PAPER NAME

0412041229

WORD COUNT

2367 Words

CHARACTER COUNT

13660 Characters

PAGE COUNT

6 Pages

FILE SIZE

389.4KB

SUBMISSION DATE

Apr 12, 2025 11:13 AM UTC

REPORT DATE

Apr 12, 2025 11:13 AM UTC

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 8% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 10% Submitted Works database

Detection of Phishing websites using Machine Learning Algorithm

D. Sowjanya, K.Sravani ,S.Naveen Sai, S.Eepsitha, T.Sai Avinash

Assistant Professor, Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Sangivalasa, Visakhapatnam, Andhra Pradesh, India.

Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Sangivalasa, Visakhapatnam, Andhra Pradesh,

***Abstract*— This project solves the urgency of today's world that phishing websites and avoiding them in an effort to protect us.Phishing is a common threat which compromises digital security. In response to the ever-changing nature of these attacks as well as the limitations of traditional rule-based solutions, we are resorting to machine learning (ML). We're employing a combination of the features from the extensive dataset that includes typical phishing indicators. Our strategy focuses on training and fine-tuning ML models to remain agile and active. Through diligent testing and tuning of our models, we're demonstrating how efficient our strategy is at rapidly identifying potential phishing sites with high success rates. Our research contributes immensely to the area of cybersecurity by offering an active defence system that can adjust to changing phishing strategies. By concentrating on the implementation of Machine Learning techniques, we attain a strong system that helps prevent fraud. The results of these methods indicate the strength of Machine Learning methods in securing the internet and allowing users and organizations to protect against risks brought about by phishing attacks. This study not only presents the possibility of ML in addressing phishing but also highlights the need for innovations in cybersecurity initiatives. The results point to the future possibilities of improvement and innovation in threat detection, ultimately leading to a more secure digital space for users globally. Overall, our project underscores the central role of machine learning in strengthening defences against phishing attacks, opening the door to improved cybersecurity practices.**

Keywords—Phishing websites, Gradient Boosting, Cybersecurity, machine learning, and website categorization.

I.INTRODUCTION

Criminals who are looking for sensitive user information. build some illegal copies of real legitimate websites and email addresses. When users click on a link that is created by the aforementioned hackers, those hackers will be able to access all of the user's personal information, such as images, bank account information, and website login credentials. While this type of threat rates high on the list of phishing threats, this project intends to implement Machine Learning (ML) as a means of detection and prevention of fraudulent sites. Using these data-based formats and Machine Learning algorithms, the purpose is to develop a system that can potentially detect phishing websites. Utilizing ML methods and data analysis combinations, this type of project will help to make user experiences safer online. As cyberattacks rise, there is a need to ensure detection and prevention measures are in place.This research will address such challenges by designing an innovative solution that integrates ML and data analysis to combat phishing attacks. By examining patterns and anomalies in web page URLs, our system seeks to identify those sites that are keeping the users safe. Through ongoing learning and adaptation, we seek to remain one step ahead of cybercriminals and keep users from becoming victims of phishing scams. Our work also takes into account the ethical considerations of our research, making sure that our approach prioritizes user privacy and security. By working towards a more secure online environment, we hope to help users navigate the digital world with confidence.

II.LITERATURE REVIEW

Andrew Jones et al. (2013) [1] This review examines the literature on phishing detection research. Phishing attacks exploit the weaknesses in the system. Typically, many cyberattacks exploit end users' flaws, making them the security system's weakest link. Since there's no universal approach to resolve all vulnerabilities, typically large numbers of techniques are needed for each specific attack. The article is meant to be a survey of recent phishing mitigative research, but also to provide a brief summary of the different high-level categories of techniques, such as detection, offensive defence, correction, and prevention to show an overview of the bigger picture of phishing detection within the mitigative process.

Alessandro Acquisti et al. (2017) [2] As technology advances,the number of complex decisions that need to be taken by users regarding their privacy and security also increases. Research has shown us how people make choices about privacy and information security, the obstacles they encounter during this process, and what strategies are needed to overcome themThis article provides a transparent review of the literature on security and private decision-making, which also addresses, consistent

with the research, the use of soft interventions to steer users toward more positive decision-making. The article outlines the benefits soft interventions, fully acknowledged, have limitations in terms of ethical thoughtfulness, design limitations and research challenge.

F.J. Overink et al. (2017) [3] People are quick to divulge personal information because they frequently trust one another, which makes them susceptible to social engineering scams. The study looked at two ways to protect people from social engineering attacks. Two of the methods involved using cues to get them to be more cautious about the threats. The second was to warn them not to give out personal details. The research asked individuals in a Dutch shopping precinct for their email address, the beginning of their bank account number, and information on their internet buying habits. It discovered that individuals were willing to provide this information. But neither the cues nor the warnings had an impact on how much people shared. Actually, the warning might have had a negative effect.

M. El-Alfy et al. (2017) [4[] Information security research on anti-phishing solutions has grown in significance. due to the devastating effects and rising frequency of phishing attacks. Information disclosure, identity theft, financial loss, and reputational manipulation are examples of security risks. Raising human awareness alone is insufficient as a countermeasure; instead, complementary technical countermeasures must be used. Despite the numerous methods that have been put forth in the literature, developing successful phishing models is a difficult undertaking, and there is currently no perfect solution to the issue. In this paper, a novel probabilistic neural network (PNN)-based method for phishing website detection , we also look into integrating PNN with K-medoid clustering. In order to determine whether the suggested strategy is feasible, we carried out.

III.METHODOLOGY

We have designed our project using a website as a medium for the target population. The website is interactive and responsive, and can be used to examine if a webpage domain is legitimate or if it is a phishing domain. This website was created using various web designing languages that include HTML, CSS, JavaScript and Flask that is a Python framework. HTML was used to create the overall structure of the website. CSS was included to enhance the website with effects to make it aesthetically pleasing and easier for the user. As this website was developed for all users, it is vital to include easy-to-use functions so users can use the website without complications.

CURRENT SYSTEM

Phishing pages usually maintain the same CSS style as their target pages, according to frameworks proposed by H. Huang et al. (2009) [5] that differentiate phishing by using page section similitude that deconstructs universal resource locator tokens to create forecast precision.

S. Marchal et al., (2017) [6] suggested This method uses real-site server log data analysis to identify phishing websites. detection of phishing websites or an off-the-shelf application. High accuracy, total independence, good language independence, selection speed, adaptability to dynamic phishing, and the ability to progress in phishing techniques are just a few of its advantageous features. It is also free.By extracting URL features from websites and using subset-based feature selection techniques, Mustafa Aydin et al. proposed a classification algorithm to identify phishing websites. This uses techniques for feature extraction and selection to identify phishing websites. Five distinct analyses—Alphabetic Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis, and Rank Based Analysis—are applied to the extracted features pertaining to the pages' URLs and the feature matrix that is produced. The majority of features are connected.

In the previous system, the machine learning methods used for comparison were the same three machine learning methods outlined in the current system: Logistic Regression, Multinomial Naive Bayes, and Gradient boost. It was shown that the Logistic Regression had superior performance over the other two. The proposed system preprocesses the same Logistic Regression model, applies the tokenization to the features, applies stemming, and the same data classification is reported. Data Processing is the act of either converting or encoding data for machine transferability. The Logistic Regression model has an accuracy of 96.63 percent, and an overall comparisons are shown.

NEGATIVE ASPECTS OF CURRENT SYSTEM

i. The current models are characterized by low latency.
ii. Current systems lack an explicit user interface.
iii. Current models cannot predict a continuous outcome.
iv. Only when the dependent or outcome variable is dichotomous does the model function.
v. The model may yield inaccurate results if the sample size is small.vi. The model may be overfit due to an existing system.

PROPOSED SYSTEM

Proposed Framework

Bagging Classifiers.

Technologies used

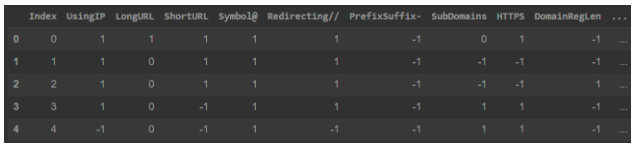
Flask, Machine Learning, Web Development.

I. Introduction

Phishing, a prevalent cyber threat, poses significant risks to online security and user privacy. Detecting and mitigating phishing attacks are imperative to safeguard users' sensitive information and digital assets. Conventional phishing detection techniques do not adequately address the changing methods of cybercriminals. Machine learning algorithms provide opportunities for improving phishing detection techniques by detecting patterns in URL features to separate benign from phishing URLs. In this paper, we present a systematic procedure for identifying phishing websites using machine learning models, including the elements of data collection, data preprocessing, algorithm execution, model evaluation, and analyzing the results.

II. Data Collection and Preparation

The study's first step involves obtaining data from Kaggle, which includes a dataset that lists 36 URL attributes (figure 1). Generative AI techniques are used to increase the size and diversity of datasets. To ensure compatibility with machine learning algorithms, data preprocessing includes standardization, normalization, and categorical variable encoding.



	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...
0	0	1	1	1	1	1	-1	0	1	-1	...
1	1	1	0	1	1	1	-1	-1	-1	-1	...
2	2	1	0	1	1	1	-1	-1	-1	1	...
3	3	1	0	-1	1	1	-1	1	1	-1	...
4	4	-1	0	-1	1	-1	-1	1	1	-1	...

Figure 1: Dataset used for analysis

III. Exploratory Data Analysis (EDA)

is performed to gain an understanding of the characteristics of a dataset and relationships between features (shown in figure 2). Feature importance maps and correlation matrices are utilized to discern influential attributes and inter-feature dependencies, aiding in feature selection.

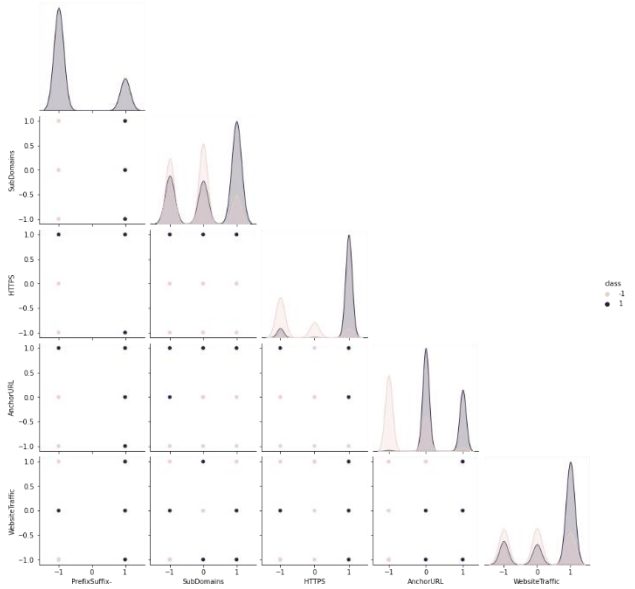


Figure 2: Pair-plot between selected features ('PrefixSuffix-', 'SubDomains', HTTPS', 'AnchorURL','WebsiteTraffic','class')

IV. Algorithm Implementation

Several models such as K-Nearest Neighbors (KNN), Logistic Regression and Gradient Boosting Classifier are implemented. Hyperparameters are fine-tuned using grid search and optimization techniques to enhance model performance.

V. Model Evaluation

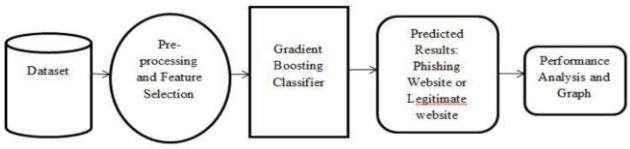
To assess the model, the dataset is divided into training and testing sets. The Accuracy Score, R2 Score, Confusion Matrix, and F1 Score are used to assess the model's performance.

The project includes the development of an interactive and responsive website designed to facilitate the identification of legitimate or phishing websites. Employing a combination of web designing languages such as HTML, CSS, JavaScript, and the Flask framework in Python, the website offers a user-friendly interface accessible to all users. HTML forms the foundational structure of the website, while CSS enhances its visual appeal and usability through the addition of effects. Notably, the website prioritizes ease of operation to ensure a seamless user experience, with meticulous attention to detail aimed at minimizing any potential user difficulties. This platform serves as a pivotal tool in extending the benefits of phishing detection to a broad user base, thereby contributing to the enhancement of online safety and security measures.

The dataset with different features is used to train the system we propose, and it should be noted that the dataset does not contain any website URLs. The dataset contains every feature that must be taken into account, including whether or not the website URL is authentic and whether or not it is a phishing website. The Gradient Boosting Classifier is used in the suggested model. Once the model has been trained using the dataset, the classifier classifies the supplied URL to the trained data. It alerts the user that the URL is a phishing site if it is; legitimate, it notifies the user that it is a safe website. Using this dataset, we would like to categorize phishing websites using an appropriate algorithm.

BENEFITS OF PROPOSED SYSTEM

1. It has a user interface
2. The model based on many features.
3. It is accurate.
4. The proposed system is generally more accurate than other nodes
5. With larger datasets , the proposed ystsem can be trained more quickly.
6. Typically, the suggested systems support categorical features.



HARDWARE SPECIFICATIONS

System: Intel Pentium i3 Processor.
Storage: 500 GB Hard Disk.
Display: 15-inch LED screen.
Keyboard and mouse are input devices.
RAM: 4 GB.

SPECIFICATIONS OF THE SOFTWARE

Windows 10 is the operating system.
Python is the writing language.
Flask is the framework.

IV.RESULTS AND ANALYSIS

Results indicate the Gradient Boosting Classifier as the most effective model, achieving 98% accuracy on the testing set and 97% on the training set. Feature importance maps and correlation matrices facilitate the selection of optimal features for the final model.

Currently with a test dataset we're testing all 3 machine learning algorithms currently we finished K Means with Training accuracy: 85% Testing accuracy: 81% Expected Results We want the model to be highly efficient so we set our goals at at least Training accuracy: 95% Testing accuracy: 95% Sample Results Evaluation Metrics to measure your algorithms' accuracy Currently we are using Confusion matrix R2 score and Accuracy score to know the model's metrics (shown in figure 3).

Training and Testing plot between learning rate and accuracy and Training and Testing plot between depth and accuracy is also shown in figures.

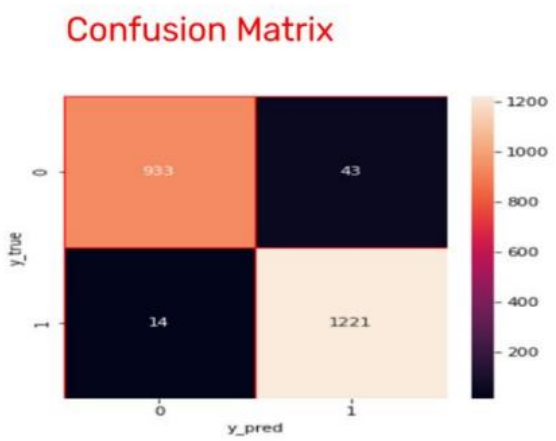


Figure 3: Confusion matrix

Figure 4: Plotting learning rate against accuracy for training and testing data

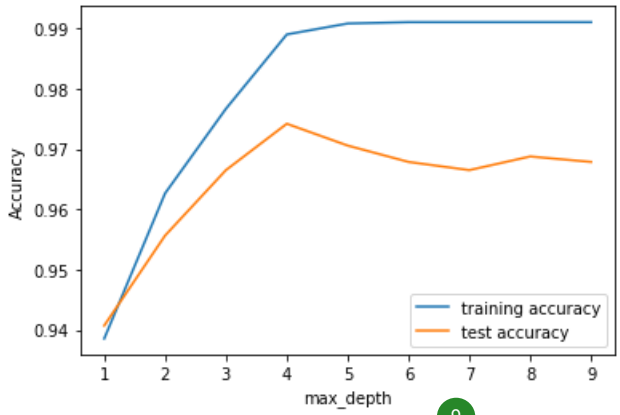


Figure 5: Plotting maximum depth against accuracy for training and testing data

V.CONCLUSION

In conclusion, the implementation of algorithms has demonstrated promising potential in fortifying our defenses against the pervasive threat of phishing attacks. Through the development and refinement of sophisticated models, this project has showcased the viability of leveraging data-driven approaches to swiftly identify and neutralize fraudulent websites. Considering the never-ending evolution of cyber threats in the world today, we must also note that the findings here emphasize the need for continual innovation and adaptation in cybersecurity. The research done here will significantly contribute to ongoing efforts to protect people and organizations online and could serve as a launching pad for more proactive threat detection in the future.

12% Overall Similarity

Top sources found in the following databases:

- 8% Internet database
- Crossref database
- 10% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	**uwe on 2023-09-24** Submitted works	3%
2	**shanlaxjournals.in** Internet	1%
3	**University of Sunderland on 2024-08-07** Submitted works	1%
4	**jetir.org** Internet	1%
5	**R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sek...** Publication	<1%
6	**Universiti Putra Malaysia on 2019-06-14** Submitted works	<1%
7	**issuu.com** Internet	<1%
8	**ijraset.com** Internet	<1%

9	Engineering Computations, Volume 29, Issue 8 (2012-11-03) Publication	<1%
10	University of Greenwich on 2023-04-12 Submitted works	<1%
11	termpaperwarehouse.com Internet	<1%
12	"Proceedings of International Conference on Recent Trends in Computi... Crossref	<1%
13	uwe on 2024-05-15 Submitted works	<1%