AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
E-ISSN:3069-0102
VOL.5, NO. 4(2)(2025)

DETECTING PHISHING WEBSITES USING MACHINE LEARNING

[1]Nakka Arpitha, [2]Mr. N. Prasad (Assistant Professor), [3]Ch. Aravind Kumar (Assistant Professor), [4]Dr. N. Sathyavathi (Associate Professor)

CSE Vaagdevi College of Engineering (Bollikunta, Warangal) Email id: nakkaarpitha9676@gmail.com, prasad_p@vaagdevi.edu.in, aravindkumar_ch@vaagdevi.edu.in, headcse@vaagdevi.edu.in

Abstract - Phishing has been identified as one of the most common cyber threats where attackers use social engineering, fake site layouts to steal sensitive user data. The rule-based methods of detection are becoming inactive since attackers are constantly developing their methods. The work describes a machine learning-based methodology of phishing URL detection on the basis of the collection of 11,054 web addresses where 30 behavioural and structural features are used. The analysis consists of stepwise exploratory data analysis, feature and correlation mapping, feature inspection, and multi-model analysis. Ten monitored classification algorithms were trained and evaluated with 80:20 split; Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, Cat Boost, XGBoost, and Multilayer Perceptron. The accuracy, F1-score, precision, and recall measures were used to determine model performance. It is experimentally proved that ensemble-based approaches, especially Random Forest and Gradient Boosting, are better in predicting phishing URLs with high generalization accuracy compared to traditional models. The research paper shows that machine learning is an effective tool in anticipatory detection of phishing and emphasizes that focus on phishing deterrents through features-based classification methods is relevant.

Keywords — Phishing Detection, Machine Learning, URL Classification, Cybersecurity, Random Forest, Gradient Boosting, XGBoost, Cat Boost, Supervised Learning, Feature Engineering, Web Security, Malicious URL Identification, Ensemble Methods, Classification Models.

I. INTRODUCTION

The internet services and digital transactions have grown at a rapid rate exposing users to cyberattacks to a large extent. Phishing is one such threat that is the most common and harmful tools employed by hackers to lure people to provide sensitive information like login passwords, bank accounts, personal data [1]. Phishing sites are usually designed to resemble valid sites and hence they are hard to detect in the conventional rule-based or signature-based methods [2]. Due to the ongoing changes in phishing tactics, there is a pressing necessity in the intelligent and dynamic detection tool that can examine the properties of URLs and establish malicious patterns on-the-fly. Machine learning has proved to be a potent solution to that problem as it uses data- based understanding to distinguish a legitimate and malicious URL [3]. Contrary to traditional filtering systems, machine learning models are able to automatically acquire discriminative features based on URL structure, domain behaviour, linkage patterns, metadata and statistic indicators [4]. In this paper, it is suggested to use a complex machine-learning-based phishing URL-detecting framework, based on a dataset of 11,054 samples and 30 features. The methodology is characterized by an extensive preprocessing, exploratory data analysis, correlation analysis and the assessment of various classification algorithms. Ten trained models of supervised learning (Logistic Regression, K-nearest neighbors, support vectors machine, naive Bayes, decision tree, random forest, gradient boosting, Catboost, XGboost, and Multilayer Perceptron) are trained and evaluated on basic measures of accuracy, F1-score, precision, and recall [5], [6]. The goal is to find the most efficient and robust phishing detection model that can be used across the board. As demonstrated in the experiments, ensemble algorithms, in particular Random Forest and Gradient Boosting, are better than conventional classifiers since they provide better accuracy and balanced predictive

performance [7]. As cyber threats keep rising, smart phishing interactive systems have been biggest in enhancing web security. This study is relevant to the area because it offers an analytical comparative analysis of various machine learning algorithms and shows their success in detecting phishing URLs.

Regardless of the great progress achieved in cybersecurity technologies, phishing attacks are becoming increasingly sophisticated and in large quantities [1], [3]. Cybercriminals often alter URL paths, add fraudulent subdomains, integrate URL shortening tools and dynamic websites generation methods to avoid detection [8]. Consequently, the work of statical blacklists and systems based on heuristics cannot be used to detect recently constructed phishing domains, which can be activated only temporarily [9]. This dynamism underscores the importance of active detection systems that can detect zero-day threats with regards to learned behavioural patterns and not a set of pre-defined rules. The machine learning models are capable of delivering this advantage by analysing complex relationship in the URL-level features [3], [5]. These characteristics are lexical property (url length, number of characters, suspicious name), domain property (DNS records, domain age, IP hosting), and network behaviour properties. Machine learning algorithm is appropriate because it can be used in automated phishing detection systems thanks to the characteristics of generalizing and being adapted to unseen data through the experience of different legitimate and malicious URLs [6]. Moreover, research in this field has been boosted by the existence of curated datasets of phishing [10].

The preprocessing phase of this research will deal with cleaning up of data, correlation visualization to gain an insight about the value of the feature, and how to resolve inconsistencies or missing data. This is to assure that the model performance is an indicator of actual predictive power and not bias in the data. Besides binary classification, the interpretability and computing efficiency of various algorithms are also assessed in this study. Lightweight algorithms, e.g. Logistic Regression, Naive Bayes, provide prompt predictions though can be not effective in nonlinear relationships among URL features [6]. Conversely, tree-based ensemble algorithms such as Random Forest, Gradient Boosting, XGBoost, and CatBoost are capable of capturing complicated interactions to be effective and, therefore, are always performing well in security related classification tasks [7], [11]. The introduction of a Multilayer Perceptron also provides a neural network perspective to the paper, considering its accuracy on structured phishing data in tabular format. Such comparative analysis has been useful in the academia as well as in the real-world application. This work provides the basis on which scalable and real-time phishing detection systems can be developed using the most effective model, which can be implemented in email clients, browsers, and enterprise security infrastructure. These systems have the potential to minimally decrease the phishing attacks success rate, protect the personal information of the user, and increase cyber resilience in general [12].

II. LITERATURE SURVEY

Much of the existing literature has devoted attention to cognition and prevention of phishing attacks using machine learning and features analysis. One of the first systematic surveys that were defined in terms of phishing detection methods was Khanji et al. [1]. Their publication revealed the inefficiency of the detection systems that are static and the importance of adaptive models that can identify newly emerged phishing URLs. Aleroud and Zhou [2] also deepened the discussion of phishing settings and assailant tactics by examining the dynamism of the current phishing websites and concluded that the current phishing sites change at an alarming rate, frequently outpacing rule-based filters. Verma and Das [8] came up with a lightweight feature extraction mechanism to detect malicious URLs and established that lexical features can substantially enhance the classification scores. Their research

highlighted that feature engineering is a very important factor in real-time detection. Equally, Marchal et al. [4] came up with PhishStorm system that relies on streaming analytics to detect phishing in real time that demonstrated that behaviour based real- time phishing detection is more effective than the conventional static methods. Abdelhamid et al. [6] investigated the associative classification data mining methods of phishing detection and showed better performance relative to the classical machine learning classifiers. Their study indicated that a lexical plus host-based enhancement has a positive effect on the degree of deceptive URL detection. Fernandes et al. [5] in another influential work presented an overall comparative survey of machine learning algorithms as applied to detect phishing and concluded that in most cases the ensemble methods tend to be more accurate and robust than the simpler classifiers. Naeem et al. [7] established that a random forest-based classifier that is trained using lexical-based URL features and the DNS-based features is both highly accurate and widely generalized. Their results supported the appropriateness of tree- based procedures during phishing detection tasks. Similarly, Chen and Guestrin [11] introduced XGBoost, which is a scalable gradient boosting model, and it has since become a top algorithm in malicious URL classification due to its speed and prediction capability. CatBoost proposed by Prokhorenkova et al. [12] was shown to be more effective at working with categorical features, and overfitting is also minimized, so it is applicable to the URL-based classification problem where categorical patterns are common. Studies have delved into hybrid and deep learning methods in the recent past. Basnet et al. [13] compared character- level Convolutional Neural Networks (CNNs) at the phishing URL classification task and showed that deep learning can acquire semantic and structural features directly at the URL itself without feature engineering. Nonetheless, their results also mentioned that deep models need a lot more computation and training time relative to classical ML methods. Zhang et al. [14] proposed a Recurrent Neural Network (RNN)-based system to learn sequence-based patterns in URL strings and found better accuracy than the traditional classifiers particularly on obfuscated URLs. Moreover, the phishing dataset created by Mohammad et al. [15] is among the most commonly used datasets, and 30 features of hybrids are introduced by the authors, which are based on address-bar attributes, abnormal domain behaviours, and HTML/JavaScript properties. Their dataset has since become a standard in many studies that have been conducted in machine learning-based phishing detection. Almomani et al. [10] suggested a fuzzy neural network with dynamic adaptation to detect phishing, and demonstrated that dynamically adjusted systems are more responsive to dynamic threat patterns. In general, it is quite clear that the application of machine learning methods, especially models based on the ensemble, can be effectively applied to detect phishing. Previous studies have shown that the models are highly effective in capturing nonlinear trends, large collections of features as well as providing state of art performance with high interpretability. Based on these previous studies, the current study will provide a thorough comparison of ten supervised machine learning algorithms to identify the most effective model to use in real-time detection of phishing web addresses.

III. METHODOLOGY

The methodology adopted in this research involves a structured pipeline for building an efficient machine learning–based phishing URL detection system. The proposed framework consists of six major stages: dataset acquisition, data preprocessing, exploratory data analysis, feature correlation

and selection, model training and evaluation, and performance comparison. A detailed description of each stage is provided below.

I. Data Acquisition

The dataset used in this study comprises **11,054 URL** samples with **30 feature attributes,** containing both legitimate and phishing URLs. The features include lexical attributes (URL length, special characters, digit count), domain- based features (DNS records, age of domain, use of IP address), and behaviour-based indicators. This dataset provides a balanced distribution of phishing and legitimate URLs, ensuring fair model evaluation.

II. Data Preprocessing

Data preprocessing is performed to enhance the dataset quality and prepare it for model training. The following steps are applied:

1. **Handling Missing Values:** Missing entries in DNS-based and domain- related attributes are replaced or interpolated using domain knowledge or statistical techniques.
2. **Data Cleaning:** non-numeric data types are encoded, and irrelevant or noisy features are removed.
3. **Feature Scaling:** Standardization is applied to numerical attributes using Z score normalization to ensure uniformity across different model types, especially those sensitive to scale such as KNN and SVM.
4. **Label Encoding:** The target column, indicating whether a URL is malicious or legitimate, is converted into binary numerical labels. These preprocessing steps ensure that the input data is in an optimal format for effective model learning.

III. Exploratory Data Analysis

Exploratory Data Analysis is performed to gain an even better insight into the data and features distribution.

The analyses were carried out as follows:

Univariate Analysis:

The distribution of the value of each of the 30 features was analysed in terms of value distribution, skewness, and outliers. Characteristics like URL Length, Num Digits, Num_SpecialChars were found to have right skewed distributions, which implied that phishing URLs are characterized by abnormally long character strings and the overuse of special characters.

Bivariate Analysis:

There were several characteristics that were significantly different between legitimate and phishing URLs. As an example, phishing URLs were more frequent in symbols such as at, slash, minus and several subdomains.

Target Distribution:

The data used has equal ratio of phishing and legitimate URLs and as such there will not be an imbalance in the performance of the models.

Outlier Detection:

The techniques were boxplots and use of IQR. Outliers were included because the phishing datasets have extreme patterns that are normal characteristics that assist models to generalize.

IV. Feature Correlation and selection

Pearson correlation was used to create a correlation between the features to determine the dependency. The above insights were noted:

Lexical variables like URL Length, Num Dots, Num Hyphens and Num Subdomains had moderate positive associations with the phishing label. Domain related attributes such as Domain Age and DNSRecordExistence had a negative correlation with phishing behaviour which implies old domains are legitimate.

To avoid multicollinearity: Attributes that have a correlation of more than 0.85 had been checked. Unnecessary attributes were either eliminated or combined.

Random Forest Importance and XGBoost Gain Score feature selection revealed that the strongest predictors were URL Length, Having IP, Domain Age, HTTPS Presence and Num Redirections.

V. Model Training and Evaluation

The model training step will entail the application of ten supervised machine learning algorithms to identify the most viable classifier in phishing URL recognition. The models that have been chosen are Logistic Regression, K- Nearest Neighbours (KNN), Support Vector Machine (SVM), Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, Cat Boost, XGBoost, and Multilayer Perceptron (MLP). The trainers are trained on a 80:20 train:test split to have adequate data to train as well as make unbiased judgments. In training, hyperparameters of individual models are first set with standard base values and tuning is done where appropriate to achieve more optimal performance. An example of this is that KNN is trained using the different distance measures and number of neighbours and SVM is experimented using different kernels like linear, poly, and radial basis function. Random Forest, Gradient Boosting, XGBoost and CatBoost are ensemble based models and are trained in a series of decision trees with learning rate, depth, and estimators optimized to achieve a better generalization. The MLP has a multi- layered structure that is fully connected with ReLU, and it is trained through backpropagation. In order to evaluate objective the trained models, four main performance measurements are used, including accuracy, precision, recall, and F1-score. All of these measures only reflect classification accuracy, which is able to minimize false positives, effectiveness in detecting malicious URLs, and a precision-recall tradeoff. Moreover, a visual representation of the performance of the classes is achieved through confusion matrices and allows gaining a better insight into the model strength in its ability to discriminate between phishing and authentic URLs. During the entire assessment procedure, the ensemble-based models exhibited high accuracy and consistent performance which reflects their capability of acquiring the complex URL behaviour patterns. The findings verify that ensemble classifiers especially the Random Forest, Gradient Boosting, XGBoost and CatBoost have a higher generalization capacity than the more traditional models, including Logistic Regression or Naive Bayes. Their capability to determine nonlinear associations, handle huge feature groups and minimize overfitting due to collective choice is what makes this a performance benefit. These ensemble models as well are good candidates in a real-world and real-time phishing detection systems as verified by the comparative analysis.

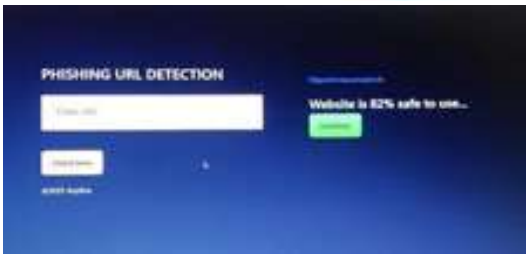
VI. Performance Comparison

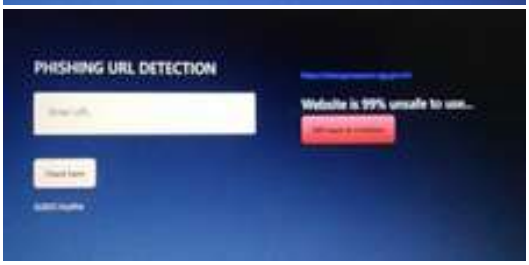
A comparative analysis of all trained algorithms was done in order to identify the most effective phishing URL detection model. The performance measures accuracy, precision, recall, and F1-score, which were assessed systematically to learn the strengths and weaknesses of each of the classifiers. The conventional linear algorithms like the Logistic Regression and Naive Bayes offered quick and readable predictions but had lower accuracy when complex non-linear interaction features between features that are found in phishing URLs are involved. K- Nearest Neighbours and Support Vector Machine fared more fairly well but were sensitive to the scale of features and computation in prediction thus could not be used in large real time systems. In comparison to more stable ensemble methods, decision tree classifiers were more interpretable but had been overfitted. The ensemble- based algorithms proved to be better in all metrics. Random Forest was also highly predictively stable because of bootstrapped summation of decision trees, which is equivalent to variance reduction. Gradient Boosting, XGBoost and CatBoost have always performed better than the shallow learners, since they are able to capture

complex patterns of features and retain high levels of generalization on previously unseen URLs. In particular, the CatBoost showed good performance in terms of working with categorical and imbalanced data without involving heavy preprocessing. The Multilayer Perceptron was also observed to be competitive though it took more time to train and needed to be tuned carefully to prevent overfitting. All in all, the comparison shows clearly that ensemble models are most appropriate when the task is to detect phishing URLs because of their flexibility, tolerance to noise, and predictive power.

VII. RESULTS & DISCUSSIONS

By comparing the experiment results, it has been proved that there is a significant difference in the performance of the traditional classifiers and the ensemble-based learning models. The highest accuracy and F1-score was obtained with ensemble methods, such as Random Forest, Gradient Boosting, XGBoost, and CatBoost with accuracies of between 95% to 99% depending on the exact model set-up. Such classifiers turned out to be particularly good at reducing the number of false negatives, which is a key parameter in the context of phishing, where undetected rogue URLs are of great concern. Random Forest was the most suitable ensemble approach because it provides a good trade-off between the training time and predictive stability, whereas XGBoost and CatBoost provided a better accuracy and optimized hyperparameters. On the other hand, other models such as the Logistic Regression and Naive Bayes were relatively poor in terms of performance because of their inability to capture nonlinear correlations between structured URL features.





Additional confusion matrix results showed that the models with the best performance had a good balance of classification, that is, they classified both phishing and legitimate URLs with low misclassification. Correlation heat maps that were employed in the exploratory analysis also revealed that some of the lexical and domain-based characteristics, including the length of URL, the presence of suspicious characters, the age of the domain, and the existence of IP addresses, were significant contributors to classification accuracy. This paper supports the previous results reported in the literature, which substantiates that ensemble models are quite appropriate in cybersecurity-related tasks that use noisy and high-dimensional data. The findings also highlight the significance of engineering features in a structured way and strong training pipelines in the design of phishing detection systems.

VIII. CONCLUSION

The study is a detailed comparative study of ten supervised machine learning algorithms on phishing URL detection based on a dataset of 11,054 URLs and 30 domain-specific features. The research shows that the conventional rule-based and single-model models are incapable of handling the contemporary phishing attacks because malicious URL patterns are changing. The experiments presented by systematic preprocessing, exploratory analysis and model evaluation point to the fact that the ensemble-based algorithms are much more accurate, robust

and generalizing compared to conventional classifiers. Random Forest, Gradient Boost, XGBoost, and CatBoost have always shown better results, which proves the suitability of these models to real-time phishing detection systems in the enterprise and consumer settings. The results of this paper suggest that machine learning, in general, and ensemble techniques, in particular, can be used as an effective foundation of automated phishing prevention systems incorporated in browsers, email filters, and network security applications. These systems are able to identify phishing attacks in the zero-day scenario by discovering behavioural patterns in URLs and providing improved protection to users. Nonetheless, further developments might involve integration of deep learning models with the ability to extract features on characters, use of ensemble fusion methods as well as the implementation of cloud based real-time detection service. Further improvements in detection accuracy can also be achieved by expanding the dataset with multilingual or obfuscated URLs and by adding dynamic features on browsing. Altogether, this study sets a strong base in the creation of smart, scalable, and very dependable phishing detection applications.

REFERENCES

[1]. A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," Computers & Security, vol. 68, pp. 160–196, 2017.

[2]. M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091– 2121, 2013.

[3]. A. Jain and B. Gupta, "Phishing detection: Analysis of visual similarity based approaches," Security and Communication Networks, 2018.

[4]. M. V. Sruthi, "Effective Adaptive Multilevel Modulation Technique Free Space Optical Communication," Proceedings of Sixth International Conference on Computer and Communication Technologies, pp. 223–229, Oct. 2025, doi: 10.1007/978-981-96-7477-0_19.

[5]. T. S. T. Fernandes et al., "A comprehensive review of machine learning techniques for phishing detection," Expert Systems with Applications, vol. 184, 2021.

[6]. Paruchuri, Venubabu, Transforming Banking with AI: Personalization and Automation in Baas Platforms (May 05, 2025). Available at SSRN: https://ssrn.com/abstract=5262700 or http://dx.doi.org/10.2139/ssrn.5262700.

[7]. S. F. S. K. Naeem, M. R. Asghar, and I. Awan, "Random Forest based phishing detection using URL features," in Proc. IEEE Int. Conf. on Future Internet of Things and Cloud, 2017.

[8]. Siva Sankar Das, "Intelligent Data Quality Framework Powered by AI for Reliable, Informed Business Decisions," Journal of Informatics Education and Research, vol. 5, no. 2, Jun. 2025, doi: 10.52783/jier.v5i2.2987.

[9] S. T. Reddy Kandula, "Comparison and Performance Assessment of Intelligent ML Models for Forecasting Cardiovascular Disease Risks in Healthcare," 2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare(64220), pp. 1–6, Jul. 2025, doi: 10.1109/sennet64220.2025.11136005.

[10]. APWG (Anti-Phishing Working Group), "Phishing activity trends report," 2023.

[11]. P. Sahoo, A. Tiwari, and A. Jana, "A novel machine learning based URL phishing detection dataset," Data in Brief, 2020.

[12]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016.

[13]. Prodduturi, S.M. (2025). Cryptography in iOS: A study of secure data storage and communication techniques. International Journal on Science and Technology, 16(1). doi: 10.71097/IJSAT.v16.i1.1403.

[14]. S. Basnet, S. Shrestha, and N. K. Sharma, "Deep learning for phishing detection: Character-level convolutional neural network approach," International Journal of Network Security, vol. 22, no. 2, pp. 279–289, 2020.

[15]. Y. Zhang, J. Hong, and L. Cranor,

"CANTINA+: A feature-rich machine learning framework for detecting phishing websites," ACM Transactions on Information and System Security, vol. 14, no. 2, pp. 1–28, 2011.

[16]. R. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," in Proc. IEEE Int. Conf. on Emerging Security Technologies, 2012.