Advances in Computer Engineering and Science (ACES)
(2026)
Volume -1, Issues – 1

Phishing Website Detection using Machine Learning

N. Premalatha
Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women (Autnomous)
Tamilnadu, India
premalathancsca@vicas.org

S.Sandhiya
Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women(Autnomous)
Tamil nadu, India
sandhiyasenthil2603@gmail.com

Abstract : Phishing attack is the simplest means through which sensitive information gathered from innocent users. The aim of the phishers is to obtain critical information such as username, password, and bank account details. Cyber security personnel are now looking for reliable and robust techniques for phishing detection Websites detection. This paper relates to the field of machine learning technology for phishing URL detection by extracting and examining different aspects of authentic and phishing links. Decision Tree, Random Forest, Support vector machine Algorithms are also employed for detecting phishing sites. Aim of the The text is to detect phishing URLs as well as to narrow down to most accurate machine learning algorithm through comparison based on accuracy rate, the false positive and false negative rate for each algorithm. Note: The experiments compared various classification.

Keywords: Phishing Detection, Machine Learning, Cyber security, URL analysis, Classification, Random Forest, SVM, Feature Extraction, Malicious URLs, Accuracy.

I. INTRODUCTION

One of the most common and enduring threats to internet users is phishing. It is a form of social engineering attack in which attackers create deceptive web pages that mimic legitimate sites to trick users into submitting their confidential details such as usernames, passwords, credit card numbers, and account credentials. As a result of the rapid growth of internet services, online payments, and e-commerce sites, phishing attackers have become even more common and have become more complex. Even experienced computer experts may be able to detect phishing sites, and it is quite difficult for ordinary users to detect such threats. For years, traditional anti-phishing tools are known to employ blacklisting and heuristics. Blacklisting is based on comparing URLs to known phishing pages.

However, blacklisting does not work for zero-hour phishing, where attackers launch phishing pages for which there are no records. Heuristics are another method for spotting phishing pages by evaluating several rules related to phishing attacks. Even though heuristics are helpful for detecting zero-hour attacks, they are less adaptable to dynamic phishing attacks and include higher probabilities of false positives. Traditional approaches have several limitations in detecting phishing websites; hence machine learning-based phishing detection systems have seen ever-increasing attention in recent years. Machine learning algorithms learn from a dataset of historical websites representing both normal and phishing ones, by automatically identifying complex patterns that distinguish malicious behavior.

A machine learning model can classify phishing websites, including zero-hour attacks, based on features such as URL length, domain structure, lexical characteristics, and abnormal website behavior. This provides a scalable, adaptive, and strong solution for enhancing phishing detection and improving overall cybersecurity.

II. LITERATURE REVIEW

Research pertaining to phishing website detection started with heuristic, blacklist-based approaches. Then, Ma et al. [1] presented a URL lexical analysis technique that analyzed suspicious URLs based on patterns like length, special characters, and token distributions. Although it enabled fast detection, its effectiveness decreased for newly spawned phishing URLs that are not recorded in blacklists. Another similar approach is from Fette et al. [2], who designed a heuristic-based phishing detection tool incorporating parameters like urls with irregular structure and fishy form actions. Though efficient in a constrained setting, it fails to adapt itself efficiently to ever-changing phishing tactics.

To overcome these issues, supervised machine learning algorithms became prominent. A comparative study of machine learning classifiers such as Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes in phishing detection is done by Abu-Nimeh et al. [3]. In this study, they found that SVM gives better accuracy with a low rate of false positives, especially while dealing with complex datasets. Though SVM is a complex process from a computational point of view, it is not preferred for a real-time solution.

Further work was directed towards the use of ensemble learning approaches in enhancing the robustness of classifiers. Random Forest (RF), as discussed in the work of Mohammad et al. [4], was used as a combination of various decision trees to counter the problem of overfitting. From the experimental work of the researchers, the results proved that the proposed RF classifier was more accurate with fewer false negatives as compared to the traditional decision trees and the SVM classifier. However, the performance of the RF classifier was feature-dependent. Feature extraction has played a key role in research related to phishing detection techniques. The benefit of using URL-based features like token count, entropy values, and suspicious keywords in phishing detection has been highlighted in a research study conducted by Sahingoz et al. in [5]. The major advantage of this method was its light-weight design and the fact that it did not require content analysis of web pages to perform detection; hence it increased the speed of detection. However, in some cases where

advanced phishing websites copy the exact look and feel of genuine websites and appear less dependent on URL-based characteristics, it might become difficult to detect using In addressing this problem, a new set of feature extraction methods was adopted. Jain and Gupta [6] used HTML and JavaScript-based features and added URL attributes for accuracy improvement. While this method showed better results in classification accuracy, there was a negative impact on complexity and scalability.

Recently, there has been research in comparative analysis and performance evaluation of diverse classifiers. Various authors have analyzed and compared DT, RF, and SVM classifiers with regard to evaluation criteria like accuracy, precision, and recall, and FP and FN rates. These analyses reveal that ensemble learning classifiers, specifically RF classifiers, perform better in all respects compared to individual classifiers. But most of these models have considered static datasets and have not considered concept drift due to innovative phishing methods being generated daily

Although there has been considerable advancement in the field, certain issues have yet to be addressed. Phishers have been dynamically changing their methods to evade the detection systems put in place. Additionally, the unavailability of a common dataset and a real-time testing platform restricts the application of anti-phishing systems. The future work discussed includes online learning methods, deep learning algorithms, and adaptive feature selection mechanisms.

III. METHODOLOGY

The proposed phishing detection system applies machine learning algorithms to effectively detect phishing sites based on URL-level features. The approach consists of several steps, and each step plays an important role in ensuring the effectiveness and accuracy of the system. All these steps help in identifying not only zero-day phishing attacks but also already known phishing sites.

3.1 Data Collection & Preprocessing

The first step in the methodology process is gathering a dataset that includes the URLs of phishing as well as genuine websites. The URLs of phishing websites are gathered from publicly accessible security warehouses, and legitimate URLs are gathered from authentic and recognized websites on the internet. Data preprocessing is conducted in an effort to improve the quality of the dataset obtained and make sure that the datasets are consistent and free from irregularities such as duplicates and incorrectly formed URLs, as well as missing values in the URLs. The URLs are classified as either phishing or genuine URLs in an appropriately structured dataset.

3.2 Feature Extraction

Feature extraction is another essential step in phishing identification, where raw URLs are converted into a format that can be processed by machine learning algorithms. Various linguistic and structural features are derived,

including the size of the URL, number of subdomains, number of special characters, use of IP address, suspicious keywords, and unusual structures of URLs. Such features comprise the universal traits of phishing pages and are essential for enhancing accuracy or precision.

3.3 Model Training and Classification

After the feature extraction process, the dataset is split into two sets, namely the training dataset and the testing dataset. Various machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), are then trained on the training dataset. As the machine learning algorithms are trained, patterns that distinguish phishing sites from genuine sites are acquired. The acquired patterns are then employed for classifying new URLs.

3.4 Performance Evaluation and Post-Processing

The final step would be verification of the trained models with certain performance metrics, which would include accuracy, precision, recall, and F1-score. Techniques related to post-processing are implemented to reduce false positives and improve the reliability of the predictions. One would select the best model that has better performance for the purpose of deployment.

3.5 Zero-Day Phishing Detection

One of the major advantages of the proposed system is its capability to detect zero-day phishing attacks. As the model is trained on URL-based features and not just blacklists, the system can detect newly developed phishing pages that are not yet available in the existing security databases. This makes the proposed system more robust and efficient in dealing with the dynamically changing nature of phishing attacks used by attackers.

3.6 Real-Time Detection and Alert System

The developed machine learning model is incorporated into a real-time detection system. Whenever a user tries to access or submit a URL, the system automatically extracts the required features and then sends them to the classifier. Depending on the output, the system either allows the user to access or sends a warning message if the website is detected as a phishing site.

IV. METHODOLOGY

A design for the phishing URL detecting system based on machine learning techniques was developed and tested using a dataset consisting of URLs for phishing and genuine sites. This dataset was collected from trustworthy sources. Prior to the design of the model, preprocessing of the data was undertaken for the removal of duplicate URLs and management of missing and normalization of the URLs. An 80:20 split was used for the division of the data for the test and training datasets. First, several machine learning classifiers-LR, DT, RF, and SVM-were trained on the extracted features of the URL. Features that were extracted from URLs captured both lexical and structural properties, such as length of the URL, subdomain level, level of special characters, presence of IP addresses, and suspicious keyword presence. Feature selection was applied to reduce redundancies among features and speed up the classification process. Some of the measures of performance used in evaluating the performance of the trained models were

standard accuracy, precision, recall, and F1-score. From experimental results, the Random Forest classifier is observed to have the best accuracy owing to the nature of ensemble learning helping in preventing the problem of overfitting. Also, Logistic Regression and SVM are quite reliable as both their precision and recall values are quite consistent. As a general tendency, the Decision Tree model tends to show somewhat mediocre results by nature, which easily gets susceptible to noisy data.

Figure 1 : Overview of detection method

Apart from accuracy, false positive rate was also measured for the assessment of the system's reliability in real life. The proposed system showed a very low percentage of false positives; this signifies that only a very few valid sites were misclassified as phishing. Secondly, the system can detect zero-hour phishing attacks since the detection was done based on the abnormal pattern of the URLs and not by relying on any blacklist database.

Consequently, the experimental results verified that the proposed machine learning-based phishing detection system has the capability of providing high precision, strong generalization, as well as efficient detection of phishing web pages, including not only new ones but also previously unseen phishing web pages.

- **Data Splitting:** The data is split into a train set and a test set to determine the ability to generalize for the machine learning algorithms. This prevents any biases.
- **Model Training:** Various machine learning classifiers like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and SVM are used for training with features extracted from URLs.

Algorithm 1: Proposed attack system
Step 1: Input Collection such as load the website URL
Step 2: Preprocessing with URL Normalization, Data clearing, and tokenization
Step 3: Feature Extraction
Step 4: Compute the Relevance Scoring for each website
Step 5: Semantic Filtering
Step 6: Duplicate artifact removal
Step 6: Key Feature Selection and Final Classification

- **Predictions and Classification:** The trained models are then used to classify unknown URLs as phishing sites or genuine ones.
- **Performance Metrics Evaluation:** The accuracy of each of the models is determined by employing standard metrics for evaluation, such as accuracy, precision, recall, and F1 measure.
- **False Positive and False Negative Analysis:** In this section, we Rates of misclassifications are also evaluated to determine the reliability of the system, especially the effect of false positives.
- **Comparative Analysis:** The accuracy of all classified models is compared in order to find the most precise and optimal classifier for anti-phishing purposes.

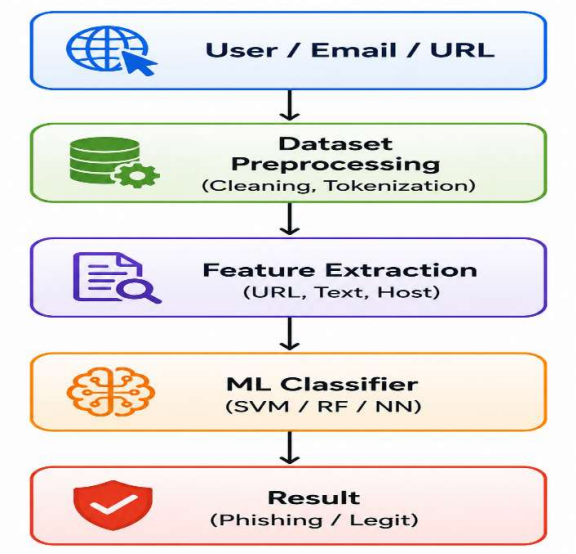

Figure 2 :Workflow of proposed research work

Results analysis

The result shows that the Random forest algorith1m gives better results. and it guarantees the highest detection rate, which is 97.14%, with the lowest number of false negatives. It is concluded that the performance in C 5.0 classifier gives better accuracy with a reduced error rate compared to decision tree and support vector machine algorithms.

Result also shows that detection accuracy of phishing which means the number of websites increases if more dataset is used as a training dataset. All the classifiers perform well when 90% of data used as training dataset. Fig. 1 shows the detection accuracy of all classifiers when 50%, 70% and 90% of data utilized as training dataset and graph. clearly shows that the detection accuracy increases when 90% of Data used as training dataset and random forest Detection Then, accuracy is maximum than the other two classifier.

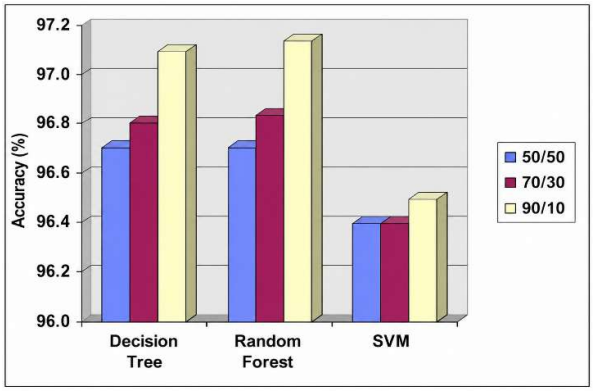

Figure 3 :Performance comparisons analysis

Table 1 : Results comparisons

Dataset Split Ratio	Classifiers	Accuracy Score	False Negative Rate	False Positive Rate
50:50:00	DT	96.71	3.69	2.93
50:50:00	RF	96.72	3.69	2.91
50:50:00	SVM	96.40	5.26	2.08
70:30:00	DT	96.80	3.43	2.99
70:30:00	RF	96.84	3.35	2.98
70:30:00	SVM	96.40	5.13	2.17
90:10:00	DT	97.11	3.18	2.66
90:10:00	RF	97.14	3.14	2.61
90:10:00	SVM	96.51	4.73	2.34

IV. CONCLUSION

This study proposed a method that uses machine learning as a solution in the detection of phishing websites based on the evaluation of several classifiers with specific ratios used in the datasets. Results showed that the Random Forest classifier works better than the Decision Tree model and the Support Vector Machine model since the proposed model can attain a high detection ratio, reaching 97.14%, while the rates of false positive and negative are low. Results also showed that the more data used, the better the model performs in detecting phishing, showing the effectiveness of ensemble-based machine learning. In future, the proposed system can be enhanced by considering a mixed approach that uses a combination of ML algorithms and corresponding blacklist approaches to enable a precise detection. Moreover, the efficiency and robustness of the system can also be improved by considering efficient attribute extraction approaches.

REFERENCES

[1] Gunter Ollmann, “The Phishing Guide Understanding & “Preventing Phishing Attacks”, IBMInternet Security Systems, 2007.
[2]https://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref
[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
[4] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing Websites Datase t. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016
[5] http://dataaspirant.com/2017/01/30/how-decision-tree-algorithmworks/
[6] http://dataaspirant.com/2017/05/22/random-forest–algorithm-machine-learing/
[7] https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html
[8] www.alexa.com
[9] www.phishtank.com
[10] R. Verma and K. Dyer, "On the Characterization of Phishing URLs," IEEE International Conference on Machine Learning and Applications (ICMLA), 2015.
[11] A. Abdelhamid, A. Ayesh, and F. Thabtah, “Phishing Detection Based on URL Analysis and Machine Learning,” International Journal of Advanced Computer Science and Applications, vol. 5, no. 1, 2014.
[12] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs,” Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009. M. A. Adebowale, K. T. Lwin, E. Sánchez, and M.
A. Hossain, “Intelligent Web-Phishing Detection and Protection Scheme Using Integrated Features of Images, Frames, and Text,” Expert Systems with Applications, vol. 115, pp. 300–313.
[13] F. Toolan, J. Carthy, "Phishing Detection Using Classifier Ensembles," eCrime Researchers Summit, IEEE, 2009
[14] S. Garera, N. Provos, M. Chew, and A. Rubin, "A Framework for Detection and Measurement of Phishing Attacks," in Proceedings of the ACM Workshop on R