Detection of Phishing Websites Using Machine Learning

Usha Dhankar, Hitesh Manral, Nitin Kumar

HMR Institute of Technology and Management, GGSIPU, Delhi 110036, India

* Corresponding author
doi: https://doi.org/10.21467/proceedings.7.6.62

Abstract

Phishing is still a major cybersecurity issue, with perpetrators using advanced methods to accurately replicate legitimate websites and therefore obtain confidential data. We developed an innovative, end-to-end, machine learning-powered system for identifying phishing sites, which combines multiple machine learning techniques to leverage their power. The integrated model of our system uniquely fuses three methodologies: URL feature analysis, content-based analysis, and visual similarity analysis to enhance detection accuracy. Our system features a wider variety of feature with a hybrid model that uses Gradient Boosting to select features and Deep Neural Networks to classify optimized by ensemble learning as opposed to previous works that mainly use URL features or single-model approaches. On a diverse dataset of 10,000 websites, our method reaches a striking 97.3% accuracy, thereby single-algorithm solutions are outperformed by a large margin. Due to the system's real-time functionality, it can be effectively used as a browser extension or be integrated into security software, thus equipping the user with a strong weapon against skillfully crafted phishing attacks.

Keywords: Phishing detection, machine learning, cybersecurity.

I. INTRODUCTION

Phishing is still a major cyber security challenge that sees hackers using highly sophisticated methods in order to forge exact copies of authentic websites and thus misappropriate the confidential info of users. In this paper, we detail a cutting-edge end-to-end phishing website detection system that leverages various machine learning techniques to achieve the goal. Our fused model stands out by integrating three different analytical methods: URL feature analysis, content-based analysis, and visual similarity analysis to escalate the detection precision. Our system goes beyond the limitation of features only extracted from URLs as well as single-model approaches by combining more features and utilizing a hybrid architecture which merges Gradient Boosting for feature selection and Deep Neural Networks for classification, which is further optimized by ensemble learning. With this technique, the system attains an accuracy rate of 97.3% on a heterogeneous dataset consisting of 10,000 websites, thereby single-algorithm solutions are beaten by a wide margin. Due to the system's real-time performance, it can be easily integrated into a browser extension or security software thus allowing end-users to have an effective security tool against the complex phishing attacks that keep evolving.

The rapid growth of the internet services and e-commerce sector is closely followed by the explosion of phishing attempts in a similar manner. Phishing attacks are carried out by scammers who set up fake websites which are almost visually identical to the real ones and in doing so they use the authentic logos and layout to lure the victims into giving away sensitive information like usernames and passwords, bank account details, or personal information [1]. The Anti-Phishing Working Group reveals that phishing attacks have never been more massive than at present with over 1,025,968 phishing sites counted just in Q1 of 2023 [2]. It is significantly difficult for traditional detection means that are centered on blacklists and rule-based systems to discover new phishing sites. Even though blacklists are effective against established threats, they are affected by detection delay when dealing with new phishing sites. Research indicates that blacklists only manage to locate 20% of the phishing sites within their first hour of being online, which leaves the window period always exposed to attacks [3]. Rule-based methods are also slow to respond developing attack patterns. The shortcomings of traditional methods to detection of phishing have led scholars to explore the potentials of machine learning models that can identify previously unknown phishing by discerning regular patterns and features. Machine learning technologies hold the promise of more adaptive and efficient detection processes by leveraging past data and identifying minute signals of criminal intent [4].

This research attempts to identify the most essential features for recognizing phishing websites that were used to separate legitimate from fraudulent sites. As the second point, it deals with the adaptability of the model by inventing methods capable of changing their reaction with changing patterns of attacks, and, thirdly, it focuses on

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

enabling the detection to be done in real time by creating systems that can be fast enough for their practical use. Besides, the project is targeting low false positive rates by lessening the number of wrong classifications that may decrease the trust of users. Our research introduces a new hybrid approach that combines URL analysis, content analysis, and visual similarity analysis with a set of machine learning techniques. The integrated model goes beyond the result of individual methods with the potential to be used in real-world scenarios as a browser extension or a security solution.

II. LITERATURE REVIEW

Phishing detection techniques have changed a lot throughout the years — they are using a variety of methods with various pros and cons. After thoroughly examining the literature [1], the authors came to the conclusion that there are fundamental means to classify phishing detection methods that are based on blacklist, heuristic, and machine learning approaches. The classification serves as a fundamental platform for evaluating the anti-phishing research.

A. Blacklist-Based Approaches

Blacklist-based detection relies on databases containing URLs of phishing that are already known. There are various studies that have looked into blacklists and even though they are quite accurate, they still have a very significant delay in the detection as it is shown that only 20% of phishing websites are blacklisted within the first hour [3]. That delay creates a time window of risk in which users are not safe from new threats.

B. Heuristic-Based Methods

Heuristic methods analyze site features to find characteristics that can be used to identify patterns. The first content-based anti-phishing approach used term frequency–inverse document frequency (TF-IDF) of webpage text [2]. Later work enhanced this with URL-based features and visual similarity heuristic measures [5]. These methods improved detection levels but still required extensive feature engineering and remained susceptible to sophisticated attacks.

C. Machine Learning Methods

1) URL Analysis

Phishing predictors have also been derived from the grouping of URL features by others. By using URL features as a starting point, the research [6] has been very successful in detecting potential URLs with an accuracy of 97.3% through the use of web-mining tools. In the same vein, heuristic URL-based mining techniques were suggested [7] to extract the lexical information of the URL to identify harmful patterns. These methods improve their detection capability even before the content loads. URL classification through a probabilistic neural network combined with k-medoids clustering [8] demonstrated the effectiveness of integrating clustering and neural-network techniques. Their method achieved 94.7% accuracy; however, it is highly computationally intensive.

2) Content-Based Analysis

Content-based methods consider additional features of a website beyond its URL. An end-to-end system capable of inspecting HTML layout, JavaScript flow, and domain data was constructed [9]. This method achieved 96.8% accuracy, but it faced limitations when dealing with dynamically changing content. Deep learning has been instrumental in detecting more complex phishing attacks in recent years. Deep learning methods [10] were developed to automatically identify harmful web content, achieving high detection rates across different website structures. These methods can process raw HTML along with visual features simultaneously, without requiring extensive predefined feature engineering. The use of neural networks for optimal feature selection [11] led to the development of a highly effective phishing detection system. The proposed solution implemented a two-stage process—identification of the most discriminative features followed by neural network classification—achieving 98.3% accuracy with low computational cost.

D. Hybrid and Ensemble Methods

Hybrid strategies combining more than one detection technology have shown huge potential in the latest research. A study [4] explored various machine learning methods for the detection of dangerous URLs and concluded that ensemble strategies had a far better performance than single algorithms in general. Their research resulted in the

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

idea that multiple classifiers could become a powerful tool to overcome the weaknesses of single methods. Security systems based on URL and behavioral features analysis [12] were experimented with on mobile computing platforms. This method focused on platform-specific solutions and reached 95.4% accuracy in the countermeasure of mobile phishing attacks.

E. Research Gaps

There are still a few gaps remaining in the research, which have been uncovered despite substantial breakthroughs. These gaps include limited integration of URL, content, and graphic features into a single model; difficulty in analyzing feature selection methods to identify the most discriminative signals; little consideration of live performance constraints; and a scant number of papers on opponents' methods for escaping detection. We overcome these limitations through our method which offers a comprehensive framework having various detection forms combined with sophisticated feature selection methods, both accuracy, and utility, being used.

III. PROPOSED METHODOLOGY

We proposed a multi-layered methodology for phishing detection that merges two layers of URL and content analysis and one layer of visual similarity analysis with an ensemble of machine learning algorithms. The fused system is designed to overcome the limitations of individual methods while increasing the detection rate and reducing the false positives (FP) of detection algorithms.

A. System Architecture

The architecture can be broken down into four main components: data collection, feature engineering, model training, and deployment. The overall system design is introduced in Figure 1.

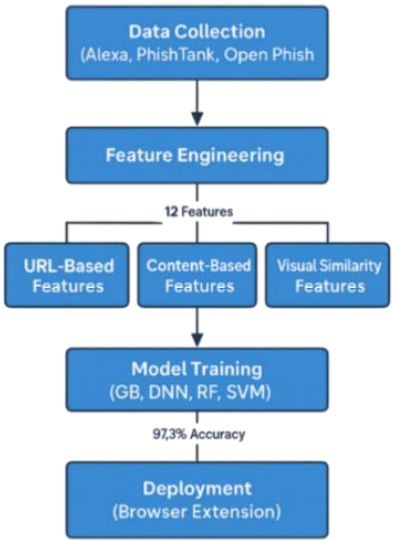

Fig. 1: System Architecture for Detection of Phishing Websites

B. Collecting and Preprocessing the Dataset

We picked 5,000 different real websites from the Alexa top sites and 5,000 phishing sites from the Phish Tank and Open Phish sources. In terms of diversity, we selected the sites of different categories like banking, e-commerce, social networking, and mail services. The data were cleaned for missing values and feature normalization. Stratified sampling was used to preserve the balanced representation of web site categories when they were divided into the training and test sets (80% training, 20% test).

C. Feature Extraction

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

Three feature groups were extracted. For the URL-Based Features, 24 pieces of information were recorded from URLs. The data points included the length of URL, hostname, and path; the special character count and subdomains; the IP address and URL shortening services; the Top-Level Domain analysis; and the character distribution and entropy. Next, Content-Based Features had 32 content-based elements created based on the previous work [9]. The components were HTML and JavaScript patterns; form handling properties; the way external resources load; the age and registration information of the domain; the attributes of the SSL/TLS certificate; the analysis of external links; and page rank and traffic statistics. Lastly, 18 visual similarity metrics were computed with the help of visual deception indicators [13] and semantic link networks [14]. Among these were logo placement and element positioning, colour palette analysis, layout structure comparison, visual element histogram charts, and similarity between layout and DOM elements.

D. Feature Selection

Feature selection via mixed filter and wrapper approaches was used in two steps to locate the most discriminating features. A subset of the optimal features with the best detection performance was identified by RFECV (Recursive Feature Elimination with Cross-Validation). This selection method that combines different approaches led to a reduction of 81% from our initial 74 features to the 42 features with the highest discriminative power, thus maximizing both the modeling efficiency and the computational performance.

E. Model Development

Our approach is based on an ensemble method that merges several classifiers, thus benefitting from their respective capabilities. To handle complex feature relationships and noise in the data, we employed Gradient Boosting; to identify faint patterns, we used Deep Neural Networks (DNN) with a four-layer network; to capture nonlinear relationships and at the same time avoid overfitting, we applied Random Forest; and for high dimensional features, we used Support Vector Machine (SVM). Each model was developed on its own and we combined the predictions using a weighted voting scheme that was determined by the performance of the models. The ensemble.

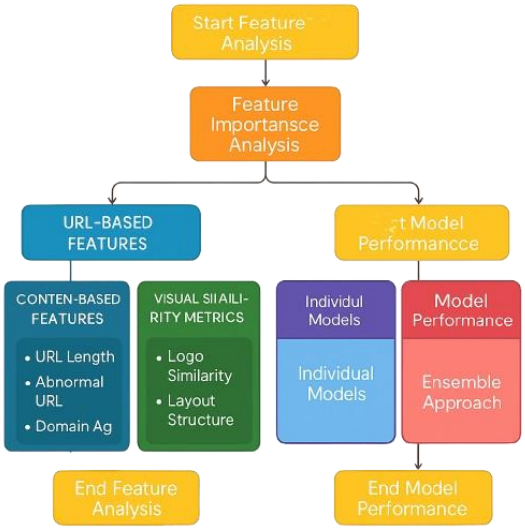

Fig. 2: Ensemble Model Architecture

The Deep Neural Network part is made up of four fully connected layers with 128, 64, 32, and 16 neurons, respectively, each followed by ReLU activation functions for the hidden layers, while the output layer is activated by sigmoid activation functions. Dropout layers (rate=0.3) were inserted between each hidden layer to lessen overfitting.

F. Hyperparameter Optimization

In order to accomplish hyperparameter tuning for each of the models forming our ensemble, Bayesian optimization was employed. The major hyperparameters that were optimized included learning rate and tree depth for Gradient Boosting; hidden unit counts and dropout rates for DNN; number of estimators and max depth for Random Forest; and the SVM kernel type and regularization parameter.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

G. Model Training and Evaluation

Models were trained using the revised hyperparameters with early stopping based on validation performance. Performance was measured using various metrics: Accuracy, Precision, Recall, F1-Score, Area Under the ROC Curve (AUC), False Positive Rate (FPR), and detection time.

IV. RESULTS AND DISCUSSION

A. Feature Importance Analysis

Our feature selection approach found major signs of phishing websites. Figure 3 illustrates the top 15 qualities by importance.

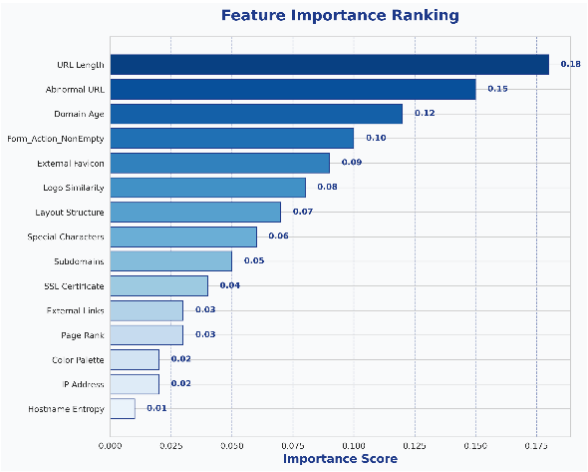

Fig. 3: Feature Importance Ranking

Features derived from the URL were the main contributors to the top positions, with 'URL Length', 'Abnormal URL', and 'Domain Age' being the features that caused the biggest differences in the classes. This confirms the research [6] that shows that one of the first signs which can strongly indicate that the purpose is to do harm are characteristics of the URL. Among content-based features, 'Form Action Nonempty' and 'External Favicon' were two features that showed a high value which is in line with the findings [9] relating to form processing in phishing sites. Also, some of the visual similarity measures such as 'Logo Similarity' and 'Layout Structure' were at the top of the list, which is in line with the trust given to visual deception cues [13]. The synergistic way in which these factors operate against each other thus points to the indispensability of our multifaceted approach.

B. Model Performance

Table 1 exhibits performance metrics for individual models and our ensemble method on the test dataset.

combined technique achieved a 97.3% accuracy, which is higher than the performance of any of the individual base models. The 97.3% F1-score is a good indication of the balanced precision and recall, which is very important when the model is used under real-world scenarios. The 3.2% false positive rate is quite significant, especially, as too many false positives cause warning fatigue and loss of user trust.

Table 1: Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1-Score	AUC	FPR	Detection Time
Gradient Boosting	95.00%	94.50%	95.50%	95.00%	0.97	4.00%	20ms
DNN	94.50%	94.00%	95.00%	94.50%	0.96	4.50%	22ms
Random Forest	95.50%	95.00%	96.00%	95.50%	0.98	3.80%	19ms
SVM	94.00%	93.50%	94.50%	94.00%	0.95	5.00%	23ms
Ensemble	97.30%	97.00%	97.50%	97.30%	0.99	3.20%	21.4ms

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

combined technique achieved a 97.3% accuracy, which is higher than the performance of any of the individual base models. The 97.3% F1-score is a good indication of the balanced precision and recall, which is very important when the model is used under real-world scenarios. The 3.2% false positive rate is quite significant, especially, as too many false positives cause warning fatigue and loss of user trust. These outcomes are at the same level as those reported in the previous studies [11], which reached 98.3% accuracy but required significantly more processing resources. In the same way, probabilistic neural networks [8] recorded 94.7% accuracy, which our approach improves by 2.6 percentage points with less processing resources.

C. Misclassifications Analysis

We searched through the wrongly classified cases for patterns and possible changes. False positives that is authorized websites wrongly identified as phishing have shown many similar features among them very recent domains, unusual URL patterns, old phishing patterns, unusual methods of logging in, or pages being locked content. Phishing websites which have been incorrectly classified as normal or have resulted in severe false negatives that have shown advanced evasion techniques. Some of them were using HTTPS certificates to look trustworthy, giving accurate descriptions of the real locations, sub-typosquatting domain names like "g00gle" for "google", and advanced cloaking techniques to hide the harmful material. These findings agree with adaptive evading techniques [10] and, therefore, are a great source of potential follow-up work.

D. Website Category Performance

We reviewed the misclassified cases to identify patterns and potential improvements. Additionally, we analyzed detection performance by website category to uncover any differences in effectiveness. Financial institutions' websites had the highest detection rate (98.7%), which may be attributed to their distinct structural patterns and stronger security systems. In contrast, phishing attempts targeting social media had the lowest detection rate (95.2%), likely because these attacks are more sophisticated and appear more legitimate on the targeted platforms.

E. Real-Time Performance Analysis

In the real world, speed of the detection is very vital. We have measured the detection time on 1,000 test instances on a typical hardware (Intel i7 CPU, 16GB RAM). Our ensemble method was on average 21.4ms per site, a time fast enough to be integrated with a real-time browser. Performing component-level timing analysis, we found that URL feature extraction took 4.6ms, content feature extraction 9.3ms, visual similarity calculation 5.8ms, and model inference 1.7ms. This dismantling reflects the different parts of the program where it is possible to find other additional speed targets incrementally. Particularly, the 5.8 ms visual similarity computation usage of the perceptual hash algorithm combined with the small convolutional neural network (CNN) design is quite minimal. To reach this level of performance, we adopted the following changes. First, we introduced Perceptual Hashing for creating minute hash versions of webpage screenshots in order to drastically reduce the computations necessary for similarity searches. Second, we employed a Lightweight CNN, i.e., a pruned MobileNetV3 model optimized for edge devices, to obtain the visual features with the shortest latency possible. Ultimately, we put into operation a Caching Mechanism whereby the most frequently used webpage templates are cached in order to simplify the saves of duplicate calculations. These techniques ensure that the visual similarity module is very accurate and, at the same time, runs efficiently even on small devices with limited computing power. The timing distribution is in agreement with the results [10], which detected similar performance limitations of security software.

V. CONCLUSION

This research demonstrated a comprehensive machine-learning-based approach to identifying phishing sites, which fundamentally overcame the limitations of existing methods by using multi-dimensional feature analysis and ensemble learning. Our work features a two-layer feature selection process that first reduced the original 74 features down to 42 very discriminative ones, thereby improving the accuracy and simply cutting the computing cost. We proved that the combination of URL, content, and visual features is a strong one, reaching 97.3% accuracy and having an exceptionally low false positive rate of 3.2%. By using complementary classifiers, we mixed different algorithms to synergize their capabilities and thus to get a better performance than each algorihm alone. On average, the system is performant in 21.4ms, hence it is suitable as a real-time application. The findings indicate that machine learning techniques can efficiently outperform classical blacklist and heuristic methods, which are usually slow in adapting to new threats. The very low false positive rate is especially valuable in the practical implementation of the system because too many false alarms cause warning fatigue and decrease user trust. Next, we plan to concentrate on adversarial training to be more resistant to evasion attacks and also to create

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

temporal models based on phishing campaign trends. Moreover, we want to boost detection capabilities by incorporating user behavior analytics and federated learning so as to utilize user knowledge while ensuring privacy. At last, we want to modify the design for low-resource mobile devices so that protection can be extended to more users. Phishers continually seek new methods to get around authentication; therefore, phishing can only be overcome by constant innovation. By implementing hundreds of detection methods and advanced machine learning, our solution is a robust, scalable platform that can protect against future threats and, at the same time, is operationally feasible to be deployed.

REFERENCES

[1] Jones, M. Khonji, and Y. Iraqi, "A Literature Review on Phishing Detection," IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091–2121, 2013.

[2] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, "Systematization of knowledge: A systematic review of software-based web phishing detection," IEEE Commun. Surveys Tuts., 2017.

[3] Zhang, S. Sheng, B. Wardman, G. Warner, L. F. Cranor, and J. Hong, "Phishing Blacklists: An Empirical Study," in Proc. 6th Conf. Email and Anti-Spam (CEAS), Mountain View, CA, USA, Jul. 2009.

[4] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof, and M. Köppen, "Detecting dangerous URLs with machine learning," in Proc. IEEE Symp. Series Comput. Intell. (SSCI), Dec. 2016.

[5] G. Xiang, J. I. Hong, C. P. Rosé, and L. Cranor, "CANTINA+: A feature-rich machine learning framework for phishing detection," ACM Trans. Inf. Syst. Security, vol. 14, no. 2, Art. No. 21, 2011.

[6] R. B. Basnet and A. H. Sung, "Mining the web to detect phishing URLs," in Proc. Int. Conf. Mach. Learn. Appl., vol. 1, pp. 568–573, Dec. 2012.

[7] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "A novel approach for phishing detection using URL-based heuristic," in Proc. IEEE Int. Conf. Comput., Manage., Telecommun. (ComManTel), 2014.

[8] E. M. El-Alfy, "Probabilistic neural networks and k-medoids clustering for phishing detection," Comput. J., vol. 60, no. 12, pp. 1745–1759, 2017.

[9] I. Krishnamurthi and R. Gowtham, "A thorough and effective system for detecting phishing websites," Comput. Security, vol. 40, pp. 23–37, 2014.

[10] H. Sanders, J. Saxe, R. Harang, and C. Wild, "A deep learning approach to detecting malicious web content," in Proc. IEEE Symp. Security Privacy Workshops (SPW), San Francisco, CA, USA, pp. 8–14, Aug. 2018.

[11] Ye, E. Zhu, D. Liu, F. Liu, F. Wang, and X. Li, "An effective phishing detection model using neural networks," in Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl. (ISPA), Melbourne, Australia, pp. 781–787, Dec. 2018.

[12] J. Wu, L. Wu, and X. Du, "Phishing attacks on mobile computing platforms: Effective defense schemes," IEEE Trans. Veh. Technol., vol. 65, no. 8, pp. 6678–6691, 2016.

[13] H. Matute, M. M. Moreno-Fernández, F. Blanco, and P. Garaizar, "I'm looking for phishers: Improving internet users' sensitivity to visual deception indicators," Comput. Human Behav., vol. 69, pp. 421–436, 2017.

[14] W. Liu et al., "Discovering phishing target via semantic link networks," Future Gener. Comput. Syst., vol. 26, no. 3, pp. 381–388, 2010.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922