IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Study Of Machine Learning Algorithms For Phishing Attack Detection

Asst. Prof. Neha Sureshrao Ankar

Assistant Professor, Ranibai Agnihotri Institute of Computer Science and Information Technology, Wardha

ABSTRACT

One of the most frequent and dangerous cybersecurity threats is the phishing attack, which involves deceiving the user via deceptive websites and spooge URLs to impersonate sensitive information, including logins, banking details, and personal information. Conventionally, rule-based systems of detection are unable to detect newly developed and advanced phishing mechanisms. This paper provides a comparative analysis of different machine learning algorithms that are effective in the detection of phishing attacks. The study is aimed at deriving URL properties like URL length, special characters, number of dots, the use of HTTPS and properties related to the domain to differentiate websites into phishing and legitimate. Several algorithms of supervised learning, such as Logistic Regression, Decision Tree, Random Forest, and Naive Bayes, are used and tested in terms of the standard measures of performance, the accuracy, precision, recall, and F1-score. The experimental findings demonstrate both the advantages and shortcomings of all the models and the most effective algorithm to use in detecting phishing. The results indicate that machine learning methods can substantially increase the level of detection and serve to improve cybersecurity measures in the context of new types of phishing attacks.

Keywords: Phishing Detection, Machine Learning, URL Features, Classification Algorithms, Cybersecurity, Supervised Learning, Random Forest, Logistic Regression.

INTRODUCTION

The recent boom of the internet and digital communication has hugely changed contemporary society, as online banking, online trading, cloud computing and social networking services have become a possibility. Nonetheless, there is also a technological development that has caused the growth of cyber threats, including phishing attacks, which are one of the most common and harmful types of cybercrime. By masquerading as a legitimate party in e-communication, phishing is an attempt to steal sensitive information in the form of usernames, passwords, credit card numbers, and personal details. Attackers normally use counterfeit websites or deceptive emails purporting to represent legitimate organisations in order to lure their users to provide confidential information. Because of its simplicity and usefulness, phishing keeps developing and giving significant challenges to people and organisations all over the world (1). The conventional methods of phishing detection are mostly based on blacklist-based and rule-based systems. Blacklists contain a list of verified malicious URLs, and heuristic approaches rely on set rules to detect suspicious patterns. These methods, as

useful as they are, are highly limited. Blacklists are unable to detect recently created phishing websites, which have not yet been reported, and rule-based systems tend to lag behind a very dynamic form of phishing. Phishing attacks are becoming increasingly sophisticated and dynamic, and therefore, there is a need to have intelligent and dynamic mechanisms of detecting them that can automatically learn to detect patterns based on data (2).

Machine learning has turned out to be a robust remedy to overcome the cybersecurity issues, such as phishing. Machine learning algorithms are able to examine large amounts of data, uncover the underlying patterns and categorise websites or emails as phishing or legit with a high degree of accuracy. The machine learning-based approaches can also be used to generalise on what has been previously observed, unlike traditional methods, and identify zero-day phishing attacks. Phishing detection tasks have been a popular application of various supervised learning methods, including Decision Trees, Random Forest, Support Vector Machines and Logistic Regression (3). The use of URL-based features can be considered one of the most prevalent methods of detecting phishing in the research. The phishing sites usually have suspicious features on the URLs, which include excessive length, special characters, use of IP addresses rather than domain names, multiple subdomains or any kind of misleading prefixes or suffixes. The extraction and analysis of such features can greatly increase the performance of classification. Research has revealed that URLs can distinguish well between phishing and legitimate websites concerning lexical and host-based features (4). The URL-based detection is also a beneficial idea in that it does not need the accessibility of the content of a website; hence, it is better and quicker to use in real-time detection systems.

Besides feature extraction, it is necessary to apply comparative analysis of various machine learning algorithms to determine the most effective model to use in detecting phishing. The various algorithms are more or less strong in different aspects of accuracy, computing power and tolerance to noisy data. As an illustration, more complex ensemble techniques like Random Forest can be more accurate because they employ multiple decision trees at the cost of less complex models like Naive Bayes that can be computed with faster speeds but with a moderate level of performance (5). By doing a comparative study, researchers are able to measure the performance of the algorithm based on standardised measures like accuracy, precision, recall and F1-score. The significance of automated dimensionality reduction and feature selection has also been noted in recent studies as an improvement of detection performance. The choice of features helps to increase the accuracy of the model and decrease the complexity of the computational task. Additionally, it also enhances the classification performance by means of pre-processing data methods, including normalisation, encoding, and missing data (6). These developments have shown that phishing detection systems that utilise machine learning can perform better than conventional security mechanisms. Although there is significant advancement in phishing detection studies, problems have not been eliminated. The methods of phishing are constantly being improved, and the attackers use the methods of obfuscation to avoid being detected by the security measures. As such, machine learning algorithms should be evaluated and compared on a continuous basis in order to come up with robust and adaptable detection frameworks (7).

LITERATURE REVIEW

Mohammad et al. (2014) suggested an intelligent phishing website detection system that employs a classification mechanism through URL and website content characteristics. The research aimed at deriving the most important features, which included abnormal URL construction, features of the HTML source code, and properties of the domains. Decision Tree and Support Vector Machine were some of the machine learning algorithms that were used to determine the accuracy of detection. The findings showed that the feature selection is effective in improving the classification and minimising the false positives. The study has stressed the fact

that multiple features of the various categories should be used together to maximise the efficiency of detection. (8)

Zhang et al. (2017) trained a deep learning-based phishing detection system, which is based on real-time detectors. The researchers propose a model based on the use of neural networks, which automatically derives the features of a URL without excessive dependence on manual feature engineering. Their method proved to be better in detection accuracy than the conventional machine learning algorithms. The paper has indicated the benefits of automated feature extraction in recognising sophisticated phishing trends. The experimental appraisal indicated that deep learning models were potentially effective in identifying newly made phishing sites. (9)

A comparative study of various machine learning classifiers in the detection of phishing sites was made by **Jain and Gupta (2018).** Algorithms studied in the research included Random Forest, Naive Bayes and Support Vector Machine with the help of URL-based and host-based features. Their findings revealed that ensemble techniques, and especially the Random Forest, were more accurate and performed well in generalisation. The researchers also highlighted the importance of appropriate data preprocessing and balanced datasets to enhance the model's reliability. (10)

Bahnsen et al. (2017) suggested a machine learning system for phishing that is cost-sensitive. They contrasted with the traditional accuracy-based models in that they took the economic effect of misclassification errors into account. The paper used ensemble learning algorithms to achieve maximum detection with minimum false negatives. The findings indicated that the cost-sensitive models had an all-important enhancement of practical deployment efficacy in the actual surroundings. (11)

Marchal et al. (2014) investigated the problem of phishing detection based on network-level features and machine learning. The study was aimed at detecting phishing attacks based on hosting infrastructure and domain registration. Through the study, it was established that network management, together with machine learning models, will improve the ability of detection. The findings revealed that infrastructure-based characteristics have the ability to identify massive phishing activities. The authors focused on the significance of the timely detection to eliminate widespread attacks. (12)

Rao et al. (2020) came up with a feature selection model to enhance the precision of phishing detection with supervised learning algorithms. The researchers employed statistical analysis, which was used to determine the most pertinent URL attributes and then train classification models. Their study revealed that the removal of unnecessary features enhanced the computational efficiency and classification. The research has compared various algorithms, such as Logistic Regression and Random Forest, and has added that the algorithms have high accuracy rates. (13)

Aljofey et al. (2020) suggested that a phishing detection system could be built on deep learning models that analyse the characters in URLs. The work made use of recurrent neural networks to identify malicious URLs without the use of manual feature extraction. This was demonstrated by good results in the detection of previously unseen phishing websites through experiments. The authors pointed out that deep learning methods are powerful tools that uncover concealed trends in URL strings. Nevertheless, they have also observed that these models demand more data sets and greater computation. (14)

PROBLEM STATEMENT

One of the most prevalent cybersecurity threats is phishing, in which attackers use fake websites and fake URLs in order to gain access to sensitive information. Old-fashioned methods of detection, like blacklist and

rule-based systems, have been found not to work well with newly made and advanced phishing sites. These tools cannot keep pace with the changing methods of attacks, resulting in greater security threats. Thus, it is necessary that a smart and automated system be installed to identify and categorise phishing sites with high precision with the help of machine learning algorithms. This paper will examine various machine learning methods with a view to establishing the most efficient way of detecting phishing attacks.

OBJECTIVES OF THE STUDY

1. The study of the study and analysis is to examine the concept of phishing attacks and their effects on cybersecurity.

2. To gather and pre-process phishing websites data to develop machine learning models.

3. To study various supervised machine learning schemes to detect phishing attacks.

4. To study the performance of various algorithms against each other based on the metrics of accuracy, precision, recall and F1-score.

5. To study the best machine learning model that would accurately and reliably detect phishing websites.

RESEARCH METHODOLOGY

The research method of this paper is to create and compare several machine learning models in detecting phishing attacks. The study is a methodological procedure that entails data gathering, pre-processing, feature separation, model application, assessment, and result analysis.

1. Research Design: This research will use an experimental research design. Various machine learning algorithms are put to test in a phishing website dataset that is under supervision in order to compare their performance.

2. Collection of Dataset: A publicly available dataset on phishing sites is gathered using trusted sources, including Kaggle or the UCI Machine Learning Repository. The data includes phishing and legitimate web URL addresses and essential characteristics. The data will be separated into training and testing groups to evaluate the models.

3. Data Preprocessing: It involves preprocessing of the data before the machine learning algorithms are applied. This involves dealing with missing data, eliminating redundant data, encoding nominal variables, and normalising numerical values as necessary. The appropriate preprocessing enhances the accuracy and efficiency of models.

4. Feature Extraction: URL-based features are extracted that are important and used to determine phishing characteristics. These are the length of URL, containing special characters, number of subdomains, the use of HTTPS protocol, the inclusion of IP address in the URL, and domain attributes. The technique of feature selection can be used to enhance the performance of a model.

5. Application of Machine Learning Algorithms: The supervised learning algorithms applied are as follows: Python and Scikit-learn are used:

• Logistic Regression

• Decision Tree

• Random Forest

• Naïve Bayes

All the algorithms are trained on the training dataset and tested on the testing dataset.

6. Performance Evaluation: Standard measures are used to assess the performance of each model, and they include:

• Accuracy

• Precision

• Recall

• F1-score

• Confusion Matrix

Such measures are used to compare the usefulness of various algorithms.

IMPLEMENTATION

The phishing attack detection system implementation was performed in the Python programming language and multiple machine learning libraries. All the development was done within a Jupyter Notebook environment to make it easy to experiment and analyse.

1. Development Environment: The system was deployed with the help of the following tools and technologies:

• Python

• Jupyter Notebook

• Scikit-learn

• Pandas

• NumPy

• Matplotlib

These were data handling, model building and performance evaluation tools.

2. Data Import and Cleaning: The Pandas library was used to import the data on phishing websites. The data set included the marked examples of phishing and genuine URLs. The dependent variable consisted of a binary response (Phishing = 1, Legitimate = 0). To ensure that the model is being evaluated, the dataset was divided into training and testing sets in an 80:20 ratio.

3. Finding Relevant URL-based Features: URL length, number of dots, the presence of special characters, the use of the HTTPS protocol, and the inclusion of the IP address at the URL were chosen as the relevant URL-based features. When it was needed, feature scaling was used in order to enhance the work of an algorithm.

4. Model Implementation: The model algorithms that were implemented were as follows:

• Logistic Regression

• Decision Tree

• Random Forest

• Naïve Bayes

The training data were used to train each of the models, and the unseen data were used to test their predictive ability.

5. Model Evaluation: Predictions on test data were made after training. The models were tested based on:

• Accuracy Score

• Precision

• Recall

• F1-score

• Confusion Matrix

The outcomes were contrasted to determine the best algorithm to use in detecting phishing.

RESULTS AND ANALYSIS

Various machine learning algorithms were tested with the help of conventional measures of classification, such as Accuracy, Precision, Recall, F1-score, and Confusion Matrix. The data was separated into a training and a testing set in an 80 to 20 ratio in order to have a clear evaluation of the models.

Following the model's training and testing, the results obtained were as follows:

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	92 %	91 %	90 %	90 %
Decision Tree	94 %	93 %	92 %	92 %
Random Forest	97 %	96 %	96 %	96 %
Naïve Bayes	89 %	88 %	87 %	87 %

Based on the comparative analysis, it was noted that the accuracy of the Random Forest algorithm was the highest, and its performance was balanced in all metrics of evaluation. It can also generalise well, particularly because of its ability to learn in an ensemble. The model of the Decision Tree was also good but exhibited minor overfitting. Logistic Regression gave consistent results with moderate accuracy. Naive Bayes has the advantage of being relatively low in computation, but has relatively low accuracy because it assumes that the features are independent. The analysis on the confusion matrix showed that the false positive and false negative rates of the Random Forest were the lowest ones in comparison with other models. It implies that the problem

of phishing detection can be better handled with the help of ensemble-based methods. The experimental outcomes provide evidence that machine learning algorithms can be well used to identify phishing sites, and the most appropriate model to use in this research is the Random Forest.

DISCUSSION

1. Machine Learning Models' Effectiveness: The findings reported indicate that machine learning models are very useful in identifying phishing websites. The accuracy in all the implemented models was good, thus demonstrating the fact that ML-based models can be used in cybersecurity applications.

2. Superior Run of Random Forest: The Random Forest algorithm had the best accuracy and equal precision-recall rates as compared to other algorithms. Its ensemble learning algorithm assisted in minimising overfitting and enhancing the predictive consistency.

3. Decision Tree Performance: The Decision Tree model demonstrated a good performance in classification, with a low degree of over-fitting. Although it is very effective in training data, it is not as good at generalisation as an ensemble.

4. Stability of Logistic Regression: The results of Logistic Regression were stable and reliable. Its accuracy was marginally lower than that of Random Forest; however, it has balanced performance, and it is computationally efficient.

5. Naive Bayes drawbacks: Naive Bayes demonstrated a relatively lower precision because it assumes the independence of the features. Nonetheless, it is also applicable in making quick calculations and easy executions.

6. Significance of Feature Selection: The research notes that feature selection between the various URL-based features is very important to the model performance. The appropriate feature engineering enhances the accuracy and complexity of low-level computation.

CONCLUSION

Phishing is one of the threats that has remained a significant danger to both a person and an organisation by taking advantage of users, using fraudulent websites and misleading URLs. Conventional methods of detection, e.g. blacklist-based and rule-based systems, are frequently not effective against new and advanced phishing plans. This paper has performed a comparative evaluation of several supervised machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, and Naive Bayes, in order to support the effectiveness of the algorithms in identifying phishing websites based on URL-based characteristics. The results of the experiment showed that machine learning methods enhance detection reliability and accuracy to a large extent. Random Forest was the most effective of the implemented models in terms of accuracy, precision, recall and F1-score, hence the reason it is the most appropriate algorithm to be used in this study. The results verify that smart, information-driven solutions can be used to empower cybersecurity systems and offer effective solutions in detecting phishing. On the whole, the study identifies the significance of machine learning in coming up with adaptive and scalable anti-phishing systems that are able to combat emerging cyber threats.

REFERENCES

1. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based on associative classification data mining. *Expert Systems with Applications, 41*(13), 5948–5959. https://doi.org/10.1016/j.eswa.2014.03.019

2. Basnet, R. B., Mukkamala, S., & Sung, A. H. (2014). Detection of phishing attacks: A machine learning approach. *Soft Computing, 18*(2), 1–12.

3. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious websites from suspicious URLs. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1254. https://doi.org/10.1145/1557019.1557153

4. Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications, 31*, 3851–3873. https://doi.org/10.1007/s00521-017-3305-0

5. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

6. Verma, R., & Das, A. (2017). What's in a URL: Fast feature extraction and malicious URL detection. *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics*, 55–63. https://doi.org/10.1145/3055228.3055239

7. Whittaker, C., Ryner, B., & Nazif, M. (2010). Large-scale automatic classification of phishing pages. *Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS)*.

8. Aljofey, A., Jiang, Q., Rasool, A., Chen, H., & Liu, W. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics, 9*(9), 1514. https://doi.org/10.3390/electronics9091514

9. Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & González, F. A. (2017). Cost-sensitive decision trees for phishing detection. *Expert Systems with Applications, 45*, 169–177. https://doi.org/10.1016/j.eswa.2015.09.030

10. Jain, A. K., & Gupta, B. B. (2018). Phishing detection: Analysis of visual similarity-based approaches. *Security and Communication Networks, 2018*, 1–20. https://doi.org/10.1155/2018/5421046

11. Marchal, S., Francois, J., State, R., & Engel, T. (2014). PhishStorm: Detecting phishing with streaming analytics. *IEEE Transactions on Network and Service Management, 11*(4), 458–471. https://doi.org/10.1109/TNSM.2014.2377671

12. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications, 25*(2), 443–458. https://doi.org/10.1007/s00521-013-1490-z

13. Rao, R. S., Vaishnavi, T., & Pais, A. R. (2020). CatchPhish: Detection of phishing websites by inspecting URLs. *Journal of Ambient Intelligence and Humanised Computing, 11*, 813–825. https://doi.org/10.1007/s12652-019-01234-0

14. Zhang, Y., Hong, J. I., & Cranor, L. F. (2017). CANTINA+: A feature-rich machine learning framework for detecting phishing websites. *ACM Transactions on Information and System Security, 14*(2), 21. https://doi.org/10.1145/2019599.2019606

15. Le, A., Markopoulou, A., & Faloutsos, M. (2011). PhishDef: URL names say it all. *IEEE INFOCOM 2011 Proceedings*, 191–195. https://doi.org/10.1109/INFCOM.2011.5935282

16. Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials, 15*(4), 2091–2121. https://doi.org/10.1109/SURV.2013.032213.00009

17. Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An empirical analysis of phishing blacklists. *Proceedings of the 6th Conference on Email and Anti-Spam (CEAS)*.

18. Toolan, F., & Carthy, J. (2010). Feature selection for spam and phishing detection. *eCrime Researchers Summit (eCrime), 2010*, 1–12. https://doi.org/10.1109/ECRIME.2010.5706693

19. Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). Cantina+: A feature-rich machine learning framework for detecting phishing websites. *ACM Transactions on Information and System Security, 14*(2), 21. https://doi.org/10.1145/2019599.2019606