Phishing URL Detection

1st Ho Yen Nhi
21520380
Faculty of Computer Science

2nd Huynh Vo Ngoc Thanh
21520449
Faculty of Computer Science

3rd Ngo Tran Tuan Anh
21520567
Faculty of Computer Science

4th Ngo Xuan Cuong
21520663
Faculty of Computer Science

***Abstract*—Phishing attacks continue to be a significant threat to cybersecurity, with attackers constantly evolving their techniques to deceive users. In this study, we address the problem of phishing URL detection through machine learning approaches. Specifically, we explore the effectiveness of Gradient Boosting Classifier and Multi-Layer perceptron (MLP)algorithms in distinguishing between legitimate and phishing URLs. We examine the input, output, and features of URLs, propose solutions leveraging these algorithms, and present comparative results to assess their performance.**

***Index Terms*—Data mining, Machine learning, Feature selection, Detection, Phishing URL, MLP, Gradient Boosting Classifier.**

I. Introduction

In today's digital landscape, cybersecurity is of paramount importance, with phishing attacks posing a significant threat to individuals and organizations alike. Phishing attacks involve the use of deceptive techniques to trick users into divulging sensitive information such as login credentials, financial details, or personal data. Phishing URLs serve as the primary mechanism through which attackers lure unsuspecting victims to fraudulent websites or pages. Recent years have witnessed a significant rise in phishing attacks and websites. According to AntiPhishing Working Group (APWG), the number of reported phishing websites reached its highest in January 2021, with a total of 245,771 websites (the highest ever recorded), gradually declining towards the end of the first quarter of 2021. However, in March 2021, there were over 20,000 phishing attacks, marking it as one of the four months with the highest number of phishing incidents in history.

Traditional methods of phishing URL detection often rely on static rule-based approaches, which may struggle to keep pace with the dynamic and sophisticated nature of modern phishing attacks. Consequently, there is a growing need for data-driven solutions that can adapt to evolving threats. Machine learning offers a promising avenue for addressing this challenge, leveraging patterns and characteristics inherent in URLs to distinguish between legitimate and malicious entities.

In this study, we focus on developing and evaluating machine learning models for phishing URL detection. Specifically, we investigate the efficacy of two algorithms, namely Gradient Boosting Classifier and MLP, in accurately identifying phishing URLs. We delve into the intricacies of the problem, outline the input, output, and features of URLs, propose solutions utilizing machine learning techniques, and present detailed experimental results and analysis.

II. Preliminaries

A. Problem formulation

The main goal of this project is to develop robust machine learning models capable of effectively distinguishing between phishing URLs and legitimate ones.

We apply data mining knowledge to address the subproblem for Phishing URL Detection. Typically, feature extraction from URL dataset. Since URLs are initially in the form of character strings, they need to be transformed into numerical representations to fit the input features of machine learning models. Data mining can help pattern recognition indicative of malicious URLs by analyzing historical data and network behaviors. In addition, data mining can also be used to detect malicious URLs by comparing them to known malicious URLs.

Given an input URL, the targeted model will provide a binary classification indicating whether the URL is benign or malicious. To achieve this, the model will analyze the characteristics of the URL and then utilize machine learning methods to determine the level of reliability of each URL.

We have also developed an application interface for users to easily access and use with an input being an unlabeled URL. The task of this app is to return one of two outcomes: phishing activity (1) or legitimate (0).

B. Feature of URLs

The below mentioned category of features are extracted from the URL data:

1. Address Bar based Features

- **Domain**: feature represents the domain name of the URL. Phishing URLs often use deceptive or misspelled domain names that resemble legitimate ones, whereas legitimate URLs typically use well-known and trusted domain names.
- **Have_IP**: if an IP address is present in the URL instead of a domain name, it's more likely to be a phishing attempt. Legitimate websites typically use domain names for easy recognition, while phishing sites might use IP addresses to obscure their true identity.
- **Have_At**: the presence of the '@' symbol in a URL is unusual and may indicate a phishing attempt. Legitimate URLs rarely contain this symbol, whereas it might be used in phishing URLs to deceive users.
- **URL_Length**: phishing URLs tend to have shorter lengths compared to legitimate ones. Shorter URLs are

easier to disguise, whereas legitimate URLs often have longer, more descriptive paths.

- **URL_Depth**: the number of clicks required to navigate from the homepage to the URL. Phishing URLs often have complex and deep paths to mimic legitimate sites, while legitimate URLs typically have simpler structures.
- **Redirection** ('//'): the presence of double slashes ('//') in a URL may indicate a redirection attempt, which is commonly associated with phishing attacks. Legitimate URLs seldom contain unnecessary redirections, while phishing URLs may use this technique to mislead users.
- **https_Domain**: phishing URLs may attempt to mimic secure websites by including 'http' or 'https' in their domain names. This can deceive users into thinking the site is legitimate, whereas legitimate websites commonly use secure protocols.
- **TinyURL**: phishing URLs often use URL shortening services to disguise their true destination. These shortened URLs can make it difficult for users to discern the actual domain and may lead to malicious websites.
- **Prefix/Suffix** (Prefix or Suffix "-" in Domain): phishing URLs may add prefixes or suffixes, such as hyphens ('-'), to their domain names to mimic legitimate websites, whereas legitimate URLs typically have straightforward domain names.

2. Domain based Features

- **DNS_Record**: legitimate websites typically have valid DNS records associated with their domain names, while phishing websites may lack proper DNS records or have suspicious records.
- **Web_Traffic**: website traffic refers to the volume of visitors accessing a particular website. Higher website traffic is generally associated with legitimate websites that offer valuable content or services. In contrast, phishing websites may have lower, especially if they are newly created or not widely known.
- **Domain_Age**: the age of a domain name reflects how long it has been registered and active. Legitimate websites often have older domains, as they have been established over time. Phishing websites may have recently created domains, which could be a red flag indicating potential malicious intent.
- **Domain_End**: when a domain expires, it becomes immediately inactive, and all associated services cease to function, legitimate websites typically renew their domain registrations to maintain continuity. Phishing websites may have short or irregular registration end periods, indicating a lack of long-term commitment or potential for abandonment.

3. HTML and JavaScript based Features

- **iFrame**: "iFrame" is an HTML tag used to embed another webpage or part of a webpage within a webpage. If a website uses iframe redirection, it may attempt to load malicious content from another source without the user's knowledge.
- **Mouse_Over**: the action of a user moving the mouse pointer over an element on a webpage. Phishing websites may customize the status bar to display false information such as disguising the destination URL or showing fake security indicators. Legitimate websites typically do not manipulate the status bar in this manner.
- **Right_Click**: the action of clicking the right mouse button. Phishing websites may disable the right-click functionality to prevent users from accessing browser features like viewing page source or opening links in new tabs.
- **Web_Forwards**: "Web forwards" is a technique used to redirect users from a legitimate-looking URL to a fraudulent one, where sensitive information can be collected or malicious activities performed. Phishing websites may use this technique to redirect users from a legitimate-looking URL to a fraudulent one, where sensitive information can be collected or malicious activities performed.

III. Proposed Method

In training stage, after collecting URLs from the open source platforms, we label them and extract the required features from the URL database. Next, analyze and preprocess the dataset by using EDA techniques and run selected machine learning algorithms (MLP, Gradient Boosting Classifier). Then, train the models based on the extracted features. The process is as shown in Figure 1.

In detection stage, a new url is extracted and then put into the trained model in training stage and this model is responsible for predicting the url is Phishing URL or Legitimate URL.

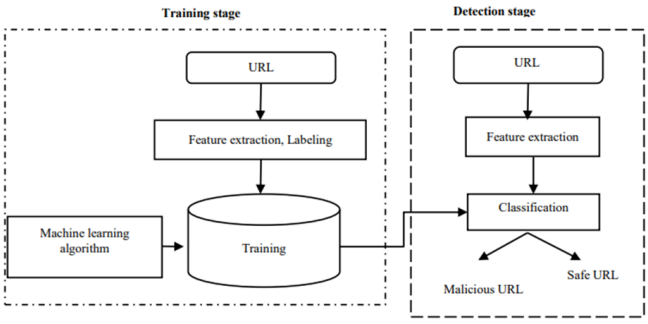

Fig. 1. Pipeline

The below section introduces two methods in machine learning, namely Gradient Boosting and Multilayer Perceptron (MLP), applied to solve the problem of phishing URL detection.

A. MLP

A Multi-Layer Perceptron (MLP) is a type of artificial neural network (ANN) designed for supervised learning. It consists of multiple layers of nodes, each connected to the next layer, forming a feedforward network. The layers are typically organized into an input layer, one or more hidden layers, and an output layer.

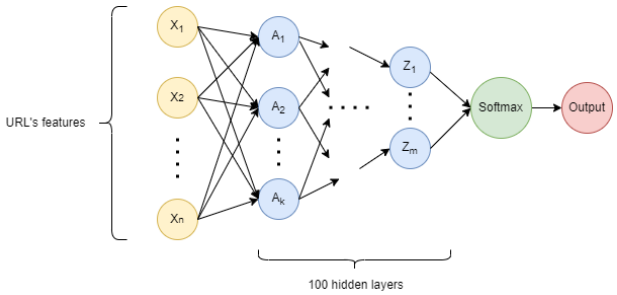

Fig. 2. MLP Architect

The architecture of our MLP model can be seen in fig 2. The proposed MLP model has 100 hidden layers followed by a ReLU activation function after each layer. To prevent overfitting by penalizing weights with large magnitudes, L2 regulation is added in the process of calculating node values with the strength of Alpha.

B. Gradient Boosting Classifier

Gradient Boosting Classifier (GBC) is a powerful machine learning algorithm used for both regression and classification tasks. It is an ensemble learning technique that combines the predictions from multiple base learners, typically decision trees, to improve predictive accuracy. GBC sequentially builds weak learners by optimizing a given loss function using gradient descent.

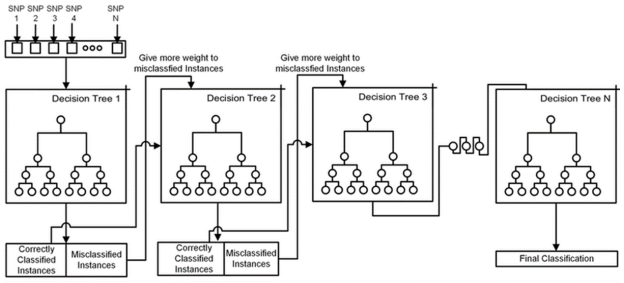

Fig. 3. GBC Architect

Fig 3. shows the architecture of our Gradient boosting. First, features are fed to the model to build a decision tree to predict its class. After the first prediction, we build the next tree to estimate the error of the previous tree. This fashion continues until the model reaches the number of trees that have been set up or the loss experienced almost no changes with the additional tree.

IV. Experiment

A. Dataset

To compare 2 methods, we conducted experiments on a large dataset of over 10000 URLs from various sources:

- The set of phishing URLs are collected from opensource service called PhishTank. This service provide a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly. From this dataset, 5000 random phishing URLs are collected to train the ML models.
- The legitimate URLs are obatined from the open datasets of the University of New Brunswick. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign url dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

B. MLP

With Multi-Layer Perceptron, we set the learning rate and Alpha at 0.001 and 0.0001 respectively and there were no changes during the training process. We trained the model with Adam optimizer over 200 epochs.

C. Gradient Boosting Classifier

To build GBC, we adopt Logloss as the loss function. This is cross-entropy loss for binary classification which is suitable for this problem. The maximum number of trees built is 100 with the maximum depth set to 4. We trained the model with a learning rate of 0.1.

V. Evaluation

A. Feature important

For analyzing feature significance in classifying benign and malicious URLs, each feature is associated with a weight corresponding to their contribution in building GBC model. The overall result is presented in Fig 4.

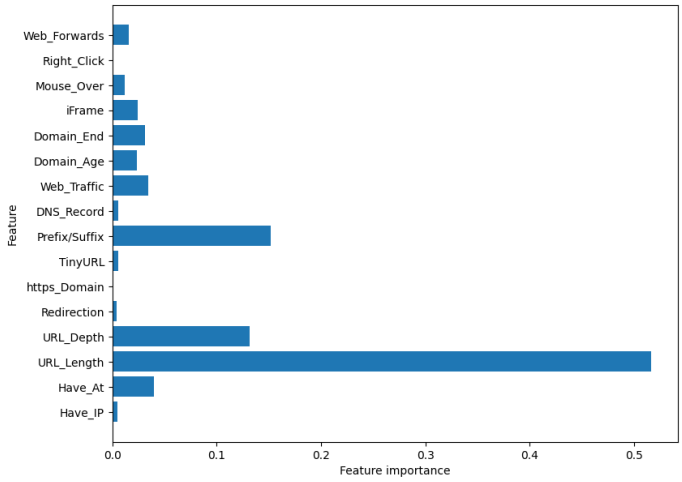

Fig. 4. Feature important

It is noticed that the URL length plays a crucial role in determining whether it's a phishing attempt, carrying significant weight with a value exceeding 0.5. This implies that even slight variations in URL length can significantly impact its classification. Notably, research cited in [7] supports the use of a threshold of 54 characters as reliable for distinguishing between legitimate URLs and potentially harmful ones.

The features URL depth and Prefix/Suffix also exert notable influence on model predictions, with weights of approximately 0.15 for both features. Deep navigation within URLs is often indicative of phishing attempts, as it deceives users into unwittingly redirecting to fraudulent websites where sensitive personal information may be stolen. Similarly, the addition of prefixes and suffixes, particularly those incorporating a dash (-), lowers user awareness towards potential phishing URLs. These observations underscore why these features hold considerable sway in guiding the decision-making process of Gradient Boosting Classifier (GBC). Another reason could be the distribution of these 2 features in the dataset. As shown in Fig 5.1, URL depth distribution nearly resembles normal distribution, meaning that GBC may have the tendency to split points around the mean of the feature values. This preference can lead to splits that align with the natural distribution of the data, potentially resulting in a more efficient separation of classes.

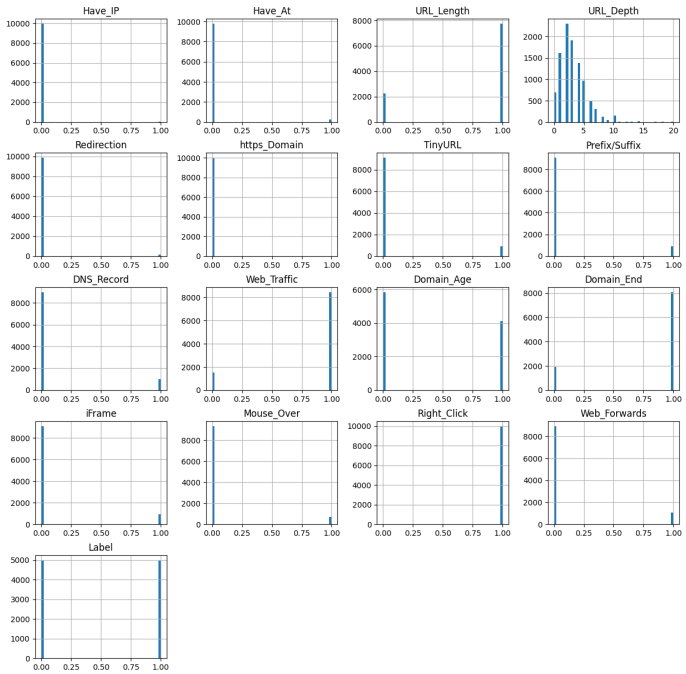

Fig. 5. EDA

In contrast, certain features make modest contributions to the classifiers, each carrying a weight of approximately 0.04. Notably, features such as Right click and HTTPS domain exhibit no influence on the model. This observation stems from the underlying data distribution across these features. As illustrated in Figure 5.1, the majority of features demonstrate imbalanced distributions, thereby constraining their impact on the classifiers.

B. Result analysis

To evaluate the performance of 2 models, we use accuracy, precision, recall and F1 score as evaluation metrics. We perform evaluation on the above dataset and provide classification results in Table below.

		Accuracy	**Precision**	**Recall**	**F1-score**
MLP	Training data	0.847	0.834	0.770	0.909
	Test data	0.846	0.834	0.783	0.903
GBC	Training data	0.866	0.856	0.800	0.920
	Test data	0.870	0.866	0.820	0.917

Both the MLP (Multi-Layer Perceptron) and GBC (Gradient Boosting Classifier) models exhibit strong performance in detecting fraudulent URLs, but with some differences.

The MLP model demonstrates stable performance with high precision, accurately predicting fraudulent websites. However, its recall values indicate potential for improvement to minimize missed fraudulent instances. The F1-score suggests a balance between precision and recall.

On the other hand, the GBC model demonstrates exceptional performance across various metrics, including high accuracy, precision, recall, and F1-score. It showcases superior performance compared to the MLP model, particularly in terms of recall and overall balance between precision and recall.

In conclusion, the GBC model appears to be the more effective choice for detecting fraudulent URLs due to its superior performance across all metrics.

VI. Conclusion

Internet users face a significant threat from phishing, and detecting malicious URLs is challenging. This study applies data mining techniques to address this problem, training MLP and Gradient Boosting Classifier models on URL features. The GBC model (accuracy = 0.87) outperforms MLP (accuracy = 0.846) due to its ability to model complex relationships with fewer computational resources. However, adding more features and fresh data is necessary to improve accuracy and reliability. Future work includes leveraging the models to develop a search engine for identifying and banning fraudulent URLs and creating a framework for autonomously detecting new phishing attacks with advanced features.

References

[1] Phishing URL Detection: A Network-based Approach Robust to Evasion
[2] Phishing URL detection using machine learning methods
[3] A Feature Extraction Approach for the Detection of Phishing Websites Using Machine Learning
[4] Phishing URL Detection using Information-rich Domain and Path Features
[5] Phishing Website URL's Detection Using NLP and Machine Learning Techniques
[6] A Web Application for Real-Time Phishing Website Detection
[7] An assessment of features related to phishing websites using an automated technique

Task assigned

	Yen Nhi	Ngoc Thanh	Tuan Anh	Xuan Cuong
Research the content and knowledge about the project	x	x	x	x
Write a report	x	x		x
Make slides for presentation	x	x		
Train model and Build app for Demo			x	
Presentation				x