INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553

# PHISHING URL DETECTION A REAL-CASE SCENARIO THROUGH LOGIN URLS

**N. Swathi[1], Potharaju Varshitha[2], Divya Palevar[3], Ruttala Manohar[4], Agre Anisha[5]**

[1]Assistant. Professor, CSE Dept, ACE Engineering College, Hyderabad, India

[2,3,4,5]Student, CSE Dept, ACE Engineering College, Hyderabad, India

## ABSTRACT.

Phishing, an internet scam, involves attackers sending deceptive messages mimicking trusted sources. These messages typically contain URLs or files intended to steal personal data or infect computers. Historically, phishing relied on mass spam campaigns targeting broad audiences. The objective was to maximize clicks on malicious links or files. Detecting such attacks employs diverse methods, including machine learning. In this approach, URLs received by users are inputted into a machine learning model, which then processes them to determine if they're phishing or legitimate. Various ML algorithms, such as SVM, Neural Networks, Random Forest, Decision Tree, and XGBoost, can classify these URLs. The proposed method focuses on employing Random Forest and Decision Tree classifiers for this purpose.The proposed approach effectively classified the Phishing and Legitimate URLs with an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers respectively.

## 1. INTRODUCTION

The internet has become an integral part of our lives, yet it also harbors opportunities for malicious activities like phishing. Phishers employ social engineering or mockup websites to trick individuals and organizations into divulging sensitive information such as account IDs, usernames, and passwords. Despite numerous detection methods proposed, phishers continuously adapt to evade detection. Machine learning has emerged as a highly effective tool in detecting these malicious activities due to the common characteristics shared by most phishing attacks.

Phishing websites, a prevalent form of social engineering, masquerade as trustworthy URLs and webpages. Unlike software vulnerabilities, they exploit human vulnerabilities. This involves luring online users to disclose sensitive information. To address this issue, datasets encompassing both phishing and legitimate URLs are compiled, from which relevant URL and website content-based features are extracted for analysis.

The performance level of each model is measured and compared.

With the deployment of the Streamlit application, users gain the capability to independently authenticate URLs.

## 2. OBJECTIVES

In phishing website detection, the process entails automatically categorizing websites into a predefined set of class values by analyzing multiple features alongside the designated class variable.

ML-based phishing techniques rely on the functionalities of websites to collect data that aids in the classification of websites for the detection of phishing sites.

While the issue of phishing may never be entirely eliminated, it can be mitigated through two approaches: enhancing targeted anti-phishing methods and educating the public about fraudulent practices.

## 3. METHODOLOGY

Review inputs and outputs for project activities. Information will be collected and prioritized. An appropriate algorithm or framework has been selected. Several estimation algorithms will be compared and the best method will be selected. Software and hardware selection will be made according to the needs. Data will be used as a process or framework

## 4. LITERATURE SURVEY

**TITLE:** PhishHaven-An Efficient Real-Time AI Phishing URLsDetection System

**Author:** Kyunghyun Han, Seong Oun Hwang, Maria Sameen

**YEAR:** 2020

**DESCRIPTION:** PhishHaven is an efficient real-time phishing URL detection system that aims to protect users from falling victim to phishing attacks. Phishing is a cybercrime where attackers impersonate legitimate websites or organizations totrick users into revealing sensitive information such as passwords, credit card numbers, or personal details.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553

**DISADVANTAGES:**

1.False Positives

2.Zero-day Attacks

3.User Awareness

**TITLE**: DEPHIDES:Deep Learning Based Phishing DetectionSystem

**AUTHOR:** Ozgur Koray Sahingoz, Ebubekir Buber, EminKugu

**YEAR: 2024**

**DESCRIPTION:** Dephides is a deep learning-based phishing detection system that uses advanced machine learning algorithms to identify and block phishing emails. It leverages thepower of deep neural networks to analyze various attributes and patterns in email content and headers to determine whether an email is a phishing attempt.

**DISADVANTAGES:**

1. Training Data Bias
2. False Positives
3. Adversarial Attacks
4. Resource Intensive

**TITLE:** D-Fence: A Flexible, Efficient, and ComprehensivePhishing Email Detection System Public

**AUTHOR:** Jehyun Lee, Farren Tang, Pingxiao Ye

**YEAR:** 2021

**DESCRIPTION:** D-Fence is a phishing email detection system that aims to provide flexibility, efficiency, and comprehensive protection against phishing attacks. It employs intelligenttechniques to analyze and identify malicious emails, specifically focusing on phishing attempts. The system offers flexibility by allowing users to customize and adapt the detection rules according to their specific needs. This enables organizations to tailor the system to their unique requirements and stay ahead of evolving phishing techniques.

**DISADVANTAGES:**

1. False Negatives
2. Training and Maintenance
3. User Education
4. Integration Challenges

## 5. PROPOSED SYSYTEM

Machine Learning and AI:

Machine learning and artificial intelligence (AI) algorithms are increasingly used to analyze medical data and identify patterns indicative of cancer. This major project focuses on the development of a robust and accurate cancer prediction system for early stage detection. Leveraging the power of Machine Learning and medical data analysis, this project aims to revolutionize cancer diagnosis by enabling timely interventions and treatments.

## 6. HARDWARE AND SOFTWARE REQUIREMENTS

### 6.1 HARDWARE REQUIREMENTS:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 14’ Colour Monitor.
- Mouse : Optical Mouse.
- Ram : 512 Mb.

### 6.2 SOFTWARE REQUIREMENTS:

- Operating system : Windows 7 Ultimate.
- Coding Language : Python.
- Front-End : Python.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553

**PACKAGES USED**

TensorFlow

TensorFlow is a popular open-source Python machine learning toolkit for creating and training deep neural networks. It has a versatile architecture and supports a variety of platforms, including CPU, GPU, and TPU. TensorFlow simplifies the implementation of complicated algorithms and models, allowing developers to create scalable and efficient machine learning systems.

Keras

Keras is a Python-based high-level neural network API that operates on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. It offers an easy-to-use interface for building and training deep learning models, letting users to easily experiment with alternative architectures and hyperparameters. Keras also provides pre-trained models as well as a huge collection of building blocks for developing sophisticated models.

Numpy

NumPy is an important Python package for scientific computation.It provides extensive support for large, multidimensional arrays and matrices, along with a wide range of high-level mathematical operations tailored for these arrays. NumPy is a popular choice for numerical operations in scientific research and data analysis due to its efficient and user-friendly interface.

Pandas

Pandas is a popular open-source Python data analysis and manipulation package. It offers sophisticated data structures and tools for working with structured data, including as data frames and series, and it allows for quick data processing, cleaning, merging, and reshaping. Pandas also supports reading and writing a variety of file types, including CSV, Excel, and SQL databases.

Matplotlib

Matplotlib is a popular Python data visualization package. It includes line graphs, scatter plots, bar plots, and histograms among its 2D and 3D displays. Matplotlib is a useful tool for data exploration and communication since it is extremely customizable and supports extensive labelling, annotations, and text formatting.

OS and time

The Python 'os' module enables interaction with the operating system. It has functions for creating and removing folders, manipulating files, and changing environment variables. The 'time' module in Python contains methods for working with time-related actions. It has functions for obtaining the current time, postponing program execution, and converting between several time formats.

**TECHNOLOGY DESCRIPTION**

Python is an interpreted high-level programming language that issimple to learn and use. It features a basic and clear syntax that makes it suitable for both beginners and professionals. Python is utilized in many different areas, such as web development, scientific computing, data analysis, and artificial intelligence.

## 7. SOURCE CODE

```
import pandas as pd
from sklearn.feature_extraction.text importCountVectorizer
from sklearn.naive_bayes import MultinomialNB
from tkinter import * df=pd.read_csv('dataset/sms.txt',delimiter='\t') cv=CountVectorizer(stop_words='english') mnb=MultinomialNB()
def mytrain():
df.columns=['label','msg'] X=cv.fit_transform(df.msg).todense()y=df.iloc[:,0].values
mnb.fit(X,y)  def mypredict():
msg=e.get() X_test=cv.transform([msg]).todense()pred=mnb.predict(X_test) if(pred[0]=='spam'):
outlbl.configure(text=pred[0],fg='red')
else:
outlbl.configure(text=pred[0],fg='green')root=Tk()
root.state('zoomed') root.configure(background='yellow') title=Label(root,text='Phishing  Email  Detection  UsingImproved  RCNN  Model  WithMultilevel',bg='yellow',font=('',20,'bold'))
```

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553

```
title.place(x=100,y=10) title1=Label(root,text='Multilevel Vectors andAttention
Mechanism',bg='yellow',font=(',20,'bold')) title1.place(x=200,y=50)
lbl=Label(root,text='Enter msg:',fg='blue',bg='yellow',font=(',20,'bold'))lbl.place(x=200,y=200)
e=Entry(root,font=(',15,'bold')) e.place(x=350,y=205)
b=Button(root,text='Predict',command=mypredict,font=(',15,'bold'))
b.place(x=400,y=250) outlbl=Label(root,bg='yellow',font=(',20,'bold'))outlbl.place(x=350,y=350)
mytrain() root.mainloop()
```

## 8. OUTPUT

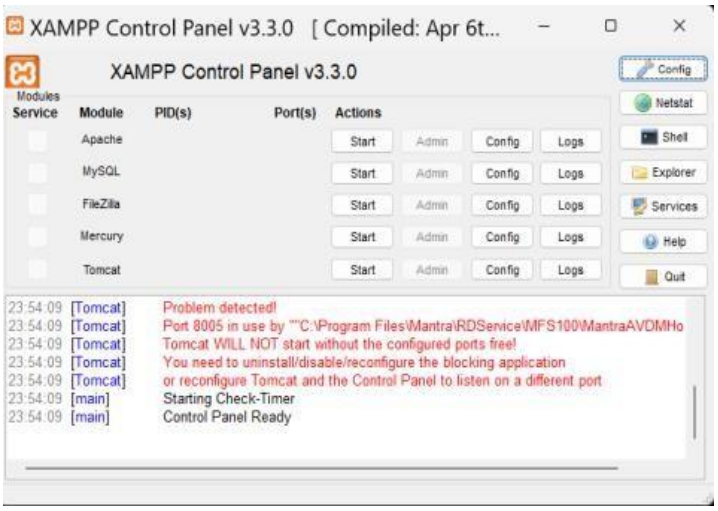



**Fig.1** XAMPP Control Panel

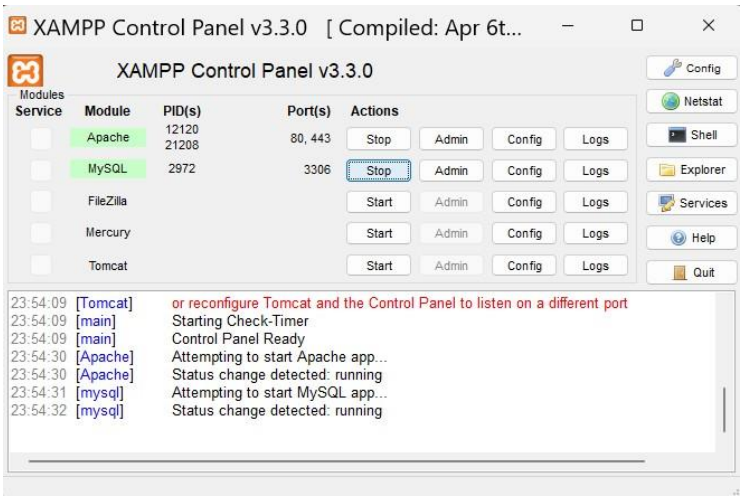
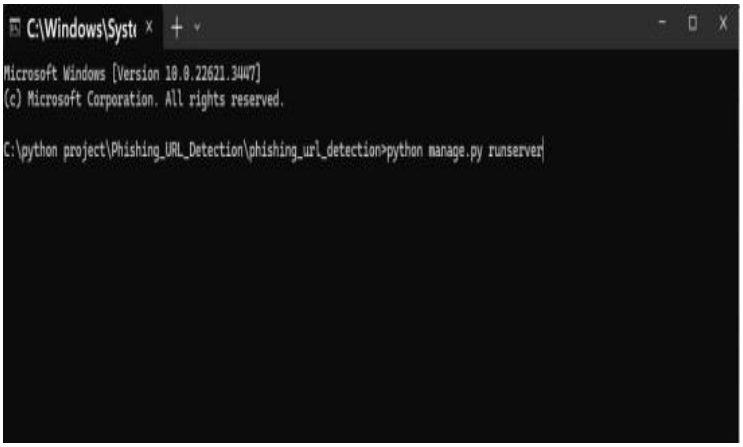



**Fig.2** .Stop Appache & Mysql




**Fig,3.** Command Fig.1 0.4.Copy Website Link

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553


**Fig.4** .Home page


**Fig. 5.**user registration page


**Fig.6** .registration status


**Fig.7** .user login details

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553

e-ISSN : 2583-1062

Impact Factor: 5.725

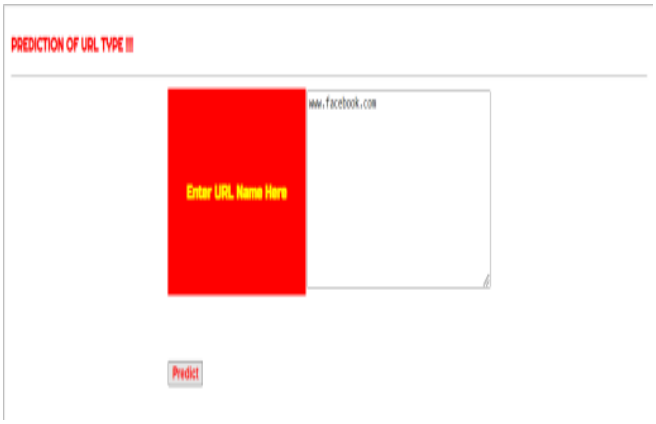



**Fig.8** predicting URL


**Fig.9** predicted URL Type


**Fig10** user profile

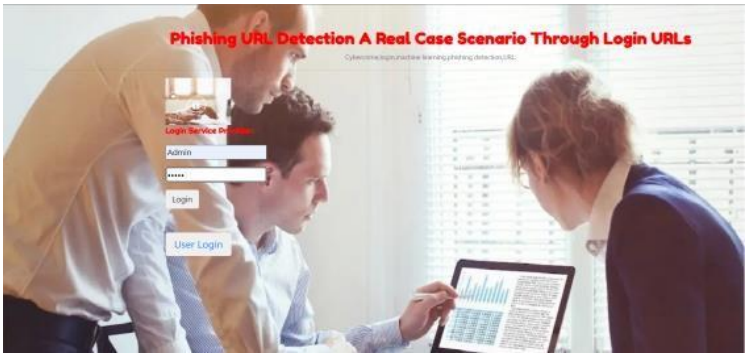



**Fig.11** admin page


**Fig.12** Train and test url dataset

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553


**Fig.13** view url datasets train and tested accuracy in bar chart


**Fig.14 .**view url datasets train and tested accuracy in pie chart


**Fig. 15** Fig.10.14.view url datasets train and tested accuracyin line chart

**Fig. 16** view predicted url

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553

Impact Factor: 5.725


**Fig.17** view url type ratio


**Fig.18** .download predicted dataset

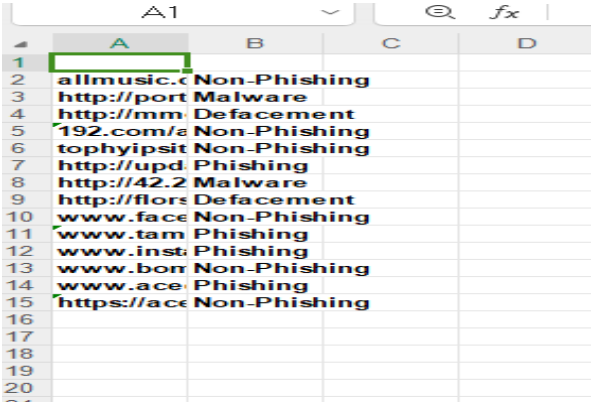



**Fig.19**.dataset


**Fig.20** view url type ratio result in pie chart


**Fig21** view url type ratio result in line chart

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 545-553


**Fig.22** view all remote users

## 9. CONCLUSION

Phishing is a crucial threat to individual's data nowadays. Detection of phishing sites is actually a tiresome task, as the outcome phishers are actually quickly enhancing. To overcome the problem, researchers and specialists dealt with lots of methods and techniques, however it led to reduced prices of detection. In conclusion, the application of machine learning algorithms we used lots of methods like Decision Tree, Random forest, Multi- layer Perceptron's, XG Boost Classifier, SVM, Light BGM Classifier, Cat Boost Classifier. Away from which our team observed that Light GBM obtained the greatest precision of around 85.5% when compared to a variety of other methods. Whereas one class SVM is the one with the lowest precision of approximately 79.6%. As previously discussed, these algorithms were used to develop the model and predict the outcomes. Furthermore, our team discovered that the versatility of machine learning algorithms of light GBM performed significantly better than other techniques or algorithms discussed previously. Overfitting of information is actually prevented, which is a key feature. As a result, the Light GBM classifier is the best fit for us to detect whether the site is phishing or not.

## 10. FUTURE SCOPE

- Real-time Detection: Developing real-time phishing detection systems capable of analyzing and flagging suspicious URLs and emails as they are received could provide immediate protection to users.
- Deployment in Email Services and Browsers: Collaborating with email service providers and web browser developers to integrate phishing detection models directly into their platforms could offer seamless protection to users, preventing them from accessing malicious content altogether.
- Continuous Model Training: Implementing mechanisms for continuous model training and updating to adapt to evolving phishing techniques and patterns is essential for maintaining detection efficacy over time.
- Evaluation on Diverse Datasets: Evaluating the proposed approach on diverse datasets collected from different sources and regions would validate its effectiveness across various phishing scenarios and demographics.

## 11. REFERENCES

[1] Patil, Srushti, and Sudhir Dhage. "An In-Depth Analysis of Phishing Detection Techniques and Systematic Development of an Anti-Phishing Framework" presented at the 5th International Conference on Advanced Computing in 2019. IEEE, 2019.

[2] Garcés, Ivan Ortiz, Maria Fernada Cazares, and Roberto Omar Andrade."Utilizing Machine Learning for Phishing Attack Detection within a Cognitive Security Framework" presented at the 2019 International Conference on Computational Science and Computational Intelligence (CSCI).IEEE, 2019.

[3] Ahmed, Abdulghani Ali, and Nurul Amirah Abdullah."Immediate Identification of Phishing Websites in Real-Time" presented at the 7th Annual Information Technology, Electronics, and Mobile Communication Conference (ICON) by IEEE in 2016.. IEEE, 2016.