Phishing URLs Detection Using Naives Baiyes, Random Forest and LightGBM Algorithms

Cik Feresa Mohd Foozy [a,b,c,1,*], Muhammad Amir Izaan Anuar [c], Andi Maslan [d], Husaini Aza Mohd Adam [e], Hairulnizam Mahdin [c]

[a] Institut Kejuruteraan Integrasi, Pusat Kecemerlangan Industri-Rail (ICoE-Rel), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[b] Information Security Interest Group (ISIG), Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[c] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[d] Faculty of Engineering and Computer Science, Putera Batam University, Batam, Indonesia
[e] Kolej Komuniti Seberang Jaya, Permatang Pauh, Pulau Pinang, Malaysia.
[1] feresa@uthm.edu.my
* Corresponding author

ARTICLE INFO

Article history
Received March 15, 2023
Revised December 17, 2023

Keywords
detection
feature selection
machine learning
phishing

ABSTRACT

In response to the increasing complexity of phishing attacks, particularly in Malaysia, this study aims to compare the accuracy and precision effectiveness of three machine learning algorithms Naive Bayes, Random Forest and LightGBM in detecting URL (Uniform Resource Locator) phishing. This research employs a comprehensive four-stages methodology including data collection, preprocessing, feature selection, and classification to analyze data for URL phishing attacks classification. The objectives are to identify phishing attack features based on dataset using and machine learning algorithms, to compare between three classification algorithms of Naïve Bayes, Random Forest, and Light Gradient Boosting Model (LightGBM), and to evaluate the model in terms of accuracy, and precision using machine learning algorithms. Through this comparative analysis, the study seeks to develop a phishing detection model, to identify the suitable features and classification algorithms for the datasets. The result accuracy, precision for NB, Random & LightGBM. The Accuracy result of Naives Baiyes is 94.24%, the result of Random Forest is 94.80% and the result of LightGBM is 95.00%.

1. Introduction

In Malaysia, the rise of online conversations and transactions has led to growing concerns about cyberattacks, particularly phishing. Phishing attacks have become a common and unpredictable form of fraud websites. This research project, “A Comparative Study of Naives Baiyes, Random Forest and LightGBM Algorithms for URL Phishing”, tackles the urgent need to effectively detect and prevent these attacks. Traditional methods struggle with scalability and adapting to new cyber threats such as URL Phishing attack. Some existing studies in URL Phishing attack detection The solution uses advanced machine learning algorithms, including supervised, unsupervised, and reinforcement learning, to improve accuracy in identifying legitimate or illegitimate phishing URLs. This approach focuses on handling large datasets and adapting to URL new phishing techniques. This research aim is to provide protection in the expanding digital world, overcoming the limitations of current URL phishing detection methods.

1.1 Problem Statement

As more aspects of our lives become digitalized, the constant threat of cyberattacks becomes ever more prevalent, with URL phishing attacks as a prime case, poses a threatening challenge URL. Phishing attacks can extract some sensitive user information like login credentials and financial data if there are no mechanisms to prevent these issues [1]. Irrelevant or redundant features can lead to decreased accuracy and increased computational costs in machine learning algorithms [2]. And a problem is the inherent inflexibility of conventional phishing detection methods [3]. Transforming URLs into feature matrices can result in losing important information and failing to capture long-distance dependencies. These rigid techniques struggle to adapt quickly to new phishing strategies, reducing their effectiveness. In contrast, using more flexible and adaptable features in phishing detection can improve accuracy. As cyber criminals employ increasingly sophisticated tactics, the inflexibility of current systems leaves individuals and organizations vulnerable, unable to respond dynamically to emerging threats. In addition, comparing three classification algorithms, Random Forest, Naïve Bayes, and Light Gradient Boosting Model (LightGBM), has its strengths and weaknesses, making it evaluate their effectiveness [4]. Three (3) algorithms, comparing their performance metrics to determine the most suitable algorithm for a specific task and making it hard to evaluate their effectiveness in different scenarios [4]-[6]. The objective of the project is as follows.

i. To identify suitable phishing features based on dataset and machine learning techniques.

ii. To compare three classification algorithms Random Forest, Naïve Bayes, and Light Gradient Boosting Model (LightGBM) by SelectKBest and Recursive Forest Elimination (RFE).

iii. To evaluate the A comparative study of Naives Baiyes, Random Forest and LightGBM algorithms for URL Phishing model in terms of accuracy, and precision.

2. Literature Review

2.1 Dataset

The use of a huge phishing dataset from the UCI machine learning website to evaluate their classification algorithm for phishing detection [7]-[8]. The references highlight the importance of utilizing diverse and representative datasets for training and testing, which 40-60 and evaluating machine learning models for phishing URL detection.

It is also essential to emphasize the importance of feature selection, as well as the challenges associated with imbalanced datasets in the context of phishing detection.

In conclusion, UCI machine learning has been utilizing to detect phishing URLs using various classification algorithms. The study utilized a large phishing dataset to evaluate the effectiveness of these models, emphasizing the importance of diverse datasets and addressing issues with imbalanced dataset. Below here Table 1 dataset.

Table 1. ANN Results for North Central states

Citation	Dataset
[8]	Dataset Phishing
[9]	Dataset Phishing
[10]	Uci-Ml-Phishing Dataset
This Research	Dataset Phishing and Uci-Ml-Phishing Dataset

2.2 Data Preprocessing

To optimize the effectiveness of machine learning algorithms like Naives Baiyes, Random Forest, and LightGBM, Preprocessing is one of important step. Various studies have the impact of different preprocessing techniques on these algorithms. The purpose of preprocessing is to handling missing data, normalization, and null of data. Below here Table 2 processing dataset.

Table 2. Processing Dataset

Citation	Normalization	Missing value	Null
[8]	×	×	×
[9]	✓	✓	✓
[10]	×	×	×
This Research	×	✓	×

2.3 Features Selection

Feature selection in machine learning is a process that involves selecting the most relevant features from the input data while eliminating irrelevant or redundant ones to enhance model performance and accuracy. Based on previous research, a few studies regarding features selection have been done by [11]. The research have been applied several features selection such as SelectKBest and Recursive Forest Elimination (RFE). The challenges associated with feature selection for URL Phishing detection, emphasizing the need to address manual feature selection to enhance accuracy and precision [12]. In conclusion, the research highlights an importance of feature extraction in A comparative study of Naives Baiyes, Random Forest and LightGBM Algorithms for URL Phishing. Below here Table 3 about features selection.

Table 3. Features Selection

Citation	Features
[8]	having_IP_Address', 'URL_Length','Shortening_Service', 'having_At_Symbol','double_slash_redirecting','Prefix_Suffix', 'having_Sub_Domain','SSLfinal_State','Domain_registration_length', 'Favicon"port','HTTPS_token','Request_URL','URL_of_Anchor','Links_in_tags', 'SFH','Submitting_to_email','Abnormal_URL','Redirect','LikelyPhishing', 'URLLength','UseOfHTTPS','ContainsSuspiciousKeyword','having_Sub_Domain', 'age_of_domain','Page_Rank','Prefix_Suffix','web_traffic','Statistical_report'
[9]	'URLLength','UseOfHTTPS','Contains SuspiciousKeyword','having_Sub_Domain','age_of_domain','Page_Rank','Prefix_Suffix', 'web_traffic','Statistical_report','having_At_Symbol','SFH','Redirect','Google_Index', Request_URL, popUpWindow, URL_of_Anchor, SSLfinal_State,URL_Length, age_of_domain,Google_Index
[10]	'Using IP address',Long URL',URLhaving @',Symbol Adding',Prefix and Suffix', Sub-Domain(s)',Misuse of HTTPs','Request URL','URL of Anchor','Server Form', 'Handler Abnormal', 'URL_Length','Redirect Page','Using Pop-up','Window', 'Hiding Suspicious Link','DNS','Record Website','web_traffic', 'Age of Domain','Disabling Right Click'
This research	Prefix_Suffix', 'having_Sub_Domain', 'SSLfinal_State', 'Request_URL', 'URL_of_Anchor', 'Links_in_tags', 'SFH', 'age_of_domain', 'web_traffic', 'Links_pointing_to_page', 'Prefix_Suffix', 'having_Sub_Domain', 'SSLfinal_State','Domain_registeration_length', 'Request_URL', 'URL_of_Anchor','Links_in_tags', 'SFH', 'web_traffic', 'Google_Index'

2.4 Classification

Classification in machine learning involves the process of categorizing data into predefined classes or categories based on the features or attributes of the data. The focus was on categorizing phishing URLs using machine learning methods. For instance, a machine learning-based phishing detection framework specifically for URLs was developed by K. L. Chiew et al., emphasizing the importance of machine learning in classifying phishing URLs [13] and N. Nagy et al. [14]. Below here Table 4 of classification.

Table 4. Classification

Citation	Classification
[8]	Naives Baiyes , Linear regression, Decision Tree, Random Forest
[9]	Random Fores,SVM,K-Means, Naives Baiyes
[10]	LightGBM
This research	Naives Baiyes, Random Forest, LightGBM

2.5 Parameter Evaluation

The parameter evaluation of the parameters of the for Naïve Bayes, Random Forest, and LightGBM algorithms in terms of Accuracy, Precision [15][16]. Below here Table 5 parameter evaluation.

Table 5. Parameter Evaluation

Citation	Parameter Evaluation
[8]	Accuracy, Precision
[9]	Accuracy, Precision
[10]	Accuracy, Precision
This research	Accuracy and Precision

3. Research Methodology

The methodology for this research consists of are through the four (4) phases such as of data preprocessing, features selection, classification and detection. Figure 1 below, shows the research methodology of this research.

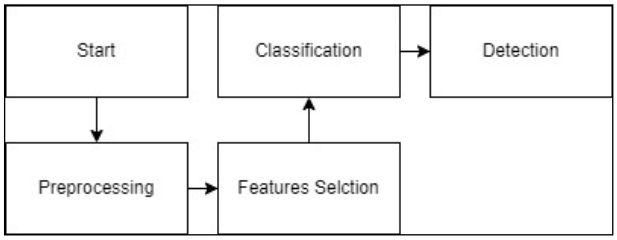

Figure 1. Research Methodology

3.1 Data Preprocessing

Data preprocessing is a critical phase in machine learning that involves transforming raw data into a clean, structured format suitable for analysis by machine learning algorithms. This process aims to improve the quality of the data, enhance model performance, and ensure accurate predictions. Data preprocessing begins with data cleaning, where missing values, outliers, and irrelevant data are identified and handled appropriately. This step ensures that the dataset is free from errors that could impact the accuracy of the model. Below here Figure 2 data preprocessing.

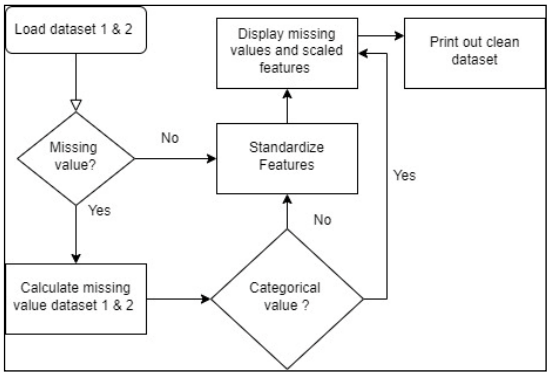

Figure 2. Data Preprocessing

3.2 Features Selection

Feature selection is a step in machine learning aimed at enhancing model performance and accuracy by choosing the most relevant features from the input data. feature selection methods have been extensively researched and applied to various tasks, showcasing their effectiveness in both established and emerging applications [17]. Additionally, methods like forward feature selection and backward feature selection are widely used to optimize machine learning algorithms such as Random Forest and Naives Baiyes [18]. Below here Figure 3 of features selection.

The provided flowchart outlines a comprehensive machine learning workflow involving data preprocessing, feature selection, and model training and evaluation. Initially, the process checks for the presence of an 'index' column in dataset 'd1' and removes it if found. Next, it separates the features (X) and the target variable (Y). The workflow then assesses whether specific indices are used for feature selection; if so, it adjusts these indices to exclude the 'id' column if it exists, otherwise, it selects features based on the provided indices. The data is subsequently split into training and testing (40-60) sets.

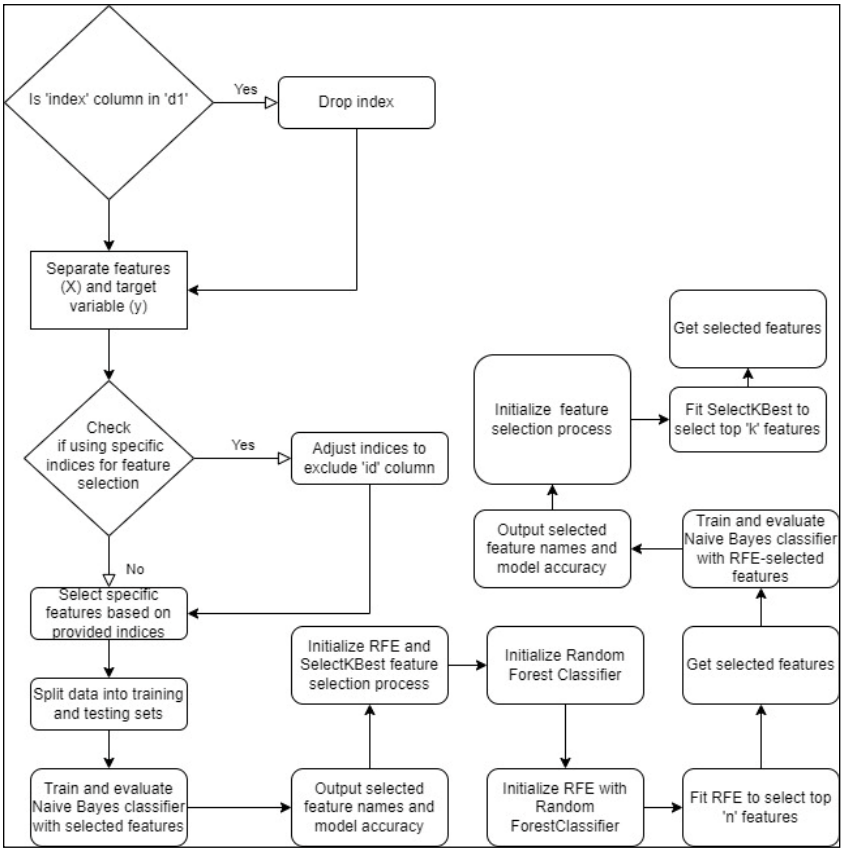

Figure 3. Feature Selection

In the model training and evaluation phase, the workflow first trains and evaluates a Naives Baiyes classifier with the initially selected features. It then starts a feature selection process using SelectKBest to identify the top 'k' features based on their scores, retrieves these features, and trains and evaluates a Naives Baiyes classifier using them. Following this, a Random Forest classifier is initialized, and Recursive Feature Elimination (RFE) is applied to select the top 'n' features. The selected features from RFE are retrieved, their names and model accuracy are output, and a Naives Baiyes classifier is trained and evaluated using these RFE-selected features. The workflow concludes by outputting the names of the selected features and model accuracy for both feature selection methods (SelectKBest and RFE), ensuring a thorough evaluation of the model's performance with the best-selected features.

3.3 Ten (10) Forth Cross Validation

Ten (10) Forth Validation in a process is for ensuring the accuracy and reliability of the outcomes. In the context of research and experimentation, the fourth validation step often involves confirming the effectiveness, efficiency, or accuracy of a proposed method or technique. Several studies provide insights into different fields where the fourth validation step plays a significant role. Figure 4, Below shows the process of 10-fold Cross validation.

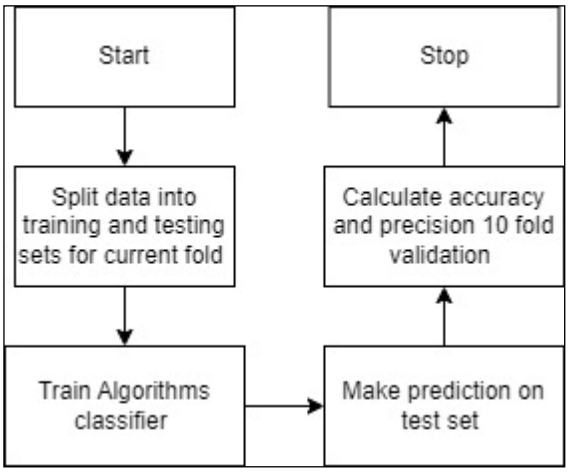

Figure 4. Ten (10) Forth Cross Validation

4. Result

Classification in machine learning is a fundamental task that involves categorizing data into predefined classes or categories based on the features or attributes of the data. Various machine learning algorithms are utilized for classification tasks, such as Naïve Bayes, and Random Forest. The classification is to build a model that can accurately predict the class labels of new, based on the patterns learned from the training data. Below here Table 6, and 7 classification Naives Baiyes, Random Forest and LightGBM.

Table 6. Naives Baiyes, Random Forest and LightGBM (Without SelectKBest and RFE)

Algorithms	phishing dataset [9] & uci-ml-phishing dataset [11]	
	Accuracy	Precision
Naives Baiyes	57.33%	56.00%
Random Forest	54.67%	49.00%
LightGBM	62.67%	61.00%

Table 7. Naives Baiyes, Random Forest and LightGBM(SelectKBest and RFE)

Algorithms	phishing dataset [9] & uci-ml-phishing dataset [11]	
	Accuracy	Precision
Naives Baiyes	94.24 %	94.51 %
Random Forest	94.80%	94.56%
LightGBM	95.00%	95.00%

The table 6 and 7 summarizes the performance metrics for three classification algorithms: Naive Bayes, Random Forest, and LightGBM. Naive Bayes achieved an accuracy of 94.24%% and a precision of 94.51%%, indicating a strong performance with just over 90% of instances correctly classified and positive predictions accurately made. Random Forest outperformed Naive Bayes with an accuracy of 94.80% and a precision of 94.56%, prediction. LightGBM further excelled, achieving both an accuracy and precision of 95.00%.

5. Conclusion

The comparative study of Naive Bayes, Random Forest, and LightGBM algorithms highlights their distinctstrengths across various applications. Naive Bayes is praised for its simplicity and efficiency in classificationtasks, while Random Forest is adept at managing large datasets with complex relationships, making it suitablefor tasks like soil texture classification and forecasting. LightGBM stands out for its speed and high performancein large-scale tasks such as predicting chemical toxicity. Research indicates that Random Forest often excels inmedical predictions, including breast cancer survival and acute kidney injury after liver transplantation. Futurestudies should explore the impact of dataset size, delve into medical applications, and develop optimizationstrategies to enhance the practical utility of these algorithms in diverse fields. For the summaries, LightGBM aremost powerful with accuracy the result of LightGBM is 95.00% and precision 94.51%. while Random Forest is94.80% and precision 94.56%, and lastly, Naives Baiyes is 94.24% and precision 95.00%.

Acknowledgement

This work was supported by the Universiti Tun Hussein Onn Malaysia (UTHM) through Tier1 (vot Q508).

Authors Contribution

Muhammad Amir Izaan Anuar did the experiments and writing the report. Dr. Cik Feresa supervised the work and Editing. Dr. Andi Maslan editing.

References

[1] P. Yang, G. Zhao, and P. Zeng, “Phishing website detection based on multidimensional features driven by deep learning,” IEEE Access, vol. 7, pp. 15196–15209, 2019, doi: 10.1109/ACCESS.2019.2892066.

[2] Q. Khanh, T. Hoang, T. Nguyen, & V. Ong, "Predicting and avoiding hazardous occurrences of stuck pipe for the petroleum wells at offshore vietnam using machine learning techniques", IOP Conference Series: Earth and Environmental Science, vol. 1091, no. 1, p. 012003, 2022. https://doi.org/10.1088/1755-1315/1091/1/012003.

[3] Yi Yong Lee, Chin Lay Gan, and Tze Wei Liew, “Phishing victimization among Malaysian young adults: cyber routine activities theory and attitude in information sharing online,”pp. 8–31, 2022.

[4] D. Zuo, L. Yang, Y. Jin, H. Qi, Y. Liu, & L. Ren, "Machine learning-based models for the prediction of breast cancer recurrence risk", BMC Medical Informatics and Decision Making, vol. 23, no. 1, 2023. doi:10.1186/s12911-023-02377-z.

[6] A. Hannousse and S. Yahiouche, “Towards Benchmark Datasets for Machine Learning Based Website Phishing Detection: An experimental study,” Oct. 2020, doi:10.1016/j.engappai.2021.104347.

[7] N. Nagy et al., “Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis,” Sensors, vol. 23, no. 7, Apr. 2023, doi:10.3390/s23073467.

[8] Y. Kang, W. Kim, S. Lim, H. Kim, and H. Seo, "DeepDetection: Privacy-Enhanced Deep Voice Detection and User Authentication for Preventing Voice Phishing,"Applied Sciences (Switzerland), vol. 12, no. 21, Nov. 2022, doi:10.3390/app122111109.

[9] S. A. Khan, W. Khan, and A. Hussain, "Phishing Attacks and Websites Classification Using Machine Learning and Multiple Datasets (A Comparative Analysis)," Intelligent Computing Methodologies, pp. 301–313, 2020, doi:10.1007/978-3-030-60796-8_26.

[10] M. Hasan, M. Jawad, A. Dutta, M. Awal, M. Islam, M. Masudet al., "Associating measles vaccine uptake classification and its underlying factors using an ensemble of machine learning models", Ieee Access, vol. 9, p. 119613-119628, 2021. doi:10.1109/access.2021.3108551.

[11] E. Oram, P. B. Dash, B. Naik, J. Nayak, S. Vimal, and S. K. Nataraj, "Light gradient boosting machine-based phishing webpage detection model using phisher website features of mimic URLs," Pattern Recognition Letters, vol. 152, pp. 100–106, Dec. 2021, doi: 10.1016/j.patrec.2021.09.018

[12] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, p. 153-158, 1997. doi:10.1109/34.574797

[13] I. I. Ismagilov, A. A. Murtazin, D. V. Kataseva, A. S. Katasev, and A. I. Barinov, "Definition of Phishing Sites Based on the Team Model of Fuzzy Neural Networks,"HELIX, vol. 10, no. 5, pp. 133–140, Oct. 2020, doi: 10.29042/2020-10-5-133-140

[14] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," Inf Sci (N Y), vol. 484, pp. 153–166, May 2019, doi: 10.1016/j.ins.2019.01.064

[15] N. Nagy et al., "Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis," Sensors, vol. 23, no. 7, Apr. 2023, doi:10.3390/s23073467

[16] J. Sun, H. Yu, G. Zhong, J. Dong, Z. Shu, & H. Yu, "Random shapley forests: cooperative game-based random forests with consistency", Ieee Transactions on Cybernetics, vol. 52, no. 1, p. 205-214, 2022, doi:10.1109/tcyb.2020.2972956

[17] M. Li, H. Chen, H. Zhang, M. Zeng, B. Chen, & L. Guan, "Prediction of the aqueous solubility of compounds based on light gradient boosting machines with molecular fingerprints and the cuckoo search algorithm", Acs Omega, vol. 7, no.46, p. 42027-42035, 2022. doi:10.1021/acsomega.2c03885

[18] Y. Saeys, I. Inza, & P. Larrañaga, "A review of feature selection techniques in bioinformatics", Bioinformatics, vol. 23, no. 19, p. 2507-2517, 2007. doi:10.1093/bioinformatics/btm344

[19] K. Sharma, "Quantum adiabatic feature selection",, 2019. doi:10.48550/arxiv.1909.08732.