
Asian University of Bangladesh
এশিয়ান ইউনিভার্সিটি অব বাংলাদেশ

# Detection of Phishing URLs for Secure Browsing: A Machine Learning based Approach

Presented By

**MD. RASHIDUL ISLAM**

CSE, AUB

Supervised By

**MD. SHAMIMUL ISLAM**

LECTURER, CSE, AUB

**Date: 06.06.2024**

# Table of Contents



# Introduction

A cyber attack is when someone tries to damage or break into computers, networks, or devices without permission. There can be different types of cyber attacks that are – Malware, Phishing, Man-in-the-Middle (MitM), Denial-of-Service (DoS), SQL Injection, Cross-Site Scripting (XSS), Social Engineering, Ransomware, DNS Spoofing etc.

Phishing is a cyber attack where attackers impersonate legitimate entities to steal sensitive information, such as login credentials and financial details, via deceptive emails, websites, or messages.

Phishing attacks can lead to financial loss, identity theft, data breaches, and severe damage to both individual and organizational reputations.

**Let's looks out about phishing method:**

- Deceptive emails
- Fake websites
- Fraudulent messages (SMS, social media)
- Spear phishing (targeted attacks on specific individuals)
- Whaling (targeting high-profile individuals)
- Clone phishing (duplicating legitimate emails with malicious links)
- Vishing (voice phishing through phone calls)

# Understanding Phishing URLs

Phishing URLs are web addresses designed to deceive users into believing they're legitimate sites, often used for stealing sensitive information like passwords, credit card numbers, or personal details. Here examples of Phishing URLs vs. Legitimate URLs -

| Feature | Phishing URL | Legitimate URL |
|---|---|---|
| Misspelled Domains | paypa1.com | paypal.com |
| Use of Subdomains | login.paypal.security.com | paypal.com/login |
| Suspicious URL Lengths | yourbank-login-update-security-info.com | yourbank.com/security |
| HTTPS Deception | https://secure.paypal.com.fake.com | https://www.paypal.com |
| Unusual Characters | amaz0n.com/login | amazon.com |
| IP Address Usage | http://192.168.1.1/bank/login | http://bank.com/login |
| Shortened URLs | bit.ly/2HkjP1Q | paypal.com/login |

# Phishing attack Statistics 2021-2023

Phishing incidents have been on the rise, with millions of attempts reported annually.


| | Breach of customer or client data | Ransomware infection (payload delivered via email) | Credential/account compromise | Loss of data/intellectual property | Other malware infections | Widespread network downtime | Advanced persistent threat | Zero-day exploit | Reputational damage | Financial penalty |
|---|---|---|---|---|---|---|---|---|---|---|
| 2021 | 29% | 32% | 27% | 32% | 22% | 25% | 23% | 20% | 27% | 22% |
| 2022 | 44% | 43% | 36% | 33% | 28% | 26% | 21% | 20% | 18% | 9% |
| 2023 | 54% | 46% | 48% | 44% | 27% | 22% | 18% | 25% | 24% | 11% |

Note: The following data was collected from statista.com on June 1, 2024.

# Now, Why is it important to detect phishing URLs?

**Detecting phishing URLs is crucial for several reasons:**

- **Financial Protection:** Detecting phishing URLs prevents users from falling victim to scams targeting financial information.
- **Privacy Preservation:** Identification helps safeguard personal data like passwords and addresses.
- **Reputation Management:** Prevents damage to a business's reputation by avoiding data breaches.
- **Malware Prevention:** Blocks access to sites distributing harmful software like ransomware.
- **Cybersecurity Awareness:** Raises awareness among users, encouraging caution online.
- **Compliance Fulfillment:** Helps organizations meet regulatory obligations for data protection.

# Literature Reviews

Feature selection is essential for optimizing machine learning models. Datasets can contain irrelevant features that hinder performance. Existing research shows promise with different types of feature selections and machine learning algorithm including -

| Author | Year | Data Set | Result |
|---|---|---|---|
| Khan et. al. [12] | 2020 | UCI and other online sources | RF method and the ANN achieves 97% accuracy. |
| Salihovic et al. [13] | 2018 | UCI phishing dataset and Spam dataset | RF with Ranker + Principal Component Optimization achieves 97.33% accuracy for Phishing detection. RF with BestFirst + CfsSubsEval Optimization achieved 94.24% accuracy. |
| Vishva et al. [14] | 2021 | UCI phishing dataset, LIAR dataset | RF classifier offers the highest classification accuracy of 97%, and LR achieves 92% accuracy for URL analysis. |
| Hutchinson et al. [15] | 2018 | UCI phishing dataset | RF achieves higher accuracy for Sets D and E with 95.5% and 96.5%, respectively. |
| Sarasjati et al. [16] | 2022 | UCI phishing dataset & Mendeley phishing dataset | For the UCI dataset, RF classification reaches an accuracy level of 88.92% whereas, for the Mendeley dataset, it reaches an accuracy level of 97.50%. |
| Al-Sarem et al. [19] | 2021 | UCI phishing dataset & Mendeley phishing dataset | Optimal stacking ensemble method improved detection accuracy to 97.16%, 98.58% and 97.39% for Datasets-1, -2 and -3, respectively. |

# Methodology

# Dataset

PhiUSIIL Phishing URL Dataset is a substantial dataset comprising 134,850 legitimate and 100,945 phishing URLs. Most of the URLs we analysed, while constructing the dataset, are the latest URLs. And this dataset have 55 features (columns) and 235796 rows. The feature explanation in PhiUSIIL dataset:

| Features | | | |
|---|---|---|---|
| URL | ObfuscationRatio | Title | HasHiddenFields |
| URLLength | NoOfLettersInURL | DomainTitleMatchScore | HasPasswordField |
| Domain | LetterRatioInURL | URLTitleMatchScore | Bank |
| DomainLength | NoOfDegitsInURL | HasFavicon | Pay |
| IsDomainIP | DegitRatioInURL | Robots | Crypto |
| TLD | NoOfEqualsInURL | IsResponsive | HasCopyrightInfo |
| URLSimilarityIndex | NoOfQMarkInURL | NoOfURLRedirect | NoOfImage |
| CharContinuationRate | NoOfAmpersandInURL | NoOfSelfRedirect | NoOfCSS |
| TLDLegitimateProb | NoOfOtherSpecialCharsInURL | HasDescription | NoOfJS |
| URLCharProb | SpacialCharRatioInURL | NoOfPopup | NoOfSelfRef |
| TLDLength | IsHTTPS | NoOfiFrame | NoOfEmptyRef |
| NoOfSubDomain | LineOfCode | HasExternalFormSubmit | NoOfExternalRef |
| HasObfuscation | LargestLineLength | HasSocialNet | |
| NoOfObfuscatedChar | HasTitle | HasSubmitButton | |

# Feature selection

In this study, using five (05) feature selection technique, which is Mutual Information (MI), Lasso Regression (L1 Regression), Random Forest, Chi-square Distribution and Information Gain (IG). This technique are so efficient and so powerful for finding best feature.

**A. Mutual Information (MI)**

Mutual Information measures the mutual dependence between two variables. It quantifies the amount of information obtained about one variable through another variable. The formula for mutual information between two variables $X$ and $Y$ is given by:

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

where $p(x,y)$is the joint probability distribution of $X$ and $Y$, and $p(x)$ and $p(y)$ are the marginal probabilities.

# Feature selection

**B. Lasso Regression (L1 Regression)**

Lasso Regression is a type of linear regression that uses L1 regularization. It not only improves the prediction accuracy but also performs feature selection by forcing the coefficients of less important features to be exactly zero. The Lasso optimization problem is defined as:

$$min\beta \left( \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where $\lambda$ is the regularization parameter, $\beta$j are the coefficients, and $N$ is the number of samples.

**C. Random Forest**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. The importance of a feature is evaluated based on the decrease in Gini impurity or entropy when the feature is used to split the data. The importance score for a feature $Xj$ can be expressed as:

$$importance(X_j) = \frac{1}{T} \sum_{t=1}^{T} \Delta_t(X_j)$$

where $T$ is the total number of trees, and $\Delta It$ ($Xj$) is the decrease in impurity due to feature $Xj$ in tree $t$.

# Feature selection

**D. Chi-square Distribution:**
The Chi-square test measures the association between categorical features and the target variable. It helps in selecting features that have a significant impact on the target variable. The Chi-square statistic is calculated as:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where $Oi$ is the observed frequency and $Ei$ is the expected frequency under the null hypothesis.

**E. Information Gain (IG):**
Information Gain evaluates the worth of a feature by measuring the reduction in entropy or uncertainty about the target variable given the feature. It is commonly used in decision tree algorithms. The Information Gain for a feature $X$ is given by:

$$IG(Y, X) = H(Y) - H(Y|X)$$

where $H(Y)$ is the entropy of the target variable $Y$, and $H(Y|X)$ is the conditional entropy of $Y$ given $X$.

# Classification Method

In this study, using four (04) classification algorithms, which is Support Vector Machine (SVM), Logistic Regression, Decision Tree and Random Forest. This algorithms are so efficient and so powerful for finding best feature.

**A. Support Vector Machine (SVM)**

SVM finds the best hyperplane to separate classes with maximum margin, using support vectors. It can handle linearly separable and non-linearly separable data by transforming it into a higher-dimensional space.

**B. Logistic Regression**

Logistic Regression is a method used for binary classification tasks. It predicts the probability of an observation belonging to a specific class. It models this probability using a sigmoid function and optimizes parameters to minimize the difference between predicted probabilities and actual class labels.

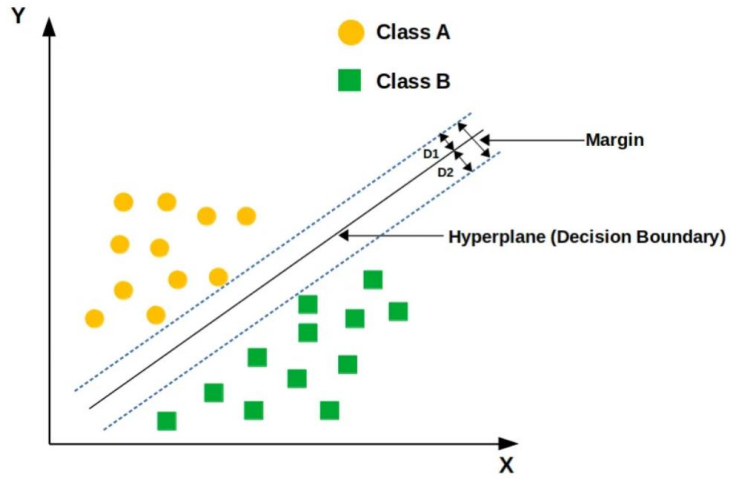


Fig: Support Vector Machine (SVM)

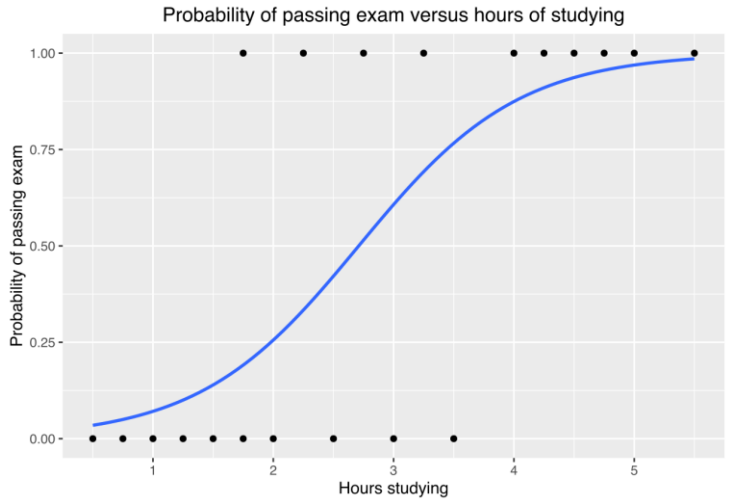


Fig: Logistic Regression

# Classification Method

**C. Decision Tree**

Decision Tree recursively partitions data based on features to create a tree-like structure for classification or regression. It selects features that best split the data to maximize purity in subsets. The result is interpretable but can overfit without proper tuning.

**D. Random Forest**

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to enhance accuracy and reduce overfitting. It randomly selects subsets of data and features for each tree, resulting in a robust and versatile model.

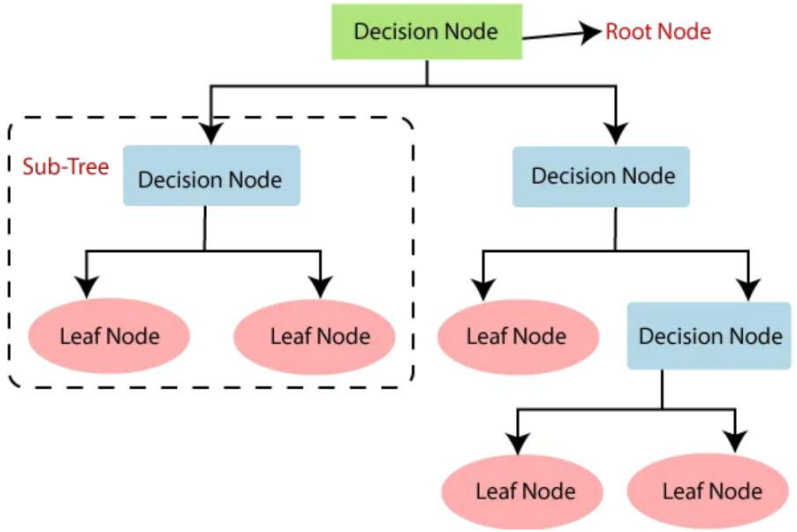


Fig: Decision Tree

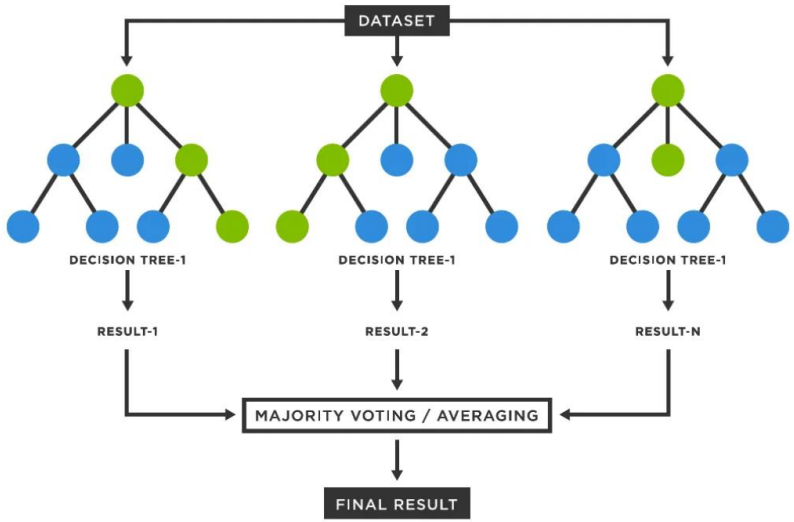


Fig: Random Forest

# Experiment and Results

| Feature Selection Method | Selected Features | Model | Accuracy | Precision | Recall | F1 Score | Feature Selection Method | Selected Features | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mutual Information(MI) | 10 | SVM | 100.00 | 100.00 | 100.00 | 100.00 | Chi-Square Distribution | 10 | SVM | 100.00 | 99.00 | 100.00 | 100.00 |
| | 20 | | 100.00 | 100.00 | 100.00 | 100.00 | | 20 | | 100.00 | 99.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 | | 30 | | 100.00 | 99.00 | 100.00 | 100.00 |
| | 40 | | 100.00 | 100.00 | 100.00 | 100.00 | | 40 | | 100.00 | 99.00 | 100.00 | 100.00 |
| | 50 | | 100.00 | 100.00 | 100.00 | 100.00 | | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 10 | LR | 98.00 | 98.00 | 98.00 | 98.00 | | 10 | LR | 100.00 | 99.00 | 100.00 | 100.00 |
| | 20 | | 99.00 | 100.00 | 99.00 | 99.00 | | 20 | | 100.00 | 99.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 | | 30 | | 100.00 | 99.00 | 100.00 | 100.00 |
| | 40 | | 99.00 | 100.00 | 99.00 | 100.00 | | 40 | | 100.00 | 99.00 | 100.00 | 100.00 |
| | 50 | | 99.00 | 100.00 | 99.00 | 100.00 | | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 10 | DT | 100.00 | 100.00 | 100.00 | 100.00 | | 10 | DT | 100.00 | 100.00 | 100.00 | 100.00 |
| | 20 | | 100.00 | 100.00 | 100.00 | 100.00 | | 20 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 | | 30 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 40 | | 100.00 | 100.00 | 100.00 | 100.00 | | 40 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 50 | | 100.00 | 100.00 | 100.00 | 100.00 | | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 10 | RF | 100.00 | 100.00 | 100.00 | 100.00 | | 10 | RF | 100.00 | 100.00 | 100.00 | 100.00 |
| | 20 | | 100.00 | 100.00 | 100.00 | 100.00 | | 20 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 | | 30 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 40 | | 100.00 | 100.00 | 100.00 | 100.00 | | 40 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 50 | | 100.00 | 100.00 | 100.00 | 100.00 | | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |

# Experiment and Results

| Feature Selection Method | Selected Features | Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| Information Gain (IG) | 10 | SVM | 90.00 | 98.00 | 84.00 | 90.00 |
| | 20 | | 90.00 | 98.00 | 84.00 | 90.00 |
| | 30 | | 90.00 | 98.00 | 84.00 | 90.00 |
| | 40 | | 90.00 | 98.00 | 84.00 | 90.00 |
| | 50 | | 90.00 | 98.00 | 84.00 | 90.00 |
| | 10 | LR | 100.00 | 100.00 | 100.00 | 100.00 |
| | 20 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 40 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 10 | DT | 100.00 | 100.00 | 100.00 | 100.00 |
| | 20 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 40 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 10 | RF | 100.00 | 100.00 | 100.00 | 100.00 |
| | 20 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 30 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 40 | | 100.00 | 100.00 | 100.00 | 100.00 |
| | 50 | | 100.00 | 100.00 | 100.00 | 100.00 |

# Experiment and Results

| Feature Selection with Lasso Regression based | Performance Metrics | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 Score |
| SVM | 91.00 | 99.00 | 86.00 | 92.00 |
| Logistic Regression | 100.00 | 100.00 | 100.00 | 100.00 |
| Decision Tree | 100.00 | 100.00 | 100.00 | 100.00 |
| Random Forest | 100.00 | 100.00 | 100.00 | 100.00 |

| Feature Selection with Random Forest based | Performance Metrics | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 Score |
| SVM | 100.00 | 100.00 | 100.00 | 100.00 |
| Logistic Regression | 100.00 | 100.00 | 100.00 | 100.00 |
| Decision Tree | 100.00 | 100.00 | 100.00 | 100.00 |
| Random Forest | 100.00 | 100.00 | 100.00 | 100.00 |

# Contribution

- In this research, we explore different types of feature selection techniques, which helps to understand the dataset and enhances the overall performances with minimum features.
- We employ several machine learning algorithms on UCI PhiUSIIL Phishing URL dataset and compare performances with them.
- Our achieved accuracy is significantly better than that reported in existing papers.

# Future work

- Ensemble techniques and ANN on  UCI PhiUSIIL Phishing URL dataset.
- And also we will use transformer based network for detecting phishing URLs on this datasets to with compare existing system performances.
- Using our model, we aim to build an browser application or web extension.

# THANK YOU

For Your Kind Patience.

For code:
https://github.com/mrashidulcom/Detection-of-Phishing-URLs-for-Secure-Browsing-A-Machine-Learning-based-Approach.git