International Journal of Intelligent Computing Systems

July-2026 | IJICS | Volume-2 Issue-2 | ISSN: 3107-7218

Feature-Based Machine Learning Approach for Detecting Phishing URLs

N.Bhavyasree[1], S.Sarojini[2]

Student[1], Assistant Professor[2]

Priyadarshini Institute of Technology & Science for Women, Chintalapudi, Andhra Pradesh, India

bhavyasreenarra@gmail.com[1], sarojinisandhu1@gmail.comsni[2]

Abstract - Phishing attacks remain a serious threat to Internet users as they continue to develop fake sites that resemble the real sites. The purpose of these deceptive websites is to trick users into giving up their personal information, including usernames and passwords, banking details, and personal data. Phishing methods are becoming more advanced, so it is a must to create effective and secure detection tools. In this study, a machine learning based method that can detect phishing websites based on URL features without accessing the phishing website content is presented. The proposed system only uses URL analysis, which limits any user from going to malicious websites and limits the risk of running malicious scripts and downloading malicious content.

The proposed model takes more than 11,000 URLs as the dataset along with several hundreds of URL based features such as domain related features, URL structure, redirection pattern, security features etc. These features are utilized to train and test a number of machine learning algorithms for phishing detection. Experimental analysis shows that ensemble learning methods such as the Gradient Boosting Classifier (GBC) performs better by providing high accuracy, precision and recall in the identification of phishing URLs. The results gained suggest that URL-based machine learning models are a promising, light-weight, and scalable method for phishing detection and can be incorporated into web browsers, email filtering systems, and cybersecurity applications to boost online security.

Keywords -Phishing URL Detection, Machine Learning, URL Analysis, Cybersecurity, Gradient Boosting Classifier, Web Security, Classification.

I INTRODUCTION

Internet and digital technologies have profoundly changed many industries such as banking, education, healthcare, e-commerce and communication. These developments have made services more accessible and convenient, but have also come with an increased threat of cyber attacks. Phishing is one of the more prevalent and harmful types of attacks, attempting to trick users into going to fake websites that look like well-known ones to get their sensitive information like their username, password, or even banking details. Malicious website identification techniques like blacklist, rule-based detection techniques have been widely used in traditional phishing detection methods. However, these methods are not usually successful in combating new or “zero day” phishing attacks since they rely on known phishing locations. Meanwhile, phishing attacks are continually evolving, and more sophisticated and adaptive methods of detection are needed to effectively identify malicious sites before the users get caught. One benefit of machine learning is that it is able to learn patterns from the labeled data and recognize suspicious URLs that do not rely on predefined rules. One of the main benefits of URL-based detection is that it examines the content of a URL without loading the web page, which helps to prevent the exposure to malicious content and speeds up the prediction process. This paper presents the creation of a machine learning based phishing url detection system based on url characteristics like length of the url, usage of ip address, https support, age of domain, and redirection properties. Many machine learning algorithms are trained and tested for URL classification – legitimate versus phishing. The experimental results show that the proposed

method is effective and accurate for detecting phishing, and it can be used to be integrated into web security applications and protection systems of web browsers.

IIRELATED WORK

Phishing, which uses technical methods to steal consumers' private information, has become a serious cybersecurity concern. Phishing-related financial and personal damages are continuously increasing [1]. Conventional detection techniques mostly rely on feature engineering, which frequently necessitates extensive processing time and prior domain expertise. A deep learning-based multidimensional phishing detection method (MFPD) was put forth to get beyond these restrictions. It integrates statistical URL features, webpage code, and textual content for more precise identification after using character-level URL features for quick categorization without the need for third-party data or pre-established rules. This two-step approach demonstrated a balance between speed and precision, achieving 98.99% accuracy and a low false positive rate of 0.59% [17].

Although hardware-based gateway anti-phishing systems provide an extra degree of security, they are sometimes expensive and ineffective because phishing attacks are constantly changing [18]. Software-defined methods utilizing fog computing have been suggested as a solution to this problem. For example, by utilizing URL and web traffic features, Fi-NFN, a neuro-fuzzy framework installed at the fog network edge, efficiently monitors and safeguards user interactions [18].Phishing attacks often take advantage of email correspondence with embedded links, making detection and mitigation extremely difficult. Because phishing attempts have a dynamic structure, traditional systems with static rules frequently fail.To address this problem, Phish Limiter was developed, a system that combines Software-Defined Networking (SDN) and Deep Packet Inspection (DPI). In order to detect threats flexibly and effectively manage network traffic, it integrates real-time inspection with phishing signature classification using an Artificial Neural Network (ANN) [19].

Despite advancements, phishing emails continue to evade cutting-edge filters by subtly altering their structure and semantics. To overcome this difficulty, a semi-automated machine learning-based feature extraction technique called SAFE-PC was created. In a university setting, SAFE-PC identified more than 70% of missed phishing emails, demonstrating greater detection performance above current solutions like Sophos and Spam Assassin.

Additionally, it facilitates online learning, continuously increasing retraining efficiency over time [20].The risk of phishing on mobile platforms has also increased. Because of hardware limitations and user behavior, some platforms are particularly vulnerable. Conventional desktop web-based solutions frequently don't work in mobile settings. This is addressed by Mobi Fish, a lightweight anti-phishing program made for Android that verifies the purported identity of accounts, apps, and websites. Mobi Fish is quite successful at identifying phishing assaults on mobile devices, according to experimental study [21].

III EXISTING SYSTEM

Phishing is a type of online cyberattack in which people are tricked into visiting phony websites that closely resemble authentic ones in order to obtain private data, including passwords, usernames, and financial information [1][3]. Attackers usually create these malicious webpages to mimic the layout and style of authentic websites, making it challenging for the regular user to tell them apart. Phishing attacks frequently involve social engineering techniques in addition to technical trickery to deceive users into disclosing personal information [11]. Because of this, safeguarding internet users against phishing and fake websites has emerged as a critical component of online security. Attackers can create convincing phony pages using sophisticated tactics, which makes it feasible for even seasoned users to become victims [9].

Emails that seem to be from reputable public or private institutions and urge recipients to click on links to update or validate their credentials are common phishing vectors [12].

Attackers may also use blogs, forums, and file-sharing websites to spread phishing information [2]. A multifaceted strategy including legal proceedings, user awareness campaigns, and technological solutions is needed to combat phishing [5][10]. Numerous detection techniques have been created to handle the increasing complexity and variety of phishing attacks as a result of the extensive use of information and communication technologies [4][14].

IVPROPOSED SYSTEM

The suggested solution uses machine learning models to analyze a variety of URL-based parameters in order to identify phishing websites. The algorithm makes use of a dataset with 30 extracted attributes from more than 11,000 URLs, each of which is classified as either phishing or authentic. The suggested approach is intended to identify phishing websites by examining a variety of URL-based properties using machine learning models [1][4][6]. These features capture different aspects of the URL, such as the presence of an IP address (using IP). The algorithm makes use of a dataset with 30 extracted attributes from more than 11,000 URLs, each of which has been classified as either phishing or authentic [15][17].

The presence of an IP address (Using IP), the use of shortening services (ShortURL), suspicious symbols like "-" (Symbol@), the use of HTTPS, the age of the domain, and other behavioral indicators like redirection patterns and external resource requests are all captured by these features [3][6][10]. Numerical values are used to encode features, usually ranging from -1 (which denotes phishing-like conduct) to 1 (which denotes genuine behavior) [1][9].

The system learns to identify patterns frequently linked to phishing assaults by using this dataset to train machine learning classifiers [4][7][17].

To obtain strong detection accuracy, high-performing models like Gradient Boosting, Cat Boost, and Multi-layer Perceptron are used [14][17][19]. This method makes the system suitable for incorporation into online browsers, email gateways, or security software since it enables the system to classify URLs in real-time and successfully reduce phishing risks [18][19][21].

VALGORITHMS USED

Gradient Boosting Classifier: This ensemble technique uses gradient descent to minimize the errors of previous trees while building trees one after the other.

Cat Boost Classifier: This gradient boosting technique performs well with little data preparation and is tailored for categorical features.

Multi-layer Perceptron (MLP): For supervised learning, this feedforward artificial neural network has one or more hidden layers.

XGBoost Classifier: This scalable and effective gradient boosting method makes use of enhanced regularization and parallelization.

Random Forest: To increase accuracy and manage overfitting, an ensemble of decision trees produces the mode of their forecasts.

Support Vector Machine (SVM): This classification algorithm determines the best hyperplane to divide data into different classes.

Decision Tree: This model, which resembles a flowchart, divides data into branches in order to make judgments based on feature values.

K-Nearest Neighbors (KNN): This non-parametric approach categorizes data according to the majority label among its k nearest neighbors.

Logistic Regression: This linear model estimates the likelihood of class membership and is used for binary classification.

Naive Bayes Classifier: This probabilistic classifier assumes feature independence and is based on Bayes' Theorem.

VI IMPLEMENTATION

The dataset has 31 characteristics and 11054 instances. Thirty of these traits are independent, and one is dependent. Label Encoder is not required because every feature is in the int datatype. The dataset does not contain any outliers. The dataset has no missing values.

Using IP: URLs that use IP addresses rather than domain names are frequently dubious.

LongURL/ShortURL: Phishing frequently uses extremely long or truncated URLs that conceal the true location.

International Journal of Intelligent Computing Systems

July-2026 | IJICS | Volume-2 Issue-2 | ISSN: 3107-7218

Symbol@: The @ symbol may cause the browser to reroute and deceive users.

HTTPS: Phishing websites frequently omit HTTPS, whereas secure websites use it.

RequestURL, AnchorURL, and LinksInScriptTags: Calculates the number of external resources or links; excessive external dependencies are questionable.

Domain Age, DNS Recording, and Website Traffic: New or obscure domains with little traffic are more likely to be phishing.

Google Index/PageRank: A measure of a website's online credibility and indexing status.

These characteristics are used by machine learning models to identify patterns that indicate whether or not a URL is likely to be phishing. They can generalize to new URLs after being trained on tagged examples.

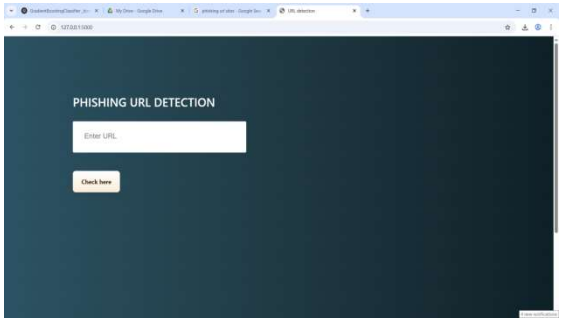

Figure 1: Web UI for Phishing URL Detection

The picture displays a web interface for a Phishing URL Detection System that was constructed with Flask, a Python web framework.

Applications of this kind are helpful for:

• Before clicking on dubious links, general users should verify them.

• Email security tools or browsers that incorporate URL verification.

• Cybersecurity teams will immediately identify threats.

A correlation heatmap of the characteristics utilized in the phishing URL detection model is shown in the figure below. The correlation coefficient between two attributes, which ranges from -1 (strong negative correlation) to +1 (high positive correlation), is displayed in each cell.

Since a feature is always completely connected with itself, all diagonal values are 1.0. Higher positive correlations are represented by brighter areas that are closer to white, whilst lower or negative correlations are shown by darker areas that are closer to black.

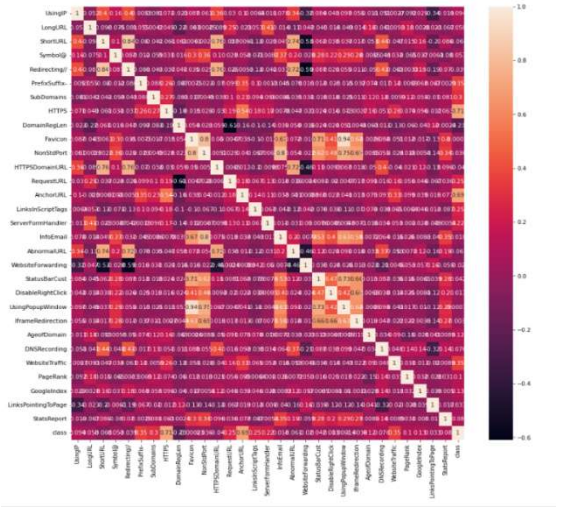

Figure 2: Heat map of the dataset with parameters

VII RESULTS AND DISCUSSION

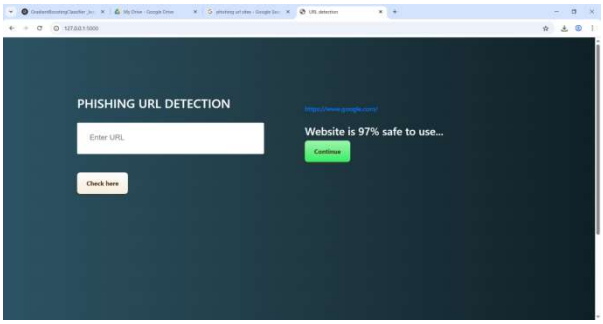

Figure 3: URL Input Interface for Phishing Detection System

The URL that was entered as a clickable link.A confidence message such as "Website is 97% safe to use" probably indicates that the model has 97% confidence in its prediction that this URL is secure.

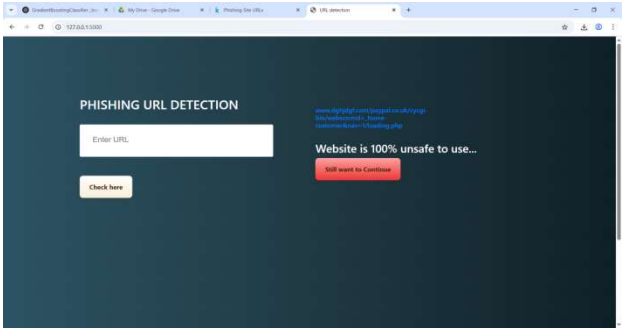

Figure 4: Prediction Result Display Showing URL Safety Confidence Score

This interface shows a real-time prediction system that evaluates a URL's security using a trained machine learning model (such as the Gradient Boosting Classifier).It assesses the danger and gives a percentage-based safety score upon URL submission, assisting users in avoiding hazardous or phishing websites.

ML Model	Accuracy	f1_score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Cat Boost Classifier	0.972	0.975	0.994	0.989
Multi-layer Perceptron	0.971	0.974	0.992	0.985
XGBoost Classifier	0.969	0.973	0.993	0.984
Random Forest	0.967	0.970	0.992	0.991
Support Vector Machine	0.964	0.968	0.980	0.965
Decision Tree	0.961	0.965	0.991	0.993
K-Nearest Neighbors	0.956	0.961	0.991	0.989
Logistic Regression	0.934	0.941	0.943	0.927
Naive Bayes Classifier	0.605	0.454	0.292	0.997

Table I. Comparative Performance of Machine Learning Classifiers

Several machine learning models were used to assess the phishing URL detection system, and the findings show a great overall performance, especially among ensemble and deep learning approaches [1][5][17]. With an accuracy of 97.4%, an F1-score of 0.977, and an outstanding recall of 0.994, the Gradient Boosting Classifier proved to be the best model, successfully identifying almost all phishing attempts while retaining a high precision rate [16][17]. The Multi-layer Perceptron (MLP) and Cat Boost Classifier come next, with recall values of 0.994 and 0.992, respectively, and accuracy exceeding 97% [9][14][17]. In phishing detection, where failure to detect a rogue URL can pose significant security implications, very high recall rates are essential [4][20].

Models like XGBoost, Random Forest, and Support Vector Machine also displayed robust performance, with accuracy ranging from 96.4% to 96.9% [4][7][16]. Specifically, the Random Forest model demonstrated exceptional precision at 0.991, suggesting a very low false positive rate [4][16]. Although Decision Tree and K-Nearest Neighbors (KNN) showed somewhat reduced accuracy, they maintained high recall scores around 0.991, demonstrating promise for usage in systems where detecting every phishing event is a priority [1][6].

Logistic regression, on the other hand, produced an F1-score of 0.941 and a relatively lower accuracy (93.4%) despite being quicker and easier [16]. With an accuracy of only 60.5% and a relatively low recall of 0.292, the Naive Bayes Classifier performed noticeably poorly, suggesting a high miss rate for phishing URLs. This is explained by the strong feature independence assumption of Naive Bayes, which does not hold true in this situation [5][10].

Overall, the findings demonstrate how well neural networks and ensemble models can detect phishing URLs [1][9][17].Strong precision helps lower false alarms, and high recall across top models guarantees that the system can decrease undetected threats. These results encourage the use of these models in practical applications including email filters, browser plugins, and cybersecurity monitoring systems [18][19][21].

When determining if a URL is phishing or not, characteristics like "HTTPS," "AnchorURL," and "Website Traffic" are more significant [10][3][6]. Gradient Boosting Classifier lowers the likelihood of harmful attachments by correctly classifying URLs up to 97.4% into the appropriate category [1][17].

VIII CONCLUSION

This study successfully built a machine learning-based system to reliably and accurately identify phishing URLs. The algorithm was able to differentiate between dangerous and valid links by extracting and evaluating important characteristics from URLs, including the use of HTTPS, IP addresses, URL shortening, domain age, and behavioral indicators. With accuracy scores above 97% and outstanding recall and precision, ensemble learning methods like Gradient Boosting

and Cat Boost fared better than the other models tested. These findings support machine learning's efficacy in combating the escalating threat of phishing attempts. By enabling users to enter URLs and instantaneously obtain predictions, the deployed web interface greatly improves usability.

End users, businesses, and cybersecurity platforms can all benefit from this system's ability to proactively detect and stop phishing threats. Real-time crawling for dynamic feature extraction and interaction with browser extensions for wider impact are possible future improvements.

IX FUTURE WORK

Even though the suggested machine learning and URL-based phishing detection system has shown encouraging outcomes, there are a number of avenues for further development and growth. Adding real-time detection capabilities that can dynamically scan URLs as users interact with websites is one such improvement that might offer instant security. Furthermore, using elements other than the URL, like website content, HTML structure, visual resemblance, and SSL certificate analysis, can help detect more complex phishing assaults and increase detection accuracy.Using deep learning methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which can automatically extract hierarchical features from URLs and webpage content, is another exciting avenue.

Better performance in identifying zero-day threats may also result from ensemble approaches that include several machine learning models or hybrid systems that combine rule-based and learning-based detection techniques.

To keep the system relevant in changing threat landscapes, the model might also be improved by using real-world datasets with regularly updated phishing and genuine URLs. Developing a mobile application or browser plugin that uses this framework can result in useful, user-facing applications. Lastly, an extra degree of protection might be offered by incorporating Natural Language Processing (NLP) tools to examine linguistic clues from the webpage or email content where the phishing URL occurs.

XREFERENCES

[1] Zuochao Dou; Issa Khalil; AbdallahKhreishah; Ala Al-Fuqaha; Mohsen Guizani, “Systematization of Knowledge (SoK): A Systematic Review of Software- Based Web Phishing Detection”, IEEE Communications Surveys & Tutorials, 2017.

[2] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious javascript code", Proceedings of the 19th International Conference on World Wide Web, pp. 281-290, 2010.

[3] Choon Lin Tan, Kang LengChiew, San Nah Sze, “Phishing Website Detection Using URL-Assisted Brand Name Weighting System”, 2014 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) December 1-4, 2014.

[4] R. B. Basnet, A. H. Sung, "Mining web to detect phishing urls", Proceedings of the International Conference on Machine Learning and Applications, vol. 1, pp. 568- 573, Dec 2012.

[5] Mohiuddin Ahmed, AbdunNaserMahmood, Jiankun Hu, "A survey of network anomaly detection techniques", J. Netw. Comput. Appl., vol. 60, no. C, pp. 19-31, 2016.

[6] LuongAnh Tuan Nguyen, Ba Lam To, HuuKhuong Nguyen1 and Minh Hoang Nguyen, “A Novel Approach for Phishing Detection Using URL-Based Heuristic”, 2014 International Conference on Computing, Management and Telecommunications (ComManTel), IEEE 2014.

[7] S. Carolin Jeeva, Elijah Blessing Rajsingh, "Intelligent phishing urls detection using association rule mining", Human-centric.

[8] S. Duffner and C. Garcia, "An Online Backpropagation Algorithm with Validation Error-Based Adaptive Learning Rate," in Artificial Neural Networks – ICANN 2007, Porto, Portugal, 2007.

[9] R. M. Mohammad, F. Thabtah and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, no. 2, pp. 443-458, 2013-B.

[10] HibaZuhair, Ali Selamat, MazleenaSalleh, “Feature selection for phishing detection: a review of research”, International Journal of Intelligent

Systems Technologies and Applications, Vol. 15, No. 2, 2016.

[11] Huang, H., Tan, J. and Liu, L. (2009) „Countermeasure techniques for deceptive phishing attack", International Conference on New Trends in Information and Service Science (NISS"09), 30 June–02 July, 2009, China, pp.636–641.

[12] Mayuri, A. and Tech, M. (2012) „Phishing detection based on visual similarity", International Journal of Scientific and Engineering Research (IJSER), Vol. 3, No. 3, March, pp.1–5

[13] Anvil, Search Engine Optimization Whitepaper, http://www.anvilmediainc.com/wpcontent/uploads/ami_seo_whitepap er_1104.pdf

[14] FadiThabtah, Rami M. Mohammad, Lee McCluskey, "A Dynamic Self- Structuring Neural Network Model to Combat Phishing", 2016 International Joint Conference on Neural Networks (IJCNN), 2016.

[**15]** UCI Machine Learning Repository, https://archive.ics.uci.edu/ml

[16] http://dataaspirant.com/2017/05/22/random-forest-algorithm-machinelearing

[**17]** Peng Yang, Guangzhen Zhao, Peng Zeng, "Phishing Website Detection based on Multidimensional Features driven by Deep Learning", IEEE Access, 2018.

[18] Chuan Pham, Luong A. T. Nguyenz, Nguyen H. Tran, Eui-Nam Huh, Choong Seon Hong, "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks", IEEE Transactions on Network and Service Management, 2018.

[19] Tommy Chin, Kaiqi Xiong and Chengbin Hu, "Phish Limiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking", IEEE Access, 2018.

[20] Christopher N. Gutierrezy, Taegyu Kimy, Raffaele Della Cortez, Jeffrey Averyyx, Dan Goldwassery, Marcello Cinquez, Saurabh Bagchiy, "Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks", Transactions on Dependable and Secure Computing, 2018.

[21] Longfei Wu, Xiaojiang Du, and Jie Wu, "Effective Défense Schemes for Phishing Attacks on Mobile Computing Platforms", IEEE Transactions on Vehicular Technology, 2018