International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

PhishNull: Enhancing Cyber Hygiene Through Supervised Machine Learning

Purva Kulkarni[1], Siddhi Jadhav[2], Tanya Gupta[3], Sangeeta Mishra[4]

[1,2,3]Student, Department of Electronics and Telecommunication, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India
[4]Associate Professor, Department of Electronics and Telecommunication, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India

Abstract
The research paper addresses the pervasive issue of phishing within the realm of Internet security, acknowledging its persistence despite the advancements in antivirus and technical safeguards. Focusing on combating this online scam, the study delves into two primary methodologies: Black Listing and Machine Learning. Opting for a Machine Learning and heuristic-based approach, the thesis conducts a comparative analysis of various Machine Learning algorithms, including Logistic Regression, alongside ensemble algorithms such as Adaboost and Gradient Boost. While initial expectations leaned towards ensemble algorithms yielding superior results, the outcomes revealed a nuanced reality. Although ensemble algorithms demonstrated promising predictive capabilities, their performance did not surpass expectations.

Keywords: Phishing, Internet security, Machine Learning, Black Listing, Heuristic-based approach, Logistic Regression, Ensemble algorithms, Adaboost, Gradient Boost, Comparative analysis.

1. Introduction

Phishing, a prevalent form of cybercrime, continues to pose significant threats to Internet users despite advancements in security measures. Traditional approaches such as blacklisting struggle to keep pace with the evolving tactics of malicious actors. As the sophistication of phishing attacks increases, researchers have turned to innovative solutions, with Machine Learning emerging as a promising avenue for detection and prevention [1]. This paper explores the efficacy of Machine Learning, coupled with heuristic-based approaches, in combating phishing attacks. By examining the performance of various Machine Learning algorithms, including logistic regression and ensemble methods such as Adaboost and Gradient Boost, this study aims to shed light on the effectiveness of different approaches in mitigating the risks posed by phishing [2].

The evolution of phishing detection methodologies reflects a shift towards leveraging advanced technologies. Abbasi, Zhang, and Chen (2019) highlight the application of statistical learning in fake website detection, while Yazhmozhi (2019) introduces a phishing website detection system based on Natural Language Processing (NLP) and Machine Learning [3]. Furthermore, Zhu, Chen, and Ye (2019) propose OFS-NN, an effective phishing website detection model integrating optimal feature selection and Neural Network techniques [4]. These studies underscore the growing recognition of Machine Learning as a powerful tool in identifying and mitigating phishing threats.

The literature surrounding phishing detection underscores the multifaceted nature of the challenge and the diverse array of solutions proposed [1]. Abbasi, Zhang, and Chen (2019) emphasize the role of statistical learning in discerning fake websites, showcasing the potential of data-driven approaches [2]. Similarly, Yazhmozhi (2019) explores the synergy between NLP and Machine Learning, demonstrating the effectiveness of linguistic analysis in identifying phishing attempts [3]. Additionally, Zhu, Chen, and Ye (2019) contribute to the literature by introducing OFS-NN, a model that prioritizes feature selection to enhance the accuracy of phishing detection algorithms [4]. These studies collectively illustrate the evolving landscape of phishing detection methodologies, with Machine Learning at the forefront of innovation.

2. Theory

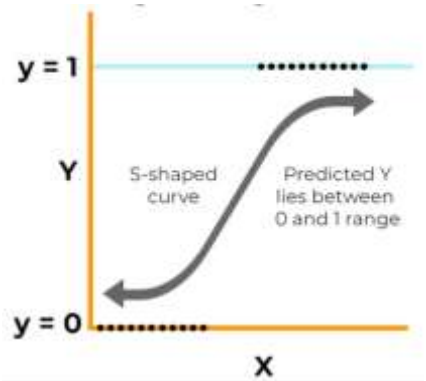
Machine Learning algorithms, characterized by their ability to learn from data and make predictions or decisions, offer a data-driven approach to phishing detection. This section explores the theoretical foundations of various Machine Learning algorithms utilized in the detection and prevention of phishing attacks. Specifically, it examines Logistic Regression, Adaboost, GradientBoost, and Stacking, elucidating their principles, mechanisms, and applications in the context of phishing detection.

Understanding the theoretical underpinnings of these algorithms is essential for evaluating their efficacy in identifying phishing attempts accurately. By elucidating the workings of these algorithms, this section aims to provide insights into their strengths, limitations, and potential contributions to enhancing cybersecurity measures against phishing attacks.

A. Logistic Regression:

Logistic Regression, a fundamental statistical method for binary classification tasks, models the probability of a binary outcome based on one or more independent variables [9]. Employing the logistic function, it maps the linear combination of features to a probability between 0 and 1, making it ideal for binary classification tasks such as phishing detection. In this context, logistic regression predicts the likelihood of a website being malicious based on various extracted features, demonstrating interpretability and effectiveness in discerning between malicious and benign entities. Additionally, logistic regression serves as a key component within ensemble learning frameworks like stacking, contributing to the overall predictive accuracy by combining its predictions with those of other base classifiers [10]. Research by Abbasi, Zhang, and Chen (2019) showcases the application of logistic regression as a component of statistical learning systems for detecting fake websites, underscoring its interpretability and effectiveness in identifying malicious online entities [11].

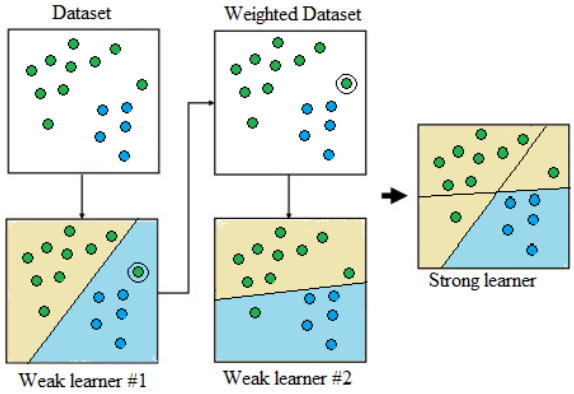
Figure 1: Illustration of Working of Logistic Regression



B. AdaBoost Classifier:

Adaboost, short for Adaptive Boosting, is an ensemble learning technique that combines multiple weak classifiers to create a strong classifier. It sequentially trains a series of weak learners, each focusing on the instances that were misclassified by the previous learner. By giving more weight to the misclassified instances, Adaboost iteratively improves the model's performance, ultimately achieving higher accuracy than any individual weak learner [3]. In the context of phishing detection, Adaboost can be applied to combine the predictions of various base classifiers, such as decision trees or logistic regression models, to enhance the overall accuracy of detecting phishing websites [4].

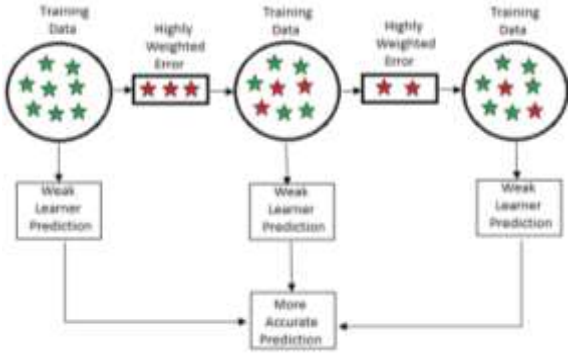
Figure 2: Illustration of Working of AdaBoost Classifier



C. GradientBoost Classifier:

GradientBoost, similar to Adaboost, is an ensemble learning method that builds a strong predictive model by sequentially combining multiple weak learners. However, instead of adjusting the weights of misclassified instances, GradientBoost minimizes a loss function by iteratively fitting new models to the residuals of the previous models. By focusing on minimizing the errors of the previous models, GradientBoost gradually improves the overall predictive accuracy [5]. In phishing detection, GradientBoost can be utilized to combine the predictions of different base classifiers, leveraging their collective strengths to accurately identify malicious websites [6].

Figure 3: Illustration of Working of GradientBoost Classifier

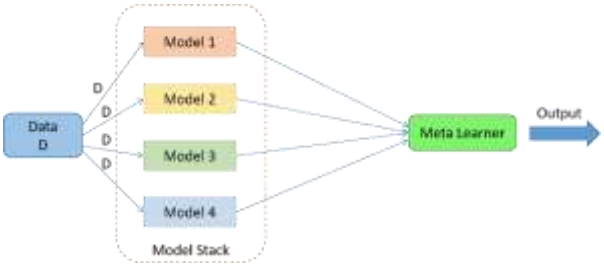


D. Stacking Classifier:

Stacking, also known as stacked generalization, is an ensemble learning technique that combines multiple base classifiers using a meta-learner. Instead of simply averaging the predictions of the base classifiers, stacking trains a meta-learner to learn how to best combine their outputs. The meta-learner takes the predictions of the base classifiers as input features and learns to generate a final prediction

based on their collective insights [7]. In the context of phishing detection, stacking can be applied to integrate the predictions of various classifiers, such as linear regression, Adaboost, and GradientBoost, to create a more robust and accurate detection model [8].

Figure 4: Illustration of Working of Stacking Classifier



3. Methodology

This study employs a systematic methodology to detect and categorize malicious URLs, focusing on phishing attacks. The project flow is delineated into several sequential steps, starting from data collection to model training and performance evaluation.

Technology Stack Used:

1. Front End Development:

- HTML
- CSS
- JavaScript
- Visual Studio Code

2. Model Training:

- Google Colab (Python)

3. Integration

- Flask (Python)
- Visual Studio Code

A. Dataset Collection and Preprocessing:

- The dataset comprises 1000 URLs from each of the two classes: benign and phishing URLs.
- Features are extracted from the URLs using libraries such as urllib and whois, yielding 19 informative features and a target class label.
- Extracted features are saved into a CSV file for further processing.

B. Feature Extraction:

- Features extracted from the URL data encompass various categories:
  1. Address Bar-based Features
  2. Domain-based Features
  3. HTML & Javascript-based Features (content-based features)
- Each category of features contributes to the detection and classification of malicious URLs, providing valuable insights into the characteristics of phishing attacks.

C. Library Setup and Data Loading:

- Python libraries including pandas, matplotlib, scikit-learn, seaborn, and numpy are imported and installed.

- The dataset is loaded into a pandas DataFrame for data manipulation and analysis.

D. Dataset Cleaning and Preprocessing:

- Data cleaning involves handling null values and removing unnecessary columns to ensure data integrity.
- Categorical data are converted into numerical format through encoding techniques to facilitate model training.

E. Exploratory Data Analysis (EDA):

- Data visualization techniques are employed to gain insights into the distribution and relationships among features.
- Graphs and plots are created to visualize patterns and identify potential correlations between features and the target class.

F. Data Splitting:

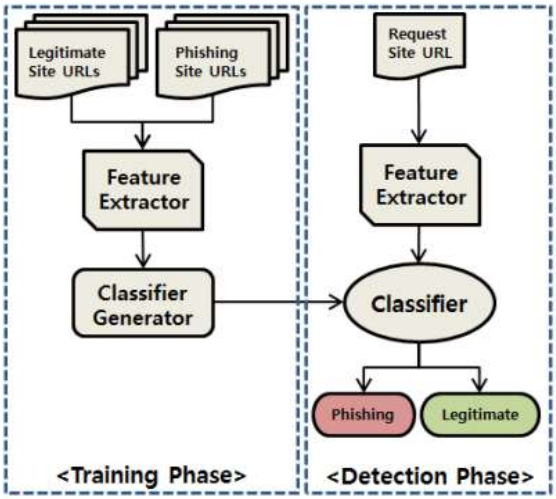
- The dataset is split into training and testing sets with a ratio of 90:10, ensuring model generalizability and performance evaluation on unseen data.

G. Model Training:

- Four machine learning models are trained on the dataset:

1. Logistic Regression
2. AdaBoost Classifier
3. Gradient Boosting Classifier
4. Stacking Classifier (comprising a combination of GradientBoosting Classifier and AdaBoostClassifier, with a meta-classifier LogisticRegression).

Figure 5: Flow of the Project



During the training phase, a classifier is generated with the help of a dataset consisting of URLs for both phishing and legitimate websites. This collection of URLs is passed on to the feature extractor. The feature extractor's job is to extract all features from these URLs. The feature extraction process depends on the features selected for the feature extractor. These extracted features serve as input and are passed to the classifier generator. The classifier generator then generates a classifier using this newly generated input and a selected machine learning algorithm. During the detection phase, when a website is visited, its URL is transmitted to the feature extractor. The feature extractor extracts the required features from

the URL of the currently visited website. These extracted features are then transmitted to the classifier. Based on the knowledge gained from its previous training, the classifier makes a decision regarding whether the website is legitimate or not. It then displays a pop-up to the user based on its results

4. Result & Discussion

The performance of various classifiers was evaluated using a dataset consisting of URLs for both phishing and legitimate websites. The classifiers were trained using features extracted from these URLs, and their performance was assessed based on metrics such as accuracy, precision, recall, and F1-score. The table below summarizes the results obtained:

Table 1: Comparison Between Algorithms

Model	Test Accuracy	Precision (Class 2)	Recall (Class 2)	F1-Score (Class 2)
Logistic Regression	90.5%	0.91	0.91	0.90
AdaBoost Classifier	89.5%	0.91	0.90	0.89
GradientBoost Classifier	90%	0.92	0.90	0.90
Stacking Classifier	91.5%	0.92	0.92	0.91

The table illustrates the performance of each classifier in terms of accuracy and metrics specifically related to the detection of phishing websites (Class 2). Among the classifiers evaluated, the Stacking Classifier achieved the highest test accuracy of 91.5% and the highest F1-score for Class 2 at 0.91.

The final output of the project includes the deployment of the Stacking Classifier for real-time website classification. Snippets of the final output showcase the classifier's ability to accurately classify websites as either legitimate or phishing based on their URLs. Users are provided with pop-up notifications indicating the classification results, enabling them to make informed decisions while browsing the web.

Figure 6: Phishing URL Detected Successfully

Figure 7: Benign URL Detected Successfully

These results demonstrate the effectiveness of the Stacking Classifier in accurately distinguishing between phishing and legitimate websites, thereby enhancing user security and privacy during web browsing. However, further research and optimization may be warranted to explore additional features and improve classifier performance further.

5. Conclusion & Future Scope

In conclusion, the evaluation of various classifiers for the detection of phishing websites based on URL features has provided valuable insights into their performance and effectiveness. The Stacking Classifier emerged as the top-performing model, achieving a high test accuracy of 91.5% and demonstrating robust precision, recall, and F1-score for identifying phishing websites. These results underscore the importance of utilizing machine learning techniques in combating cyber threats and enhancing web security. By leveraging sophisticated algorithms and feature extraction methods, it is possible to develop reliable systems capable of effectively distinguishing between legitimate and malicious websites, thereby safeguarding users from potential cyberattacks.

Looking ahead, there are several avenues for further research and development in the field of website classification and cybersecurity. One potential area of exploration is the integration of more advanced feature extraction techniques, such as deep learning-based approaches, to capture subtle patterns and nuances in website URLs. Additionally, incorporating contextual information and behavioral analysis into the classification process could enhance the accuracy and robustness of detection systems, enabling them to adapt to evolving cyber threats. Furthermore, exploring ensemble methods and hybrid approaches that combine multiple classifiers could lead to even more powerful and resilient detection systems. By continuing to innovate and refine our methodologies, we can stay ahead of cyber adversaries and ensure a safer and more secure online experience for users worldwide.

6. References

1. Abbasi, Z. Zhang, and H. Chen, "A statistical learning based system for fake website detection," in Proceedings of the IEEE.
2. V. Yazhmozhi, "Natural language processing and Machine learning based phishing website detection

system," in Proceedings of the IEEE, 2019.
3. E. Zhu, Y. Chen, and C. Ye, "OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network," in Proceedings of the IEEE, 2019.
4. A. Abunadi, O. Akanbi, and A. Zainal, "Feature extraction process: A phishing detection approach," in Proceedings of the IEEE, 2013.
5. M. Aburrous, M. Hossain, K. Dahal, and F. Thabt, "Predicting phishing websites using classification mining techniques with experimental case studies," in Proceedings of the IEEE, 2010.
6. A. Alswailem, B. Alabdullah, N. Alrumayh, and D. Alsedrani, "Detecting Phishing Websites Using Machine Learning," in Proceedings of the IEEE, 2019.
7. M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," in Proceedings of the IEEE, 2015.
8. M. Aydin, I. Butun, K. Bicakci, and N. Baykal, "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs," in Proceedings of the IEEE, 2020.
9. M. Chatterjee and A. Siami, "Detecting Phishing Websites through Deep Reinforcement Learning," in Proceedings of the IEEE, 2019.
10. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the SIGKDD Conference.
11. A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine learning," in Proceedings of the IEEE, 2017.
12. K. Firdous, "Hybrid Client Side Phishing Websites Detection Approach," in Proceedings of the IEEE, 2014.
13. X. Han, N. Kheir, and D. Balzarotti, "Phisheye: Live monitoring of sandboxed phishing kits," in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2016.
14. J. Hong, "The state of phishing attacks," Communications of the ACM, 2012.
15. Y. Huang, Q. Yang, and J. Qin, "Phishing URL Detection via CNN and Attention-Based Hierarchical RNN," in Proceedings of the IEEE, 2019.
16. A. K. Jain and B. Gupta, "Phishing detection: Analysis of visual similarity based approaches," in Security and Communication Networks, vol. 2017.
17. A. K. Jain and B. B. Gupta, "Comparative analysis of features based machine learning approaches for phishing detection," in Proceedings of the IEEE, 2017.
18. J.-L. Lee, D.-H. Kim, and C.-H. Lee, "Heuristic-based Approach for Phishing Site Detection Using URL Features," in Proceedings of the IEEE, 2015.
19. N. Megha and K. R. Babu, "An Intelligent System for Phishing Attack Detection and Prevention," in Proceedings of the IEEE, 2019.
20. T. Nathezhtha, D. Sangeetha, and V. Vaidehi, "WC-PAD: Web Crawling based Phishing Attack Detection," in Proceedings of the IEEE, 2019.
21. N. S. and V., "WC-PAD: Web Crawling based Phishing Attack Detection," in Proceedings of the IEEE, 2019.
22. N. and L. A. Tuan, "A novel approach for phishing detection using URL-based heuristic," in Proceedings of the IEEE, 2014.

23. A. Oest, Y. Safaei, A. Doupe´, and G.-J. Ahn, "PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques Against Browser Phishing Blacklists," in Proceedings of the IEEE, 2019.
24. A. Park, R. N. Quadari, and H. H. Tsang, "Phishing Website Detection Framework Through Web Scraping and Data Mining," in Proceedings of the IEEE, 2017.
25. S. Patil and S. Dhage, "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework," in Proceedings of the IEEE, 2019.