URL Phishing Detection System Utilizing Catboost Machine Learning Approach

Lim Chian Fang[1†], Zakiah Ayop[1†], Syarulnaziah Anawar[1†], Nur Fadzilah Othman[1†], Norharyati Harum[1†], Raihana Syahirah Abdullah[1†]
chianfang990905@gmail.com zakiah@utem.edu.my syarulnaziah@utem.edu.my
fadzilah.othman@utem.edu.my norharyati@utem.edu.my raihana.syahirah@utem.edu.my
Information Security Forensics and Computer Networking (INSFORNET),
Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka (UTeM)

Summary
The development of various phishing websites enables hackers to access confidential personal or financial data, thus, decreasing the trust in e-business. This paper compared the detection techniques utilizing URL-based features. To analyze and compare the performance of supervised machine learning classifiers, the machine learning classifiers were trained by using more than 11,005 phishing and legitimate URLs. 30 features were extracted from the URLs to detect a phishing or legitimate URL. Logistic Regression, Random Forest, and CatBoost classifiers were then analyzed and their performances were evaluated. The results yielded that CatBoost was much better classifier than Random Forest and Logistic Regression with up to 96% of detection accuracy.

Key words:
Phishing; URL; CatBoost; Logistic Regression; Random Forest.

1. Introduction

The development of various websites including online banking, education, and social media is driven by the rapid growth of the internet in recent years. Phishing attacks have increased significantly and are now widely regarded as the most serious new internet crime, potentially causing people not to trust in e-business. As a result, phishing creates adverse effect on internet banking, e-business, organizational revenues, client partnerships, and overall market operations [1].

Phishing is a type of identity fraud that utilizes social engineering strategies as well as complex attack vectors to obtain financial details from unsuspecting users [2]. Users are tricked by phishers who use social engineering techniques to lure them to enter a phishing website whereby the website requires personal or financial information. When web users are tricked to access a phishing website, they are duped into completing the website's goal.

Since phishing webpages focus on banks and companies, , web users' detection of web phishing attacks has become critical. However, various advanced techniques used by attackers to confuse web users, has made the detection of a phishing website difficult. To identify phishing webpages, several traditional strategies focus on set blacklists and whitelists databases. These methods, however, are ineffective since a new website can be created in a matter of seconds. As a result, these strategies fail to determine if a new website is phishing or not in real-time [3].

Machine learning algorithms have gained popularity in recent years as a way to improve the generality of malicious URL detectors. [2]–[4]. Compared to other approaches, the machine learning approach performs better in terms of accuracy and performance [2]. However, new attack schemes emerge which manipulate all the security features available [1]. This paper compared the performance of machine learning algorithms by using URL-based features and training sets of known attacks. Then, the best performance algorithm to categorize new phishing sites was selected.

The remainder of this paper is organized as follows. Section 2 describes the background and state-of-art techniques, its advantages, and limitations. In Section 3, we go over the specifics of our proposed phishing detection system. Section 4 discusses the outcomes of the suggested system's evaluation. Finally, Section 5 is a closure which provides an overview of the study and addresses the scope of future research.

2. Related Works

Phishing URL detection has been solved using several approaches. The machine learning approach is explained by detailing the fundamental principles.

Machine learning uses data and expertise to automate the construction of analytical models. Machine learning is a subfield of AI that concentrates on computers that can study data, recognize patterns, and make decision with very little human input. It aims to find patterns in data and then make predictions based on these often-complex patterns to

Manuscript received September 5, 2021
Manuscript revised September 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.9.39

answer questions, identify and analyze trends, and assist in problem-solving.

Mahajan et al. [5] proposed detecting phishing websites by using machine learning algorithms. The objective of this paper was to identify phishing URLs and to select the best machine learning algorithm by comparing each algorithm's accuracy rate, false-positive rate, and false-negative rate. First, features are extracted from the URLs such as the existence of IP address, @ symbol, the number of dots in the hostname, and others. Then, Dataset is split into training testing set in 50:50, 70:30, and 90:10 ratios respectively. Phishing websites are detected by using Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) algorithms. As a result, the RF algorithm is the most precise detection which achieves 97.14% of accuracy and the lowest false positive rate. Machine learning models perform better with the existence of more data as training data.

Patil et al. [6] proposed Detection and Prevention of Phishing Websites by using Machine Learning Approach. First, URL is compared by using Blacklist and Whitelist Approach. If URL is found in Blacklist and Whitelist then the URL is a phishing website. If URL is not found in Blacklist and Whitelist, then, the features of URL are extracted by using Heuristic and Visual Similarity Approach. Next, to predict the result, researchers use machine learning algorithms such as LR, DT, and RF to analyze these various features of URLs and webpages. The system emphasizes the efficiency gained by combining heuristic features, visual features, and a blacklist and whitelist approach with machine learning techniques. The results show that the LR and DT achieve 96.23% accuracy whereas RF achieves 96.58% of accuracy.

Mao et al. [7] analyzed the pros and cons of many machine-learning-based phishing detection techniques, as well as their performance on a real-world dataset such as SVM, DT, RF, and AdaBoost (AB). The classifier is SVM, and four measures related to the gamma parameter in the SVM algorithm are tested; accuracy, precision, f1, and recall. The results show that the accuracy is 96.95%. DT classifiers are evaluated in relation to the parameter max_depth in the DT algorithm. The results indicate the accuracy of DT is 93.68%. AB classifiers are evaluated regarding the parameter n_estimators in the AB algorithm. The accuracy is 94.5%. RF is the classifier and four measures related to the parameter estimators are evaluated. Based on results, the accuracy of RF is the highest among the four-machine learning which is 97.31%.

Several types of machine learning algorithms which are supervised and unsupervised machine learning are included. Supervised learning known as all data is tagged to build algorithms so that algorithms can predict the results and classify data accurately whereas unsupervised learning is defined by using untagged data so that algorithms learn patterns from input data. Supervised learning classification is effective in spam detection, revenue forecasting, and fraud detection. The supervised learning process improves by continuously measuring the system's outputs and fine-tuning it to get similar to the target accuracy. Ali [3] stated that a supervised learning algorithm examines the training phishing sites datasets and generates a predictor that can accurately categorize the right category of unknown datasets and detect freshly generated phishing websites rapidly. The use of unsupervised learning has various drawbacks. Because the exact outcome is not known in advance, unsupervised learning models may produce less accurate results than supervised learning models. So, supervised learning is the most suitable method to detect phishing URLs.

Among all other classification algorithms, the RF approach offers the best results with the highest accuracy. Several studies have shown that by implementing an RF classification model, more than 95% assault detection accuracy may be achieved. However, the Boosting algorithms are able to produce better experimental results compared to RF in terms of accuracy and other parameters for detecting phishing URLs. XGBoost (XGB) provides great accuracy (up to 97%) in a short amount of time, and the XGB classifier is the most accurate of all the classifiers. Together with accuracy, the Boosting algorithm is also consistent in terms of precision [8]. Boosting is a collection of algorithms and the main goal is to turn weak learners into strong ones. Boosting algorithms can enhance the model's prediction accuracy by a significant number of features. CatBoost (CB) is the latest boosting algorithm in machine learning and CB can increase the model's performance in Medicare Fraud Detection compared to XGB. In terms of AUC, the categorical feature for XGB and CB enhances performance, and CB's performance is statistically significantly higher. CB and XGB achieve nearly identical AUC on a purely numerical dataset. However, the XGB model having a faster training time compared to the CB model [9]. CB can perform better than XGB in fraud detection. CB model that has been trained makes prediction much quicker than XGB. Research about CB-based phishing URL detection is conducted to compare detection accuracy with LR and RF.

Due to the pandemic, most government and corporate activities, educational activities, companies, and non-commercial activities have shifted online. People are increasingly using the internet to conduct daily jobs. As a result, having a comprehensive phishing attack detection system with better accuracy and response time has become more critical. [2].

3. Methodology

3.1 Dataset

The dataset for this study was obtained from the UCI repository [10]. This dataset was mainly gathered from the MillerSmiles archive, Phish Tank archive, and Google searching operators which consisted of 31 columns and 11055 URLs (4898 legitimate, 6157 phishing). The value presented in each attribute was -1, 0, and 1. -1 represented legitimate, 0 represented suspicious and 1 represented phishing. The dataset split into two parts, 80% of the training dataset and 20% of the testing dataset, so that the machine learning algorithms learn from the data and make predictions [3].

3.2 Comparison

Comparison between 3 machine learning algorithms such as LR, RF, and CB were conducted in order to determine which machine learning algorithms achieved the highest accuracy in phishing URLs detection based on the dataset. The confusion matrix was used to measure a classification algorithm performance. The confusion matrix compared actual data with the predictions of the machine learning classifier, hence, giving a complete picture of the classification model' performance as well as their inconsistencies.

1) Logistic Regression (LR)

LR is a classification algorithm used to categorize data into a binary set of classes. The output of LR, the probability value that may be assigned to two or more distinct groups is transformed by the logistic sigmoid function. This method converts any real value into a range between 0 and 1. The equation of Sigmoid Function is shown below [11]:

$$f(x) = \frac{1}{1+e^{-(x)}} \tag{1}$$

2) Random Forest (RF)

RF predicts by averaging the outputs of several randomized decision trees built during the training phase. The Gini Index is used to determine the branching of nodes on a decision tree during classification. The equation of the Gini Index is shown below [12]:

$$Gini = 1 - \sum_{i=1}^{c} (Pi)^2 \tag{2}$$

3) CatBoost (CB)

CB uses Ordered Target Statistic (OTS) and Order Boosting (OB). CB's use of OTS and OB makes it a great option for datasets with categorical data because OTS and OB provided unusual training data, so CB can systematically changed its estimate for the unusual data. OTS and OB used random permutations of the training examples to combat the prediction shift caused by a specific type of target leakage found in all existing gradient boosting algorithms. The base predictor in CatBoost was binary decision trees. The following is an equation of the estimated output [13].

$$Z = H(x_i) = \sum_{j=1}^{J} c_j 1_{\{x \in R_i\}} \tag{3}$$

3.3 Classification

The most accurate machine learning algorithm was used to create a URL Classifier. The features of the URL were extracted once URL was entered. The features were divided into 4 parts namely address bar-based features, abnormal-based features, HTML and JavaScript features, and domain-based features. After extracting the features of the URL, the machine learning model was loaded to classify the URL as phishing or legitimate. Table 3.1 shows the feature extraction which includes feature group, features factor indicator, and feature value.

Table 1: Feature extraction

Feature Group	***Features Factor Indicator***	***Features Values***
Address Bar based Features Right	With the IP Address	{ -1,1}
	Long URL	{-1,0, 1}
	Using the “Tiny URL”	{ -1,1}
	Using “@” in the URL	{ -1,1}
	Redirecting using	{ -1,1}
	Adding Prefix or Suffix	{ -1,1}
	Sub Domain and Multi-Sub Domains	{-1,0, 1}
	SSLfinal_State	{-1,0, 1}
	Domain Registration Length	{ -1,1}
	F avicon	{ -1,1}
	Using Non-Standard Port	{ -1,1}
	“HTTPS” Token	{ -1,1}
Abnormal Based Features	Request URL	{ -1,1}
	URL of Anchor	{-1,0, 1}
	Links in tags	{-1,0, 1}
	Server Form Handler	{-1,0, 1}

	Email Submission	{ -1,1}
	Abnormal URL	{ -1,1}
HTML and JavaScript-based Features	Website	{ -1,1}
	Status Bar Customization	{ -1,1}
	Disabling Right	{ -1,1}
	Using Popup	{ -1,1}
	IFrame Redirection	{ -1,1}
Domain based Features	Age of Domain	{ -1,1}
	DNS Record	{ -1,1}
	Website Traffic	{-1,0, 1}
	PageRank	{ -1,1}
	Google Index	{ -1,1}
	Number of Links Pointing to Page	{-1,0, 1}
	Statistical Reports	{ -1,1}

4. Results in Discussion

4.1 Evaluation Metrics

The confusion matrix is a metric for evaluating machine learning classification performance where the output can be more than two classes. There are four values in the confusion matrix. The four values are used to calculate accuracy, precision, recall, and f1 score [7].

True Positive (TP) known as the positive value is predicted correctly. False Positive (FP) known as the negative value is predicted incorrectly where the true value is negative, but it predicts positive value. False Negative (FN) known as the positive value is predicted incorrectly where the true value is positive, but it predicts as negative value. True Negative (TN) defines as the negative value is predicted correctly.

Accuracy is the number of correctly predicted data over the total amount of data. The equation of accuracy is shown below.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Precision defined as the percentage of the actual positive classes divided by the total of predicted positive classes in the classifier. The precision formula is shown as follows:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall represents the percentage of the correctly predicted positive values to the actual count of positive data present in the dataset. The recall formula is shown below.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1-score is the weighted average of precision and recall. The formula of F1-score is shown below.

$$F1 - score = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} \quad (7)$$

4.2 Comparative Analysis

The dataset was carried out on three machine learning classifiers. LR, RF, and CB were used for training and testing and the outcomes of each classifier were experimentally compared with evaluation metrics such as accuracy, precision, recall, and f1 score acquired for both training and testing.

Table 2 summarizes the overall result of phishing URL detection based on the UCI repository dataset. From 11055 URL dataset, 2211 URL were tested in order to predict the result. CB achieved the highest accuracy score which is 97.87%, compared to RF and LR. CB also obtained the highest precision, recall, and F1-score in phishing and legitimate URL, with an average of 98%, as compared to RF and LR. CB classifier outperformed all other performance indicators such as precision, recall, and f1-score. As a result, a CB classifier was used as the final URL classification model due to its best performance in terms of accuracy, precision, recall, and F1-score.

Table 2: Comparative Analysis Between Three Machine Learning Algorithms

Classi fiers	Evaluation Parameter						
	AC	PC		RC		F1-score	
		PU RL	LU RL	PU RL	LU RL	PU RL	LU RL
LR	93.22%	92 %	94 %	92 %	94 %	92 %	94 %
RF	96.43%	97 %	96 %	95 %	97 %	96 %	97 %
CB	97.87%	98 %	98 %	97 %	99 %	98 %	98 %

AC: Accuracy; CB: CatBoost; LURL; Legitimate URL; LR: Logistic Regression; PURL: Phishing URL; PC: Precision; RF: Random Forest; RC: Recall;

From the result, the CB model outperforms the LR and RF. This may be due to the LR model overfits on the training dataset, exaggerating the prediction accuracy on the training dataset, thus, avoiding the model from correctly predicting test results. As this approach has been sensitive to outliers, including data in the dataset that falls outside of the normal range may lead to inaccurate results. When classes are distinct, the estimation method in LR becomes incorrect because of a logistic function that forces the derivatives to be endless and, therefore, leading to computationally unstable [14].

RF does not have a problem when classes are distinct. Instead, when adequate tree pruning methods are utilized, it assists in the reduction of computations. So, the performance of RF is better than LR. The performance of RF is lower than CB because RF may lack readability due to the grouping of decision trees and failed to neglect the importance of each feature. Data with categorical features with varying amounts of features can be a major issue because the RF method prefers those with more values, posing a risk of incorrect prediction [15].

CB eliminates the need for intensive hyper-parameter adjustment because the default parameter in CB produces a good result. It also decreases the risk of overfitting, resulting in more flexible and accurate models [16]. When determining the tree structure, CB uses a strategy to calculate leaf values, that greatly reduces overfitting. CB can evaluate then select the important feature. One of the methods used in CB is Prediction Values Change. It shows the prediction changes on average when the feature value changes. The feature becomes more important when an adjustment in its value causes a large change in the expected value. This is the default method of calculating feature importance for non-ranking metrics. CB provides a novel technique for analysing category features. Some of the most prevalent approaches for encoding categorical data, such as one-hot encoding, result in an infeasibly large number of additional features in the case of features with high cardinality [17]. As a result, CB's accuracy would be superior for data with categorical attributes compared to RF.

Table 3 shows the total time for training and testing the dataset.

Table 3: Training and Testing Time

Classifier	Time(s)
Logistic Regression	5.03
Random Forest	3.19
CatBoost	54.86

Based on table 3, CB spent far more time than LR and RF. CB was the most time-consuming models with 54.86 seconds. RF required the least time for training and testing which was 3.19 seconds whereas LR was between CB and RF with 5.03 seconds. CB was the most time-consuming because CB built 1000 trees by default. RF only operated on a subset of variables in the model, it is fastest to train. As a result, CB consumed more computational resources than RF and LR.

5. Conclusion

This study evaluates the performance of machine learning algorithms in detecting phishing URLs. As a result, a new phishing URL detection model CatBoost-based URL classifier is implemented in this project. The CatBoost algorithm has the best performance, followed by the Random Forest and Logistic Regression. Furthermore, this study will assist web users by allowing them to detect and keep alert to phishing websites in real-time, resulting in a more secure network experience. In other words, this study can be applied in the security domain where cybersecurity authorities can use it to prevent users from visiting phishing websites and develop powerful security mechanisms that can detect and prevent phishing domains from reaching users. However, CatBoost requires more time for training and testing the dataset which consumes more computational resources. The effectiveness of the CatBoost algorithm has been demonstrated through its higher performance over competing algorithms. In terms of future works, Apache Spark framework can be used to improve the Sickit-learn library called Sk-dist. Sk-dist has overcome the limitation of Sickit-learn library such as time-consuming and lagging model training [18].

Acknowledgments

This publication has been supported by Center of Research and Innovation Management (CRIM), Universiti Teknikal Malysia Melaka (UTeM). The authors would like to thank UTeM and INSFORNET research group members for their supports.

References

[1] A.-P. W. Group, “Phishing Activity Trends Report,” 2021.

[2] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, “A comprehensive survey of AI-enabled phishing attacks detection techniques,” *Telecommun. Syst.*, pp. 1–16, 2020.

[3] W. Ali, “Phishing website detection based on supervised machine learning with wrapper features selection,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 72–78, 2017.

[4] A. K. Jain and B. B. Gupta, “PHISH-SAFE: URL features-based phishing detection system using machine learning,” in *Cyber Security*, Springer, 2018, pp. 467–474.

[5] R. Mahajan and I. Siddavatam, “Phishing website detection using machine learning algorithms,” *Int. J. Comput. Appl.*, vol. 181, no. 23, pp. 45–47, 2018.

[6] V. Patil, P. Thakkar, C. Shah, T. Bhat, and S. P. Godse, “Detection and prevention of phishing websites using machine learning approach,” in *2018 Fourth international conference on computing communication*

control and automation (ICCUBEA), 2018, pp. 1–5.

[7] J. Mao *et al.*, “Phishing page detection via learning classifiers from page layout feature,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2019, no. 1, pp. 1–14, 2019.

[8] S. Masurkar and V. Dalal, “ENHANCED MODEL FOR DETECTION OF PHISHING URL USING MACHINE LEARNING.”

[9] J. Hancock and T. M. Khoshgoftaar, “Performance of catboost and xgboost in medicare fraud detection,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020, pp. 572–579.

[10] “UCI Machine Learning Repository: Phishing Websites Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/phishing+websites. [Accessed: 13-Sep-2021].

[11] L. Khairunnahar, M. A. Hasib, R. H. Bin Rezanur, M. R. Islam, and M. K. Hosain, “Classification of malignant and benign tissue with logistic regression,” *Informatics Med. Unlocked*, vol. 16, p. 100189, 2019.

[12] Y. Liu and H. Wu, “Prediction of Road Traffic Congestion Based on Random Forest,” in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, 2017, vol. 2, pp. 361–364, doi: 10.1109/ISCID.2017.216.

[13] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, “CatBoost model and artificial intelligence techniques for corporate failure prediction,” *Technol. Forecast. Soc. Change*, vol. 166, p. 120658, 2021.

[14] “Advantages and Disadvantages of Logistic Regression.” [Online]. Available: https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/. [Accessed: 13-Sep-2021].

[15] “Random Forest Pros & Cons - HolyPython.com.” [Online]. Available: https://holypython.com/rf/random-forest-pros-cons/. [Accessed: 13-Sep-2021].

[16] D. Mwiti, “Fast Gradient Boosting with CatBoost | by Derrick Mwiti | Heartbeat,” 16-Jun-2020. [Online]. Available: https://heartbeat.fritz.ai/fast-gradient-boosting-with-catboost-38779b0d5d9a. [Accessed: 13-Sep-2021].

[17] A. Nahon, “XGBoost, LightGBM or CatBoost — which boosting algorithm should I use?,” 30-Dec-2019. [Online]. Available: https://medium.com/riskified-technology/xgboost-lightgbm-or-catboost-which-boosting-algorithm-should-i-use-e7fda7bb36bc. [Accessed: 13-Sep-2021].

[18] M. S. O. Djedíden, H. Reguieg, Z. M. Maaza, and others, “A distributed intrusion detection system based on apache spark and scikit-learn library,” *J. Appl. Phys. Sci.*, vol. 5, no. 1, pp. 30–36, 2019.

Lim Chian Fang received diploma of IT from Universiti Teknikal Malaysia Melaka in 2020 and currently is a degree student of the Universiti Teknikal Malaysia Melaka. Her research interest includes malicious code pattern analysis, linear regression for malware analysis and security application development.

Zakiah Ayop holds BSc. in Computer Science (2000) from UTM and MSc in Computer Science (2006) at UPM. Currently she is a senior lecturer in Faculty of Information and Communication Technology (FTMK), Universiti Teknikal Malaysia Melaka (UTeM). She is a member of the Information Security, Digital Forensic, and Computer Networking research group. Her research interest are Information System, Internet of Things (IoT) and Network and Security.

Syarulnaziah Anawar holds her Bachelor of Information Technology (UUM), Msc in Computer Science (UPM), and PhD in Computer Science (UiTM). She is currently a Senior Lecturer at the Faculty of Information and Communication Technology, UTeM. She is a member of the Information Security, Digital Forensic, and Computer Networking (INSFORNET) research group. Her research interests include human-centered computing, participatory sensing, mobile health, usable security and privacy, and societal impact of IoT.

Nur Fadzilah Othman received a degree in Computer Engineering in 2008 and master’s in educational technology in 2011 at Universiti Teknologi Malaysia (UTM). In 2017, she obtained her PhD in the field of Information Security at Universiti Teknikal Malaysia Melaka (UTeM). She started her career as a senior lecturer at the Faculty of Information Technology and Communication, UTeM from March 2018. She is an active researcher and has been written and presented a number of papers in conferences and journals.

Norharyati Harum holds her bachelor’s in engineering (2003), MSc. in Engineering (2005) and PhD in Engineering (2012) from Keio University, Japan. She is currently a senior lecturer at Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). Her interests in research area are Internet of Things (IoT), Smart Applications, Embedded System, Wireless Sensor Network, Next Generation Mobile Communication, Radio Frequency Planning and Signal Processing. She is an accomplished inventor, holding patents to radio access technology, copyrights of products using IoT devices.

Raihana Syahirah Abdullah is currently a senior lecturer at the Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. She received her PhD in Network Security from Universiti Teknikal Malaysia Melaka (UTeM). Her Research Interest include intrusion detection, network security, malware analysis and design.