INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING TECHNOLOGY AND APPLIED SCIENCES (IJIETAS)

Machine Learning Based Phishing Website Detection System

[1]Dhruv Maheshwari

[1]UG Student, Department of Computer Science and Engineering Sharda University Greater Noida, India, dhruvmaheshwari1281@gmail.com

ABSTRACT

Attacks that are phishing related to digital financial services, mobile wallets, and online banking have risen dramatically hence putting the security and integrity of the users at a risk. Phishing sites are designed to look like legitimate online banking platforms so as to steal personal information, such as financial and user logins. Traditional phishing defense systems like rules-based and blacklist-based systems have a significant false-positive rate and are useless when negotiating new and untested phishing links.

The article Phinance AI offers an analyzed phishing URL detector system based on machine learning with a particular application to banking and financial websites. The system's lightweight and real-time feature enables light detection that does not require webpage content or webpage HTML examination. The combined dataset of 483, 745 URLs obtained via publically available sources was used as the training and evaluation dataset. Lexical, structural, and entropy related characteristics were elicited because it obtained forty-one URL-based features. Several monitored machine-learning models, i.e., Random Forest, AdaBoost, XGBoost, and Gradient Boosting classifiers, were also trained and compared.

The overall classification accuracy of 96.27% and heightened accuracy of 97.58% when dealing with banking related URLs give the indication that the Gradient Boosting classifier performs at a high level but with a high level of precision and a low false negative occurrence. It is a real-time phishing URL detector interface based on a web interface and API that uses Flask and Python. The results indicate that the use of machine-learning-based URL analysis is an effective and scalable process of detecting phishing with banking inclination.

Keywords: Phishing Detection, Machine Learning, URL Analysis, Cybersecurity, Gradient Boosting

ARTICLE HISTORY
Received: 15 May 2026, **Accepted:** 28 May 2026, **Published:** 12 June 2026

CITATION
Maheshwari, D., (2026). Machine Learning Based Phishing Website Detection System, *International Journal of Innovations In Engineering Technology And Applied Sciences (IJIETAS),* 2(1), 21-28.
DOI: https://doi.org/10.64764/ijietas.v2.i1.04

1 Introduction

The rapid growth of online financial services, mobile wallets, online banking has grown exponentially, and phishing attacks have become one of the biggest cybersecurity threats as a consequence. Phishing websites take the appearance of credible banking websites, thus fooling users into sharing sensitive personal data including financial personal information, one-time passwords and logins. Confidence among the users of virtual financial systems is destroyed and these attacks incur major losses of money.

The conventional phishing detection models - rule based and blacklist based models are limited in detecting phishing URLs that were created recently and require under regular manual updates. In addition to this, the type of systems will tend to have high levels of false-positives, which is especially harmful in banking settings where potentially legitimate sites will be mistakenly blocked.

In recent scholarship, machine-learning-based methods are shown to overcome these drawbacks by detecting previously unknown phishing URLs more accurately and learning discriminative behaviour on the basis of URL properties [1], [2]. It is based on this that the current paper proposes the Phinance AI; a machine-

learning system based phishing URL detector optimized precisely to the domain of banking and financial websites, which is able to achieve high-detection rates with a minimal number of false positives.
The major contributions of this work include (i) banking-oriented phishing URL predictor, (ii) an all-encompassing feature engineering based on the URL properties, and (iii) a comparative evaluation of a number of ensemble machine-learning models.

As recent research studies have found out, phishing attacks against financial institutions have been more advanced in that they are often using deceptive domain structures coupled with URL obfuscation schemes to bypass traditional phishing detection techniques [1], [8]. As such, the ever-growing need is intelligent detection systems which can be used in real time financial settings without depending on web-page content assumptions which are computationally intensive.

2. Related Work

The detection of phishing has been researched widely in a range of different methodologies, and that includes complex machine-learning methods to blacklist methods. The previous methods of detection were quite formal, as they were based on a blacklist of suspect sites (that were described to be harmful before), meaning that they allude to phishing sites that have been reported in the past but not those that have been recently created or have zero-day attacks [3], [10].

They later designed heuristic and rule-based techniques to analyze suspicious keywords and URL layouts; however, the techniques were often associated with large rates of false-positives and, to such an extent, necessitated manual updates on a regular basis [4], [11].

With the advent of machine-learning classifiers, phishing detection has been rethought as a classification issue that capitalizes on URL-based, content-based and hybrid attributes. It has been demonstrated through empirical research that other algorithms like the Random Forest, Support Vector Machine and Gradient Boosting classifier are able to effectively distinguish between phishing and legitimate URLs through the identification of trends in URL properties [9], [11]. They have also been strong in ensemble learning methods, which combine a number of weak learners to enhance the overall detection efficiency and strength [12].
The system proposed is purely focused on URL-based functionalities to consume lightweight and real-time phishing detection, and it does not require the complexity of webpage content analysis that many modern systems imply and demand is expensive.

Despite the high detection accuracy demonstrated by deep-learning approaches based on Convolutional Neural Networks and Long Short-term Memory networks, their application in the real-time applications is hindered by the need to use huge training ranges and considerable resources of computing power [7]. On the other hand, scalable phishing detection systems can utilize the trade-off between the accuracy, interpretability, and computational efficiency balancing available to ensemble machine-learning models with engineered URL-based features. However, the high computational and training cost and complexity associated with deep-learning models might rule them out of lightweight security systems, which require application of fast phishing recognition.

In recent studies, the obstacles in implementing deep-learning models in detecting phishing in real time have also been noted. Though the convolutional and recurrent neural network based methods have potential to achieve high detection accuracy, their practical implementation in lightweight security systems would be constrained by the fact that future computation resources and training data would be huge, [7], [8]. Scalable phishing detection in banking and financial applications, however, requires a more effective interpretable, and scaled machine-based solution using ensemble machine-learning resources with engineered URL-based features.

There has also been much research concerning the balancing act of the computational efficiency and the detection accuracy in phishing detection systems. Even though content-based and hybrid techniques analyze web-page HTML, JavaScript, and visual items, they have a high cost of extra processing load and privacy issues [8]. Unfortunately, URL based phishing detection methods can be competitive and lightweight, and therefore able to be deployed in real time, as demonstrated by many studies [2], [6]. These results support the use of ensemble learning models alongside the use of URL only feature engineering in order to scaleably detect phishing on banking and financial environments.

More recent research has explored feature engineering strategies of phishing. Some of the most commonly used variables in the identification of suspicious patterns in malicious URL are the lexical and the hostbased URL attributes. These features include the length of the URL, number of subdomains, and the use of special characters, domain age, and so on. Mechanical-learning models that are trained on these designed characteristics have shown successful performance on detecting phishing without having a complex computational cost [9], [11]. The scheme based on features is also more desirable in implementing an efficient system in real time, because of the fact that, there is no need to download the web-page content to carry out analysis.

A second impactful path of research deals with using ensemble techniques of learning. Ensemble techniques involve using the output of two or more classifiers and, therefore, contribute to the accuracy of detecting phishing-pages. There is empirical evidence that ensemble methods such as Gradient Boosting, Random Forest, and XGBoost generally produce better outcomes than individual classifiers in the area of phishing - page detection [12], [13]. The models are also effective on the hitherto unseen phishing pages as they can discover intricate feature associations.

3. Dataset and Feature Engineering

The data used to test and train the proposed system was a composite which contained 483,745 URLs which were collected based on publicly available sources. The dataset is a combination of the Mendeley Phishing Dataset and the UCI PhiUSiIL Phishing URL Dataset hence containing the phishing and legit URLs respectively. In order to maintain the distribution of classes during testing, 80:20 stratified traintest split was used. The datasets used to train and test the proposed system are summarised in Table 1.

Table I. Dataset Description

Dataset	Source	Samples
Mendeley	Mendeley Data	247,950
UCI PhiUSIIL	UCI ML Repository	235,795
Total		**483,745**

A total of 41 URL-based features were obtained with each URL. These features, as shown in Figure 1, are of structural, lexical, domain based, and entropy related features with such features as the length of the URL, the count of digits, the count of special characters, the number of subdomains and the measure of randomness.

```
...   Top 20 Most Important Features:
                                  feature  importance
                               url_length    0.233229
                            domain_length    0.169173
                  number_of_digits_in_url    0.119755
                average_subdomain_length    0.075464
                    number_of_subdomains    0.071333
                       entropy_of_domain    0.056869
                          entropy_of_url    0.052008
           number_of_special_char_in_url    0.034294
                  number_of_slash_in_url    0.024089
                   number_of_dots_in_url    0.023123
                number_of_dots_in_domain    0.022680
               number_of_digits_in_domain    0.022198
                              path_length    0.022043
                number_of_hyphens_in_url    0.010360
             number_of_hyphens_in_domain    0.009433
           number_of_questionmark_in_url    0.008402
                  number_of_equal_in_url    0.007566
           number_of_digits_in_subdomain    0.004441
        having_repeated_digits_in_domain    0.004215
  number_of_special_characters_in_domain    0.003653
```

Fig. 1. URL feature categories used in the system

The balance of the detection accuracy and the computational cost of URL-based machine-learning models offers a beneficial compromise in detecting phishing tactics, which is suggested by current surveys on phishing detection. These methods take advantage of lexical and structural features that are directly derived out of the URL path and can be identified quickly with privacy-friendly features on the content as opposed to content-based methods that require webpage HTML and visual frameworks to identify such attributes [13]. In turn, the use of URL-based will especially apply to real-time cybersecurity systems used in the banking and financial environment.

Where content based analysis is hindered by privacy or performance factors, previous research has shown that URL based features in isolation could be used to successfully identify phishing sites and normal web sites [2], [6]. Feature engineering enables lightweight identification with competitive precision which brings in lexical, structural, and entropy-based features.

The extensive use of URL-based feature engineering in the study of phishing-detection can be explained by its effectiveness and low implementation cost [2], [6]. Figure 2 visualises the relative significance of the extracted features of the URL.

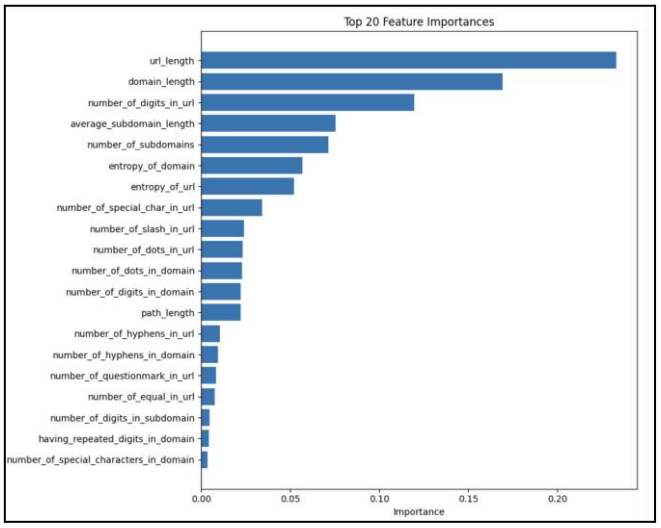

Fig. 2. URL feature importance distribution

4. Proposed System and Methodology

Phinance AI system identifies phishing URLs through conciliated machine learning. The general flow of work includes checking the URLs, extracting the features, model prediction, and the conclusion in form of classification. Random Forest, AdaBoost, XGBoost and Gradient Boosting trained and tested classifiers were used in this study.

The reason behind choosing the Gradient Boosting classifier as the final model is its overall high performance when compared to the rest of the algorithms tested. Gradient Boosting, an ensemble algorithm construction that uses weak learners, in turn decision-trees, in a cumulative fashion to build a strong predictive model, trains successive models to reduce the errors of the predecessor models. The algorithm proceeds in an iterative method of minimizing a loss measure by training new models to reduce the residual variation of old models, thus modeling the interactions of features that are not easily explained by simple mechanisms in addition providing very precise predictions to problems of classification like phishing offenses [21], [24].

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Where Fm(x) refers to the revised prediction model at iteration m, Fm-1(x) refers to the previous model, hm(x) refers to the newly trained weak learner to minimize the residual errors, and g m refers to the learning rate used to control the weight of the revised model. The successive addition of these weak learners gradually boosts prediction accuracy as the Gradient Boosting algorithm does.

A domain reputation heuristic was added to minimize the false positives of the legitimate banking URL, which enhances the accuracy of detection in the financial context. The trained model was used as a web application, based on Flask, providing the real-time analysis of the URL in a web interface and a REST API. The table below (figure 3) illustrates the general system design of the proposed phishing detection model.

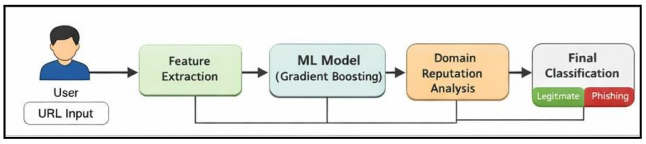

Fig. 3. System architecture of Phinance AI

This architecture increases the modularity and maintainability during the deployment since the layers of feature extraction, model inference and user interaction are segregated. According to these principles of scalable machine-learning systems, this segregation allows the detection engine to be updated or retrained independently without affecting the user-facing interface [1]. The latter design decision also supports the fact that the system can be applied to operational financial security situations.

5. Result and Analysis

The performance assessment used standard classification measures. The Gradient Boosting classifier reached an overall accuracy of 96.27, a precision of 98.22, a recall of 93.49, and F1 95.79, and thus outperformed all other evaluated models. This ability of Gradient Boosting to capture complex interactions between features as well as its ability to fight overfitting makes it very popular in the area of cybersecurity classification [5], which is especially helpful in phishing detection where minor differences in URL template structure can dramatically change the classification rate.

In addition to this, feature importance analysis sheds more light on how the model trained operates. According to the empirical findings, the strength of special characters, the length of URL, and the subdomains have a significant effect on the phishing detection effectiveness. Such properties often represent typical ways of obfuscating a program used by malicious parties to masquerade as legitimate banking environment. By using such salient patterns, the Gradient Boosting model can be confidently used to detect malicious and legitimate URLs. There are a number of conventional classification metrics that were taken into account to determine the performance of the proposed phishing detection system. These measures, including accuracy, precision, recall, F1 -score, and RO-CU sub-AU, are regularly used to present the performance of machine-learning classifiers in the domain of cybersecurity. Whereas, recall and precision represent the information about the accuracy of the model in identifying phishing URLs correctly and minimizing false positives, accuracy is the model in general. The fact that ROC-AUC measures the ability of the classifier to distinguish between real and fake URLs at different levels of decision criteria, but the F1-score is a more balanced measure of accuracy and recall. Table 2 below gives the aggregate performance measures of the chosen Gradient Boosting classifier.

Table 2. Overall Model Performance Metrics

Metric	Value (%)
Accuracy	**96.27**
Precision	**98.22**
Recall	93.49
F1-Score	95.79
ROC-AUC	98.26

Figure 4 is the confusion matrix of the Gradient Boosting classifier. The high precision indicates a low false-positive rate which is essential when working with banks and finance.

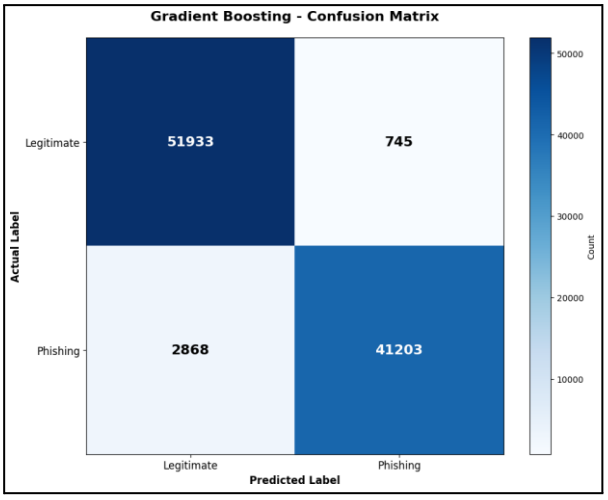

Fig. 4. Confusion matrix of the Gradient Boosting classifier

Additional testing of URLs associated with the banking domain demonstrated an increased detection rate of 97.583, which highlights the effectiveness of the suggested approach to financial cybersecurity. Gradient Boosting was always more than the Random Forest, AdaBoost, and XGBoost classifiers in benchmarking with other models. Table III presents a comparative performance analysis of the various machine-learning models and is in line with recent phishing identification research [1], [5].

Table 3 Comparison Of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	**0.9627**	**0.9822**	**0.9349**	**0.9580**	**0.9827**
Ensemble	0.9597	0.9825	0.9280	0.9545	0.9812
XGBoost	0.9546	0.9791	0.9199	0.9486	0.9795
Random Forest	0.9542	0.9777	0.9204	0.9482	0.9795
AdaBoost	0.7513	0.8805	0.5252	0.6580	0.8065

Another point supporting the usefulness of ensemble learning methods in phishing detection is the comparison based on the evaluation, which is presented in Table 3. Gradient Boosting showed a strong ability to record more complex relationships between URL features because its accuracy and F1-score were good compared to the compared models. The relevance of tree-based ensemble techniques to the issue of cybersecurity classification is also supported by the fact that the competitive performance of models like Random Forest and XGBoost is high.

The gain on URLs attached to banking can be explained by the fact that the model is good at identifying domain-specific patterns and applying heuristics based on domain reputation. Similar results are observed during earlier studies on the topic of financial phishing detection, where domain-based models provide less false positives to reputable banking sites [4].
The suggested system is quite practical with regards to the integration into active cybersecurity environments. Since the detection mechanism uses only URL-based features, it can be analyzed quickly with no necessity of having a full-body webpage content loaded, which significantly alleviates the processing load and allows it to operate at light speed in application in areas like email filters, web gateways, and browser extension security software.

The soundness of the suggested system was tested through analysis of the behavior of the system on the URL patterns which have not been seen before, as well as accuracy-based analysis. The ensemble-based learning paradigm of Gradient Boosting classifier makes use of this fact to reduce the variance of weak learners and also to be more stable to enable effective generalization by combining multiple learners [5]. In phishing detection, this feature has the added benefit of phishers often altering the URL structures to beat the heuristics of static detection. The same results have been reported in the previous literature on the robustness of ensemble models to dynamic patterns of phishing attacks [1], [7]. Figure 5, which is also contained in appendix 2, compares the performance of different machine-learning models tested in the study.

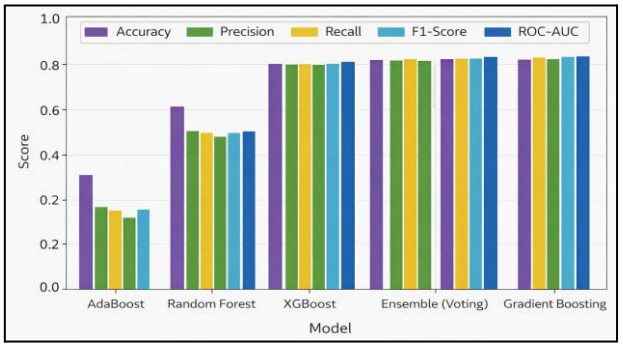

Fig. 5. Performance comparison of machine learning models

Additionally, the empirical results support the need to strike a balance between the level of detection and computation in real-time cybersecurity applications. Although the high precision is required, the fact that the system will be able to handle URLs in a fast manner without the need to examine the entire webpage content makes it most appropriate when it comes to deployment in real-life scenarios like email filtering applications, browser extensions, and security gateways in enterprises. The lack of such ability clearly shows that the Gradient Boosting model is more robust in working with data it has never seen before, which is yet another example of its generalization ability that ensures that it will perform well even with changing phishing methods. These findings support the feasibility and effectiveness of this system in the context of financial cybersecurity in the real world.

6. Conclusion

The machine-learning-based phishing universal web URL detector, called PhinanceAI, was rendered on banking and financial websites and introduced in this paper. The system had low false-positive rates, but high detection accuracy was ensured by using Gradient Boosting classifier and designed URL features. The experimental results indicate that URL based machine-learning analysis is a cost-effective and scalable way of detecting phishing attack in the financial sphere. The suggested methodology demonstrates that the need to detect phishing in the financial sector can be achieved through the combination of feature-engineering and ensemble-learning. Cyber threats continue to advance to be more indispensable in machine-learning-based security systems that protect users and organizations against phishing attacks. The further work can be aimed at adding more external characteristics and implementing the system in the production facilities.

References

1. S. Marchal, J. François, R. State, and T. Engel, “PhishStorm: Detecting Phishing with Streaming Analytics,” IEEE Transactions on Network and Service Management, 2014.
2. A. Safi et al., “A Systematic Literature Review on Phishing Website Detection,” Computers & Security, Elsevier, 2023.
3. C. Whittaker, B. Ryner, and M. Nazif, “Large-Scale Automatic Classification of Phishing Pages,” NDSS Symposium, 2010.
4. M. Aburrous et al., “Intelligent Phishing Detection System for E-Banking,” Expert Systems with Applications, Elsevier, 2024.
5. A. Jain and B. Gupta, “Machine Learning Based Phishing Detection Using URL Features,” IEEE Conference on Computer Communications Workshops, 2022.
6. UCI Machine Learning Repository, “PhiUSIIL Phishing URL Dataset.” Available: https://archive.ics.uci.edu