WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detection of Phishing Website Using Machine Learning

Perla Hari Priya

GMR Institute of Technology

ABSTRACT:

Nowadays many people rely on Internet for various tasks such as online banking, shopping, bill payments, and trading. Phishing is one of the best and most successful techniques for hackers to cheat us and get our personal information like username, password, and bank account details. Often people struggle to differentiate between legitimate and phishing websites. An anti-phishing machine strategy can assist in distinguishing a real website from a phishing website using multiple sources such as URLs, search engines, page content, and so on. Phishing websites are detected using machine learning. Machine learning algorithms have shown promise in improving the accuracy and efficiency of detecting phishing websites. The URLs received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it's phishing or legitimate. In this study, various machine learning algorithms are explored for phishing detection including Random Forest, Decision tree, Gaussian Naive Byes, SVM, and Neural Networks. The study's purpose is to detect phishing URLs as well as to select the best machine learning algorithm by analyzing each algorithm's accuracy rate, false positive and false negative rate.

Keywords: Phishing, Machine learning, legitimate, URL, website.

Introduction:

Phishing websites pose a significant threat to online security, as they are designed to deceive users into disclosing sensitive information such as login credentials, financial details, or personal data. Detecting these malicious websites promptly is crucial to safeguarding individuals and organizations from cyber attacks and data breaches.They look just like real websites, making it hard to tell them apart. Anti-phishing tactics include both education and technology resistance. This study focuses on the technical defense strategies suggested in recent years.Traditional methods are list-based solutions that collect valid, legitimate websites for a whitelist or verified phishing websites for a blacklist and widely distribute the list to prevent other users from being attacked.These methods effectively prevent users from reusing the same phishing website URL, reducing the number of affected users and losses. However, these methods have a significant drawback: they cannot detect new phishing URLs. As a result, some innocent users will be attacked before the link is blacklisted. Some scientists proposed rule-based methods to recognize new fake websites. This method entailed security expert experience and website analysis of phishing sites. The World Wide Web Consortium (W3C) standard URL includes the protocol, subdomain, domain name, port, path, query, parameters, as well as fragment.Specifically, rules are generated from URL components, such as whether the domain name is similar to other legitimate domains. In these rules, some must request third-party services to obtain information, such as the domain's registration date. When the rules were published in technical articles, phishers discovered them and devised new phishing URLs that did not match the rules.Along with the development of machine learning techniques, various machine learning based methodologies have emerged for recognizing phishing websites to increase the performance of predictions. Phishing detection is a supervised classification approach that uses labeled datasets to fit models to classify data. In this study, we will delve into the key components of phishing website detection using machine learning models such as Decision Tree, Random Forest, SVM, KNN, GBM, exploring the features and datasets used for training, the algorithms employed for classification, and the evaluation metrics that measure the model's effectiveness.

Literature Survey:

There have been numerous papers published that introduce and compare various solutions for detecting phishing websites. Shahrivari et al. discusses the use of machine learning techniques for detecting phishing websites and compares the results of multiple machine learning methods.One of the most difficult aspects of their research was the scarcity of phishing datasets.They implemented and evaluated twelve classifiers on the phishing website dataset, which contains 6157 legitimate websites and 4898 phishing websites.The findings encourage future research to add more features to the dataset, which could improve the performance of these models; thus, it could combine machine learning models with other phishing detection techniques, such as List-Base methods, to achieve better performance [1].

Ahammad et al. conducted a research on Phishing URL detection using machine learning methods. This paper proposes using machine learning algorithms to detect malicious URLs, which could improve detection accuracy over traditional blacklisting methods. The dataset contains 3000 URLs, 1500 of which are malicious and 1500 of which are benign. The model could be used to create a search engine and identify new types of phishing attacks [2].

Kulkarni el at. developed methods to differentiate between phishing websites and legitimate ones using machine learning techniques.Over 90% of the time, the classifiers used in the study were successful in distinguishing between real and fake websites.The limitation is that they only considered a small data set of 1353 URLs, with 9 features for each URL. Additionally,they can assess classifiers using a large data set containing thousands of URLs and extract a greater number of features that may be important in decision making [3].

Kiruthiga et al. presented various algorithms and approaches to detect phishing websites in the survey. The paper makes no mention of the size or diversity of the datasets used to train and test the machine learning models.After reviewing the papers, they concluded that the majority of the work was done using well-known machine learning algorithms such as Nave Bayesian, SVM, Decision Tree, and Random Forest. For detection, some authors proposed new systems such as Phish Score and Phish Checker [4].

Manuel Sanchez-Paniagua et al. compares machine learning and deep learning techniques for detecting phishing URLs via URL analysis. It creates a more representative dataset for a real-world scenario by using URLs from the login page in both classes (phishing and legitimate).Their work provides researchers with an updated dataset PILU-90K on which to train and test their approaches.Their approach has the main advantage of having a low false-positive rate when classifying this type of URL. TFIDF combined with N-gram and LR algorithm produced the best results with a 96.50% accuracy among the various evaluated models [5].

Aljabri et al. Presented a research on Machine learning and deep learning models that are used to detect malicious URLs.The paper makes a significant contribution to the field by conducting extensive feature engineering and analysis to identify the best features for predicting malicious URLs. The paper makes no mention of the preprocessing techniques used on the dataset.Using the Nave Bayes (NB) classifier, the paper achieves a high accuracy of 96% in detecting malicious URLs [6].

Singh et al. conducted a survey on phishing website detection based on machine learning techniques.The main objective of the survey is to raise reader awareness of phishing attacks and encourage phishing prevention.Due to the constant race between researchers and phishers, phishing attacks are difficult to detect and prevent.Machine learning for detecting phishing websites can aid in the identification of fraudulent web links.The paper discusses several studies that used machine learning models to achieve high accuracy in phishing detection.Continuous research and development of new techniques is required to stay ahead of phishers and improve the effectiveness of phishing detection and prevention [7].

Abdul Karim et al. conducted a survey on Phishing detection system through hybrid machine learning based on URL.The current paper focuses on machine learning-based phishing detection systems and presents a hybrid machine learning approach that employs a variety of models, including decision trees, linear regression, random forest, naive Bayes, gradient boosting classifier, K-neighbors classifier, support vector classifier, and a proposed hybrid LSD model.The proposed approach achieves its goal with effective efficiency. Future phishing detection systems should combine list-based machine learning-based systems to more efficiently prevent and detect phishing URLs [8].

Yi Wei et al. conducted a research on Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection. This paper compares the performance of various machine learning and deep learning methods for detecting phishing websites. The paper discusses the usefulness of ensemble machine learning methods in anti-phishing techniques, emphasizing their advantages in detection accuracy and computational consumption, as well as their ability to handle a reduction in the number of features in the dataset. In the future, they will validate their findings on larger datasets with more features and instances [9].

Ubing et al. discusses anti-phishing techniques such as prevention and detection, with a focus on the structure and components of a URL, feature selection methods, ensemble learning, and existing phishing detection technologies.By employing a feature selection algorithm and integrating it with an ensemble learning methodology based on majority voting, the research contributes to improving the accuracy of phishing website detection.The experimental results show that the proposed model has a promising accuracy rate of up to 95%, which is higher than current technologies for detecting phishing websites [10].

Mahajan et al. focuses on detecting phishing websites using machine learning algorithms such as Decision Tree, Random Forest, and Support Vector Machine.The authors compare each algorithm's accuracy rate, false positive rate, and false negative rate to determine which one is best for detecting phishing URLs.The Random Forest algorithm detects phishing websites with 97.14% accuracy and the lowest false positive rate, making it the most effective algorithm.The paper emphasizes the importance of using more data as training data because classifiers perform better with more training data, resulting in higher detection accuracy [11].

Junaid Rashid et al. proposes an effective machine learning-based phishing detection technique that integrates with the Support Vector Machine (SVM) classifier to accurately distinguish 95.66% of phishing and legitimate websites while utilizing only 22.5% of the novel functionality.The proposed method employs the "StringtoWordVector" Weka function to translate each URL into carrier-specific words, which are then used to detect phishing websites [12].

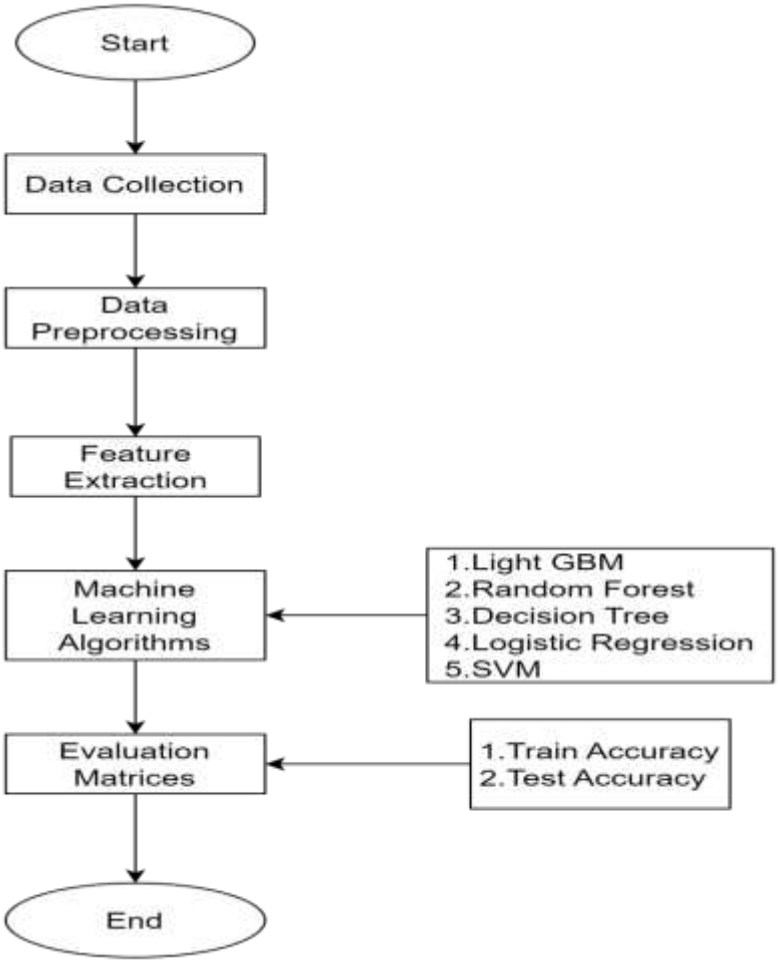
Mourtaji et al. describes a new hybrid solution for detecting and preventing phishing URLs.The blacklisted method, lexical and host method, content method, identity method, identity similarity method, visual similarity method, and behavioral method are all used in the paper to detect phishing URLs.The hybrid solution is effective at analyzing URL stress from various perspectives, resulting in high efficiency and accuracy.The paper makes no mention of the computational resources or time needed for the proposed solution to analyze URLs [13].

SsDeshpande et al. examined some of the traditional approaches to phishing detection, specifically blacklist and heuristic evaluation methods, as well as their drawbacks. They ran two machine learning algorithms against the 'Phishing Websites Dataset' and analyzed the results.In the future, they plan to

build the phishing detection system as a scalable web service with online learning so that new phishing attack patterns can be easily learned and to improve the accuracy of their models with better feature extraction [14].

Alswailem et al. Developed an intelligent system for detecting phishing websites using machine learning techniques.They have a database of 16000 phishing and legitimate URLs.They investigated all 36 features in order to reduce computation time and provide high performance with the fewest combinations of powerful features.When the system detects a phishing website, it automatically notifies the user, providing real-time protection.Their main goal is to achieve a higher accuracy with a smaller number of features, therefore they removed the features and achieve the same accuracy with 26 features [15].

Methodology:



Dataset:

To train an excellent model, it must be designed with a large dataset. Because of the increased number of phishing attacks in recent years, there is a greater need for high-quality and new datasets for improved model performance.PhishTank is the dataset that has high popularity in phishing detection models. It is a Cisco Talos Intelligence Group (Talos) blacklist with approximately 7 million phishing URL records. Two million records have been verified as phishing, according to their website. Furthermore, 11000 of the two million URLs are online, while the rest are offline, because the phishing URLs are removed after the attacks. If a URL is available online, PhishTank can be used to view its webpage.

Feature Extraction:

The process of identifying the characteristics that distinguish phishing and legitimate web pages is referred to as feature extraction. This procedure is crucial because the features chosen affect the accuracy and speed of phishing detection.To detect phishing effectively, the properties describing web pages must take into account the strategies and practices commonly used by attackers to create phishing web pages, as well as the characteristics that distinguish these pages from their legitimate counterparts.They outlined that the main sources of content are page URLs and source codes.Various techniques, such as lexical analysis, information retrieval, statistical techniques, Natural Language Processing, and heuristics are used to characterize the identified properties and encode them into meaningful features.

Feature Selection:

Feature selection refers to the process of selecting the best features from among those extracted to classify a page as legitimate or phishing. This process is critical for making models parsimonious, avoiding overfitting, improving accuracy, and reducing computation requirements, especially when there are a large number of features. The papers develop traditional feature selection techniques, such as filter methods and wrapper methods, in order to identify the best subset of features for phishing website detection.

Machine Learning Models:

Phishing detection is viewed as a binary classification problem, with the primary goal of determining whether a web page is legitimate or phishing. Many different supervised learning algorithms have been considered in the literature to solve this problem and derive the corresponding models.They used machine learning algorithms such as the Random Forest Classifier, Decision Tree, SVM, and Light GBM to apply the machine learning model to all of the features generated by the feature extraction module.

1) SVM:

A support vector machine (SVM) is a supervised machine learning algorithm that is defined by a hyperplane that separates different classes.It is commonly used in classification problems, where it provides the highest accuracy between two classes. Each data item is represented as a point in n-dimensional space (where n is the number of features in the dataset), with the value of each feature corresponding to the value of a specific coordinate.Then, they performed classification by locating the hyperplane that best distinguishes the two classes, and the classes in this dataset are zero and one, which are easily separated by the hyperplane, as this algorithm works best for two classes.

2) Decision Tree:

The decision tree algorithm is a well-liked machine learning technique in which a tree structure represents the model logic. The decision tree has nodes that represent features, stems that show feature values and possibilities, and a final node that displays the outcome. Performance is typically better with the simpler tree structure. Very deep tree growth probably results in overfitting training datasets.

3) Random Forest:

An ensemble of decision trees for regression and classification is called a random forest. Random forests classify or average the output of individual trees during training processing to address the overfitting problem.As a result, random forests usually perform more accurately than decision tree algorithms.

4) KNN:

A k-nearest neighbors algorithm (KNN) is a non-parametric classification algorithm that makes predictions by calculating the distance between the target and its nearest neighbors. For continuous data, there are some methods for calculating the distance with respect to the Euclidean distance, and for discrete values, there are some methods for calculating the distance with respect to the Hamming distance. It lacks a training process in particular, and each prediction will take a long time. As a result, this algorithm is in general unsuitable for real-time scenarios.

5) CNN:

A feedforward deep learning algorithm known as a convolutional neural network (CNN) is widely used in image classification. A CNN's regular architecture consists of multiple layers, including the input layer, hidden layers, and output layer. Convolutional layers, pooling layers, and fully connected layers are common in hidden layers.For images, it typically employs 2D or 3D convolutional layers, whereas 1D convolutional layers have shown success with text and sequence data, particularly in time-series analysis.

6) Naive Bayes:

The naive Bayes classifier is one of the most basic and effective machine learning classification algorithms, and it aids in the development of a fast machine learning classifier capable of making quick predictions from a given dataset. A naive Bayes classifier is a probabilistic statistical algorithm with robust independence features that is based on Bayes' theorem. A conditional probability theory is Bayes' theorem. It's also known as simple Bayes or independence Bayes.

7) Light GBM:

Light GBM is a framework based on decision trees that uses GOSS or Gradient-based one-side sampling and EFB or Exclusive Feature bundling. It is used for increasing the accuracy of the algorithm as well as improving memory usage. In light GBM, data and features are downsampled, using GOSS and EFB, to reduce complexity of the histogram building process. As it is based on decision trees, it uses trees and splits the trees based on leaf unlike other boosting algorithms where the tree grows level by level and the leaf does not change so there is less loss than other boosting algorithms.

Performance Evaluation:

During the testing process, performance was evaluated.The original dataset would be divided into training and test data, with 80% and 20% being used for training.When evaluating the classifier's performance on the testing dataset, four statistical numbers were used: the number of correctly identified positive data points (TP), the number Of correctly identified negative data points (TN), the number of negative data points labeled as positive by the classifier (FP), and the number of positive data points labeled as negative by the model (FN).

Accuracy measures model performance in terms of the number of accurate predictions

made by the model. Precision measures the positive rate of the model to the extent to which the model predicts the positive values and indicates the extent to which the model classifies the phishing URLs. The recall is the portion of positive data points labeled as such by the model among all truly positive data points. The F1 score is the harmonic mean of precision and recall, where the F1 score reaches its best value.

Results:

The internet is a massive network-based industry populated by hackers, attackers, and cyber criminals. Civilians, business people, industries, and every market that relies on the Internet and networks require security to prevent phishing and protect their customers, as well as to ensure the security of their own systems. In this study, we have learned the key components of phishing website detection using machine learning models such as Decision Tree, Random Forest, SVM, KNN, exploring the features and datasets used for training, the algorithms employed for classification, and the evaluation metrics that measure the model's effectiveness. The models have been evaluated using a variety of metrics, including accuracy, precision, recall, and F1 score.

S.NO	Author	Dataset	Method	Accuracy	Precision	F1-score	Recall
1	Shahrivari	Kaggle	XGBoost	98.32%	98.7%	97.6%	98.1%
2	Ahammad	Phish Tank	Light GBM	89.5%	-	-	-
3	Kulkarni	UCI	Decision Tree	91.5%	-	-	-
4	Kiruthiga	Phish Tank	Random Forest	98.4%	-	-	-
5	Manuel Sanchez	PIU-60k	TF-IDF	96.93%	96.57%	96.58%	96.93%
6	Aljabri	Malicious and Bengin Webpages Dataset	Naive Bayes	96.01%	96.5%	93.9%	92.2%
7	Junaid Rashid	UCI	SVM	95.66%	-	-	-
8	Mourtaji	Alexa	CNN	97.945%	-	98.591%	-

Discussion:

Machine learning for phishing website detection has the potential to be more accurate and effective than traditional methods such as blacklists or heuristic-based systems.Despite many significant advances in malicious URL detection using ML approaches over the last decade, there are still many critical and pressing unresolved problems and difficulties.Because of a lack of awareness about phishing, phishing attacks are more successful. Therefore, one of the main challenges is security, specifically how to encourage users to protect themselves against phishing.The data sample size was also a limitation. As a result, further evaluation and validation of ML models for detecting malicious URLs using sufficient samples with an acceptable ratio of normal and malicious URLs are required.More research is needed to investigate designing a lightweight model and running it on small computers in order to evaluate and improve the model's performance.

Conclusion:

With the increasing number of web domains, there has been an increase in the number of malicious URLs commonly used by cybercriminals to inject malicious code into victims devices, risking system confidentiality, integrity, and availability. As a result, there is a pressing need for detection methods to evolve and recognize the increasingly sophisticated methods used by attackers to target victims. Identifying ways to use intelligent methods to address this problem has grown into a significant research area. In conclusion, employing machine learning for detecting phishing websites represents a promising avenue in cybersecurity. The study focuses on feature engineering and the development of various models, including RF, NB, CNN, KNN, DT, and SVM, to analyze and compare in order to achieve a high level of classification accuracy. By leveraging advanced algorithms, machine learning can analyze patterns and behaviors to identify potential phishing threats. While these methods have shown success in recognizing known phishing tactics, it is important to acknowledge the ever-evolving nature of cyber threats. Continuous refinement and adaptation of machine learning models are essential to keep pace with the dynamic strategies employed by attackers. Despite some challenges, the integration of machine learning contributes significantly to enhancing the overall effectiveness of phishing detection, offering a valuable layer of defense in safeguarding users against deceptive online practices.

References:

1. Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. arXiv preprint arXiv:2009.11116.
2. Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, M. D. K. J. (2022). Phishing URL detection using machine learning methods. Advances in Engineering Software, 173, 103288.
3. Kulkarni, A. D., & Brown III, L. L. (2019). Phishing websites detection using machine learning.
4. Kiruthiga, R., & Akila, D. (2019). Phishing websites detection using machine learning. International Journal of Recent Technology and Engineering, 8(2), 111-114.

5. Manuel Sanchez-Paniagua, Eduardo Fidalgo Fernandez,Enrique Alegre,WesamAlNabki,Victor Gonzalez-Castro(2022).Phishing URL Detection: A Real-Case Scenario Through Login URLs,Journal Article.
6. Aljabri, M., Alhaidari, F., Mohammad, R. M. A., Samiha Mirza, Alhamed, D. H., Altamimi, H. S., & Chrouf, S. M. B. (2022). An Assessment of Lexical, Network, and Content-Based Features for Detecting Malicious URLs Using Machine Learning and Deep Learning Models. Computational intelligence and neuroscience, 2022, 3241216.
7. Singh, C. (2020, March). Phishing website detection based on machine learning: A survey. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 398-404). IEEE.
8. Abdul Karim,Mobeen Shahroz,Khabib Mustofa,Samir Brahim Belhaouari ,S.Ramana Kumar Joga(2023).Phishing Detection System Through Hybrid Machine Learning Based on URL.IEEE,vol. 11, pp. 36805-36822
9. Rasha Zieni,Luisa Massari,Maria Carla Calzarossa(2023).Phishing or Not Phishing? A Survey on the Detection of Phishing Websites.IEEE.
10. Ubing, A. A., Jasmi, S. K. B., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M.(2019). Phishing website detection: An improved accuracy through feature selection and ensemble learning. International Journal of Advanced Computer Science and Applications, 10(1).
11. Mahajan, R., & Siddavatam, I. (2018). Phishing website detection using machine learning algorithms. International Journal of Computer Applications, 181(23), 45-47.
12. Junaid Rashid,Toqeer Mahmood,Muhammad Wasif Nisar,Tahir Nazir(2020). Phishing Detection Using Machine Learning Technique.2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH).IEEE.
13. Mourtaji, Y., Bouhorma, M., Alghazzawi, D., Aldabbagh, G., & Alghamdi, A. (2021). Hybrid rule-based solution for phishing URL detection using convolutional neural network. Wireless Communications and Mobile Computing, 2021, 1-24.
14. SsDeshpande, A., Pedamkar, O., Chaudhary, N., & Borde, S. (2021).Detection of phishing websites using Machine Learning. International Journal of Engineering Research & Technology (IJERT), 10(05).
15. Alswailem, A., Alabdullah, B., Alrumayh, N., & Alsedrani, A. (2019, May). Detecting phishing websites using machine learning. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-6).IEEE.