|  |  |
| --- | --- |
| **Symbiosis Institute of Geoinformatics**  **Symbiosis International (Deemed University)**  5th Floor, Atur Centre, Gokhale Cross Road, Model Colony, Pune – 411016 |  |

**Name: Swachchha Roy**

**PRN: 23070243054**

**Internal Guide: Dr. Rajesh Dhumal**

**LITERATURE REVIEW**

**Title: Detection of Phishing Websites using Machine Learning**

**Literature Review**​

This document, titled **“A systematic literature review on phishing website detection techniques” by Asadullah Safi and Satwinder Singh** is a systematic literature review on phishing website detection techniques. It compares dissimilar approaches such as List Based, Visual Similarity, Heuristic, Machine Learning and Deep Learning. This review was performed on 80 scientific papers published in the last five years including algorithms, datasets and research questions related to phishing website detection. Machine learning techniques had the most applications with 57 studies using them. The PhishTank website was the primary source for phishing datasets which were explored in 53 studies while Alexa’s site was used for downloading legitimate datasets in 29 studies. In terms of Machine Learning methods Random Forest Classifier was used by thirty one articles. Convolution Neural Network (CNN) achieved the highest accuracy in detecting phishing websites at a rate of 99.98%.

"**Detecting Phishing Domains with Machine Learning" written by Shouq Alnemari and Majid Alshammari** is an article that focuses on the menace of phishing and suggests employing machine learning to identify them. The researchers used the UCI Phishing Domains dataset to evaluate four machine learning models; SVM, ANN, RF, and DT. The RF model has the highest detection accuracy at 97% of all other works in this area. It also hints on previous researches conducted on phishing domain detection models such as ensemble learning techniques and analysis of HTML page source code. More complicated machine learning algorithms for phishing domain detection are yet to be explored in future work.

Machine learning techniques are demonstrated in **Detecting Phishing Websites Using Machine Learning Aniket Garje, Namrata Tanwani, Sammed Kandale, Twinkle Zope, Prof. Sandeep Gore** document as a means of discovering phishing websites. It refers to phishing as a type of cyber-security threat which often involves theft of personal information like passwords and credit card numbers. Phishing website detection uses machine learning algorithms like KNN, Naive Bayes, Gradient Boosting, Decision Trees; the paper additionally contains an overview of various research papers that have been dedicated to this issue. The study also contains tables that show the confusion matrices for different algorithms used and their associated ROC curves. According to the conclusion drawn by the document, f1 score is best with decision trees being concluded as better for detecting phishing websites.

An overview of phishing detection techniques is given in the article **Phishing Detection: A Literature Survey Mahmoud Khonji, Youssef Iraqi, Senior Member, IEEE, and Andrew Jones**. It includes topics such as expanding user awareness and using extra software that are used to detect phishing attempts. In addition to this, it goes into list-based detection methods particularly through the use of whitelist and blacklist techniques; machine learning based detection approaches. The proposed approach improves accuracy of phishing detection by extracting and analyzing various characteristics on which suspects’ websites can be identified. For example, it introduces eight new features combined with existing ones to generate a feature vector for each webpage. During its detection phase, XGBoost classifier has been used for building strong classifiers for phishing detection purposes. The authors have also assessed how different features perform under the same classifiers thereby illustrating why their approach works better than others do within this context. The work ends by stating where the method does not work very well at present as well as possible future studies on identifying phishing attacks.

In the document **Applications of deep learning for phishing detection: a systematic literature review by Cagatay Catal, Görkem Giray, Bedir Tekinerdogan, Sandeep Kumar, Suyash Shukla** The document contains a systematics literature review (SLR) for the deep learning-based phishing detection techniques. References are categorized under regular or primary references. Regular references are derived from the basic research areas covering the DLS (Deep Learning Techniques based detection of Phishing). They are broadly classified as follows: a. Deep Learning b. Machine Learning a. Convolutional Neural Networks b. Recurrent Neural Networks c. Feature Engineering d. Detection Models Primary Studies: These studies covered the area of DLS based phishing Detection. They are pertaining to the above mentioned four regular references. Research Objectives: This part details out the research objectives of the SLR. This presents the method followed in this SLR.

The document **An effective detection approach for phishing websites using URL and HTML features by Ali Aljofey, Qingshan Jiang, Abdur Rasool, Hui Chen, Wenyin Liu3, Qiang Qu and Yang Wang** gives an overview of phishing detection techniques which are increasing awareness of the user and additional software. Also, list-based detection methods, as long as whitelist or blacklist methods, are discussed here and also machine learning-based detection techniques. Thus, the proposed technique suggests that in order for there to be better identification rates with regard to cases such as website forgery there should be some extraction and examination done on particular elements constituting those pages minus warning signs among others. To do this job well requires extraction of various features that together make up suspected webpages including those that have not yet surfaced formerly called novel document features belonging into each candidate set : liveness text in form fields based Image-only approach VKDZ distinctiveness. Introduce eight new features, combining them with those that already exist.

The article **Detecting phishing websites using machine learning technique by Ashit Kumar Dutta** considers a research study for detecting either malicious or legitimate URLs using Machine Learning methods, specifically the RNN-LSTM approach. ​ The proposed technique is called LURL. Compared with other state-of-the-art URL detectors, it had better accuracy and F1-score. ​ Details are provided in this document pertaining to methodology with equations and algorithms used in data collection, preprocessing, training, and testing. It also covers the history of phishing attacks, classification of phishing attack techniques, and related research works. At the end, the paper is concluded by presenting the need for ML-based anti-phishing techniques and specifying the future directions. Also, one can find the reference list on Phishing Detection and Machine Learning with respect to URL-based phishing detection, machine learning algorithms, feature selection techniques, and use of deep learning for malicious URL detection.

The document **Phishing Websites Detection Using Machine Learning** by **P. Amba Bhavani, Chalamala Madhumitha, Pinnam Sree Likhitha, Chanda Pranav Sai**  find the phishing URL by comparing different machine learning algorithms in terms of accuracy, false positive and false negative. The document further includes a literature survey, detailing how the respective methods used till now i.e., online toolbars, data mining algorithms and deep learning approaches. Results section will provide the accuracy of CNN LSTM, CNN Bi-LSTM, Logistic regression and XGBoost machine learning models. This discussion will shed a light on shortcomings of previous methods and benefit using machine learning-based approaches. ​

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr. No. | Author | Dataset | Methods | Result | Publish Year |
| 1 | Asadullah Safi, Satwinder Singh | PhishTank website,  Alexa website | Machine Learning techniques, Heuristic techniques, Visual Similarity, CNN. | Convolutional Neural Network (CNN) achieved the highest accuracy of 99.98% in detecting phishing websites. | 11th January 2023 |
| 2 | Shouq Alnemari, Majid Alshammar | Phishing dataset from UCI machine learning repository | Artificial Neural networks (ANN), SVM, Decision Tree classifier, Random Forest. | Random forest achieved 97%. | 2023 |
| 3 | Aniket Garje, Namrata Tanwani, Sammed Kandale, Twinkle Zope, Prof. Sandeep Gore | Phishing dataset from UCI machine learning repository | Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest | The Random Forest algorithm showed the highest accuracy of 98.4%, 98.59% recall, and precision of 97.70% | September 2023 |
| 4 | Mahmoud Khonji, Youssef Iraqi, Senior Member, IEEE, and Andrew Jones | Kaggle | machine learning and clustering algorithms (k-NN, C4.5, SVM, k-means, DBSCAN), similarity-based detection, Bayesian approach, hybrid features, Google Safe Browsing API, Bayesian classification, NLP techniques | Automatic Detection of Phishing Target from Phishing Webpage: FP rate of 3.4% and FN rate of 8.56%. ​  Detecting DNS-poisoning-based phishing attacks from their network performance characteristics: FP rate of 0.7% and FN rate of 0.6%. ​  Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach: FP rate of 0-0.02% and FN rate of 0-1.95%. ​ |  |
| 5 | Cagatay Catal, Görkem Giray, Bedir Tekinerdogan, Sandeep Kumar,  Suyash Shukla | 18 different datasets for deep learning-based phishing detection, with PhishTank being the most used dataset | Deep Neural Networks (DNN),  Convolutional Neural Networks (CNN),  Recurrent Neural Networks (RNN),  Long Short-Term Memory Networks (LSTM),  Autoencoders  Generative Adversarial Networks (GANs)​ | Achieved an accuracy ranging from 94% to 99.34%. | 2020 |
| 6 | Ali Aljofey,  Qingshan Jiang,  Abdur Rasool,  Hui Chen,  Wenyin Liu,  Qiang Qu,  Yang Wang | The dataset used in the study includes 60,252 webpages, with 32,972 benign webpages and 27,280 phishing webpages. | URL character sequence features, various hyperlink information, and textual content of the webpage to train the XGBoost classifier. The features are extracted from the HTML source code and URL without relying on third-party services. | The proposed approach achieved an accuracy of 96.76% with a false-positive rate of 1.39% on the custom dataset. | 2022 |
| 7 | Ashit Kumar Dutta | PhishTank and Crawler dataset | Utilizes a combination of URL and HTML features evaluated through various machine learning classifiers. | The proposed method achieved accuracy of 98.48%. | 2022 |
| 8 | P. Amba Bhavani, Chalamala Madhumitha, Pinnam Sree Likhitha, Chanda Pranav Sai | Kaggle | CNN LSTM, Logistic regression | CNN LSTM 57.85% and Logistic Regression 91.89% |  |

Proposed Methodology:

**Dataset Collection and EDA:** A large collection of false and legitimate URLs are to be collected and perform necessary EDA and preprocessing to clean and prepare the data for further processing.

**Feature Extraction:** Identify and extract relevant features from websites such as Address Bar-based features, Domain based features and HTML based features which can help in distinguishing the phishing websites from the legitimate ones.

**Model development:** Develop and train model based on the preprocessed data using machine learning algorithms like Decision tree, random forest, SVM and others. Experiment with different algorithms to determine the most effective approach.