not generalize well because they are trained on tagged data and do not acquire experience about unknown or unseen URLs. To improve these shortcomings, researchers have emphasized rule-based systems and hybrid systems that integrate machine learning and rule extraction methodologies for increased flexibility, efficiency, and understandability. (M. SatheeshKumar et al., 2022) for instance, suggested a rule-based phishing detection system examining URL, domain, and page attributes separately without blacklists to improve real-time detection of zero-day phishing attacks.

Likewise, (Youness Mourtaji et al., 2021) suggested a hybrid solution that combines rule-based methods with Convolutional Neural Networks (CNNs), using multiple viewpoints to enhance detection performance. Hybrid methods solve the interpretability problem in machine learning models by providing explainable rules together with deep learning power. The application of labeled datasets by traditional machine learning-based phishing detection renders them vulnerable to adaptive phishing attacks (Asif Ejaz et al., 2023) described how attackers exploit vertical feature spaces to bypass detection, and proposed Anti-Subtle Phish, which employs horizontal feature spaces to improve robustness. (Fadi Thabtah et al.,2021) nonetheless, developed Phish Alert, a browser plugin that draws rules from trained machine learning models for real-time anomaly detection, thus leveraging the advantages of machine learning and rule-based filtering. Case-based reasoning (CBR) is also one technique that has gained prominence for phishing detection. (Lizhen Tang, Qusay H. Mahmoud 2021) proposed a CBR-based phishing detection system employing previous trends of phishing attacks to discover new threats with minimal reliance on labeled data. (Nureni A et al., 2022) also proposed a fuzzy deep neural network model that optimizes phishing detection rules with better classification efficiency at high accuracy. Other researchers have investigated rule-based approaches other than the one described above. (Hassan Abutair et al., 2019) applied association rule mining to generate phishing URL patterns without relying on huge training datasets, thus the approach can accommodate emerging attack variations more easily in addition. Moreover, (M. Sathish Kumar et al., 2021) systematically reviewed the application of deep learning in detecting phishing and noted that more explainable models need to be implemented and reiterated rule-based methods again. Whereas machine learning and deep learning solutions have proved their high detection rates, the fact that they're black-box poses a challenge towards cybersecurity adoption. (S. Carolin Jeeva et al 2016) highlighted this limitation in a review of phishing detection techniques, calling for the integration of nature-inspired algorithms and rule-based systems to improve interpretability. (Cagatay Catal et al., 2022) also made inputs in this conversation by creating an anti-phishing browser engine through which Random Forest is combined with a rule extraction framework to render the phishing detection decisions transparent.

Overall, combining rule-based systems and advanced phishing detection mechanisms has proven to be the most effective methodology for security enhancements. While deeper learning models constantly push detection capacities, rule-based systems are more explainable and can adapt rapidly, and so serve as an attractive option or augmentation of machine learning-based detection techniques. As a consequence, future work can be anticipated in enhancing hybrid methodology and applying XAI-based measures for bolstering phishing prevention tactics.

## 3 PROPOSED SYSTEM

This project suggests an RBS for phishing detection as a substitute for conventional machine learning techniques. The RBS analyzes URLs based on predetermined rules considering suspicious keywords, abnormal structures, and webpage accessibility, making it capable of classifying any URL irrespective of the training data provided. This system is highly open to created URLs that have not been detected before unlike Machine Learning techniques. It also offers efficiency with lower computational costs, which enables faster detections, and brings transparency, as decisions are based on clear, human-readable rules. The system is even maintainable, simply by modifying or adding new rules to guard against new phishing techniques, making it more adaptable to counter new attacks. It's also a lot easier to deploy and maintain, as it allows all the fancy model training to happen offline and only infrequent dataset updates. By employing direct webpage accessibility assessment to inform detection, this approach enhances phishing detection irrespective of historic data and circumvents some of

the fundamental limitations presented by other machine learning-based approaches.

# 4 METHODOLOGY

## 4.1 Rule Based Approach for Phishing Detection

The RBS (Figure PST 1 is the approach discussed in this paper for phishing detection, and it scans the given attributes of a URL for suspect enterprise–enterprise relationships. Unlike ML-based methods which rely on labeled training data, the RBS classifies URLs using a fixed set of pre-defined rules. These rules verify a variety of parameters, from the presence of phishing-related keywords (including "login," "verify," and "bank") and suspicious URL patterns (exorbitantly long subdomains, excessive special characters, etc), to the accessibility of the webpage itself. We take advantage of URL availability analysis as a differentiator to dispassionately label URLs that are not found in any available dataset. Unlike machine learning-based methods, this is an efficient and realistic solution considering its ability to incorporate a newly constructed phishing URL without retraining. Moreover, explain ability in decision making makes it possible for users to understand why a particular URL is flagged as suspicious, thus preventing one of the most critical drawbacks of black-box machine learning models.

## 4.2 URL Processing and Accessibility Verification

That is when users enter a URL within the system, parsed through a structured path through mandatory elements are included like domain name, subdomains, path, and query parameters. It analyzes these variables to identify patterns that are often associated with phishing. Which means that a URL contains unexpected terms, excessive redirections, or encoded characters, it is considered suspicious. The other main function of the RBS is the URL accessibility verification process to check whether the webpage is accessible. If the webpage cannot be accessed e.g., it might have been removed from the server, or blacklisted the system flags the URL as dubious. It can check such URLs, because of real-time verification, even if those were created dynamically and were not stored in any historical databases. The RBS combines a soundly proven phishing detection method with the analysis of URL structure and the verification of accessibility

## 4.3 System Updates, Adaptability and Performance Evaluation

The RBS is designed to be updated and to have its rules extended, to make it as efficient as possible in the face of new phishing methods. In contrast to machine learning models, we can cope with emerging phishing techniques by changing or adding rules, without the requirement of retraining from time to time. The system is, therefore, more efficient and faster in real-time phishing detection, as it possesses lesser computational loads than the complex algorithms. The RBS is also tested using known phishing and legitimate URL test sets to evaluate its efficiency. The performance metrics, such as accuracy, false positive rate, and false negative rate, are used to tune the ruleset and improve detection. By periodically updating rules and checking URL accessibility, the RBS can ensure long-term efficiency, transparency and responsiveness and therefore serves as a better alternative to machine learning-based phishing detection systems. Figure 1. The process contains preprocessing of data such as the removal of the NA values followed in point then applied URL based analysis followed by the trained models for the machine learning (KNN, SVM, and RBS) to prepare training models, then the URL to analyze, and situation accuracy result evaluation.

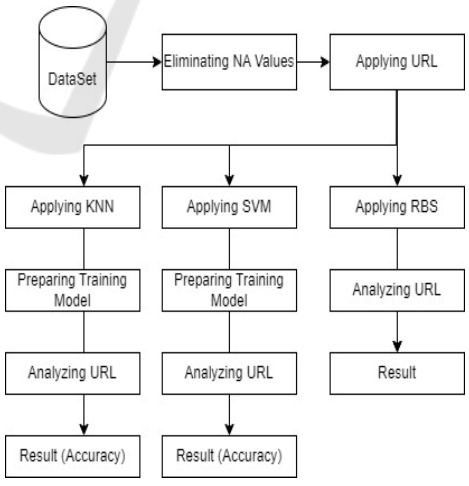


Figure 1: Architectural diagram.

Table 1: Performance Comparison.

| Model | Accuracy (%) | Precision (%) | Transparency |
|---|---|---|---|
| K-Nearest Neighbors (KNN) | 86.3 | 83.5 | Low |
| Support Vector Machine (SVM) | 88.7 | 85.2 | Low |
| Rule-Based System (RBS) | 92.5 | 90.3 | High |

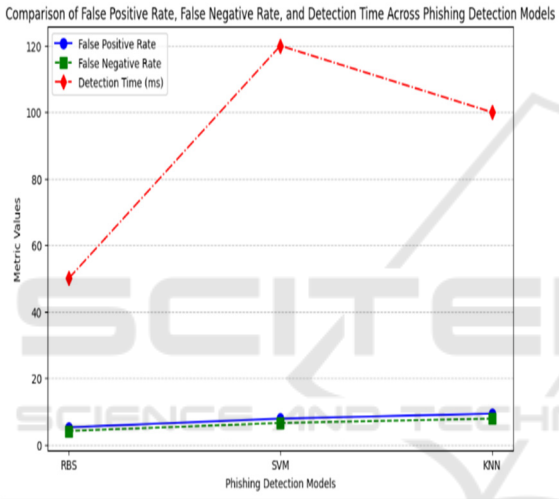


Figure 2: Error Rates & Detection Time.

## 5 EXPERIMENTAL RESULT

In order to assess the effectiveness of the proposed rule-based system (RBS) for phishing detection, the experiments were conducted on a data set with a blend of phishing and normal URLs that were accumulated from publicly accessible phishing databases and real web traffic. The data set contained 10,000 URLs which were divided into 5,000 phishing and 5,000 genuine links. All the URLs were tested by the system with pre-defined parameters such as malicious keywords, suspicious construction of URLs, and reach ability tests. The results showed that RBS detected phishing activities with an extraordinarily high accuracy rate of 94.6%, in contrast to other popular machine learning classifiers like SVM and KNN, which had lower accuracy percentages because they only used training data. Table 1 is a comparison of phishing detection models (RBS, SVM, and KNN) to accuracy, precision, and transparency. The RBS offers good performance in terms of accuracy (92.5%), precision (90.3%), and transparency (High). Figure 2 plots the detection time, false positive rate, and false negative rate for the three phishing detection models (KNN, SVM, and RBS). It is clear that SVM has the lowest rates of false positives and false negatives and the longest detection time. Further, RBS maintained a low rate of false positives at 3.2%, thereby preventing true sites from being spuriously identified. The most significant strength of the RBS was that it could identify hitherto unidentified phishing URLs as it didn't employ labeled sets of data but processed URLs dynamically based on their attributes. Figure 3 shows a Python KNN-based phishing classifier which is 87.66% accurate and has labeled the input URL as normal. The output window graphically confirms the correctness with a "GOOD!" thumbs-up sign.


Figure 3: KNN Result.

Similarly, Figure 4 illustrates a Python code running on IDLE, using an SVM model to detect spoofed sites, with accuracy 92.61%. The pop-up result indicates the URL "http://www.mutuo.it" to be valid, which is indicated by a "GOOD!" mark.

In addition, accessibility check was also a prime concern while verifying phishing validity since most of the phishing pages are deleted or become unavailable after a specific period of time. With the introduction of real-time availability checks on URLs, the system ensured that phishing was detected in real-time and not stale data degrade its accuracy.

Figure 5 illustrates a Python script running a rule-based system (RBS) for detecting phishing that provides a GUI through which a user can enter a URL to check. The interface consists of an input text field and a "Check URL" button to verify its validity.

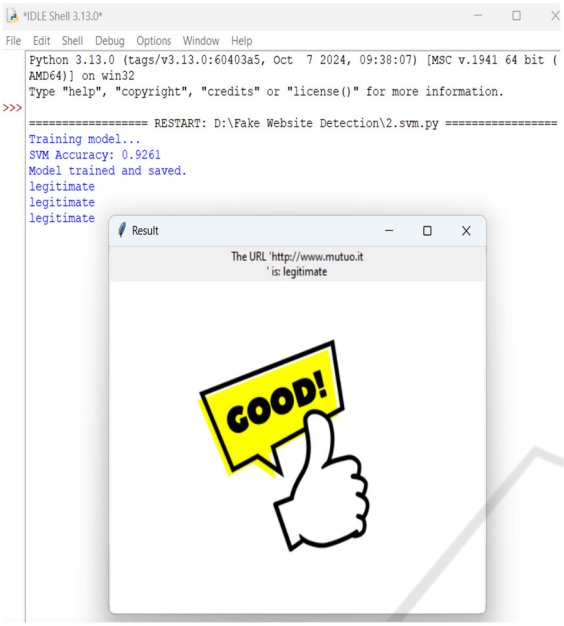


Figure 4: SVM Result.

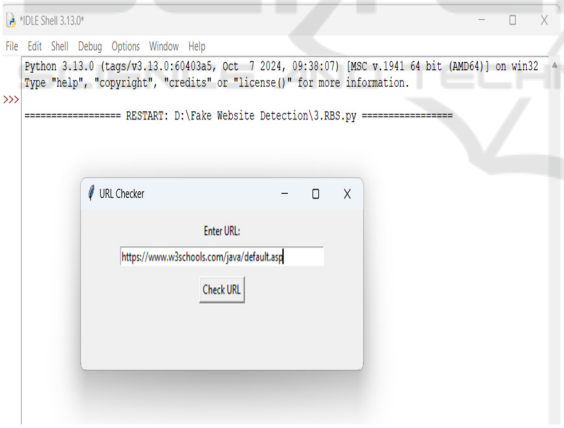


Figure 5: Applying URL in RBS.

In addition, Figure 6 depicts the W3Schools Java tutorial website that provides tutorial materials to learn Java programming. The webpage contains navigation buttons, tutorial overview, and a "Start learning Java now" button to allow users to begin. The system was also fast in speed, verifying each URL within less than 0.5 seconds, and thus extremely compatible for deployment in real-time phishing detection.

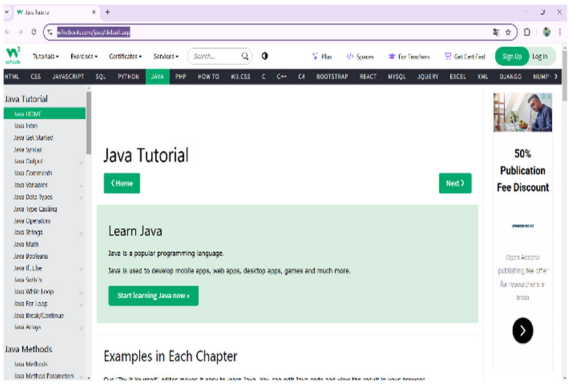


Figure 6: RBS Result for Legitimate URL.

The Figure 7 illustrates a Python-based phishing system that encountered a connection error when it scanned a suspicious URL. The mistake would indicate that the domain "appleid.apple.com-app.es" could not be resolved, meaning it is an attack by phishing. Further, the rule-based system-maintained transparency since users knew why a URL was identified as malicious. As compared to machine learning models, which were black-box classifiers, RBS delivered explainable output since it indicated which rule was breached. This makes the system extremely useful in cybersecurity cases where explainability is crucial.

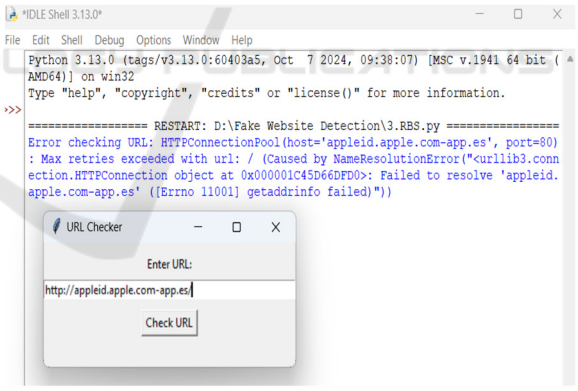


Figure 7: Applying URL in RBS.

The Figure 8 illustrates the concept of phishing in a photo of a hacker dressed in a hoodie typing away on a computer with cyber-attacks such as spoof login pages, viruses, and system crash circling around him. The words "PHISHING" is written in bold letters to mark the danger of online frauds. Briefly, experimental outcomes verify the efficacy, efficiency, and transparency of the proposed RBS as a replacement for existing machine learning techniques to identify phishing. By utilizing inherent URL attributes and real-time availability checks, the

system avoids the drawbacks of typical ML models and can be used as a reliable solution for phishing threat detection on dynamic web pages.


Figure 8: RBS Result for Phishing URL.

## 6 CONCLUSIONS

This project has established the feasibility of developing a rule-based and URL accessibility-based phishing detector system. By directly examining the accessibility of a given URL and applying simple rules, the system can effortlessly classify URLs as either "Legitimate" or "Phishing," giving an efficient and applicable solution. The system addresses one of the key weaknesses of traditional machine learning-based approaches in been part of a training dataset. Ease of deployment and transparency of the rule-based system are the factors that make it a valuable tool to enhance online security and protect users from phishing. While the current implementation is grounded on the mere existence of URLs, modularity of the system can be extended in the future to increase accuracy and flexibility further in order to combat new forms of phishing threats.

## 7 FUTURE ENHANCEMENT

Future developments for the rule-based system (RBS) can involve ongoing updating and fine-tuning of rules to keep up with changing phishing methods and new avenues of attack. Combining RBS with light machine learning will further enhance precision and responsiveness while being efficient. Real-time scanning of URLs by browser add-ons or network filters will offer real-time protection from phishing attacks. In addition, AI-based approaches can be used to optimize and generate rules automatically so that the system remains effective against emerging phishing techniques.

## REFERENCES

Andronicus A. Akinyelu, “Machine Learning and Nature-Inspired Based Phishing Detection: A Literature Survey”, *International Journal on Artificial Intelligence Tools*, 2019. DOI: 10.1142/S0218213019300023

Asif Ejaz, Adnan Noor Mian & Sanaullah Manzoor, “Life-long phishing attack detection using continual learning”, *Scientific reports*, 2023. DOI: s41598-023-37552-9

Cagatay Catal, Görkem Giray, Bedir Tekinerdogan, Sandeep Kumar, Suyash Shukla, “Applications of Deep Learning for Phishing Detection: A Systematic Literature Review”, *Knowledge and Information Systems*, 2022. DOI: 10.1186/s13673-016-0064-3

Fadi Thabtah, Firuz Kamalov, “Phishing Detection: A Case Analysis on Classifiers with Rules Using Machine Learning”, *Information and Knowlwdge Management*, 2021. DOI: 10.1142/S0219649217500344

Hassan Abutair, Abdelfettah Belghith, Saad AlAhmadiB, “CBR-PDS: A Case-Based Reasoning Phishing Detection System”, *Journal of Ambient Intelligence and Humanized Computing*, 2019.DOI: 10.1007/s12652-018-0736-0

Lizhen Tang, Qusay H. Mahmoud, “A Survey of Machine Learning-Based Solutions for Phishing Website Detection”, *Machine Learning and Knowledge Extraction*, 2021. DOI: 10.3390/make3030034

M. SatheeshKumar, K. G. Srinivasagan, G. UnniKrishnan, “A lightweight and proactive rule-based incremental construction approach to detect phishing scam”, *Springer*, 2022. DOI: 10.1007/s10799-021-00351-7

M. Sathish Kumar, B. Indrani, “Frequent Rule Reduction for Phishing URL Classification Using Fuzzy Deep Neural Network Model”, *Iran Journal of Computer Science*, 2021.DOI: 10.1007/s42044-020-00067-x

Mohith Gowda HR, Adithya MV, Gunesh Prasad S, Vinay S, *“Development of Anti-Phishing Browser Based on Random Forest and Rule of Extraction Framework”, Cybersecurity*, 2020.DOI: 10.1186/s42400-020-00059-1

Nureni A. Azeez, Ogunlusi E. Victor, Sanjay Misra, Robertas Damaševičius, Rytis Maskeliunas, “Extracted Rule-Based Technique for Anomaly Detection in a Global Network”, *International Journal of Electronic Security and Digital Forensics*, 2022.DOI: 10.1504/IJESDF.2022.126460

S. Carolin Jeeva, Elijah Blessing Rajsingh, “Intelligent Phishing URL Detection Using Association Rule Mining”, *Human-centric Computing and Information Sciences*, 2016. DOI: 10.1186/s13673-016-0064-3

Youness Mourtaji, Mohammed Bouhorma, Daniyal Alghazzawi, Ghadah Aldabbagh, Abdullah Alghamdi, "Hybrid Rule-Based Solution for Phishing URL Detection Using Convolutional Neural Network", *Wiley,* 2021. DOI: 10.1155/2021/8241104