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

Phishing Attack Using Ensemble Machine Learning

Padmaja Sunil Mengane[1], Dr. Sarika Jadhav[2]

[1,2]Computer engineering/Padmabhooshan Vasantdada Patil Institute of Technology (Pvpit)/Savitribai Phule Pune University/India

Abstract

The number of phishing websites and other new kinds of URL-based phishing threats continues to increase as more people get online. What are we supposed to do? These URL-based threats are phishing websites that use URLs to lure careless or unsuspecting online users and gather sensitive information like usernames and passwords or other personal and financial information. As the number of these phishing websites increase, existing traditional detection systems (e.g., blacklists) are not able to timely and effectively detect these new generation phishing sites. There is a need to devise smarter mechanisms that are more active and adaptive to counter the changing trendy dangerous behaviors.

In this study, we build a model that uses some new, clever, and innovative detection mechanisms to solve phishing detection based on the URL of websites. URL-based phishing detection will be based on some features extracted using Natural Language Processing and some clever detection will be based on some supervised learning techniques. Of these Random Forest and AdaBoost are two of the chosen to figure out the best of the two when compared. Initially, data is collected from external data repositories (for example, Kaggle and UCI). Then preprocessing steps (data cleaning, data monumenting, and removal of some stop words) based on some clever supervised learning techniques (with mostly the stemming process by using some Porter) are employed. Feature extraction and reduction are used to eliminate redundancy and minimize the high-dimensional data and, therefore, the high overload of the classifiers. Smart detection will use a lot of clever and adaptive supervised learning techniques, and will be based on optimization of features. These will be evaluated using some confusion matrix metrics accuracy, precision, recall, and the F-measure. Finally, based on the experiments the Random Forest classifiers perform better and achieve some 93.89% accuracy as compared to AdaBoost which achieves only some 92.67% accuracy. The study concludes that hybrid machine learning techniques outperform traditional methods regarding the efficiency of detecting phishing URLs. The system that was proposed is scalable and dependable, allowing for the detection of real-time phishing, and is reflected in an improvement for user safety and a reduction in threats.

Keywords: Phishing Detection, Machine Learning, Random Forest, AdaBoost, NLP, URL Classification

I. Introduction

Phishing is a common form of cyber crime today that exists somewhat illegally because it allows the thief to gather information using a trick. The trick is creating a false website that resembles a legitimate website, making the user unknowingly give away sensitive information that could be used to access bank

accounts to steal a person's money. These attacks are increasing rapidly due to the ease of generating fake URLs and the lack of awareness among users. [1]

Currently, the methods for detecting phishing attacks with a black list as well as heuristic-based methodologies have failed to provide a clear picture of phishing URLs that are recently created. A blacklist, which is primarily made of blacklisted websites as a result of attacks, works due to a previously created black website, and thus does not work due to blacklisted websites of 'today.' [2] A heuristic system is a method that is based on a black or white rule, and is not subject to a black or white rule, as well as a changing system. [3] Because of these factors, more adaptable, and intelligent systems are being developed to allow phishing URLs to be detected with a certain degree of sophistication and with little reliance on users.

Ml, or machine learning, has become one of the most effective and powerful systems, and for this reason, almost all aspects of phishing detection rely on learning. [4] Utilizing the algorithms employed in supervised learning, which are breaking borders in differentiated cyber security, are used in the detection of phishing [5]. Using these supervised algorithms, with the understanding of blacklisted URLs and white listed URLs of the databases being used, features were created, which in turn, generated a rationale for the system to classify and resolve the issue.

Studies have shown that combining different methods of machine learning provides enhanced accuracy [6]. The selection and elimination of redundant and unneeded information greatly aids the improvement of model performance [7]. Many of these techniques have explored the feature extraction process and the use of advanced clustering methods, as well as the use of genetic algorithms [8], [9]. Scanning the field of deep and learning casts even greater improvements on the aegis of the detection of attempts of phishing through the capabilities of the identification of patterns of increased complexity [10]. In order to move forwards with the objectives sustained by this theory, the authors have proposed and presumed the use of Random Forest and AdaBoost, which are complete and efficient machine learning frameworks, to meet the objectives of the described treatment.

The system is expected to achieve high accuracy and reliability through the processes of classification, feature extraction, and the deep learning methods of preprocessing and classification. The system also anticipates the difficult task of achieving detection of phishing through the framing of the system to solve a binary classification problem; the two classes being abundant URLs and deficient URLs..

II. LITERATURE SURVEY

According to [1] introduces a hybrid framework for detecting phishing attacks embedded in deceptive online advertisements. The authors combine multiple machine learning models to improve accuracy and minimize false positives. The approach analyzes both URL-based and content-based attributes to identify suspicious ad patterns. Experimental outcomes show the hybrid model’s superior detection capability compared to traditional classifiers. The system demonstrates robustness in identifying evolving phishing strategies in online advertising.

According to [2] a machine learning-based strategy for identifying phishing attacks with high reliability. A variety of classification algorithms were evaluated using datasets containing legitimate and malicious URLs. The results indicate that ensemble-based models provide better detection performance than standalone algorithms. The research emphasizes the significance of data preprocessing and feature extraction in boosting detection accuracy. The approach offers a scalable solution adaptable to real-world cybersecurity applications.

According to [3] integrates natural language processing (NLP) and deep learning for phishing attack detection. It focuses on analyzing the linguistic features of emails and web content to uncover deceptive intent. The study employs deep neural networks to understand contextual patterns that traditional methods often miss. Results show that NLP-driven models can outperform classical machine learning classifiers. The method enhances text-based phishing detection through automated semantic understanding.

According to [4] authors explore a machine learning-based solution to identify phishing attacks effectively. The system relies on supervised algorithms trained on features derived from phishing and legitimate websites. By applying feature selection and classification techniques, the model achieves high detection accuracy. The paper highlights how ML approaches outperform static rule-based systems. The research underscores the adaptability of machine learning in handling dynamic phishing trends.

According to [5] presents a comprehensive analysis of phishing detection techniques leveraging machine learning. It categorizes existing models based on the types of features and algorithms utilized. The study compares the strengths and weaknesses of different ML methods, including decision trees, SVM, and deep learning. It identifies major challenges like data imbalance, feature redundancy, and evolving phishing tactics. The paper concludes by suggesting hybrid and ensemble models as promising future directions.

According to [6] evaluates various machine learning classifiers and feature sets to improve phishing attack detection. It systematically tests algorithms such as Random Forest, SVM, and Gradient Boosting on multiple phishing datasets. The findings reveal that optimal feature selection significantly boosts performance and reduces overfitting. The study also explores the trade-offs between detection speed and accuracy. The authors propose an enhanced ML pipeline that balances efficiency and robustness.

According to [7] work develops a machine learning-based framework for detecting phishing websites efficiently. The approach extracts critical URL and page-level features to train models that differentiate legitimate sites from fraudulent ones. Several classifiers are tested, and ensemble methods show the best performance. The paper also analyzes the scalability of the approach for large-scale deployment. The results validate the model's practical applicability in real-world internet security systems.

According to [8] authors enhance phishing URL detection by integrating a genetic algorithm for feature selection with machine learning classifiers. Their approach automatically identifies the most influential features that contribute to phishing patterns. Experiments conducted on benchmark datasets reveal that the optimized feature subset improves classification accuracy. The study demonstrates that evolutionary algorithms can efficiently reduce computational costs. This hybrid GA-ML system offers a more adaptive and precise detection framework.

According to [9] paper presents an improved K-Means clustering algorithm tailored for phishing attack detection. The enhanced algorithm refines the clustering process to better separate malicious and legitimate data points. By combining unsupervised learning with optimized distance metrics, the method achieves higher detection reliability. The study also benchmarks its performance against standard clustering techniques. Findings confirm that the modified K-Means approach provides superior precision and recall in phishing identification.

According to [10] employs deep learning models for detecting phishing websites through automated feature extraction. CNN and LSTM architectures are utilized to learn intricate relationships in URL and web page data. The approach eliminates manual feature engineering, enabling end-to-end phishing classification. Experimental evaluations on diverse datasets confirm remarkable accuracy and

generalization ability. The research concludes that deep learning significantly advances the state-of-the-art in phishing detection.

III. METHODOLOGY

This part of the research describes the URL Detection Workflow of the phishing website detection system. The detection system is constructed using the hybrid machine learning technique, which combines the Natural Language Processing (NLP) techniques for URL classification into Legitimate or Phishing via a supervised learning technique. The entire process is divided into the following well defined stages: system architecture, data collection, data preprocessing, feature extraction, feature selection, model training, classification, and evaluation of model performance. These stages enhance the system's detection accuracy and efficiency.

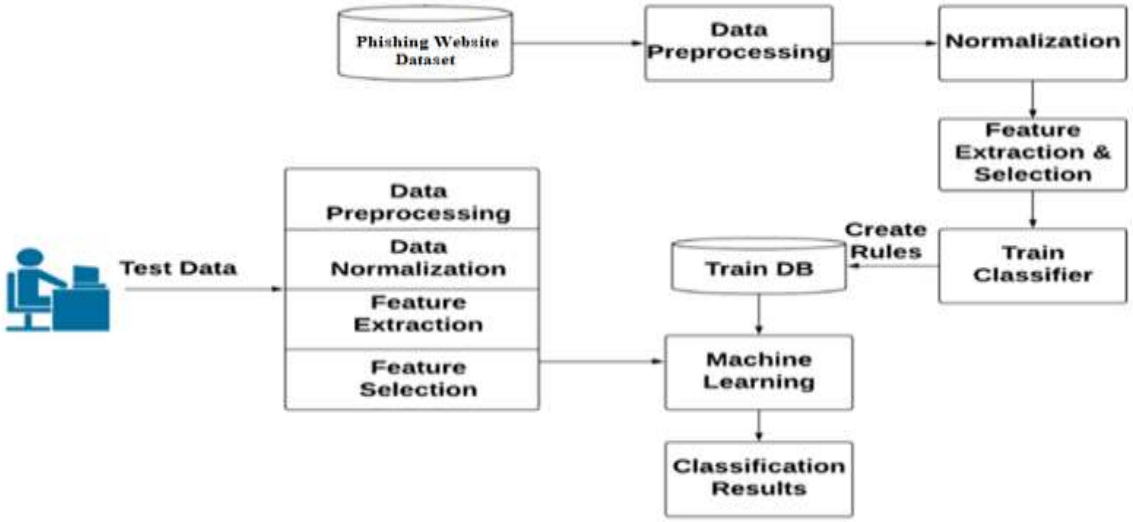

Figure 1 System Architecture

Data Collection

Data collection serves as the starting point in the development of the proposed system. To build an accurate and trustworthy phishing detection model, high-quality datasets are crucial. To develop an effective phishing detection model, different datasets have been collected from various credible sources in this research.

1. The Kaggle Phishing Dataset
2. The UCI Machine Learning Repository
3. Real-time URL datasets from crawling different sources over the web

Data Preprocessing

Data preprocessing falls among the most integral parts of the machine learning model building process. The raw URL datasets have considerable noise, presence of missing entries, and inconsistencies that would drastically deteriorate the overall performance of a model. URL datasets that are raw need to be, through preprocessing, made to be as clean and as structured as to be eligible for feature extraction and subsequently be used for URL classification.

Feature Extraction

Feature extraction aims to identify useful markers from the given raw URL data. These markers should ideally help one distinguish between phishing websites and authentic ones. Natural Language Processing (NLP) techniques are employed to offer numerical values for URL text, which are transformed into other learning model features.

Important URL features include:

- URL length
- Special characters
- Domain age and registration
- HTTPS and security
- Keywords like “login”, “verify”, “update”, “secure”, and others

Indicators of phishing schemes in a URL include a mix of high URL length, unsettling special characters, and keywords that are imitative of a legitimate site.

Feature Selection

Feature selection is a vital step that optimizes a model's performance by selecting the most meaningful and critical features and removing the rest. This decreases the complexity of the computations and increases the accuracy of the classification.

The benefits of feature selection include:

- Better accuracy of the model
- Reduced training time
- Lower cost of computations
- Improved ability to generalize

Implemented Feature Selection techniques involve:

- Threshold-Based Selection
- Hybrid Selection Approach

Implemented Feature Selection techniques help the classification stage by removing the features with little meaning and by improving the optimization and performance of the model.

Classification Algorithms

The formation of a system involves the application of different classification algorithms of which the machine learning algorithms categorize the URL’s under Phishing and legitimate.

Random Forest (RF)

Random forests use multiple decision trees to provide the final classification through voting. It is one of the most preferred algorithms since it is multivariate and handles more data in classification.

- Uses multiple decision trees
- Reduces the overfitting phenomenon
- Provides more accuracy and stability
- Handles missing data effectively

AdaBoost

This algorithm of boosting combines more of the weak classifiers to provide a single classification through voting. Once more, AdaBoost tends to pay more attention to the instances that have been unsatisfactorily classified.

- Provides a more combined view of multiple weak learners
- Amplifies the attention-to- focus wrap and present class
- Finer Accuracy
- Noisy data responsive

The choice for these algorithms is due to their criteria of being able to provide the more efficient approach to Phishing detection.

Performance Evaluation Metrics

To the effectiveness of the classification model, the traditional evaluation metrics based on the confusion matrix have been employed for the proposed model.

Accuracy: The number of true predictions divided by the total number of predictions made gives the measure of accuracy.

Precision: Out of the predicted phishing URLs, the fraction of them which are actually phishing is denoted by precision.

Recall: Recall is the measure of fraction of actual phishing URLs which have been correctly tagged.

F-measure: The balanced score of a model is denoted by the F-measure which is the harmonic mean of the precision and the recall.

IV. RESULTS

We evaluate how well a phishing website URL detection system does by testing two ML algorithms, Random Forest (RF) and AdaBoost, by standard measurement criteria such as accuracy, precision, recall, and F1-score. With these measurement criteria, it is especially clear that both models are useful for detecting phishing URLs. It is also indicated that Random Forest, in comparison to ADA Boost, performs better for most of the measurement criteria. This comparison also gives insights on what algorithm may be best employed for a real-time phishing detection system.

Table 1: The performance results of the implemented models are summarized below

Algorithm	Accuracy	Precision	Recall	F1-score
RF	93.89%	0.893	0.994	0.941
AdaBoost	92.67%	0.536	0.997	0.697

Table 1 shows how well Random Forest (RF) and AdaBoost algorithms perform in phishing detection of websites in terms of accuracy, precision, recall, and F1 score. RF looks to have the best overall performance, achieving 93.89% and seeing accuracy outperform the other algorithms while achieving a large majority of URLs in the dataset. F1 shows high precision of 0.893, earning the model trust of users as the majority of phishing URLs could be detected. F1 of 0.994 is high and suggests the model hardly misses any phishing URLs. The F1 score was 0.941 meaning that Random Forest shows a large performance between precision and recall, indicating that the model is both stable and deservedly trusted. Looking at the performance of AdaBoost in this context, the scores show slightly poorer results,

achieving accuracy of 92.67% and performing overall slightly poorer than Random Forest. Due to that, the F1 score of 0.536 was low, indicating that legitimate URLs are more likely to be incorrectly classified as phishing URLs. While AdaBoost's recall was balanced at 0.997, maintaining high sensitivity and detecting almost all phishing URLs, the F1 score of 0.697 was low as it shows ongoing poor performing results.

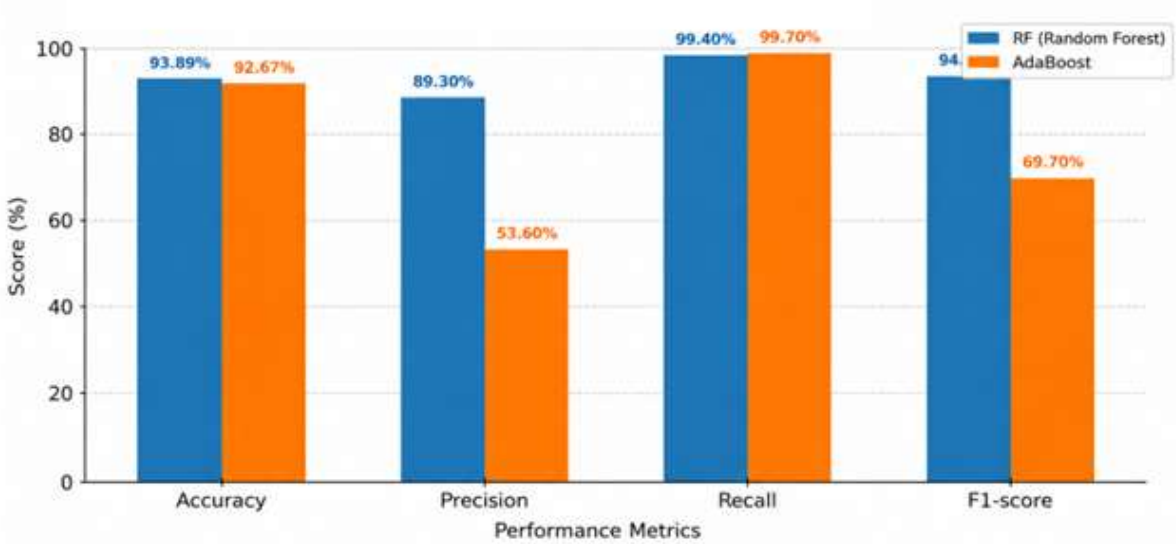

Figure 2: Comparison of Random Forest and AdaBoost Performance Metrics for Phishing URL Detection

Figure 2 shows a contrast analysis of the performance of URL phishing web pages of two of the popular classifiers of machine learning algorithms, Random Forest (RF) and AdaBoost. The charts are categorized as multiple bar chart focus on graphs consisting of two measured classifiers, and for the performance of charts, the main focus is on the graphs that are measured for accuracy, precision, recall, and the F-measure. The X-axis shows the measured classifiers, Random Forest and AdaBoost. The Y-axis is used for the measured metrics, and for Precision, Recall, and F-measure, the range is set between 0-1, and for Accuracy, a percentage, the maximum range is set at 100. Charted results with Random Forest show that all measured metrics rendered results between 0-1 with Accuracy at 93.89%, Precision at 0.893, Recall at 0.994, and the F-measure at 0.941. These results show that Random Forest performance is well balanced, having high Detection Capability with low Error Rates. The charted results with AdaBoost showed Accuracy at 92.67%, Precision at 0.536 (very low), Recall at 0.997 (extremely high), and a reduced F-measure at 0.697. These results show that, while AdaBoost is highly tuned for detecting phishing URLs, it is less reliable due to an increased number of false positives**.**

V. CONCLUSION

The research describes a hybrid approach using machine learning for detecting phishing website URLs. With the unending rise of phishing attacks, the world needs advanced systems that detect harmful URLs in real time. Traditional approaches work only with previously documented phishing URLs, thus lacking the ability to detect recently generated phishing websites. The systems combine natural language processing with various machine learning approaches. Incorporating preprocessing and feature selection improves the model performance. This occurs by filtering and eliminating the noise and redundant

information present in the data. For the purpose of this research study, the Random Forest and AdaBoost classification algorithms were utilized. The experimental results demonstrated that Random Forest has a performance increase over AdaBoost in the areas of accuracy, precision, and F1-score. Random Forest also had a performance increase in relative computational efficiency. Though AdaBoost had an increase in relative recall, it had a relative loss in precision. Thus, Random Forest was selected the best and most optimal algorithm for phishing URL detection for the purposes of this study. The research describes the design of a hybrid approach that combined multiple information processing and machine learning algorithms to detect potential phishing. The hybrid approach has potential for use in real world applications. For the purpose of future work, the design of the detection systems would improve through the use of deep learning real time deployment. Additional large scale datasets, along with more advanced machine feature selection algorithms, will increase the systems performance..

REFERENCES

1. Shaukat, Muhammad Waqas, et al. "A hybrid approach for alluring ads phishing attack detection using machine learning." Sensors 23.19 (2023): 8070.
2. Choudhary, Tarun, et al. "A machine learning approach for phishing attack detection." Journal of artificial intelligence and technology 3.3 (2023): 108-113.
3. Benavides-Astudillo, Eduardo, et al. "A phishing-attack-detection model using natural language processing and deep learning." Applied Sciences 13.9 (2023): 5275.
4. Salahdine, Fatima, Zakaria El Mrabet, and Naima Kaabouch. "Phishing attacks detection a machine learning-based approach." 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2021.
5. Borate, Vishal, et al. "A comprehensive review of phishing attack detection using machine learning techniques." International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) 4.3 (2024).
6. Kapan, Sibel, and Efnan Sora Gunal. "Improved phishing attack detection with machine learning: A comprehensive evaluation of classifiers and features." Applied Sciences 13.24 (2023): 13269.
7. Gandotra, Ekta, and Deepak Gupta. "An efficient approach for phishing detection using machine learning." Multimedia security: algorithm development, analysis and applications. Singapore: Springer Singapore, 2021. 239-253.
8. Kocyigit, Emre, et al. "Enhanced feature selection using genetic algorithm for machine-learning-based phishing URL detection." Applied sciences 14.14 (2024): 6081.
9. Al-Sabbagh, Abdallah, et al. "An Enhanced K-Means Clustering Algorithm for Phishing Attack Detections." Electronics 13.18 (2024): 3677.
10. Zara, Ume, et al. "Phishing website detection using deep learning models." IEEE Access 12 (2024): 167072-167087.