DOMAIN ARMOR

A FIELD PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE ENGINEERING

BY

22071A6609-PRIYANKA
22071A6620-SUPRITHA
22071A6642-CHAITRA
23075A6605-BHAVATILYA

UNDER THE GUIDANCE OF

DR.N.SANDHYA
PROFESSOR and Head
CS-AIML &IOT Department

DEPARTMENT OF AIML&IoT ENGINEERING

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING & TECHNOLOGY
PRAGATHI NAGAR, NIZAMPET (S.O),
HYDERABAD - 500 090

DECEMBER- 2023

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

Estd.1995

An Autonomous Institute, NAAC Accredited with 'A++' Grade
NBA Accreditation for B.Tech. CE, EEE, ME, ECE, CSE, EIE, IT, AME and M.Tech. STRE, PE, AMS and SE programmes
Approved by AICTE, New Delhi, Affiliated to JNTUH
Recognized as "College with Potential for Excellence" by UGC
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India.
Telephone No: 040-2304 2758/59/60, Fax: 040-23042761
E-mail: postbox@vnrvjiet.ac.in, Website: www.vnrvjiet.ac.in

Department of AIML&IoT Engineering

CERTIFICATE

This is to certify that the **Field Project** report entitled "**DOMAIN ARMOR"** being submitted by **PRIYANKA,SUPRITHA,CHAITRA,BHAVATILYA,** Regd. No. **22071A6609, 22071A6620, 22071A6642, 23075A6605**in partial fulfillment for the award of **BACHELOR OF TECHNOLOGY** in **AIML&IoT ENGINEERING** to the Jawaharlal Nehru Technological University Hyderabad at **VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING & TECHNOLOGY, HYDERABAD**, is a record of bonafide work carried out by him/her under our guidance and supervision.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Signature of Supervisor
DR.N.SANDHYA
PROFESSOR&HOD
AIML&IoT Engineering
VNR VJIET, Hyderabad

Signature of HOD
Dr. SANDHYA
Professor & HOD
AIML&IoT Engineering
VNR VJIET, Hyderabad

PLAGIARISM CERTIFICATE

APPROVAL CERTIFICATE

Field project evaluation for the dissertation work entitled **"DOMAIN ARMOR"** being submitted by PRIYANKA-22071a6609,SUPRITHA-22071a6620,CHAITRA-22071a6642,BHAVATILYA-230751a6605 is conducted on ……. and the work is approved for the award of **BACHELOR OF TECHNOLOGY** in **AIML&IoT ENGINEERING.**

PROJECT REVIEW COMMITTEE

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING & TECHNOLOGY
(An Autonomous Institution, Accredited by NAAC with 'A++' grade and NBA)
Pragathi Nagar, Nizampet (S.O.),
Hyderabad - 500090
Telangana
DEPARTMENT OF COMPUTER SCIENCE(AIML) ENGINEERING

DECLARATION

I hereby declare that the Field Project report entitled **"DOMAIN ARMOR",** submitted for B.Tech. degree is my original work and project has not formed the basis for the award of any degree, associateship, fellowship or any similar titles.

Signature of the student
PRIYANKA
(22071A6609)
SUPRITHA
(22071A6620)
CHAITRA
(22071A6642)
BHAVATILYA
(23075A6605)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

SG	Sample Graph

ACKNOWLEDGEMENTS

Over a span of one and a half years, VNRVJIET has helped us transform ourselves from mere amateurs in the field of Computer Science into skilled engineers capable of handling any given situation in real time. We are highly indebted to the institute for everything that it has given us. We would like to express our gratitude towards the principal of our institute, **Dr. Challa Dhanunjaya Naidu** and the Head of the Computer Science & Engineering Department, **Dr.N.Sandhya** for their kind co-operation and encouragement which helped us complete the project in the stipulated time. Although we have spent a lot of time and put in a lot of effort into this project, it would not have been possible without the motivating support and help of our project guide **Mrs.N.Sandhya.** We thank her for her guidance, constant supervision and for providing necessary information to complete this project. Our thanks and appreciations also go to all the faculty members, staff members of VNRVJIET, and all our friends who have helped us put this project together.

B.PRIYANKA
22071A6609
B.Tech. (CSE-AIML)

G.SUPRITHA
22071A6620
B.Tech. (CSE-AIML)

N.CHAITRA
22071A6642
B.Tech. (CSE-AIML)

M.BHAVATILYA
23075A6605
B.Tech. (CSE-AIML)

ABSTRACT

The project DOAMIN ARMOR elementally focusses on differenciating real domains from the phishing ones which imitate look and feel of genuine ones using ML algorithms. In the current generation we have seen how dangerous AI has become .With this we can also estimate the upcoming danger due to phishing domains. It is an online threat where the sites or domains perfectly imitate the authentic websites. One best example of this is Trolling, which has caused a severe problem to the society. Now that we can use the recent advances in machine learning and detect these phishing domains and assist audience towards much better security.

Key Words: Specify 3 to 4 keywords separated by comma

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO THE PROJECT:

Phishing Domains are the websites which mimic the real websites and take very important and personal data like usernames, passwords, bank account details, login credentials, personal address, social relantionships etc from the users and in return morph or blackmail them. This is a cybercrime where many people are made innocent and the other party gets benefit. This is a result of combining both social engineering and technical tricks. This kind of obtaining sensitive information from the users without proper authorization is a crime and should be strictly punished.

1.2 BACKGROUND:

The plain background about he phishing domains is that it is a major cyber threat which exploites humans soo desperately and gather sensitive information. Traditional methods of catching the culprit manually has become so difficult because of increasing tactics of the criminals. As we know machine learning has become one of the emerging leads in today's generation it can be used to solve such social problems. It basically involves clasifyimg the domains based on certain patterns and features. The project not only tries to increase the accuracy but also it focusses on the adaptability to the emerging threats. This provides a perfect secure layer towars the emerging world.

Fig. 1.1: User Input

1.3 TOOLS, EQUIPMENTS AND TERMINOLOGY USED:

Tools & equipment:

Software requirements:

1.Programing language: Python is usually chosen because it is basically a more robust language with simple syntaxes. Also it has various libraries installed in it.

2.Machine Learning Algorithms: Because we are developing an ML model we need to come across various algorithms for which we may use SCIKIT-LEARN,KERAS,TENSORFLOW ,PYTORCH etc.

3.Data:One major important step in setting up a ml model is to collect relevant data regarding the theme of the project. This to used to draw useful conclusions and basic insights about the project. Also data is used to train the model. For the analysis of data we can use pandas, Numpy, Matplotlib and seaborn.

4.Development Environment: A proper integrated development environment(IDE) is all we need to properly run and execute the code. Juptyter notebook is a nice choice though.

Hardware requirements

1.Storage Space:As this is a project that involves differenciating domains it usually requires a lot of storage space in order to accommodate large data sets.

2.Internet Connection : Because we are collecting, using and building a model from scratch we need the help of web browsers and also AI assistants for which we need proper internet connection.

TERMINOLOGY USED:

Phishing ,Feature extraction, Supervised learning, Unsupervised learning, Ensemble learning, Hyperparameters, Cross-validation, ROC Curve(receiver operating characteristic),feature importance.

1.4 OUTLINE OF THE PROJECT REPORT:

- Import libraries and modules like randomforests classifier, train-test-split, urlparse
- Now extract feature from given url
- Using a sample dataset, label the phishing and legitimate domains
- Now apply feature extraction and labelling to the current user input
- Split the data using train-test-split and start training the model using random forests

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW:

1>A.Amrutha Rose and Eligious Kavaivani.C,Variants of phishing attacks and their detection techniques. A combination of supervised and unsupervised machine learning techniques was used to detect known and unknown attacks.
The major drawback of this research is it did not discuss Deep Learning techniques.

2>Norah Alrumayh and Dr.Aram Alsedranisa,Detecting Phishing Websites Using Machine Learning and the focus is to pursue a higher performance classifier to train it.
This work analyzed only 18 studies and did not include some machine learning approaches.

3>Gururaj Harinahalli Lokesh and Goutham Bore Gowda, Phishing website detection based on effective machine learning approach and this selection which contains the metadata of URLs and use the information to determine if a website is phishing or not.

4>Arshadet al, Different types of phishing and anti-phishing techniques are used. This analyzed that machine learning techniques approches have high Accuracy.
This is based on only 20 studies.

5>Catal etal, The main ideology is to synthesize, assess, and analyses Deep Learning techniques for phishing detection. Best preformance was given by Hybrid DL algorithms.
This work only discusses Deep Learning related studies for phishing detection.

2.2 REVIEW OF THE LITERATURE:

1> 2019,Various types of phishing attacks and the recents and the recent approches to prevent the attacks are discussed. Using ml a framework to detect and overcome phishing domains is proposed.

2> 2019,the system is based on a machine learning method, particularly supervised learning. Random Forest technique is implemented due to its good performance in classification. Studying the features of phishing websites and choose the better combinations to train.

3> 2020,this paper provides the applicability of ML techniques in identifying phishing attacks and report their positives and negatives. They have designed a phishing Classification system which extracts features that are meant to defeat common phishing detection approaches and can also

make use of numeric along with the comparative study of machine learning techniques like random forest, svm classifier and wrapper-based features.

4> 2021,This papers examines different types of phishing and anti-phishing techniques are presented. This includes phone phishing, email, phishing and manipulation are the frequently used.

5> 2022,According to this article, the most used machine learning algorithms was DNN and HYBRID Dl algorithms.42 studies applied ml algorithms out of 43 studies.

2.3 PROBLEM STATEMENT:

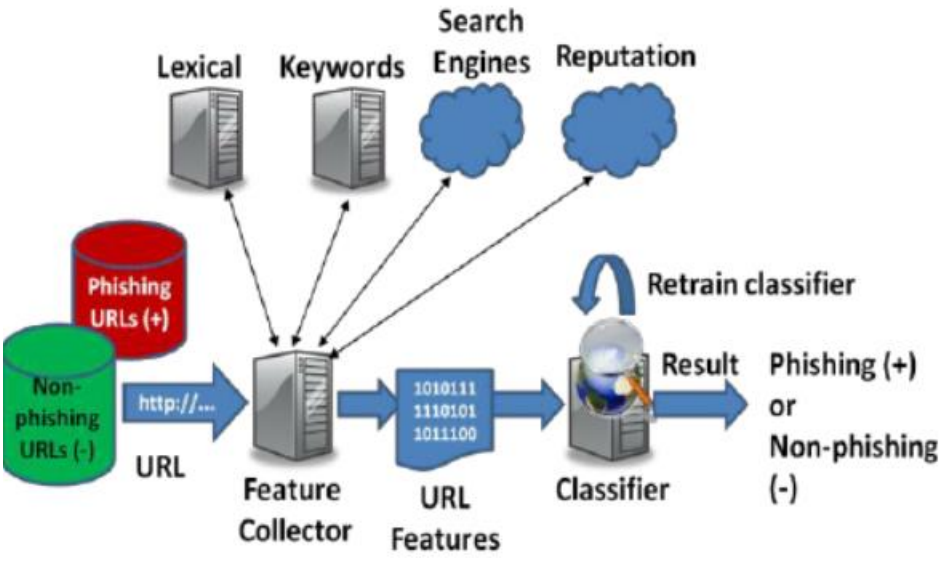
DETECTING PHISHING DOMAINS USING ML ALGORITHMS

2.4 PROJECT OBJECTIVES:

- Early detection of phishing attacks with improved accuracy and efficiency
- Adaptability to emerging threats and real time detection
- Cost effective security measures with proper user education and awareness
- Reduced impact of phishing attacks &contribution towards development

2.5 SUMMARY:

These summaries collectively gives a thorough exploration of phishing detection methods from conventional methods to machine learning approaches. Covering various studies, they check how well the conventional methods work to teach people about phishing and using legal actions and then they delve into the new technology which uses machines to learn and to make decisions. However, some studies miss out on the most advanced techniques, Deep learning, this also can highlight the limitations on certain studies. This together gives a detailed look at how we can stop phishing ,and the importance of people teaching about it, so that people can know how the machine learning is helpful and pointing out deep learning can make catching phishing even better.


Lexical
Keywords
Search Engines
Reputation
Phishing URLs (+)
Non-phishing URLs (-)
http://...
URL
Feature Collector
1010111
1110101
1011100
URL Features
Retrain classifier
Classifier
Result
Phishing (+) or Non-phishing (-)

CHAPTER 3

DEVELOPMENT OF THE PROJECT

3.1 METHODOLOGY OF THE PROJECT:

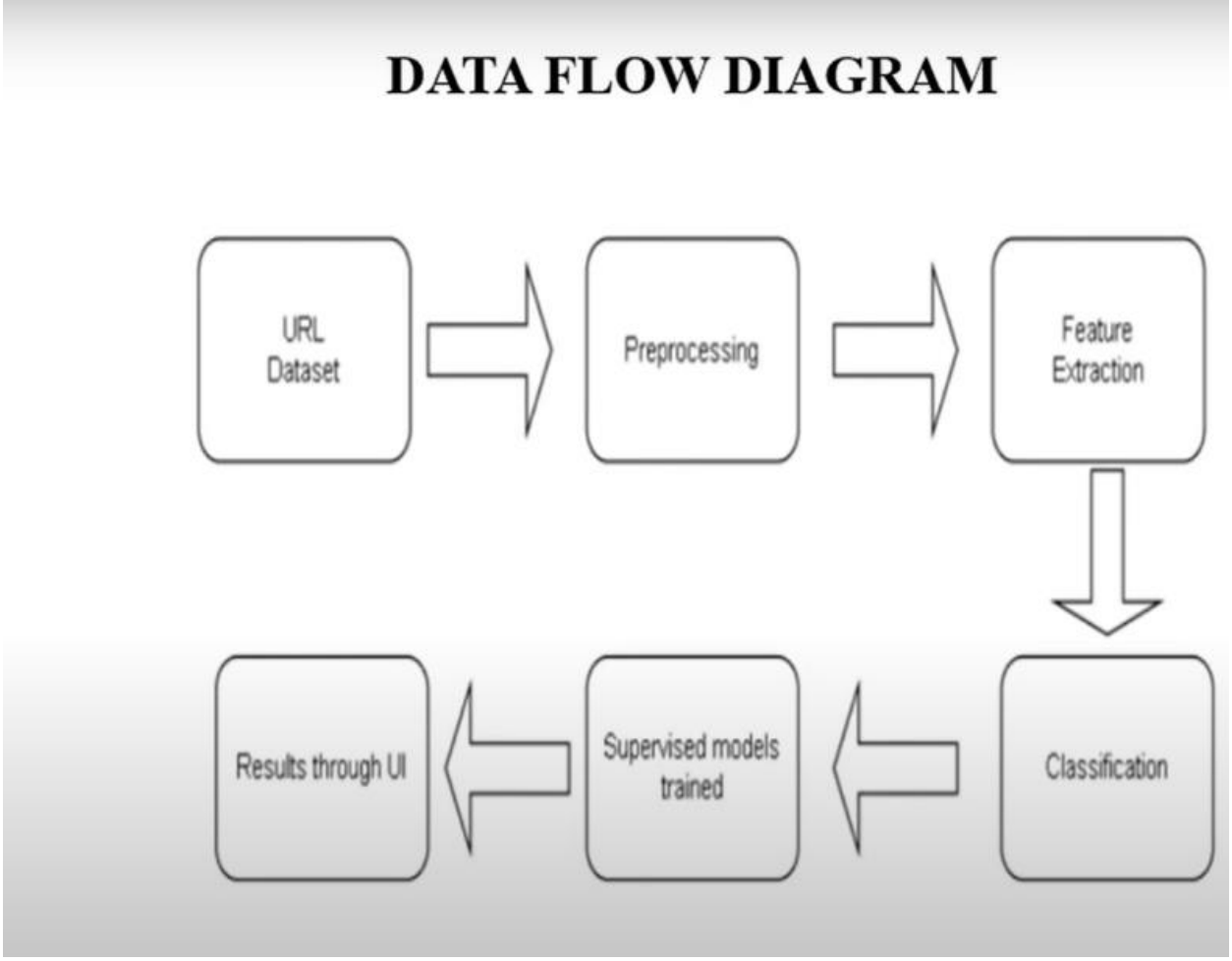
- **Data collection& preprocessing**: The first step in building a model is to gather the relevant data from various sources. In this case we need a wide range of both phishing and real domains which includes almost all of the characteristics such as the url length, structure, special characters and other content features. Now cleaning and pre-processing the data is also very much needed to ensure the consistency and accuracy. This includes removing the unnecessary noices, handling missing data, formats and converting categorical features into suitable format so that it can be given as input to the ml algorithm.

- **Feature extraction &Labeling**: Now we require necessary features which are used to distinguish real and fake websites. also labelling a domain as legitimate or phishing is also equally important as this step is crucial for supervised learning.

- **Splitting dataset and selecting model**: After we get the datasets it is essential to divide to training, validation and testing. Training is used to train and validation is used to properly align the parameters and testing to used evaluate. Now after splitting choose a ml algorithm that is a perfect fit for the need. Some popular ml algorithms include random forests ,decision trees, neural networks, k nearest neighbour and support vector machines.

- **Training and evaluation of the model**: Training the model to evaluate the problem so that it can recognize patterns. Once its done tune hyperparameters in order to optimize the performance. If needed we can adjust the ratings, strengths and also depths. also analyse the important features in detecting the phishing domains

- **Deployment &monitoring**: Now its time to make our model into a real world entity where it can take user inputs and evaluate it in return. Regular monitoring of the model's performance is essentially important



3.2 DEVELOPMENT OF THE PROJECT:

Here in this code we took five of the insights from the data collected.They are as follows:

1.presence of special characters

2.presence of dots

3.presence of slashes

4.valid ip address

5.presence of “https”

Code:

```
import re
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from urllib.parse import urlparse

def extract_aspects(url):
    aspects = []
    aspects.append(len(url))
    aspects.append(bool(re.search(r"\b(?:\d{1,3}\.){3}\d{1,3}\b", url)))
    aspects.append(url.count('/'))
    aspects.append(bool(re.search(r"[!@#$%^&*(),.?\":{}|]", url)))
    aspects.append(1 if urlparse(url).scheme == 'https' else 0)
    aspects.append(urlparse(url).netloc.count('.'))
    aspects.append(1 if urlparse(url).scheme == 'http' else 0)
    return aspects

phishi_urls = ["http://example-phishing-site.com", "http://phishy-url.net", "http://evil-site.org"]
legiti_urls = ["http://legit-site.com", "http://trusted-url.net", "http://secure-site.org"]

phishi_labels = [1] * len(phishi_urls)
legiti_labels = [0] * len(legiti_urls)

all_urls = phishi_urls + legiti_urls
all_labels = phishi_labels + legiti_labels

P = [extract_aspects(url) for url in all_urls]
Q = all_labels

P_train, P_test, Q_train, Q_test = train_test_split(P, Q, test_size=0.2, random_state=42)

model = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced')
model.fit(P_train, Q_train)

while True:
    user_url = input("Enter URL: ")
```

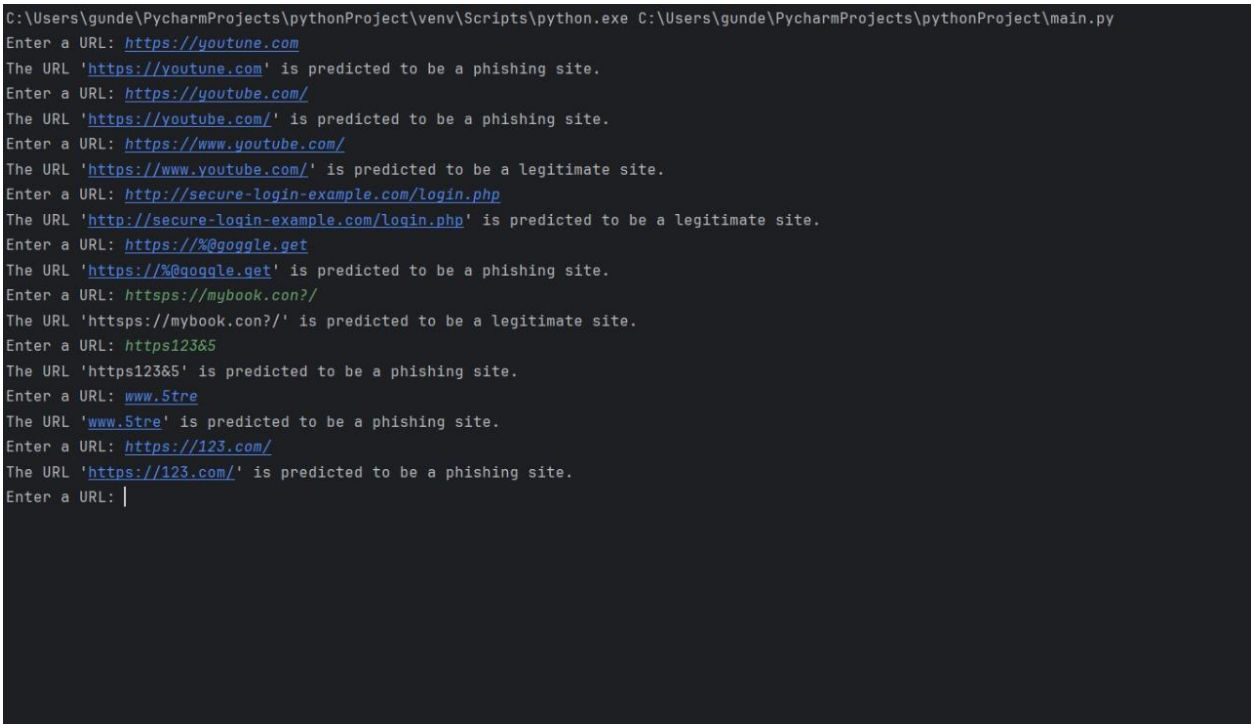
```
    if user_url.lower() == 'exit':
        break

    user_aspects = extract_aspects(user_url)

    predict = model.predict([user_aspects])

    if predict[0] == 1:
        print(f"URL '{user_url}' is a phishing site.")
    else:
        print(f"URL '{user_url}' is a legitimate site.")
```

3.3 RESULTS:



CHAPTER 4

CONCLUSIONS

4.1 CONCLUSIONS:

Phishing attacks remain one of the most significant pitfalls to cyber security, criminals using decreasingly sophisticated styles to deceive and steal sensitive information.We use machine learning,data analysis and natural language processing to identify and block phishing websites.We get very good performance in ensemble classifiers namely Random Forest on computation and accuracy.The main idea behind ensemble algorithms is to combine several weak learners into a stronger one,this is perhaps the primary reason why ensemble based laerning is used in practice for most of the classification problem.

These circumstances lead to future works to add more features to the dataset,which could improve the performance of these models.By continuously covering network business its possible to identify phishing attacks in real-time and take applicable action to help them.

4.2 RECOMMENDATIONS:

- it is recommended to train the data less and test more.A rough ratio of 1:4 for test:train is highly appreciable.
- Also it is advised to increase the considered constraints like login pages,patent rights

REFERENCES

https://www.sciencedirect.com/science/article/pii/S1319157823000034

https://www.mdpi.com/2076-3417/13/8/4649

https://link.springer.com/article/10.1007/s10586-022-03604-4

. "Phishing Websites Detection Using Machine Learning Algorithms."
By K. S. Rajasekaran, D. V. Malleswari
Source:International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2015.

Source:2018 3rd International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 2018.

Link: [IEEE Xplore](https://ieeexplore.ieee.org/document/8367543)

"Phishing Detection: A Machine Learning Approach."
by S. Yadav, A. Gupta
Source:International Journal of Computer Science and Information Technologies, 2015.
Link:
[ResearchGate](https://www.researchgate.net/publication/276282077_Phishing_Detection_A_Machine_Learning_Approach)

Link [IEEE Xplore](https://ieeexplore.ieee.org/document/8278844)

"Detecting Phishing Websites: A Machine Learning Approach."
by L. A. Bhat, M. S. M. Qadri
Source: 2018 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018.
Link:[IEEE Xplore](https://ieeexplore.ieee.org/document/8474051)

"Phishing Website Detection Using Machine Learning Techniques."
by N. K. Priyanka, G. Padmavathi
Source:2017 IEEE Calcutta Conference (CALCON), 2017.
Link: [IEEE Xplore](https://ieeexplore.ieee.org/document/8286708)

Abdelhamid N,thabtah F,Abdel-jaber H Phishing detection:a recent intelligent machine learning comparision based on models content and features.In Beijing,china,2018.

HariKrishna NB,Vinayakumar and Soman KP on "approach towards Phishing email detection;2018.

S.Mishra and D.Soni,"Smishing Detector:A security model to detect smishing through SMS content analysis and URL behaviour analysis,"(in english).Future Generation Computer Systems-the international Journal of Escience,Article vol.1o8,pp.8o3-815,jul 2020.

https://www.tessian.com/blog/phishing-statistics-2020/