Journal Pre-proof

Mutual Information based Logistic Regression for phishing URL detection

Vajratiya Vajrobol, Brij B. Gupta, Akshat Gaurav

PII: S2772-9184(24)00010-9
DOI: https://doi.org/10.1016/j.csa.2024.100044
Reference: CSA 100044

To appear in: *Cyber Security and Applications*

Received date: 20 December 2023
Revised date: 30 January 2024
Accepted date: 20 February 2024

Please cite this article as: Vajratiya Vajrobol, Brij B. Gupta, Akshat Gaurav, Mutual Information based Logistic Regression for phishing URL detection, *Cyber Security and Applications* (2024), doi: https://doi.org/10.1016/j.csa.2024.100044

Mutual Information based Logistic Regression for phishing URL detection

Vajratiya Vajrobol[a,b], Brij B. Gupta[c] and Akshat Gaurav[d]

[a]Institute of Informatics and Communication, University of Delhi, India
[b] International Center for AI and Cyber Security Research and Innovations. Asia University, Taiwan.
[c]Department of Computer Science and Information Engineering, Asia University, Taichung 413, Taiwan.
[d]Ronin Institute, Montclair, NJ, USA

ARTICLE INFO

Keywords:
Phishing
URL analysis
Logistic Regression
Mutual Information
Feature selection

ABSTRACT

Phishing is a cybersecurity problem that hackers employ to deceive individuals and organizations. Phishing is dynamic in nature; the hackers change several tricks to deceive the victims in multiple ways. It is important to track the tricks of hackers with recent technology. This study makes a notable contribution to enhancing cybersecurity defences by offering insights that aid in the detection and mitigation of phishing threats. Specifically, the study's analysis of URLs using mutual information and logistic regression techniques yielded a remarkably high accuracy rate of 99.97%, surpassing previous efforts. The identification of the most informative features for distinguishing phishing attempts provides valuable intelligence for cybersecurity professionals, enabling them to bolster defenses and stay ahead of evolving phishing tactics.

1. Introduction

Phishing attacks happen all the time on the internet, which shows how important it is to have good ways to spot fake URLs (Zahra et al., 2022; Atat et al., 2017; Wu et al., 2016; Devi and Bharti, 2022). For phishing attacks, which are known for having methods that change all the time, URLs that look like something else are used. This study is about using advanced techniques to stop hacking risks that are always changing because the authors know how important URL-based tracking is for cybersecurity (Asiri et al., 2023; Gupta and Panda, 2022; AlShaikh et al., 2024; Dwivedi, 2022; Wen et al., 2014). Figure 1 shows the steps that are used in phishing.

Phishing tricks are always changing so that they cannot be found in the normal ways(Almomani et al., 2022; Tembhurne et al., 2022; Leghari and Ali, 2023; Lei et al., 2021). This study shows that URL research is an important part of protecting yourself from online threats(Xu et al., 2015; Sharma and Sharma, 2022; Mani et al., 2021). In this study, fake URLs are looked at in detail. The goal is to make security better by making it easy to tell the difference between safe and dangerous URLs. The study uses both logistic regression and mutual information (MI)-based feature selection, which works well, to make this happen.

To finding fake URLs, this study adds to the field by using Mutual Information and Logistic Regression, a complex feature selection method. The study's goal is to make the recognition method more accurate and easier to understand. This is what our contribution looks like:

- Developing a novel method like a combination of mutual information and logistic regression for detecting phishing URLs
- A comparative study of different feature sets for detecting phishing URLs

The study is organised with an introduction, then parts on methods and earlier studies. The methods part goes into great depth about the dataset, how the data was prepared, and the methods used to choose the features. Also explained in this part are the formula and assessment method used to spot phishing. In the parts that follow, the study's results are given, and the study ends with an overview of the results.

2. Literature survey

Recently, deep learning and machine learing are used to solve many realylife problems(Jain et al., 2022; Gupta et al., 2022; Jain and Gupta, 2022; Liu et al., 2022). Several studies have used URLs to find signs of scam(Mahdi et al., 2022; Aoun Barakat et al., 2021; Hamza et al., 2022; Melki et al., 2021; Mezher et al., 2022).

Karim et al. (2023) use a machine learning-based method on a phishing URL dataset to look at the growing danger of phishing attacks. A mixed LSD model, decision trees, and linear regression are some of the methods used in the study to stop phishing efforts very effectively. Other models do not do as well as the LSD model, which combines logistic regression, support vector machine, and decision tree .

Abdul Samad et al. (2023) use machine learning models to make it easier to spot phishing URLs and investigate how complicated phishing attacks are. These models look at things like URLs, web content, and features that come from outside the site. It has been shown through tests that different setting factors, like data balance, hyperparameter optimization, and feature selection, make the accuracy much better. The study finds that the fine-tuned factors work better together than in earlier studies. This shows that machine learning techniques have improved over time. Random Forest

tiya101@south.du.ac.in (V. Vajrobol); bbgupta@asia.edu.tw (B.B. Gupta); akshat.gaurav@ronininstitute.org (A. Gaurav)
ORCID(s): 0000-0002-5796-9424 (A. Gaurav)

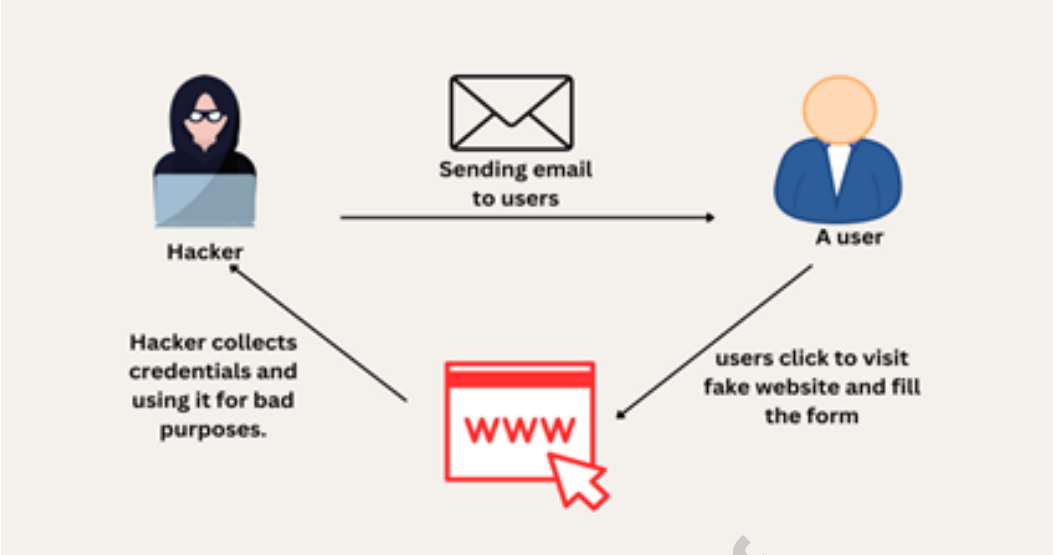

Figure 1: Phishing process

and Gradient Boosting are both 97.44% and 97.47% accurate for Dataset-1, which is pretty good. On the other hand, for Dataset-2, Extreme Gradient Boosting and Gradient Boosting are both 98.27% and 98.21% correct.

Nagy et al. (2023) say that experts should use Python's multiprocessing and multithreading abilities to make models work better. The dataset, which has 54,000 training records and 12 000 testing records, goes through five tests. The first one is linear, and the next four use parallel execution. We tried four models: random forest, naive bayes, convolutional neural network, and long short-term memory. All of them did very well and sped up the process. A comparison study shows that simultaneous processing is a good way to make models work better at finding phishing attacks.

Biswas et al. (2024) created a hybrid approach that uses explainable AI methods to look at the cyber-risks that come from phishing attacks that are linked to each other. The framework has several steps, such as finding skilled phishers, figuring out how likely it is that phishing attacks will happen despite IT security measures, sorting URLs into groups using machine-learning classifiers, guessing the joint distribution using an exponential-beta model, and figuring out how much money is expected to be lost using Archimedean Copula. Based on how risk-averse a company is, the approach helps them figure out how to best spend in cybersecurity and cyber-insurance.

Almomani et al. (2021) focused on detecting phishing websites by extracting semantic features such as URL & Domain Identity, Abnormal Features, HTML and JavaScript Features, and Domain Features. Using 16 machine learning models and 10 selected semantic features, GradientBoostingClassifier and RandomForestClassifier achieved the highest accuracy at around 97%. Conversely, GaussianNB and stochastic gradient descent (SGD) classifier showed the lowest accuracy at 84% and 81%, respectively, compared to other classifiers.

(Guendouz and Amine, 2022) introduces FWA-FS, a novel Android malware detection method utilizing the fireworks algorithm for feature selection. By employing static analysis and permissions extracted from APK files as feature vectors, the proposed technique effectively distinguishes between benign and malicious applications. Experimental results demonstrate significant performance improvements, with an average accuracy increase of 6% for KNN and 25% for NaÃŕve Bayes classifiers.

Alshdadi et al. (2021) addresses web spam detection by incorporating overlooked features from tools like the lighthouse plugin, Ahrefs, and social media platforms. Utilizing a dataset from Google Webmaster Tools, machine learning models are applied to enhance detection accuracy. The study demonstrates improved performance, particularly with support vector machine (SVM) classifiers, surpassing NaÃŕve Bayes, C4.5, AdaBoost, and LogitBoost methods. Such advancements signify a significant step forward in combating the threat of malicious domain activity on the web.

Expanding on the findings of the prior study, this investigation attempts to close the knowledge gap on feature significance in the context of phishing attempts. The importance of traits in identifying phishing attempts has not yet been thoroughly investigated, even though previous research has examined many facets of cybersecurity and machine learning models. We hope to further our understanding of the critical elements affecting the precision and efficacy of machine learning algorithms in identifying and averting phishing attempts by undertaking a thorough investigation. This study will provide insightful information that will advance the cybersecurity sector.

3. Methods

The study commences with the acquisition of a phishing dataset in Figure 2., followed by data preprocessing steps

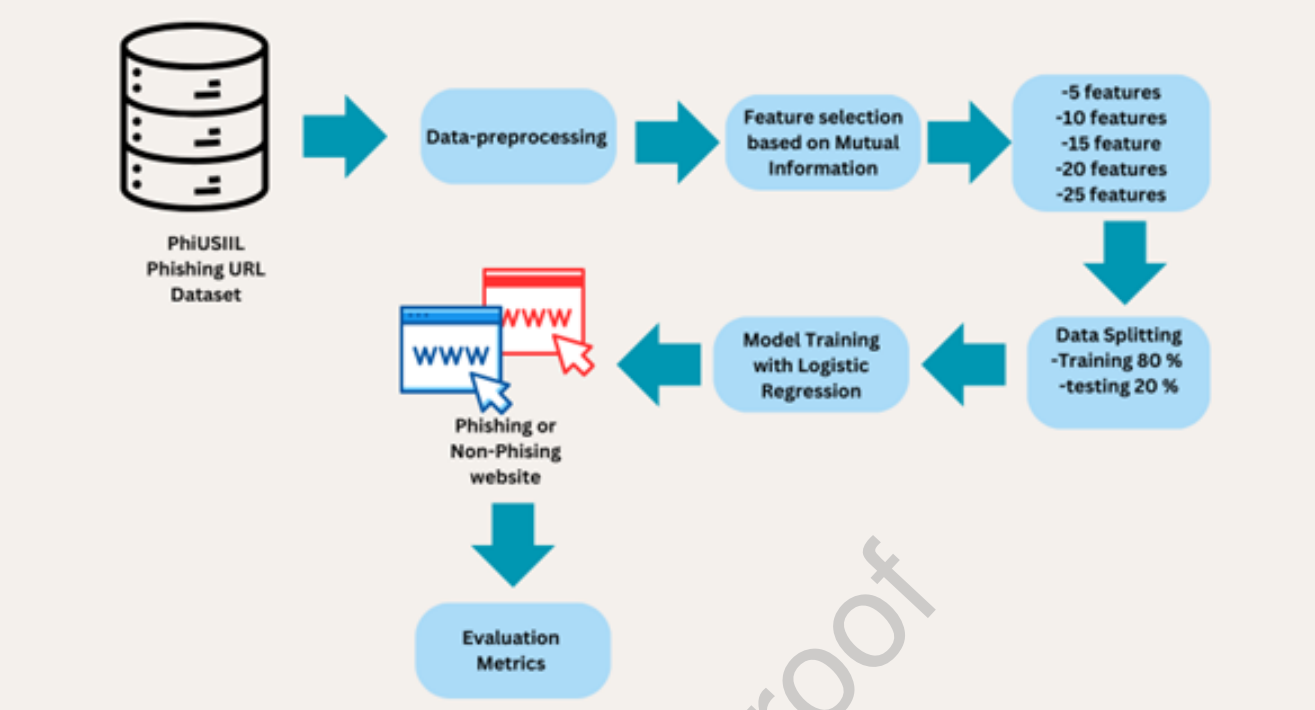

Figure 2: Detection of Phishing using URL framework

that involve tasks such as eliminating string columns and applying min-max scaling. Subsequently, feature selection techniques, specifically Mutual Information, are employed to select varying numbers of features, including 5, 10, 15, 20, and 25. After this, the dataset is partitioned into training and testing sets in an 80:20 ratio. The Logistic Regression model is then trained on the training data, and predictions are made to classify websites as either phishing or non-phishing. Finally, evaluation metrics are applied to assess the model's performance.

3.1. Dataset

The PhiUSIIL Phishing URL Dataset, sourced from Mendeley (Prasad and Chandra, 2024), is a comprehensive compilation that includes 134,850 legitimate URLs and 100,945 phishing URLs. Emphasising the integration of the most recent URLs, the dataset reflects a focus on current trends during its construction. Feature extraction includes a thorough analysis of both the webpage source code and the URL. Significant features, including CharContinuationRate, URLTitleMatchScore, URLCharProb, and TLDLegitimateProb, are derived from pre-existing features within the dataset. Further dataset details are elaborated in Table 1:

3.2. Data-preprocessing

In the data preprocessing stage, we opted to exclude features that contained string-based values and instead concentrated solely on numerical types such as float or int. Upon inspection, it became evident that the columns displayed diverse data ranges. To mitigate this discrepancy, we employed min-max scaling. This technique transforms the numerical values within each column to a standardized range, usually between 0 and 1. This normalization fosters uniformity across the dataset, thereby simplifying the modeling process and ensuring that all features contribute equally to the analysis, regardless of their original scales.

3.3. Mutual Information feature selection technique

A useful machine learning technique called Mutual Information (MI) feature selection calculates the statistical dependence of each feature on the target variable. MI assists in identifying the most useful features by assessing the amount of information learned about the target variable by knowing the values of different features. In practical terms, Mutual Information can be implemented in classification tasks with the 'mutual_info_classif' function provided by the scikit-learn module in Python. A dataset is typically divided into training and testing sets, and MI scores are calculated for every feature in the workflow. Features are ranked in descending order of informativeness using the scores that are obtained. The final set of features utilised for modelling includes the top-ranked features, which show strong relationships with the target variable (Amiri et al., 2011; Vinh et al., 2012).

The mathematical equation for Mutual Information (MI) between two random variables X and Y is expressed in eq. (1)

$$I(X;Y) = \left(\sum_{x \in X} \sum_{y \in Y} p(x,y) \bullet \frac{p(x,y)}{p(x) \bullet p(y)} \right) \quad (1)$$

Where:

Table 1
The feature explanation in PhiUSIIL dataset

No.	Features	Description
1.	TLD (Top Level Domain)	The final part of domain name
2.	URL length	The URLs of phishing frequently exhibit a tendency to be lengthier compared to legitimate URLs.
3.	IsDomainIP	A URL or IP address has been used as a domain name.
4.	NoOfSubDomain	hackers frequently employ visual similarity techniques to deceive users. They establish subdomains that closely resemble those of legitimate websites.
5.	NoOfObfuscatedChar	Displays the count of obscured characters in a URL.
6.	IsHTTPS	HTTPs is safe.
7.	No. of digits, equal, qmark	The presence of digits or symbols, such as '=', '?', or '%', in a URL elevates the likelihood of it being a phishing URL.
8.	LargestLineLength	A code may be longer, a technique used by hackers to conceal their activities.
9.	HasTitle	Most real websites provide page titles.
10.	HasFavicon	Most real websites include their website logo in the favicon tag.
11.	IsResponsive	The real websites are designed to be responsive.
12.	NoOfURLRedirect	Phishing sites may redirect users towards a different page.
13.	HasDescription	Reputable websites incorporate page descriptions utilising the 'description' meta name for each of their pages.
14.	HasExternal FormSubmit	Phishing sites frequently employ HTML forms to gather user information.
15.	HasCopyrightInfo, HasSocialNet	copyright information and links to social networking profiles may be included in real websites.
16.	NoOfPopup, Noof iFrame	pop-ups or iframes may be used in phishing websites.
17.	HasPasswordField, HasSubmitButton	HTML enables users to input data and submit it to other URLs. For instance, HTML tags like 'passwordfield' or 'submitbutton' .
18.	HasHiddenFields	Phishing websites may employ hidden fields as a technique to capture sensitive information.
19.	Bank, Pay, Crypto	Keywords such as "bank," "pay," or "crypto" in a webpage may indicate that the site is requesting sensitive financial information from the user.
20.	NoOfImage	Threat actors can employ screenshots of legitimate websites to craft phishing websites, enhancing their appearance and making them appear more legitimate to users.
22.	NoOfJS	JavaScript can be embedded in HTML to create interactive web pages. Phishing websites may use JavaScript .
23.	NoOfSelfRef, NoOfEmptyRef, NoOfExternalRef	Hyperlinks (href) are clickable links that enable users to navigate between web pages or direct them to external sites. Phishing websites may utilise hyperlinks that seemingly lead to legitimate pages but redirect users to phishing pages.
24.	CharContinuationRate	To calculate the 'CharContinuationRate' of a URL, we identify the longest sequences of alphabets, digits, and special characters and sum their lengths. Subsequently, the total length of these sequences is divided by the overall length of the URL.
25.	URLTitleMatchScore	a metric used to quantify the dissimilarity between the URL and the webpage title. It measures the extent to which the content of the URL aligns with the content indicated by the webpage title.
26.	URLCharProb	refers to the URL Character Probability. It is a metric used to understand the pattern of each alphabet and digit in a URL .
27.	TLDLegitimateProb	Top-Level Domain Legitimate Probability. It is a metric used to assess the likelihood that a URL is legitimate based on its top-level domain (TLD)

I (X; Y) is the Mutual Information between random variables X and Y

p (x, y) is the joint probability density function of X and Y

p(x) and p(y) are the marginal probability density function of X and Y respectively.

In the context of feature selection, X often represents a feature, and Y represents the target variable. The MI quantifies the reduction in uncertainty about the target variable given the knowledge of the feature's values.

3.4. Logistic Regression

Logistic regression is a statistical method employed to solve binary classification issues by estimating the probability of an observation belonging to one of two classes. Logistic regression builds on the linear regression by employing the logistic function to transform the outcome into a limited range of values between 0 and 1. The logistic regression model is defined by a linear combination of input features, with each feature being assigned a weight and a bias term. It is commonly employed for a broad range of tasks such as anomaly detection (Rakshitha and Jayarekha, 2022; Mahindru and Sangal, 2021).

3.5. Evaluation Metrics

The evaluation metrics used in this experiment are accuracy, which is the ratio of the number of correctly identified cases to the total number of cases. When you divide the number of true positives by the number of false positives and true positives, it is called precision and recall, which divides the number of true positives by the total number of true positives and false negatives. The F1 score is another important metric. It gives a fair picture of how well a model is doing by considering both false positives and false negatives.

4. Results

The feature selection process using Mutual Information (MI) resulted in subsets of features with varying sizes (5, 10, 15, 20, and 25 features) in Table 2. The analysing of the selected features for each subset:

- Subset with 5 Features:

  1. 'URLSimilarityIndex'
  2. 'LineOfCode'
  3. 'NoOfExternalRef'
  4. 'NoOfImage'
  5. 'NoOfSelfRef'

This minimal set of features may capture fundamental characteristics related to URL similarity, code metrics, external references, images, and self-referential links.

- Subset with 10 Features:

  6 'NoOfJS'
  7 'LargestLineLength'
  8 'NoOfCSS'
  9 'HasSocialNet'
  10 'LetterRatioInURL'

In addition to the previous set, this expands to include features such as JavaScript count, CSS count, presence of social networks, and letter ratio in the URL.

- Subset with 15 Features:

  11 'HasCopyrightInfo'
  12 'HasDescription'
  13 'IsHTTPS'
  14 'NoOfOtherSpecialCharsInURL'
  15 'DomainTitleMatchScore'

This further incorporates features related to copyright information, web page descriptions, HTTPS status, special characters in the URL, and domain title match score.

- Subset with 20 Features:

  16 'HasSubmitButton'
  17 'SpacialCharRatioInURL'
  18 'TLDLegitimateProb'
  19 'URLTitleMatchScore'
  20 'IsResponsive'

This extends the feature set to include the presence of submit buttons, spatial character ratio in the URL, top-level domain (TLD) legitimacy probability, URL title match score, and responsiveness.

- Subset with 25 Features:

  21 'DegitRatioInURL'
  22 'NoOfDegitsInURL'
  23 'CharContinuationRate'
  24 'NoOfiFrame'
  25 'NoOfEmptyRef'

The final set includes information about the number of digits in the URL, the presence of iframes, the character repetition rate, and the number of empty references.

- The gradual addition of more features makes it possible to show more of a website's characters.
- Features include URL structure, code implements, content characteristics, security signs, and how responsive a website is, among other things.

  - The chosen features show possible signs of hacking behavior and give the Logistic Regression model useful data during the training and prediction stages.

The purpose of this study is to help us understand how each group of features is useful and important for finding phishing emails. After the training and testing steps, the model's usefulness can be checked again using the right evaluation criteria.

There is a clear trade-off between the number of features and the performance measures in the feature selection using Mutual Information (MI) for phishing URL detection. With only 5 features, the model got very high scores for accuracy, precision, recall, and F1-score, which were all reliably around 99.97% in Table 3. This says that a small set of features can pick out the most important patterns that show phishing URLs, leading to a nearly perfect classification. All the measures have high values, which shows that the chosen features are good at telling the difference between safe and dangerous URLs. This shows how good the MI-based feature selection method is at telling the difference.

Table 2
The feature selection description

Number of features	Feature selection using MI
5	'URLSimilarityIndex', 'LineOfCode', 'NoOfExternalRef', 'NoOfImage', 'NoOfSelfRef'
10	'URLSimilarityIndex', 'LineOfCode', 'NoOfExternalRef', 'NoOfImage', 'NoOfSelfRef', 'NoOfJS', 'LargestLineLength', 'NoOfCSS', 'HasSocialNet', 'LetterRatioInURL'
15	'URLSimilarityIndex', 'LineOfCode', 'NoOfExternalRef', 'NoOfImage', 'NoOfSelfRef', 'NoOfJS', 'LargestLineLength', 'NoOfCSS', 'HasSocialNet', 'LetterRatioInURL', 'HasCopyrightInfo', 'HasDescription', 'IsHTTPS', 'NoOfOtherSpecialCharsInURL', 'DomainTitleMatchScore'
20	'URLSimilarityIndex', 'LineOfCode', 'NoOfExternalRef', 'NoOfImage', 'NoOfSelfRef', 'NoOfJS', 'LargestLineLength', 'NoOfCSS', 'HasSocialNet', 'LetterRatioInURL', 'HasCopyrightInfo', 'HasDescription', 'IsHTTPS', 'NoOfOtherSpecialCharsInURL', 'DomainTitleMatchScore', 'HasSubmitButton', 'SpacialCharRatioInURL', 'TLDLegitimateProb', 'URLTitleMatchScore', 'IsResponsive'
25	'URLSimilarityIndex', 'LineOfCode', 'NoOfExternalRef', 'NoOfImage', 'NoOfSelfRef', 'NoOfJS', 'LargestLineLength', 'NoOfCSS', 'HasSocialNet', 'LetterRatioInURL', 'HasCopyrightInfo', 'HasDescription', 'IsHTTPS', 'NoOfOtherSpecialCharsInURL', 'DomainTitleMatchScore', 'SpacialCharRatioInURL', 'HasSubmitButton', 'TLDLegitimateProb', 'URLTitleMatchScore', 'IsResponsive', 'DegitRatioInURL', 'NoOfDegitsInURL', 'CharContinuationRate', 'NoOfiFrame', 'NoOfEmptyRef'

Table 3
The performance of Logistic Regression with different features by Mutual Information

Feature selection with MI based	**Accuracy**	**Precision**	**Recall**	**F1-score**
5 features	0.9997	0.9997	0.9997	0.9997
10 features	0.9744	0.9744	0.9733	0.9739
15 features	0.9963	0.9960	0.9965	0.9963
20 features	0.9936	0.9934	0.9935	0.9935
25 features	0.9935	0.9933	0.9934	0.9933

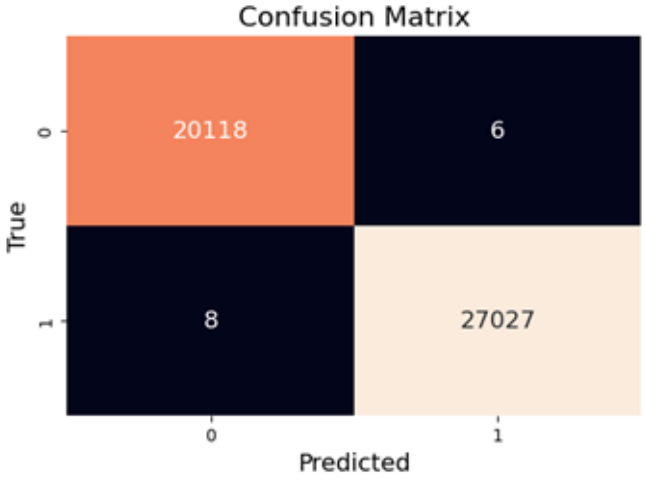

Figure 3: confusion matrix of Logistic Regression with 5 features

But as the number of traits goes up to 10, the model's performance clearly gets worse. The accuracy is still high at 97.44%, but the precision, memory, and F1-score have all gone down a little. This drop shows that adding more features after a certain point might make the model less able to generalise well by adding noise or unnecessary information.

-Among the five features - 'URLSimilarityIndex', 'LineOfCode', 'NoOfExternalRef', 'NoOfImage', and 'NoOfSelfRef' - the highest contributing factor to phishing appears to be the 'URLSimilarityIndex'. This conclusion is drawn based on its strong predictive power as observed in the feature selection analysis.

The 'URLSimilarityIndex' likely contributes significantly to phishing detection because it measures the similarity of a URL to known phishing URLs or patterns. Phishing URLs often mimic legitimate websites, and detecting such similarities can be crucial in identifying potential phishing attempts.

The confusion matrices examined are depicted in Figure 3 and 4. An intriguing pattern is seen between the models with 5 and 10 characteristics. With a model including 5 fea-

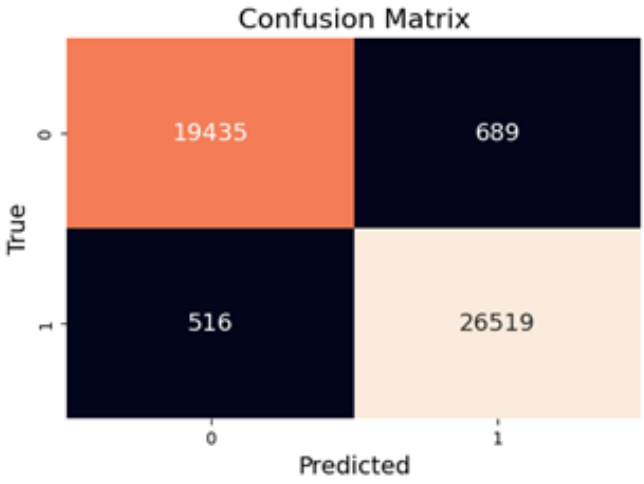

Figure 4: Confusion matrix of Logistic Regression with 10 features

Table 4
The comparison with previous studies

Publications	Task	Performance
Alsharaiah et al. (2023)	phishing-website detection	Accuracy of 98.64%
Biswas et al. (2024)	phishing attacks detection	Accuracy of 94.612
Jalil et al. (2023)	phishing URL detection on Kaggle dataset	Accuracy of 96.25%
Prasad and Chandra (2024)	phishing URL detection with PhiUSIIL dataset	Accuracy of 99.24%
Our study	phishing URL detection with PhiUSIIL dataset	Accuracy with 99.97 %

tures, the diagonal of the confusion matrix displays a greater number of accurately recognized samples. Consequently, the model, derived from a limited number of attributes identified by Mutual Information, effectively classifies instances as either authentic or counterfeit websites. The accuracy of the 5-feature model is evident in its robust performance, as indicated by the high count on the vertical axis.

Conversely, the model with 10 features exhibits a reduced count on the diagonal of the confusion matrix, indicating that it accurately classified a smaller number of samples compared to the model with 5 features. Based on this discovery, the 10-feature model may possess a larger selection of chosen traits, but it may also provide more intricacy that might impede the ability to distinguish between genuine and counterfeit websites.

Table 4. shows that a previous study compared to our study on the detection of phishing using URL based. As we can see from the fact that the research used different datasets, Alsharaiah et al. (2023) achieved a score of 98.64% in identifying phishing websites in their task. Biswas et al. (2023) investigated the detection of phishing attempts and reported a success rate of 94.612% for their approach. Jalil et al. (2023) employed a Kaggle dataset to examine the detection of phishing URLs and achieved a success rate of 96.25%. Prasad and Chandra (2024) used the PhiUSIIL dataset to detect fraudulent URLs, achieving an impressive success rate of 99.24%. However, our study, which also uses the same PhiUSIIL dataset, is more accurate than earlier ones (99.97%). This puts our study at the top of the list for finding fake URLs, showing how well our method works. Our results also show that cybersecurity steps against phishing risks could be improved and made even better.

5. Conclusions

This study investigated how feature selection based on Mutual Information (MI) can be used in Logistic Regression to find fake URLs. We can see from the results that this method works well for finding a small group of factors that significantly affect the Logistic Regression model's correctness, precision, recall, and F1-score. When MI is used as a feature selection parameter, the Logistic Regression model is easier to understand because it finds the most useful features.

A confusion matrix analysis is part of the complete test that shows how well the Logistic Regression model works when trained on features picked by MI. Amazingly, 99.97% accuracy, precision, recall, and F1-score can be reached with just 5 characteristics: "URLSimilarityIndex," "LineOfCode," "NoOfExternalRef," "NoOfImage," and "NoOfSelfRef." This is almost perfect accuracy. This is an example of how well MI picks out the key differences between real and fake URLs. The study also shows a small trade-off as the number of features increases. This suggests that larger feature sets may make the model more complicated, which could affect how well it works.

Finally, the Mutual Information-based Logistic Regression model looks like it could be a useful and easy-to-understand way to find fake URLs. More people are learning about cybersecurity thanks to this study. It also lays the groundwork for more research and better feature selection methods in the field of URL categorizing. The method has been shown to be accurate and reliable, which shows that it could be used in real life to improve threat detection and increase defences against phishing efforts.

Acknowledgement

This research work is supported by National Science and Technology Council (NSTC), Taiwan Grant No. NSTC112-2221-E-468-008-MY3.

References

Abdul Samad, S.R., Balasubaramanian, S., Al-Kaabi, A.S., Sharma, B., Chowdhury, S., Mehbodniya, A., Webber, J.L., Bostani, A., 2023. Analysis of the performance impact of fine-tuned machine learning model for phishing url detection. Electronics 12, 1642.

Almomani, A., Al-Nawasrah, A., Alomoush, W., Al-Abweh, M., Alrosan, A., Gupta, B.B., 2021. Information management and iot technology for safety and security of smart home and farm systems. Journal of Global Information Management (JGIM) 29, 1–23.

Almomani, A., Alauthman, M., Shatnawi, M.T., Alweshah, M., Alrosan, A., Alomoush, W., Gupta, B.B., 2022. Phishing website detection with semantic features based on machine learning classifiers: a comparative study. International Journal on Semantic Web and Information Systems (IJSWIS) 18, 1–24.

AlShaikh, M., Alsemaih, W., Alamri, S., Ramadan, Q., 2024. Using supervised learning to detect command and control attacks in iot. International Journal of Cloud Applications and Computing (IJCAC) 14, 1–19.

Alsharaiah, M., Abu-Shareha, A., Abualhaj, M., Baniata, L., Adwan, O., Al-saaidah, A., Oraiqat, M., 2023. A new phishing-website detection framework using ensemble classification and clustering. International Journal of Data and Network Science 7, 857–864.

Alshdadi, A.A., Alghamdi, A.S., Daud, A., Hussain, S., 2021. Blog backlinks malicious domain name detection via supervised learning. International Journal on Semantic Web and Information Systems (IJSWIS) 17, 1–17.

Amiri, F., Yousefi, M.R., Lucas, C., Shakery, A., Yazdani, N., 2011. Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications 34, 1184–1199.

Aoun Barakat, K., Dabbous, A., Tarhini, A., 2021. An empirical approach to understanding users' fake news identification on social media. Online Information Review 45, 1080–1096.

Asiri, S., Xiao, Y., Alzahrani, S., Li, S., Li, T., 2023. A survey of intelligent detection designs of html url phishing attacks. IEEE Access .

Atat, R., Liu, L., Chen, H., Wu, J., Li, H., Yi, Y., 2017. Enabling cyber-physical communication in 5g cellular networks: challenges, spatial spectrum sensing, and cyber-security. IET Cyber-Physical Systems: Theory & Applications 2, 49–54.

Biswas, B., Mukhopadhyay, A., Kumar, A., Delen, D., 2024. A hybrid framework using explainable ai (xai) in cyber-risk management for defence and recovery against phishing attacks. Decision Support Systems 177, 114102.

Devi, S., Bharti, T.S., 2022. A review on detection and mitigation analysis of distributed denial of service attacks and their effects on the cloud. International Journal of Cloud Applications and Computing (IJCAC) 12, 1–21.

Dwivedi, R.K., 2022. Density-based machine learning scheme for outlier detection in smart forest fire monitoring sensor cloud. International Journal of Cloud Applications and Computing (IJCAC) 12, 1–16.

Guendouz, M., Amine, A., 2022. A new wrapper-based feature selection technique with fireworks algorithm for android malware detection. International Journal of Software Science and Computational Intelligence (IJSSCI) 14, 1–19.

Gupta, B.B., Tewari, A., Cvitić, I., Peraković, D., Chang, X., 2022. Artificial intelligence empowered emails classifier for internet of things based systems in industry 4.0. Wireless networks 28, 493–503.

Gupta, T., Panda, S.P., 2022. Cloudlet and virtual machine performance enhancement with clara and evolutionary paradigm. International Journal of Cloud Applications and Computing (IJCAC) 12, 1–16.

Hamza, A., Javed, A.R.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z., Borghol, R., 2022. Deepfake audio detection via mfcc features using machine learning. IEEE Access 10, 134018–134028.

Jain, A.K., Gupta, B., 2022. A survey of phishing attack techniques, defence mechanisms and open research challenges. Enterprise Information Systems 16, 527–565.

Jain, A.K., Gupta, B.B., Kaur, K., Bhutani, P., Alhalabi, W., Almomani, A., 2022. A content and url analysis-based efficient approach to detect smishing sms in intelligent systems. International Journal of Intelligent Systems 37, 11117–11141.

Jalil, S., Usman, M., Fong, A., 2023. Highly accurate phishing url detection based on machine learning. Journal of Ambient Intelligence and Humanized Computing 14, 9233–9251.

Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S.B., Joga, S.R.K., 2023. Phishing detection system through hybrid machine learning based on url. IEEE Access 11, 36805–36822.

Leghari, I., Ali, S., 2023. Artificial intelligence techniques to improve cognitive traits of down syndrome individuals: An analysis. International Journal of Software Science and Computational Intelligence 15, 1–11.

Lei, W., Wen, H., Wu, J., Hou, W., 2021. Maddpg-based security situational awareness for smart grid with intelligent edge. Applied Sciences 11, 3101.

Liu, R.W., Guo, Y., Lu, Y., Chui, K.T., Gupta, B.B., 2022. Deep network-enabled haze visibility enhancement for visual iot-driven intelligent transportation systems. IEEE Transactions on Industrial Informatics 19, 1581–1591.

Mahdi, A., Farah, M.F., Ramadan, Z., 2022. What to believe, whom to blame, and when to share: exploring the fake news experience in the marketing context. Journal of Consumer Marketing 39, 306–316.

Mahindru, A., Sangal, A., 2021. Fsdroid:-a feature selection technique to detect malware from android using machine learning techniques: Fsdroid. Multimedia Tools and Applications 80, 13271–13323.

Mani, N., Moh, M., Moh, T.S., 2021. Defending deep learning models against adversarial attacks. International Journal of Software Science and Computational Intelligence (IJSSCI) 13, 72–89.

Melki, J., Tamim, H., Hadid, D., Makki, M., El Amine, J., Hitti, E., 2021. Mitigating infodemics: The relationship between news exposure and trust and belief in covid-19 fake news and social media spreading. Plos one 16, e0252830.

Mezher, A.H., Deng, Y., Karam, L.J., 2022. Visual quality assessment of adversarially attacked images, in: 2022 10th European Workshop on Visual Information Processing (EUVIP), IEEE. pp. 1–5.

Nagy, N., Aljabri, M., Shaahid, A., Ahmed, A.A., Alnasser, F., Almakramy, L., Alhadab, M., Alfaddagh, S., 2023. Phishing urls detection using sequential and parallel ml techniques: Comparative analysis. Sensors 23, 3467.

Prasad, A., Chandra, S., 2024. Phiusiil: A diverse security profile empowered phishing url detection framework based on similarity index and incremental learning. Computers & Security 136, 103545.

Rakshitha, Jayarekha, P., 2022. Detection of phishing attacks on online collaboration tools using logistic regression, in: Security and Privacy in Cyberspace. Springer, pp. 157–164.

Sharma, R., Sharma, N., 2022. Attacks on resource-constrained iot devices and security solutions. International Journal of Software Science and Computational Intelligence (IJSSCI) 14, 1–21.

Tembhurne, J.V., Almin, M.M., Diwan, T., 2022. Mc-dnn: Fake news detection using multi-channel deep neural networks. International Journal on Semantic Web and Information Systems (IJSWIS) 18, 1–20.

Vinh, L.T., Lee, S., Park, Y.T., dâĂŹAuriol, B.J., 2012. A novel feature selection method based on normalized mutual information. Applied Intelligence 37, 100–120.

Wen, H., Tang, J., Wu, J., Song, H., Wu, T., Wu, B., Ho, P.H., Lv, S.C., Sun, L.M., 2014. A cross-layer secure communication model based on discrete fractional fourier fransform (dfrft). IEEE Transactions on emerging topics in computing 3, 119–126.

Wu, J., Guo, S., Li, J., Zeng, D., 2016. Big data meet green challenges: Big data toward green applications. IEEE Systems Journal 10, 888–900.

Xu, Z., Zhu, G., He, C., Shunxiang, Z., Du, Y., 2015. Weighted indication-based similar drug sensing. International Journal of Software Science and Computational Intelligence 7, 74–88.

Zahra, S.R., Chishti, M.A., Baba, A.I., Wu, F., 2022. Detecting covid-19 chaos driven phishing/malicious url attacks by a fuzzy logic and data mining based intelligence system. Egyptian Informatics Journal 23, 197–214.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐~~The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:~~