IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE IDENTIFICATION FOR URL

Guide: Mrs. Nanda MB

Student :Lavanya K ,Vaishnavi M

ABSTRACT

Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Although phishing was first used in 1996, it hasbecome the most severe and dangerous cybercrime on the internet. Phishing utilizes email distortion as its underlying mechanism for tricky correspondences, followed by mock sites, to obtain the required data from people in question. Differentstudies have presentedtheir work on the precaution, identification, and knowledge of phishing attacks; however, there is currently no complete and proper solution for frustrating them. The proposed project is based on the phishing URL-based dataset extracted from the famous dataset repository, Which contains phishing and legal URL features collected from 11000+ website datasets in vector form. After preprocessing, several machine learning algorithms are implemented and designed to block phishingURLs and provide security to the user. This project uses machine learning models like Decision Tree (DT), Linear Regression (LR), Random Forest (RF), Naive Bayes(NB), Gradiant Boosting Classifier (GBM), K-Neighbors Classifier (KNN), Support Vector Classifier. Does. (SVC), and the proposed hybrid LSD model, which is a combination of logistic regression, support vector machine and decision tree (LR + SVC + DT) with soft and hard voting to protect against phishing attacks with high accuracy and efficiency.

Introduction of theproject

1. Introduction

Phishing is a cyber attack that lures a victim by using technical luring to collect personal or confidential information. It started with the American online attack in 1995. Later, phishers individuals or teams conducting phishing attacks—moved to more profitable targets, such as online banking and e-commerce services. Financial gain is the primarymotivating factor for phishers; However, fame and notoriety are also interesting psychological aspects of phishing.Phishing is an Internet threat and occupies a top position in the cyber threat

landscape. It has become a major cyber threat to the financialsectors and has spread to many sectors. Recent Data shows that the number of phishing attacks in 2020 has doubled compared to previous years. Approximately 84% of phishing sites recorded in late 2020 used his SSL protection.This indicates that HTTPS is not currently an important feature in detecting phishing attack In fact, the phishing attack half-life lasted less than a day, and new phishing attacks emerged. This trend is also emerging rapidly due to the dynamic nature of phishing attacks. fishing date is proposed as a representation learning-based phishing detection solution using Website URL and HTML content. It combines two deep networks i.e. the Long-term Recurrent Convolutional Network (LRCN) and the Graph ConvolutionalNetwork . It performed well during the experiments and achieved 96.42% detection accuracy with a real-worldpublic dataset.

1 Phishing Detection:- A phishing detection solution that automatically selects Features from a website's raw URL and HTML content.

2 The first GNN based phishing detection approach that uses raw HTML content.

2. Objectives

- To improve the accuracy in phishing website detection using machine learning.
- To implement the real time phishing website detection.
- To build an automated phishing website detection.
- To combine URL statistical features, webpage code features, webpage text features to improve the accuracy in phishing website detection.

3. Scope

The Internet provides a great opportunity for attackers to engage in criminal activities such as online fraud, malicious software, computer viruses, ransomware, worms, intellectual property rights, denial of service attacks, money laundering, vandalism, electronic terrorism, and extortion .Hacking is a major destroyer of the Internet throughwhich any person can hack computer information and use it in different ways to harm others. Immorality, which harms moral values, is a major issue for the youngergenerations. Detecting these websites rather than websites that appear simple and secure, will help people. Therefore, an awareness of these websites is necessary.Virusescan damage an entire computer network and confidential information by spreading to multiple computers. It is not suitable to use unauthorized websites on the internet. Phishing detection is required for all of these aspects to secure our computer system sowe have motivatedto choose this topic.

1. Automatic detection of phishing website.
2. Comparing to current technology this project saves time of phishing websitedetection.
3. Comparing to current mechanism this project gives high accuracy in phishing website detection.

4. Applications

- **User Protection:**
  Preventing identity theft and unauthorized access.
- **Financial Security:**
  Safeguarding online banking and transactions.
- **Corporate Security:**
  Protecting businesses from data breaches.
- **Email Security:**
  Blocking malicious links in emails.
- **Brand Protection:**
  Safeguarding brand reputation.
- **Data Privacy:**
  Ensuring private information isn't compromised.
- **Regulatory Compliances:**
  Meeting data protection regulations.
- **Education:**
  Raising awareness and promoting vigilance.
- **Law Enforcement:**
  Identifying andtaking down malicious sites.

Chapter2

Literature survey

OVERVIEW

A literature survey or a literature review in a project report shows the various analyzes and research conducted in the area of interest and the results already published, keepingin mind the various parameters of the project and the scope of the project. literature surveyis Mainly done to analyze the background of the current project which helps in findingout the flaws in the existing system and guides what unsolved problems we can work on. Therefore, the following topics not only reflect the background of the project but also highlight the problems and shortcomings that led to proposing solutions and working onthis project. A literature review is the text of an academic paper that contains current knowledge, including actual research results. The literature review uses secondary sources and does not report new or original experimental studies regularly, those related to academically oriented literature such as theses; For dissertations and peer-reviewed journal articles, a literature review usually precedes the methodology and results. However, this is not always the case. Literature reviews are also common in research proposals and prospectuses (documents approved in advance of a research proposal). Itsmain aim is to place current research within the following framework:

Literature reviews are the basis of research in virtually any field. Academic field.

A literature survey includes the following:

- Existing theories about a subject that are universally accepted.
- Books written on the subject, both general and specific.
- Research done in the field is usually arranged from oldest to newest.
- Challenges faced and ongoing work, if available.

Literature survey describes the existing work on a given project.

It relates to problem associated with the existing system and also gives the user a clearknowledge of how to deal with the existing problems and how to provide solutions forthe existing problems.

Objectives of Literature Survey

- Learning definitions of concepts.
- Access to the latest approaches, methods and theories.
- Searching for research topics based on existing research.
- Focus on your area of expertise – Even if another field uses similar terms, theyusually mean completely different things.
- It improves the quality of the literature survey to exclude sidetracks.

1. Survey on:

1.Title: Machine Learning Techniques for detection of website phishing: A review for promises and challenge
Author :A. Odeh, I. Keshta, and E.Abdelfattah ,Year:2021

Abstract: Website phishing is a cyber attack that targets online users and steals sensitiveinformation. Information such as login credentials and bank account details. Attackers trickusers by Display masked web pages as legitimate or trustworthy to get the content you want.

data. Many solutions to phishing website attacks have been proposed.Blacklists or whitelists and machine learning (ML)-based techniques. This paper presents a state-of-the-art technique for detecting phishing websites using ML techniques. this research Identifysolutions to website phishing problems based on ML techniques. Largely The tested approach focuses on traditional ML techniques. Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Ada Boosting are powerful ML techniques It hasbeen investigated in the literature. The research paper also identifies deep learning-basedtechniques. Improves phishing website detection performance compared to traditional MLTechnique Challenges with ML techniques identified in this study include inefficienciessuch as: ML techniques for overfitting, low accuracy, and insufficient training available data. This study suggests that Internet users should be wary of phishing to avoid cyber-attacks. This paper also focuses on proposing phishing automation solutions.

Methodologies: Random Forest (RF), Support Vector Machine (SVM), Naive Bayes(NB), and Ada Boosting

Advantages:

A comprehensive review of phishing detection techniques, highlighting the effectivenessof ML and deep learning.

Disadvantages:

Challenges like overfitting and dependence on sufficient training data may limit the efficiency o ML techniques, and the papers focus on user awareness may overlook socio-technical aspects of phishing

prevention.

2. Title: A Survey of URL-Based Phishing Detection Author: E. S. Aung, C. T. Zan, and H.Yamana Year: 2022

Abstract: Cyber phishing is considered to be of the form theft of personal information in which phishers, also known as attackers, lure users into surrendering sensitive data such ascredentials, credit card and bank account information, financial details and other behavioraldata. Are. Phishing detection is becoming an important research area, receiving more attention as the number of phishing attacks increases. Furthermore, This is considered a type of because the attacker has invented something different. Advances in technology havemade detection a top concern for developers. There are several phishing detection schemes built into their architecture, such as whitelist-, blacklist-, content-, visual similarity, and ingeneral URL-based. Each has its individual advantages and drawbacks. In this survey paper, we emphasize on URL-based phishing detection techniques, because we consider the URL to be a significant criterium in preventing phishing attacks. Moreover, examining URL- based features can also encouragefaster processing than other approaches.In this work, we aim to understand the structure of URL-based features and surveying their diverse detection techniques and mechanisms. We then analyzethe performance based on the combinations of URL features on different datasets. Finally, wesummarize our findings to promote better URL-based phishing detection systems.

Methodologies: Naïve Bayes, Support Vector Machine, Random Forest, Convolutional Neural Network.

Advantages:

URL-based phishing detection allows for preventing attacks.

Disadvantages:

The diversity of phishing techniques and constant innovation by attackersposechallenges in developingeffective URL-based detection systems

3. Title: A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites

Author: M. D. Bhagwat; P. H. Patil; T. S. Vishwanath,Year:2021

Abstract: Now a days detecting and finding some phishing websites in real time is really a
Dynamic and nuanced topic involving many variables and requirements. fuzzy logic Strategies can be an important way to detect and test phishing websites Due to the ambiguities involved in detection. Instead of precise principles, fuzzy logic provides a more intuitive way of dealing with quality variables. An approach to ambiguity resolution and intelligent phishing website detection model will be proposed in phishingwebsite assessment. This approach is based on intuitive reasoning and machine learningalgorithms that define Various factors on phishing website. A total of 30 characteristicsor features of a phishing website The features can be used to detect phishing with highaccuracy. is a real-time phishing dataset The one used is downloaded from the UCI Machine Learning Repository.

Methodologies: Fuzzy logic

Advantages:

Fuzzy logic-based approach offers an intuitive method for detecting phishing websites.

Disadvantages:

Dependency on machine learning algorithm may introduce complexity.

4.Title: Detecting Phishing Websites Through Domain and Content Analysis Author: Cristian Pascariu; Ioan C. Bacivarov,Year:2021

Abstract: Many security incidents within the digital space focus on deceiving human usersthrough the use of phishing websites that are built to capture and steal credentials. Phishingwebsites are designed to look similar to legitimate websites. In this paper the authors focus ondeveloping a new solution for detecting specific phishing websites based on domain & content analysis. A phishing website is detected when it's domain is similar to a legitimate service or includes the legitimate service name as a subdomain. Content analysis is performed to detect if the suspicious website contains elements of an authentication form used to steal credentials

Methodologies: Text matchingAdvantages:

It classifies the phishing websites efficiently.

Disadvantages:

In this paper they are considering only the URL attributes to classify the phishing website.This is not considering web HTML contents.

5. Title: A Lightweight Phishing Website Detection Algorithmby MachineLearning Author: Chenyu Gu,Year: 2021

Abstract: With the rapid development of the Internet, phishing websites now show thecharacteristics of short life cycle and low construction cost, which leads to a large amount of databrought by the detection of phishing websites for URL (uniform resource locator). It will also lead to increased retrieval time and decreased detection speed. In order to deal

with diverse, complex and hidden phishing websites, this paper proposes a lightweight framework for detecting phishing websites. We first choose the faster Minhash signature tomatch URLs. On one hand, similarity detection is employed if the websites is suspicious. On the other hand, based on machine learning, the phishing websites can be finally determined by intention detection without similar sites.

Methodologies: faster minhash signature Advantages:

It classifies the phishing websites efficiently.

Disadvantages:

Dependence on Minhash signature for URL matching may compromise accuracy in detecting diversephishing websites ,especially in cases where the phishing sites lack similarity with known instances.

Chapter3

SYSTEM REQUIREMENTANDSPECIFICATIONS

1. Introduction to requirement and specification

The System Requirements Specification (SRS) is a central report that outlines the establishment of Product development process. It provides a framework as well as records the requirements of HASADepicting its important attractions. SRS is essentially an association supervised (in composition)

at a specific time regarding the frame work requirements and conditions of a customer or potentialcustomer (usually) before any actual configuration or repair work. This is two-way security approach that guarantees that both the client and the association understand the options

Requirements from that point of view at a given time.

SRS talks about the object, but not about the enterprise that built it, so SRS works

A basis for subsequent improvement of the completed item. However, the SRS may need to bereplaced

Creation gives an establishment to proceed with the evaluation. Simply put, programming Requirements determination is the initial stage of product improvement action.

SRS means understanding the thoughts in the minds of customers - information,

In a formal collection – the product of a pre-requisite stage. Then there is the production of the platform state of formally determined needs, which are ideally finished and stable, while

The data has none of these properties.

Functional Requirements

1. Create a desktopapplication using python Tkinter framework.
2. User should load the website dataset.
3. System will extract the features from the dataset.
4. User should enter the URL as input
5. System will apply majority voting algorithm to train and detect the phishing URLs.
6. Application should accurately detects phishing URL from input automatically.

3. 3 Non Functional Requirements

These are requirements that are not functional in nature, i.e. they are limitations within which

The system must work.

The program must be self-contained so that it can be easily moved from one computer to another. it is
It is assumed that a network connection will be available on the computer on which the program is located.
Capacity, scalability and availability.
The system will achieve 100 percent availability at all times.

- The system will be scalable to support additional clients and volunteers.

- **Maintenance.**
  As far as possible, the system should be optimized for maintenance, or ease of maintenance. It Coding standards, naming conventions, use of class libraries can be achieved through documentationAnd abstraction.
  Randomness, Verifiability and Load Balancing.
- As far as possible, the system should be optimized for maintenance, or ease of maintenance. It maybe possible
- Achieved through documentation using coding standards, naming conventions, class libraries, etc.
- It should have randomness to check nodes and load should be balanced.

3.3 HARDWARE REQUIREMENTS

Processor (CPU):
Modern multi-core processor (e.g., Intel Core i7, AMD Ryzen 7).

Graphics Processing Unit (GPU):
Powerful NVIDIAGPU for accelerated deep learning tasks (e.g., GeForce RTX, Quadro series).

Random Access Memory (RAM):
Minimum 16 GB RAM (32 GB or more for larger projects).

Storage:
Solid State Drive (SSD) for fast data handling.

SOFTWARE REQUIREMENTS

- **Operating System:**
  Windows 64-bit
- **Technology**
  Python
- **IDE**
  PythonIDLE
- **Tools**
  Anaconda
- **Python Version**
  Python 3.6
- **GUI (Graphical User Interface)**

Tkinter framework.

Chapter4

Methodology

3.1 Existing System

In the past two decades, academia and the industry researched better detection approaches to combat phishing attacks, but it seemed challenging due to the nature of phishing attacks however,there are now differentiated solution that protects the privacy of Internet users against phishing. These solutions could be clustered into several approaches, and these approaches could be categorized into machine learning and non- machine

learning, as shown in above Table.

Category	Approach	Limitations / Remarks
Machine Learning	Supervised Learning • A model trains from known phishing and legitimate data	• It depends on a set of features (i.e. URL features) • A learning algorithm uses to adjust the weights of these features to achieve optimum performance
	Reinforcement Learning • An agent is used to gather its experience in web surfing for sequential decision making	• Agent produces an action (i.e. access or block a website) from a set of website features • The correctness of an agent's action is measured to have an effective learning process
Non-machine Learning	User Awareness • This technique trains Internet users to access the Internet services safer to protect them from phishing attacks	• A machine-centric approach • Game-based education has been found as an effective method when improving the user-awareness • Expecting users to get educated about technological things like phishing is not practical
	Blacklisting & Whitelisting • Blacklisting contains a list of phishing website URLs • Whitelisting is a list of legitimate website URLs	• It requires exact matching of the website URLs • It fails when detecting zero-day attacks since those may not include in the lists • Practical difficulties exist to have an up-to-date list
	Rule-based Heuristics • A technique that uses a set of rules when detecting phishing attacks	• Domain expertise is essential to constructing high-end rules • The rules need frequent updates to keep alive • The cost of updating rules is high
	Visual Similarity • This technique uses the visual appearance of the web page in phishing detection	• It depends on a threshold value, and difficult to find the optimum value • It uses visual features such as text, HTML tags, CSS and images • Maintaining an up-to-date database is challenging; therefore, it fails to detect zero-day attacks • The detection time is relatively high
	Data Mining • A technique that extracts data to discover hidden phishing patterns in a given dataset to implement predictive models	• It is not categorised under machine learning since the model does not learn over time • It focuses on finding new and interesting phishing patterns without getting a specific goal from the domain • The best features selection is a challenging task • The findings are used with rule-based heuristics to enhance the classification accuracy

Limitations:

- Most of the machine learning based existing system is implemented only using URL features.
- Most of the non machine learning based existing system is based on blocklist and non blocklist URLs are manually updated by administrator, So It is not accurate and time consuming process.
- Most of existing system is based on specific rule based so it is not suitable for new patterns of phishing attacks.

2. Proposed System

In the proposed study, ML algorithms were used with the features of the URL to solveclassificationproblems. Effective features for training purposes were selected based on aneffective phishing detection mechanism. The proposed system presents a differentiated phishing detection approach called Phishing Detection that exercises representation learning Techniques.

It combines two separate models named URL Detection and HTML Det that were modelled using ML techniques to process URLs and HTML contents. Phishing Detection performs well at present. However, the solution should be retained occasionally to be more effective in futureattacks, and the retaining process may not be costly due to the advantages of the representation learning technique. The pop-up message is activated through user interactions, ensuring a non-intrusive yet effective means of communication. This feature is designed to provide users withadditional insights, guidance, or important updates regarding the system's

functionality. By incorporating this pop-up message feature, we aim to enhance the overall user engagement and communication aspect of the Phishing Detection System, contributing to a more informed and secure online experience.

Advantages

- Testing on a dataset containing millions of phishing URLs and legitimate URLs,the accuracy reaches 98.99%, and the false positive rate is only 0.59%.
- It will support large amount of data.
- It will analysis the phishing website based URLand webpage contents.
- It takes less computational time.

System Architecture:

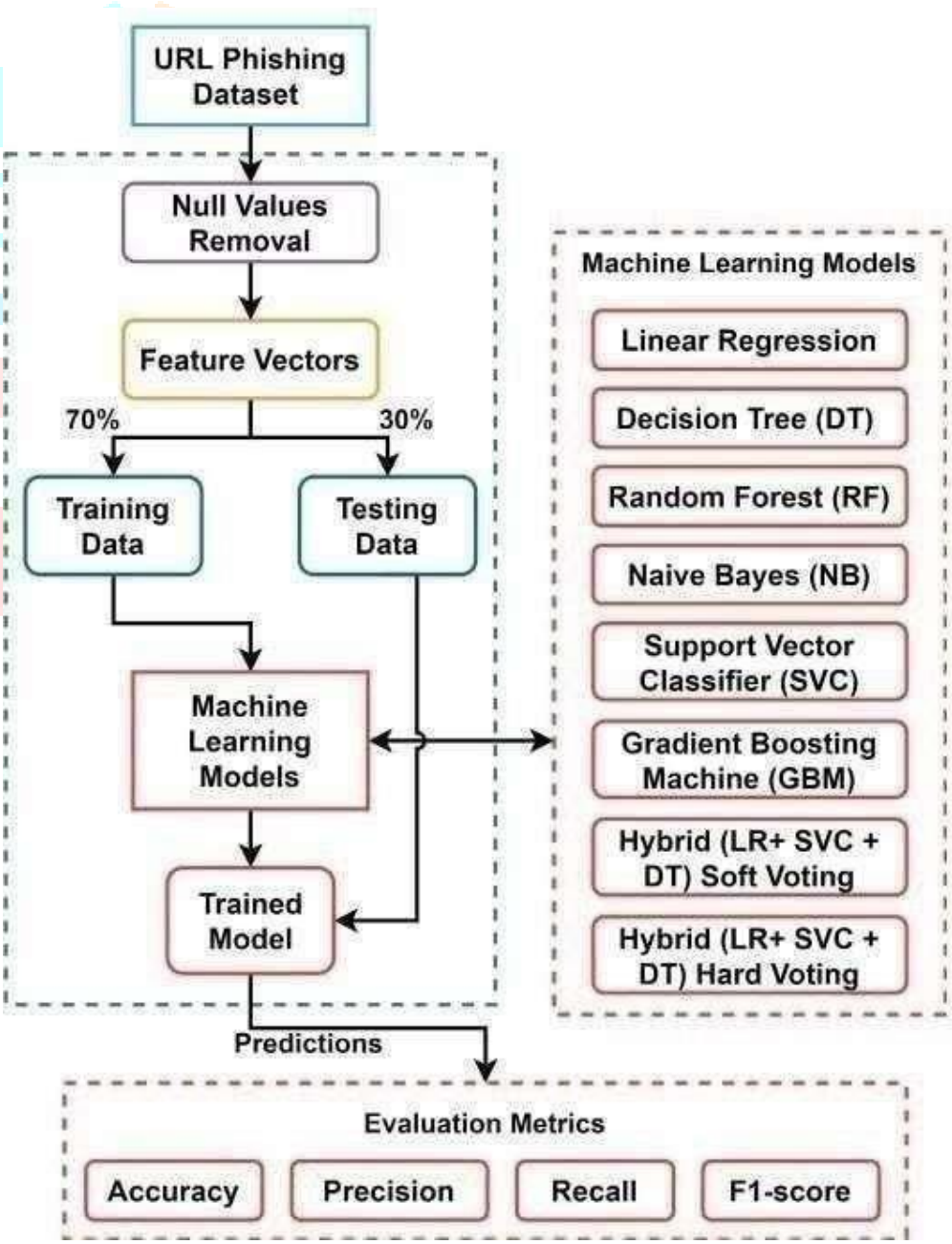


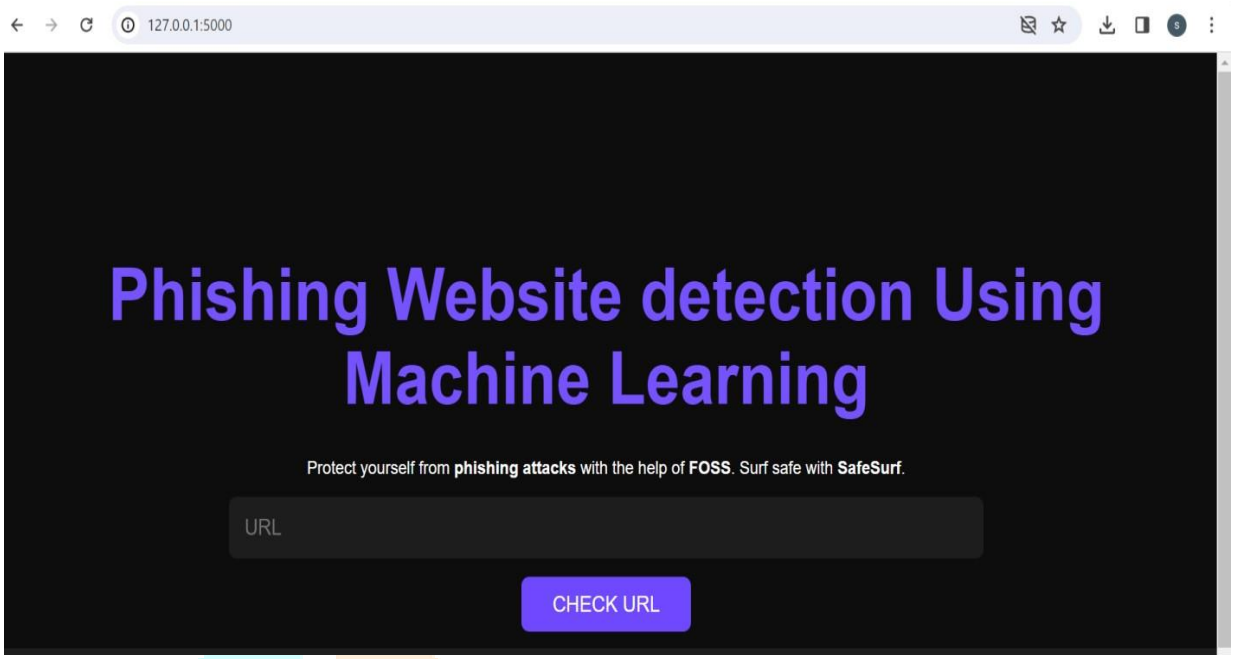
Fig 4.1 System Architecture

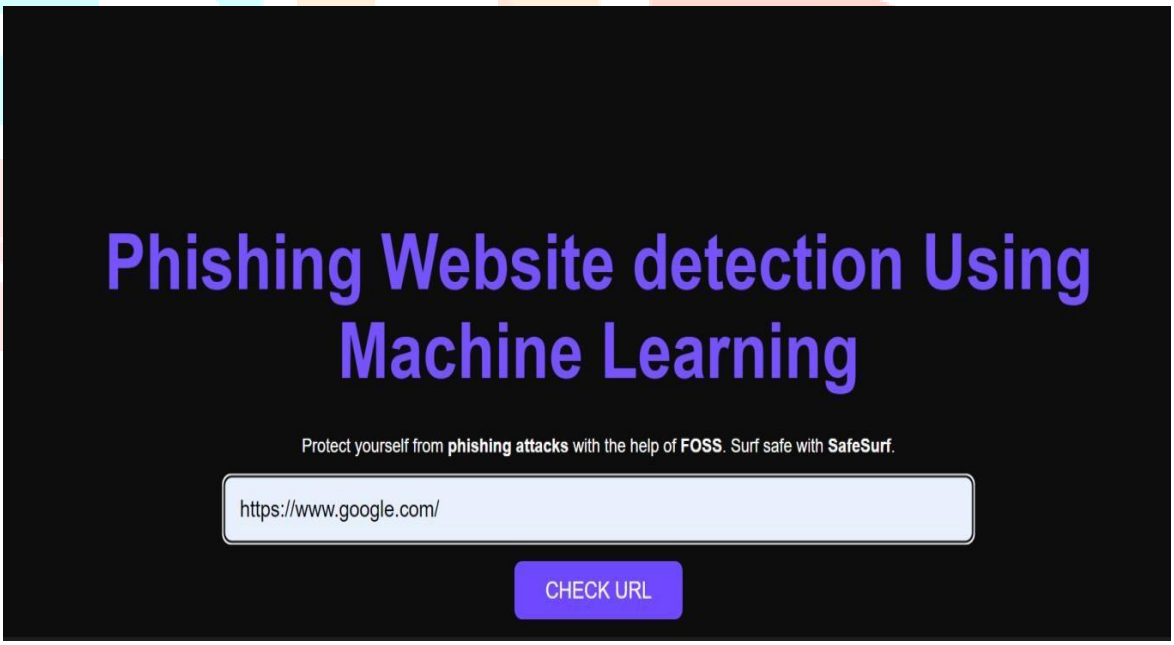
Phishing detection based on URLs proposed in this study. The classification of phishing URLswas implemented using machine learning algorithms. Cybercrimes are growing with the growth of Internet architecture worldwide, which needs to provide a security mechanism to prevent an attackerfrom getting confidential content by breaching the network through fake and malicious URLs. A phishing dataset was used to perform the experiments. The dataset is in the form of data vectors thatrequire null-value removal to remove unnecessary empty values. Multiple ML algorithms, such asdecision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB),gradient boosting machine (GBM), support vector classifier (SVC), K-neighbors classifier, andthe proposed hybrid model (LR+SVC+DT) LSD with soft and hard voting were used based on functional features .in fig 4.1

To improve the prediction results, a cross validation technique with grid search hyper- parameter tuning based on canopy feature selection was designed using the proposed LSD hybridmodel. Finally, predictions were made to classify the phishing URLs and evaluate their performance in terms of accuracy, precision, recall, specificity, and F1- score.

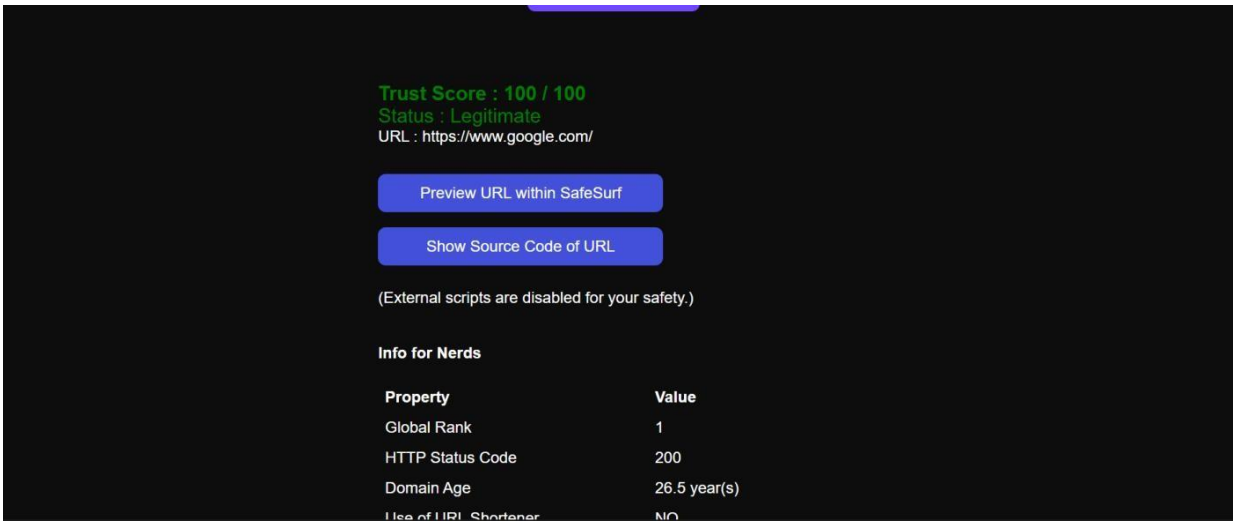
Snapshot

CHAPTER 5RESULT

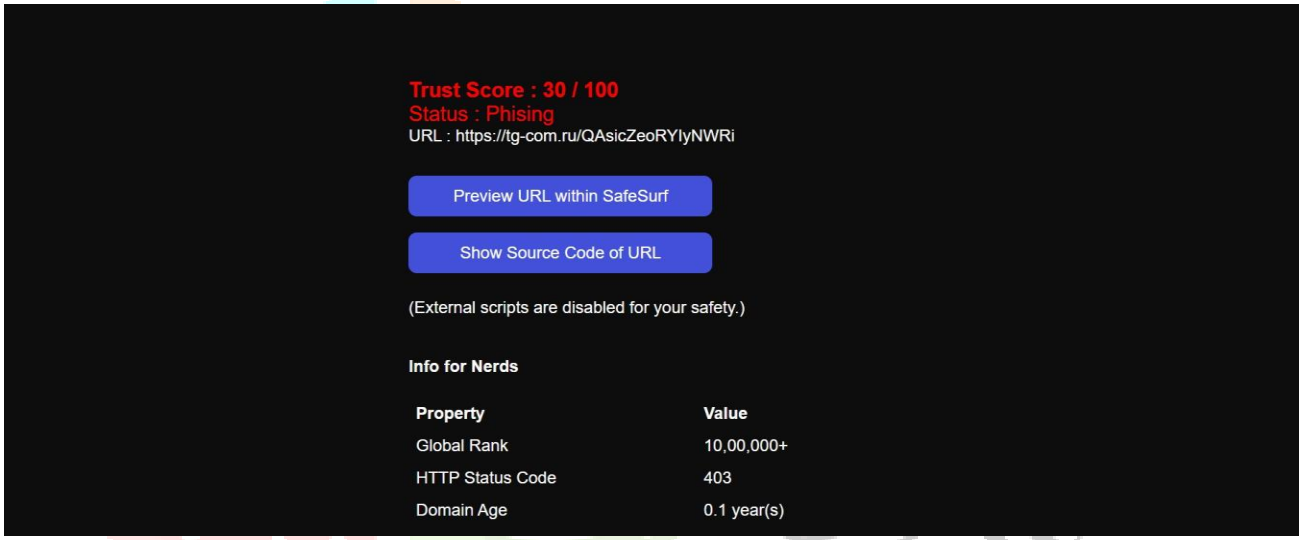

127.0.0.1:5000
Phishing Website detection Using Machine Learning
Protect yourself from phishing attacks with the help of FOSS. Surf safe with SafeSurf.
URL
CHECK URL

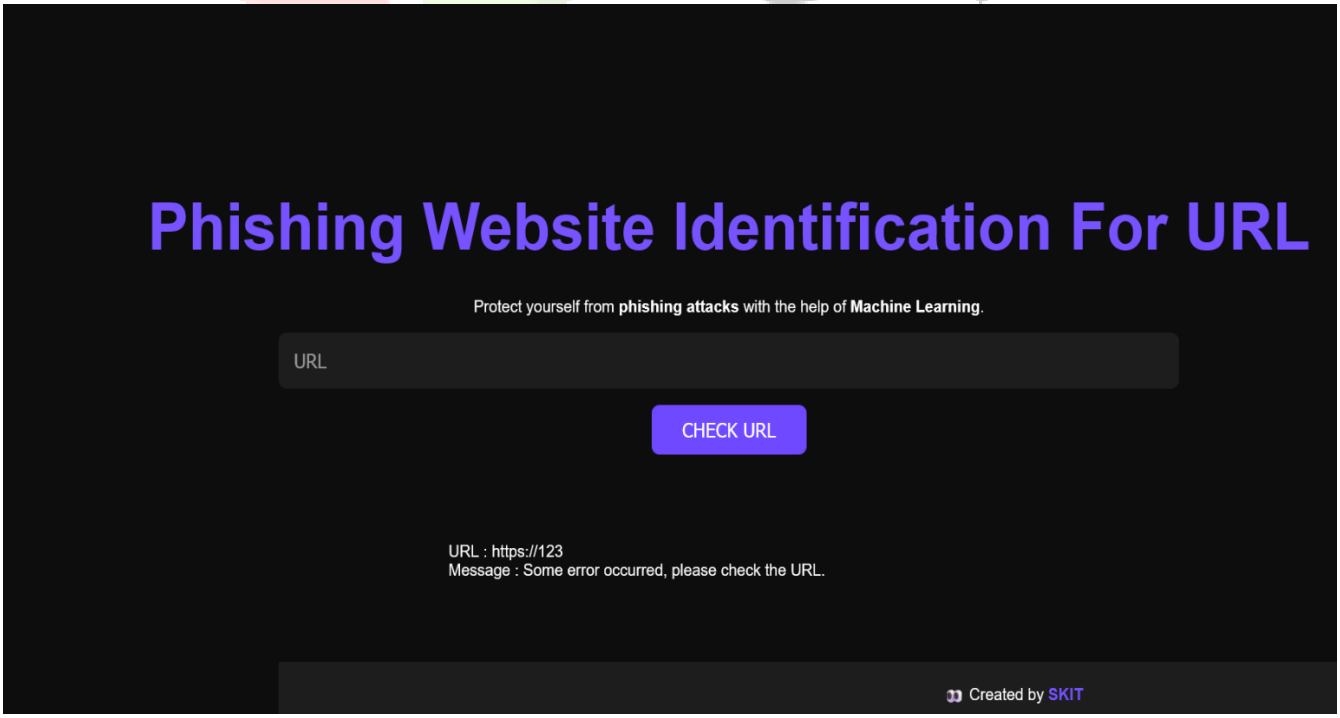

Phishing Website detection Using Machine Learning
Protect yourself from phishing attacks with the help of FOSS. Surf safe with SafeSurf.
https://www.google.com/
CHECK URL

INPUT


Trust Score : 100 / 100
Status : Legitimate
URL : https://www.google.com/
Preview URL within SafeSurf
Show Source Code of URL
(External scripts are disabled for your safety.)
Info for Nerds
Property
Value
Global Rank
1
HTTP Status Code
200
Domain Age
26.5 year(s)

RESULT


Trust Score : 30 / 100
Status : Phising
URL : https://tg-com.ru/QAsicZeoRYIyNWRi
Preview URL within SafeSurf
Show Source Code of URL
(External scripts are disabled for your safety.)
Info for Nerds
Property
Value
Global Rank
10,00,000+
HTTP Status Code
403
Domain Age
0.1 year(s)


Phishing Website Identification For URL
Protect yourself from phishing attacks with the help of Machine Learning.
URL
CHECK URL
URL : https://123
Message : Some error occurred, please check the URL.
Created by SKIT

References:

1. Phish Sim: Aiding Phishing Website Detection With a Feature-Free Tool Authors: Rizka Widyarini Purwanto; Arindam Pal; Alan Blair; Sanjay Jha. Published in: IEEE Transactions on Information Forensics and Security (Volume: 17).

2.A Deep Learning-Based Framework for Phishing Website Detection Authors: Lizhen Tang; Qusay H. Mahmoud. Published in: IEEE Access (Volume: 10) 2021

3. Multilayer Stacked Ensemble Learning Model to Detect Phishing Websites Authors:Lakshmana Rao Kalabarige; Routhu Srinivasa Rao; Ajith Abraham; Lubna Abdelkarei..Published in: IEEE Access (Volume: 10), 2022

4. PDGAN: Phishing Detection With Generative Adversarial Networks Authors: Saad Al- Ahmadi; AfrahAlotaibi; Omar Alsaleh. Published in: IEEE Access (Volume: 10), 2022.

5. Uncovering the Cloak: A Systematic Review of Techniques Used to Conceal Phishing Websites Authors: Wenhao Li; Selvakumar Manickam; Shams Ul Arfeen Laghari; Yung-Wey Chong Published in: Published in: IEEE Access (Volume: 11),2022.