A PROJECT REPORT ON

ANONYMITY AND CONFIDENTIALITY IN WEBSITE USING ML

SUBMITTED IN PARTIAL FULFILLMENT FOR AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

IN

CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

BY

Aditya Mishra (2101321530003)
Anshuman Soni (2101321530007)
Anuj Mishra (2101321530008)
Shivam Singh (2101321530045)

UNDER THE GUIDANCE OF MR. UMASHANKER SHARMA

DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

GREATER NOIDA INSTITUTE OF TECHNOLOGY, GREATER NOIDA

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

Department of CSE (Artificial Intelligence and Machine Learning)

Session 2024-2025

Project Completion Certificate

Date: 02/01/2025

This is to certify that Mr./Ms. **ADITYA MISHRA** bearing Roll No.2001321530003 student of 4TH year CSE(AI & ML) has completed major project program (KCS-753) with the Department of CSE (Artificial Intelligence and Machine Learning) from 14-Sept-24 to 02-Jan-2025.

He/she worked on the Project Titled "Anonymity and Confidentiality in Website using ML" under the guidance of Mr. Umashanker Sharma.

This project work has not been submitted anywhere for any diploma/degree.

Mr. Umashanker Sharma

Assistant Professor, CSE(AI&ML) **Project Coordinator/HoD-CSE(AI&ML)**

Department of CSE (Artificial Intelligence and Machine Learning)

Session 2024-2025

Project Completion Certificate

Date: 02/01/2025

This is to certify that Mr./Ms. **ANSHUMAN SONI** bearing Roll No.2001321530007 student of 4TH year CSE(AI & ML) has completed major project program (KCS-753) with the Department of CSE (Artificial Intelligence and Machine Learning) from 14-Sept-24 to 02-Jan-2025.

He/she worked on the Project Titled "Anonymity and Confidentiality in Website using ML" under the guidance of Mr. Umashanker Sharma.

This project work has not been submitted anywhere for any diploma/degree.

Mr. Umashanker Sharma

Assistant Professor, CSE(AI&ML) **Project Coordinator/HoD-CSE(AI&ML)**

Department of CSE (Artificial Intelligence and Machine Learning)

Session 2024-2025

Project Completion Certificate

Date: 02/01/2025

This is to certify that Mr./Ms. **ANUJ MISHRA** bearing Roll No.2001321530008 student of 4TH year CSE(AI & ML) has completed major project program (KCS-753) with the Department of CSE (Artificial Intelligence and Machine Learning) from 14-Sept-24 to 02-Jan-2025.

He/she worked on the Project Titled "Anonymity and Confidentiality in Website using ML" under the guidance of Mr. Umashanker Sharma.

This project work has not been submitted anywhere for any diploma/degree.

Mr. Umashanker Sharma

Assistant Professor, CSE(AI&ML) **Project Coordinator/HoD-CSE(AI&ML)**

Department of CSE (Artificial Intelligence and Machine Learning)

Session 2024-2025

Project Completion Certificate

Date: 02/01/2025

This is to certify that **Mr./Ms. SHIVAM SINGH** bearing Roll No.2001321530045 student of 4TH year CSE(AI & ML) has completed major project program (KCS-753) with the Department of CSE (Artificial Intelligence and Machine Learning) from 14-Sept-24 to 02-Jan-2025.

He/she worked on the Project Titled "Anonymity and Confidentiality in Website using ML" under the guidance of Mr. Umashanker Sharma.

This project work has not been submitted anywhere for any diploma/degree.

Mr. Umashanker Sharma

Assistant Professor, CSE(AI&ML) **Project Coordinator/HoD-CSE(AI&ML)**

ABSTRACT

Phishing attacks are among the simplest yet most effective ways for cybercriminals to obtain sensitive information from unsuspecting users. These attacks target critical data such as usernames, passwords, and financial information, posing serious threats to individuals and organizations alike.

In response to this growing threat, cybersecurity experts are leveraging advanced detection techniques to identify phishing websites. This project harnesses machine learning technology to analyze features of legitimate and malicious URLs, offering a robust solution.

Three algorithms—Decision Tree, Random Forest, and Support Vector Machine—are implemented to evaluate performance. The primary goal is to identify the most reliable model by comparing metrics such as accuracy, false positive, and false negative rates, ensuring an efficient approach to combat phishing threats.

ACKNOWLEDGEMENT

I have made efforts in this project. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. I am highly indebted to Mr. Arun Kumar Rai for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. I would like to express my gratitude towards my parents & members of GNIOT for their kind cooperation and encouragement which helped me in the completion of this project. I would like to express my special gratitude and thanks to the industry people for giving me such attention and time. My thanks and appreciations also go to my colleague in developing the project and the people who have willingly helped me out with their abilities.

Date: 02/01/2025

Aditya Mishra (2101321530003)
Anshuman Soni (2101321530007)
Anuj Mishra (2101321530008)
Shivam Singh (2101321530045)

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

Phishing websites are carefully crafted online traps designed to steal sensitive information from unsuspecting users. They often mimic legitimate sites, luring victims to divulge login credentials, financial data, or other valuable personal details.

Nowadays Phishing has become a main area of concern for security researchers because it is not difficult to create a fake website which looks so close to a legitimate website. Experts can identify fake websites but not all the users can identify the fake website, and such users become the victim of phishing attack.

Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US $2 billion per year because their clients become victim to phishing. In the 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because of lack of user awareness. Since phishing attacks exploit the weaknesses found in users, it is difficult to mitigate them, but it is important to enhance phishing detection techniques.

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including fast flux, in which proxies are automatically generated to host the webpage; algorithmic generation of new URLs; etc.

A major drawback of this method is that it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high.

To overcome the drawbacks of blacklist and heuristics-based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of many algorithms which require past data to decide or predict future data. Using this technique, algorithms will analyze various blacklisted and legitimate URLs and their features to accurately detect phishing websites including zero- hour phishing websites.

I. i. **LITERATURE SURVEY**

[1] Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh: Using six criteria based on URL parameters such as the subdomain, principal domain, Page rank, Alexa rank, path domain, and Alexa reputation, this article suggests a novel method for identifying phishing websites. The method focuses on evaluating how closely a phishing site's URL resembles the URL of a reliable website and also takes into account the site's ranking as a crucial component in determining its validity. The approach was tested using data from PhishTank and DMOZ, and the authors showed that it could identify over 97% of phishing sites**.**

[2] Irfan Siddavatam, Rishikesh Mahajan: Introduced the use of decision trees , SVM, Random Forest to classify phishing websites based on features like URL length, domain age, presence of special characters, and security certificate information. Their approach demonstrated improved detection rates compared to traditional methods but faced challenges in identifying phishing sites.

[3] S. Arvind Anwekar, V. Agrawal : In this study, the authors focused on extracting features from URLs, in addition to other features such as the age of the SSL certificate and the universal resource locator of the anchor, IFRAME, and website rank. They collected URLs of phishing websites from PhishTank and URLs of benign websites from the Alexa website. Using a combination of the random forest (RF), decision tree (DT), and support vector machine (SVM), contributed to improving the detection mechanism for phishing websites and achieved a high noticeable detection accuracy of 97.14%, with a low rate of false positives at 3.14%. The results also showed that the classifier's performance improves with more training data.

[4] N. Choudhary b, K. Jain, S. Jain : This study emphasizes the significance of only using attributes from the URL. Both the Kaggle and Phishtank websites make it easy to get the dataset used in this study. The researchers used a hybrid approach that com-bined Principal Component Analysis (PCA) with Support Vector Machine (SVM) and Random Forest algorithms to reduce the dataset's dimensionality while keeping all im-portant data, and it produced a higher accuracy rate of 96.8% compared to other tech-niques investigated.

[5] A. Lakshmanarao, P. Surya, M Bala Krishna: This thesis collected a dataset of phishing websites from the UCI repository and used various Machine learning techniques, including decision trees, AdaBoost, support vector machines (SVM), and random forests, to analyze selected features (such as web traffic, port, URL length, IP address, and URL_of_Anchor). The most effective model for detecting phishing web-sites was chosen, and two priority-based algorithms (PA1 and PA2) were proposed. The team utilized a new fusion classifier in conjunction with these algorithms and at-tained an accuracy rate of 97%. when compared to previous works in phishing website detection

.

[6] L. Tang, Q. Mahmoud: The proposed approach in the current study uses URLs collected from a variety of platforms, including Kaggle, Phish Storm, Phish Tank, and ISCX-UR, to identify phishing websites. The researchers made a big contribution since they created a browser plug-in that can quickly recognize phishing risks and offer warn-ings. Various datasets and machine learning techniques were investigated, and the pro-posed RNN-GRU model outperformed SVM, Random Forest (RF), and Logistic Re-gression with a maximum accuracy rate of 99.18%. On the other hand, the suggested method is not always accurate in identifying if short links are phishing risks.

[7] A. Kulkarni & L. Brown: A machine learning system was created to categorize websites based on URLs from the University of California, Irvine Machine Learning Repository. Four classifiers were used: SVM, decision tree, Naive Bayesian, and neural network. The outcome of experiments utilizing the model developed with the support of a training set of data demonstrates that the classifiers were able to successfully differentiate authentic websites from fake ones with an accuracy rate of over 90%. Limitations include a small dataset and all features being discrete, which may not be suitable for some classifiers.

[8] Tyagi; J. Shad; S. Sharma; S. Gaur Gagandeep Kaur: The research taken into account focuses on the use of various machine learning algorithms to identify if a web-site is legitimate or a phishing site based on a URL. This study's most important con-tribution is the creation of the Generalized Linear Model (GLM), a brand-new model. This model combines the results of two various methods. With a 98.4% accuracy rate, the Random Forest and GLM combination produced the best results for detecting phishing websites.

[9] M. Karabatak and T. Mustafa: The objective of this research is to assess the effectiveness of classification algorithms on a condensed dataset of phishing websites obtained from the UCI Machine Learning Repository. The paper investigates how data mining and feature selection algorithms affect reduced datasets through experiments and analysis, finally selecting the methods that perform the best in terms of classifica-tion. According to the results, some classification strategies improve performance while others have the opposite impact. Ineffective classifiers for condensed phishing datasets included Lazy, BayesNet, SGD Multilayer Perceptron, PART, JRip, J48, RandomTree, and RandomForest. However, it was discovered that KStar, LMT, ID3, and R.F.Clas-sifier were efficient. Lazy produced the highest classification accuracy rate of 97.58% on the compressed 27-feature dataset, whereas KStar performed at its best on the same dataset.

[10] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng: A phishing detection model that applies Bagging, AdaBoost, SMO, and Random Forest algorithms to learn and test phishing detection strategies is offered as a contribution to this work. The model is based on features from URLs and extracts multi-level statistical characteristics, seman-tic features of word embedding, and semantic features from Chinese web content. Legal URLs from DirectIndustry online instructions and phishing data from the Anti-Phishing Alliance of China (APAC) are included in

the dataset used to test the algorithm. The study's findings suggest that a fusion model that primarily employed semantic data to identify phishing sites with high detection efficiency had the best performance, leading to a new contribution with an F-measure of 0.99%. Keep in mind that this approach is specific to Chinese websites and is language-dependent.

[11] W. Fadheel, M. Abusharkh, and I. Abdel-Qader : The present study utilized datasets from the UCI machine learning repository, including Domain, HTML, Address Bar, and URLs, the main contribution was conducting a comparative analysis of the impact of feature selection on detecting phishing websites. The KMO test was applied in the study to evaluate the dataset using (LR) and (SVM) classification algorithms. The test was conducted based on a correlation matrix to analyze the performance. Re-sults showed that LR with the KMO test achieved an accuracy of 91.68%, while SVM with the KMO test yielded an accuracy of 93.59%.

[12] A. Ahmed and N. A. Abdullah: The research team developed a software program known as Phish Checker, which is designed to distinguish between legitimate and phishing websites. The proposed approach focuses on identifying phishing attacks by analyzing the URLs and domain names of suspected phishing websites to determine their authenticity. Data was collected from the Yahoo and PhishTank directories and the results indicate that PhishChecker has an accuracy rate of 96% for identifying phish-ing websites. However, it should be noted that this method is based on heuristics and its effectiveness is reliant on the availability of certain discriminative elements that aid in identifying the type of website. Additionally, the study only examines the validity of URLs in determining website authenticity.

[13] Ankit Kumar Jain & B. B. Gupta: The proposed strategy utilizes an Innovative methodology for defending counteract phishing attempts by incorporating a URL and DNS matching module with a white list of trusted websites that are automatically up-dated based on each user's browsing history. This method offers quick retrieval speeds, high rates of detection, and alerts users to not disclose personal information when at-tempting to access a website, not on the white list. Additionally, hyperlink properties are utilized to verify the validity of a website by retrieving hyperlinks from the source code and applying them to the phishing detection method. The performance of this strategy was evaluated using data from reputable sources such as Stuffgate, Alexa, and PhishTank and achieved an accuracy rate of 89.38 %.

[14] M. Aydin and N. Baykal: Throughout this experiment, phishing websites were detected using subset-based feature selection methods based on URL attributes. A da-taset comprising both legitimate and phishing URLs was obtained from Google and PhishTank, and multiple features were retrieved from URLs. The usefulness of two classification algorithms—Naive Bayes and Sequential Minimal Optimization (SMO)—for identifying phishing websites was investigated in this study. The results showed that SMO performed better than Naive Bayes for phishing detection, with an accuracy rate of 95.39%. The SMO algorithm also had another benefit in that it made use of more chosen features overall. The accuracy rate of the Naive Bayes method, in contrast, was 88.17% while using the same quantity of chosen features.

II. OBJECTIVES

1. PROBLEM STATEMENT

Phishing websites have become a significant threat in the digital landscape, leading to financial losses and identity theft for individuals and organizations. These fraudulent websites often mimic legitimate ones, tricking users into disclosing sensitive information such as passwords, credit card numbers, and personal details. Traditional methods for detecting phishing sites, such as blacklisting, are limited in scalability and fail to detect new or dynamically generated phishing URLs in real-time.

The goal of this project is to develop a machine learning-based system to detect phishing websites by analysing features such as URL structure, domain age, security certificates, and website content. By leveraging supervised learning techniques, the system should be able to classify websites as phishing or legitimate with high accuracy, based on a set of labelled data.

2. KEY OBJECTIVES

1. Collect and preprocess data, including features from both phishing and legitimate websites.

2. Apply feature engineering to identify key characteristics distinguishing phishing websites.

3. Train multiple machine learning models and compare their performance.

4. Develop an efficient and scalable solution that can identify phishing websites in real-time.

5. Evaluate the model's accuracy, precision, recall, and F1-score using appropriate test datasets.

The successful implementation of this project will result in a robust and efficient system capable of helping users and organizations safeguard against phishing attacks.

III. METHODOLOGY

1. DATA COLLECTION

a. Data Collection:

Phishing Website Data: Gather datasets from publicly available sources such as PhishTank, UCI repository, or Kaggle, which contain records of both phishing and legitimate websites.

Legitimate Website Data: Collect data from trustworthy sources for a balanced comparison with phishing websites.

b. Data sets:

- 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. To
  download the data: https://www.phishtank.com/developer_info.php. 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, https://www.unb.ca/cic/datasets/url-2016.html. 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.

c. Data Sets Sample:

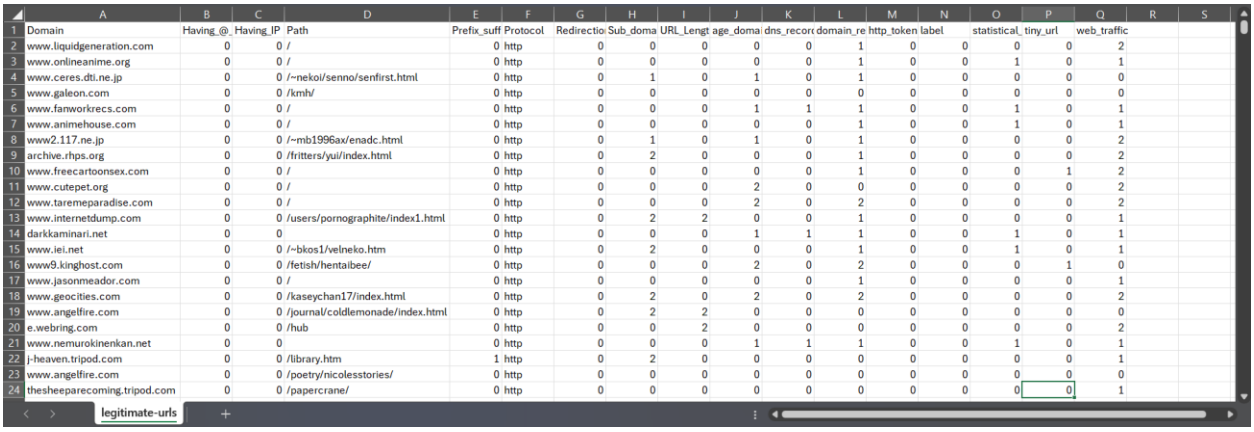
Sample of Phishing websites csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Domain	Having_@_	Having_IP	Path	Prefix_suff	Protocol	Redirectio	Sub_doma	URL_Lengt	age_domai	dns_record	domain_re	http_token	label	statistical_	tiny_url	web_traffic
2	asesoresvelfit.com	0	0	/media/datacredito.co/	0	http	0	0	0	0	0	1	0	1	0	1	1
3	caixa.com.br.fgtsagendesaqueconta.com	0	0	/consulta8523211/principal.php	0	http	0	1	1	0	0	1	0	1	1	0	1
4	hissoulreason.com	0	0	/js/homepage/home/	0	http	0	0	0	0	0	1	0	1	0	0	1
5	unauthorizd.newebpage.com	0	0	/webapps/66fbf/	0	http	0	0	0	0	0	1	0	1	1	0	1
6	133.130.103.10	0	1	/23/	0	http	0	2	0	1	0	1	0	1	0	0	1
7	dj00.co.vu	1	0	/css/	0	http	0	0	2	1	1	1	0	1	1	0	0
8	133.130.103.10	0	1	/21/logar/	0	http	0	2	0	1	0	1	0	1	0	0	1
9	httpssicredi.esy.es	0	0	/servico/sicredi/validarclientes/mobi/index.php	0	http	0	2	2	1	1	1	1	1	1	0	1
10	gamesaty.ga	0	0	/wp-content//yh/en/	0	http	1	0	2	1	0	1	0	1	0	0	1
11	luxuryupgradepro.com	0	0	/ymailNew/ymailNew/	0	http	0	0	0	0	0	1	0	1	0	0	1
12	133.130.103.10	0	1	/1/	0	http	0	2	0	1	0	1	0	1	0	0	1
13	133.130.103.10	0	1	/24/sicredi/psmlld/31/paneid/index.htm	0	http	0	1	2	1	0	1	0	1	0	0	1
14	smscaixaacesso.hol.es	0	0		0	http	0	0	0	1	1	1	0	1	1	0	1
15	133.130.103.10	0	1	/7/SIlBC/siwinCtrl.php	0	http	0	1	0	1	0	1	0	1	0	0	1
16	tinyurl.com	0	0	/kjmmw57	0	http	0	0	0	0	0	0	0	1	0	1	0
17	wrightlandscapes.org	0	0	/no/T/Y1.html	0	http	0	0	0	1	0	1	0	1	1	0	1
18	mautic.eto-cms.ru	0	0	/themes/goldstar/mtbonline/newmandt/	1	http	0	0	2	0	0	1	0	1	0	0	1
19	ginatringali.com	0	0	//al/alibaba21012015/alibaba21012015/666/in	0	http	1	0	2	0	0	0	0	1	1	0	1
20	staticmail.000webhostapp.com	0	0	/	0	https	0	0	0	0	0	0	0	1	0	0	0
21	umeda.com.br	0	0	/bba/BOA/home/	0	http	0	0	0	1	0	1	0	1	1	0	1
22	krishworldwide.com	0	0	/BackUp/under/js/ayo1/index.html	0	http	0	0	2	0	0	1	0	1	0	0	1
23	yahoo.co.in	0	0	/email_open_log_pic.php	0	http	0	2	1	1	0	1	0	1	0	0	2
24	www.avcc.ac.in	0	0	/fonts/1/wropboxp/login.html	0	http	0	1	0	1	0	1	0	1	0	0	2

phishing-urls

Fig.III.1.c.i. phishing_urls.csv

Sample of Legitimate websites csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Domain	Having_@_	Having_IP	Path	Prefix_suff	Protocol	Redirectio	Sub_doma	URL_Lengt	age_doma	dns_recor	domain_re	http_token	label	statistical_	tiny_url	web_traffic		
2	www.liquidgeneration.com	0	0	/	0	http	0	0	0	0	0	1	0	0	0	0	2		
3	www.onlineanime.org	0	0	/	0	http	0	0	0	0	0	1	0	0	1	0	1		
4	www.ceres.dti.ne.jp	0	0	/~nekoi/senno/senfirst.html	0	http	0	1	0	1	0	1	0	0	0	0	0		
5	www.galeon.com	0	0	/kmh/	0	http	0	0	0	0	0	0	0	0	0	0	0		
6	www.fanworkrecs.com	0	0	/	0	http	0	0	0	1	1	1	0	0	1	0	1		
7	www.animehouse.com	0	0	/	0	http	0	0	0	0	0	1	0	0	1	0	1		
8	www2.117.ne.jp	0	0	/~mb1996ax/enadc.html	0	http	0	1	0	1	0	1	0	0	0	0	2		
9	archive.rhps.org	0	0	/fritters/yui/index.html	0	http	0	2	0	0	0	1	0	0	0	0	2		
10	www.freecartoonsex.com	0	0	/	0	http	0	0	0	0	0	1	0	0	0	1	2		
11	www.cutepet.org	0	0	/	0	http	0	0	0	2	0	0	0	0	0	0	2		
12	www.taremeparadise.com	0	0	/	0	http	0	0	0	2	0	2	0	0	0	0	2		
13	www.internetdump.com	0	0	/users/pornographite/index1.html	0	http	0	2	2	0	0	1	0	0	0	0	1		
14	darkkaminari.net	0	0		0	http	0	0	0	1	1	1	0	0	1	0	1		
15	www.iei.net	0	0	/~bkos1/velneko.htm	0	http	0	2	0	0	0	1	0	0	1	0	1		
16	www9.kinghost.com	0	0	/fetish/hentaibee/	0	http	0	0	0	2	0	2	0	0	0	1	0		
17	www.jasonmeador.com	0	0	/	0	http	0	0	0	0	0	1	0	0	0	0	1		
18	www.geocities.com	0	0	/kaseychan17/index.html	0	http	0	2	0	2	0	2	0	0	0	0	2		
19	www.angelfire.com	0	0	/journal/coldlemonade/index.html	0	http	0	2	2	0	0	0	0	0	0	0	0		
20	e.webring.com	0	0	/hub	0	http	0	0	2	0	0	0	0	0	0	0	2		
21	www.nemurokinenkan.net	0	0		0	http	0	0	0	1	1	1	0	0	1	0	1		
22	j-heaven.tripod.com	0	0	/library.htm	1	http	0	2	0	0	0	0	0	0	0	0	1		
23	www.angelfire.com	0	0	/poetry/nicolesstories/	0	http	0	0	0	0	0	0	0	0	0	0	0		
24	thesheeparecoming.tripod.com	0	0	/papercrane/	0	http	0	0	0	0	0	0	0	0	0	0	1		

Fig.III.1.c.ii. legitimate_urls.csv

2. Feature Extraction

Feature Extraction consists of:
Extract significant features from website URLs and webpage content that differentiate phishing websites from legitimate ones.
Key features to consider:

- URL-based features: Length of URL, presence of special characters, use of IP address instead of a domain name, presence of suspicious keywords.
- Domain-based features: Domain age, domain name length, registration information, presence of SSL certificates.
- Page-based features: Website content, use of HTTPS, form handling, third-party resources.

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

1. Address Bar based Features

In this category 9 features are extracted.

2. Domain based Features

In this category 4 features are extracted.

3. HTML & Javascript based Features

In this category 4 features are extracted.
So, all together 17 features are extracted from the 10,000 URL dataset.

Feature Extraction for Phishing Website Detection:

1. Using URL Shortening Services

- **Objective**: Identify phishing attempts using shortened URLs.
- **Method**: Detect URL shortening services like "bit.ly" or "tinyurl.com" through keyword matching. These services obscure the true destination, often used by attackers.

2. Existence of HTTPS Token

- **Objective**: Avoid misleading security indicators in the domain.
- **Method**: Flag URLs containing “HTTPS” in places other than the protocol (e.g., "httpssecure.com").

3. Abnormal URL

- **Objective**: Validate the legitimacy of domains.
- **Method**: Cross-check domains against WHOIS databases to verify registration details. Abnormal or unregistered domains raise suspicion.

4. Google Index

- **Objective**: Confirm the URL is indexed by search engines.
- **Method**: Use Google search APIs to verify if the URL exists in Google's database. Non-indexed sites are more likely to be phishing.

5. Website Traffic

- **Objective**: Assess website popularity.
- **Method**: Analyze Alexa rank data. Legitimate sites usually have higher traffic, while phishing sites rank low or are absent.

6. Domain Registration Length

- **Objective**: Detect short-lived domains.
- **Method**: Phishing domains often register for under a year. WHOIS checks identify short registration durations.

7. Age of Domain

- **Objective**: Identify newly created domains.
- **Method**: WHOIS records reveal the domain's age. New domains are a common indicator of phishing attempts.

8. DNS Record

- **Objective**: Verify domain validity.
- **Method**: DNS queries ensure the existence of legitimate records. Missing or invalid records suggest phishing activity.

9. Statistical Report

- **Objective**: Leverage threat intelligence databases.
- **Method**: Match URLs and IPs against lists of known phishing entities. Regularly updated datasets enhance accuracy.

10. Long URLs

- **Objective**: Detect unusually long URLs.
- **Method**: URLs exceeding a standard character limit (e.g., >75 characters) are flagged as suspicious.

11. @ Symbol in URL

- **Objective**: Identify domain obfuscation techniques.
- **Method**: The "@" symbol redirects users to different pages and is typically used in phishing URLs.

12. Double Slashes (//) in Path

- **Objective**: Detect abnormal URL formatting.
- **Method**: URLs with "//" in unexpected places (other than "http://") are flagged as suspicious.

13. Subdomains

- **Objective**: Identify excessive subdomains.

- **Method**: Count the number of dots in the URL. Phishing URLs often use multiple subdomains to confuse users (e.g., "login.bank.com.fake.com").

14. IP Address Usage

- **Objective**: Detect direct IP addresses in URLs.
- **Method**: URLs using raw IPs instead of domain names are flagged, as legitimate sites typically avoid this practice.

3. Technologies Used

Detecting phishing websites using machine learning algorithms typically involves using supervised learning techniques. Some common machine learning algorithms and methods used for phishing website detection include:

1. Logistic Regression: This algorithm is commonly used for binary classification tasks like phishing detection, where the output is either phishing or legitimate.
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

2. Decision Trees: Decision trees can be effective for phishing detection as they can capture complex relationships between features. Ensemble methods like Random Forests and Gradient Boosting Machines (GBM) can also be used for improved performance.
A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.

3. Support Vector Machines (SVM): SVMs are good for binary classification tasks and can be trained to distinguish between phishing and legitimate websites based on various features.
Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm, each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane. The Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then construct separating line which bisects and perpendicular to the connecting line. To classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and nonlinear data, to solve this problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

4. K-Nearest Neighbors (KNN): KNN is a simple and intuitive algorithm that can be used for phishing detection by considering the similarity between features of websites.
The k-nearest neighbors (KNN) algorithm is a machine learning algorithm that uses proximity to

classify or predict a data point's grouping. It is a supervised learning algorithm that works by comparing a data point to a set of data it was trained on.

5. Random Forest: Random Forest (RF) is a machine learning algorithm that combines the results of multiple decision trees to produce a single result.
Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on the concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of trees gives high detection accuracy. Creation of trees is based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree.
Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random Forest algorithm also uses Gini index and information gain methods to find the best splitter. This process will continue until random forest creates n number of trees. Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction.

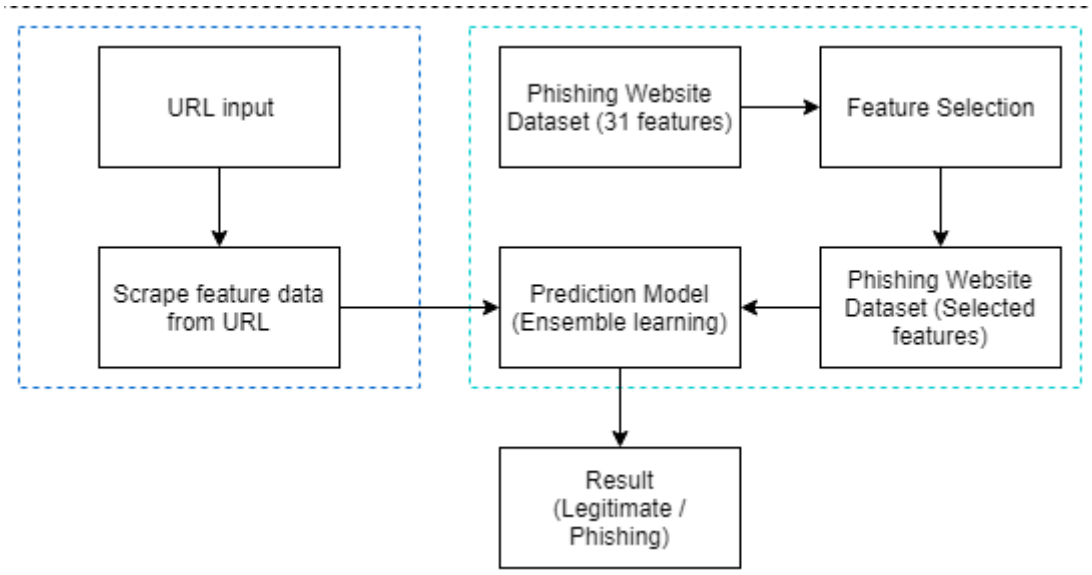


Fig.III.3.i.. Flowchart of the Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning

4. PROCEDURE

1. Data Collection:
- Collect URL data from different sources, including legitimate and phishing websites. Files like legitimate_urls.txt and 1000-phishing.txt provide the URLs used for training.

2. Feature Extraction (Class FeatureExtraction):
- **Protocol**: Extract whether the URL uses "http" or "https".
- **Domain**: Extract the domain name of the URL.
- **Path**: Extract the path from the URL.

- **Having IP**: Check if the domain contains an IP address.
- **URL Length**: Measure the length of the URL to classify it as suspicious or legitimate.
- **@ Symbol**: Check for the presence of "@" in the URL.
- **Redirection**: Detect if the URL contains a "//" after the protocol.
- **Subdomains**: Count the number of subdomains in the URL.
- **Shortening Services**: Detect URLs shortened using services like bit.ly, goo.gl.
- **Web Traffic**: Use Alexa rank to check the popularity of the domain.
- **Domain Registration Length**: Extract domain registration and compare the validity.
- **Age of Domain**: Calculate the age of the domain based on registration date.
- **DNS Records**: Check DNS information for additional classification.
- **Statistical Report**: Conduct a statistical analysis on the URL to detect anomalies.
- **HTTPS Token**: Check if the URL starts with "https://" for security validation.

3. Data Labeling:

- URLs are labeled as:
  - 0: Legitimate
  - 1: Phishing
  - 2: Suspicious

4. Model Training:

- Use RandomForestClassifier or DecisionTreeClassifier for the model to classify URLs as phishing or legitimate based on extracted features.

5. Web Interface:

- The app.py file likely serves the model using a web framework (e.g., Flask) and includes UI files like home.html, _navbar.html, and about.html for user interaction.

6. Model Saving and Loading:

- After training, save the model to a file like RandomForestModel.sav for future use.

7. Prediction:

- For new URLs, the system extracts features, processes them, and uses the trained model to classify the URL as phishing or legitimate.

5. Testing

In the **Phishing Website Detection** project, testing the model involves the following steps:

1. **Data Splitting**: The dataset is divided into training and testing sets, typically using an 80-20 or 70-30 ratio. This ensures that the model is evaluated on unseen data.

2. **Model Evaluation**: After training the classifier (e.g., RandomForest or DecisionTree), the model is tested using the testing dataset. Performance is evaluated using metrics such as:

   - Accuracy
   - Precision
   - Recall
   - F1-Score

3. **Cross-validation**: To ensure robustness, cross-validation techniques like k-fold cross-validation may be applied to assess the model's performance across multiple subsets of the data.

4. **Confusion Matrix**: A confusion matrix helps visualize the classification results, showing true positives, false positives, true negatives, and false negatives.

5. **Performance Visualization**: The performance metrics, such as accuracy and F1-score, are plotted to assess how well the model is generalizing.

6. **Model Tuning**: Hyperparameters of the classifier can be adjusted (e.g., number of trees in Random Forest) to improve the model's performance on the test set.

6. Advantages & Features of Proposed Tool

Advantages:

1. **Real-time Detection:**

   The system can analyse websites in real-time, providing instant feedback on whether a website is legitimate or a phishing attempt.

2. **High Accuracy:**

   Machine learning models can achieve high accuracy by learning from vast datasets of phishing and legitimate websites, minimizing false positives and negatives.

3. **Adaptive to New Threats:**

   Unlike traditional blacklist-based systems, the tool can detect newly created or dynamically generated phishing websites by recognizing patterns and features associated with phishing attempts.

4. **Scalability:**

   The tool can be easily scaled to analyse many websites simultaneously, making it suitable for both individual users and organizations.

5. **Low Maintenance:**

Once trained, the system can operate with minimal human intervention, only requiring periodic retraining with new data to stay updated with the latest phishing trends.

6. **Platform Independence:**

   Can be integrated into various platforms such as browsers, antivirus software, or even mobile applications, providing flexible usage across different systems.

7. **Cost-Effective:**

   Reduces the need for expensive manual oversight and cybersecurity teams dedicated to phishing detection, lowering the overall cost for businesses.

8. **Reduced Human Error:**

   Automated detection significantly reduces human errors, which can occur during manual inspection of potentially suspicious websites.

Key Features:

1. **Feature Extraction:**

   Extracts critical attributes such as domain age, URL length, HTTPS status, IP address, and content features like the presence of suspicious scripts or forms.

2. **User-Friendly Interface:**

   The tool provides a simple and intuitive interface, allowing even non-technical users to check the legitimacy of websites with ease.

3. **Continuous Learning:**

   Incorporates feedback loops to improve detection accuracy over time by learning from newly identified phishing websites.

4. **Multi-Model Support:**

   Utilizes multiple machine learning algorithms (e.g., Decision Trees, Random Forest, Support Vector Machines) to ensure the best possible performance, selecting the optimal model based on the dataset.

These advantages and features make the proposed phishing detection tool robust, accurate, and efficient in protecting users and organizations from phishing attacks.

7. System Requirements

a. Software used :-

1. Installed Python.
   Libraries needed - pandas, numpy, urlparse, urlencode, BeautifulSoup, whois, urllib.request, time, socket, HTTPError, datetime, prange, Flask, Pickle

2. Installed IDE to use Python.
   Installed Visual Studio Code - Insiders

b. Hardware used :-

1. A Laptop/Desktop
   Connected to the Internet and has a browser.
   Processor: Preferably 1.0 GHz or Greater.
   RAM: 512 MB or Greater.

2. Single Network Connection
   So that other devices can connect through network URL.

IV. RESULTS AND DISCUSSION

IV. i. RESULTS

The project focuses on detecting phishing websites using machine learning techniques, specifically evaluating the effectiveness of feature-based classification.

1. Feature Extraction
    a. A total of 14 features were extracted from URLs, including length, domain age, web traffic rank, and subdomain analysis, ensuring a comprehensive analysis of phishing behavior.
    b. Datasets of legitimate and phishing URLs were combined and preprocessed, removing inconsistencies to ensure reliable model training.
2. Model Training and Testing
    a. The dataset was split into 70% training and 30% testing subsets to maintain model integrity.
    b. The Random Forest Classifier outperformed Decision Tree and SVM algorithms in identifying patterns within the feature set.
3. Feature Importance Analysis
    a. Key features such as URL length and web traffic rank emerged as the most impactful for classification.
    b. A bar plot visually represented feature rankings, providing insights into their relative contributions to prediction accuracy.
4. Performance Metrics
    a. Comparative analysis of algorithms highlighted significant differences in false positive and false negative rates, emphasizing the need for balanced evaluation.
    b. The Random Forest Classifier demonstrated robust results, reflecting its capacity for handling feature-rich datasets.

IV. ii. DISCUSSION

1. Model Accuracy and Relevance
    a. The implemented models effectively distinguished phishing from legitimate URLs, leveraging well-defined feature extraction techniques.

    b. The study underscored the importance of specific features, like domain age and HTTPS presence, in enhancing detection reliability.

2. Scalability and Robustness

    a. The system demonstrated scalability, processing large datasets without compromising efficiency.

    b. The Random Forest model's adaptability and ability to manage overfitting were pivotal in ensuring consistent performance.

3. Challenges and Limitations

    a. The reliance on static URL features posed challenges in detecting sophisticated, dynamically generated phishing attacks.

    b. Real-world applications may require periodic updates to feature sets as phishing tactics evolve.

4. Future Enhancements

    a. Integration of dynamic behavioral analysis, such as page content scrutiny, to complement URL-based detection.

    b. Development of real-time monitoring systems to deploy the detection mechanism in active cybersecurity frameworks.

This comprehensive exploration validates the potential of machine learning in phishing detection, setting a foundation for future advancements in cybersecurity.

V. CONCLUSIONS

This project aims to enhance detection methods to detect phishing websites using machine learning technology. We get very good performance in ensemble classifiers namely Random Forest, XGBoost both on computation duration and accuracy. The main idea behind ensemble algorithms is to combine several weak learners into a stronger one. This is the primary reason ensemble-based learning is used in practice for most of the classification problems.

It is worth mentioning that there is no guarantee that the combination of multiple classifiers will always perform better than the best individual classifier in the ensemble classifiers. The results motivate future works to add more features to the dataset, which could improve the performance of these models, hence it could combine machine learning models with other phishing detection techniques like example List-Base methods to obtain better performance.

REFERENCES

[1] Marwa Abd Al Hussein Qasim, Dr. Nahla Abbas Flayh, “Phishing Website Detection Using Machine Learning: A Review” June 2023Wasit Journal of Pure sciences 2(2):270-2812(2):270-281

[2] Rishikesh Mahajan, Irfan Siddavatam ,”Phishing Website Detection using Machine Learning Algorithms” , International Journal of Computer Applications, Volume 181 - Number 23 Year of Publication: 2018

[3] S. A. Anwekar and V. Agrawal, "PHISHING WEBSITE DETECTION USING MACHINE LEARNING ALGORITHMS."

[4] S. Jain, "Phishing Websites Detection Using Machine Learning," Available at SSRN 4121102.

[5] A. Lakshmanarao, P. S. P. Rao, and M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021: IEEE, pp. 1164-1169.

[6] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Web-site Detection," IEEE Access, vol. 10, pp. 1509-1521, 2021.

[7] A. D. Kulkarni and L. L. Brown III, "Phishing websites detection using machine learn-ing," 2019.

[8] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur, "A novel machine learning ap-proach to detect phishing websites," in 2018 5th International conference on signal pro-cessing and integrated networks (SPIN), 2018: IEEE, pp. 425-430.

[9] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018: IEEE, pp. 1-5.

[10] X. Zhang, Y. Zeng, X.-B. Jin, Z.-W. Yan, and G.-G. Geng, "Boosting the phishing detection performance by semantic analysis," in 2017 IEEE international conference on big data (big data), 2017: IEEE, pp. 1063-1070.

[11] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature selection for the prediction of phishing websites," in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017: IEEE, pp. 871-876.

[12] A. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites," in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Con-ference (IEMCON), 2016: IEEE, pp. 1-6.

[13] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," EURASIP Journal on Information Security, vol. 2016, no. 1, pp. 1-11, 2016.

[14] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," in 2015 IEEE Conference on Communications and Network Security (CNS), 2015: IEEE, pp. 769-770.