M.Tech Project Report

MCO1PO2

on

MACHINE LEARNING APPROACHES FOR PHISHING WEBSITE DETECTION

BY

MD SAIF ALI (32013112)

Under the Supervision of

Dr. ANKIT KUMAR JAIN
(Assistant Prof.)

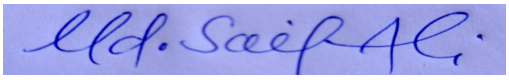
DEPARTMENT OF COMPUTER ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
KURUKSHETRA – 136119, HARYANA (INDIA)

May-June, 2020

CERTIFICATE

I hereby certify that the work which is being presented in this M.Tech Project (**MCO1PO2)** report entitled **"MACHINE LEARNING APPROACHES FOR PHISHING WEBSITE DETECTION",** in partial fulfillment of the requirements for the award of the **Master of Technology in Computer Engineering,** is an authentic record of my own work carried out during a period from January, 2020 to May, 2020 under the supervision of Dr.ANKIT KUMAR JAIN ,Assistant Professor, Computer Engineering Department.

The matter presented in this project report has not been submitted for the award of any other degree elsewhere.



Signature of Candidate

MD SAIF ALI (32013112)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Date:10-05-2021

Signature of Supervisor

Dr. ANKIT KUMAR JAIN
(Assistant Professor)

TABLE OF CONTENTS

Abstract

Phishing is one of the major problems faced by cyber-world and leads to financial losses for both industries and individuals. Detection of phishing attack with high accuracy has always been a challenging issue. At present, Machine Learning based techniques are very useful for detecting phishing websites efficiently.

Phishing is the fraudulent attempt to obtain sensitive information such as username, password, bank account details, and credit card details for malicious use. Phishing frauds might be the most popular cybercrime used today. There are various domains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. The webmail and online payment sector was targeted by phishing more than in any other industry sector. Machine Learning is efficient technique to detect phishing. It also removes drawback of black-list approach. We perform detailed literature survey and proposed new approach to detect phishing website by features extraction and machine learning algorithm.

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared.

1.Introduction

Phishing is an online identity theft, which can deceive Internet users into revealing their secret information and credentials, e.g., login id, password, credit card number, etc

Phishing is one of the major computer security threats faced by the cyber-world and could lead to financial losses for both industries and individuals.

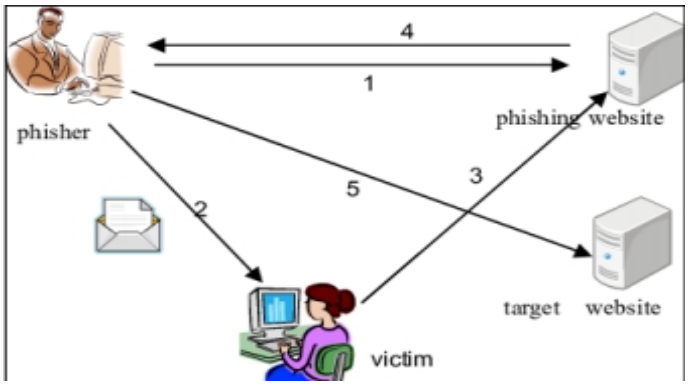
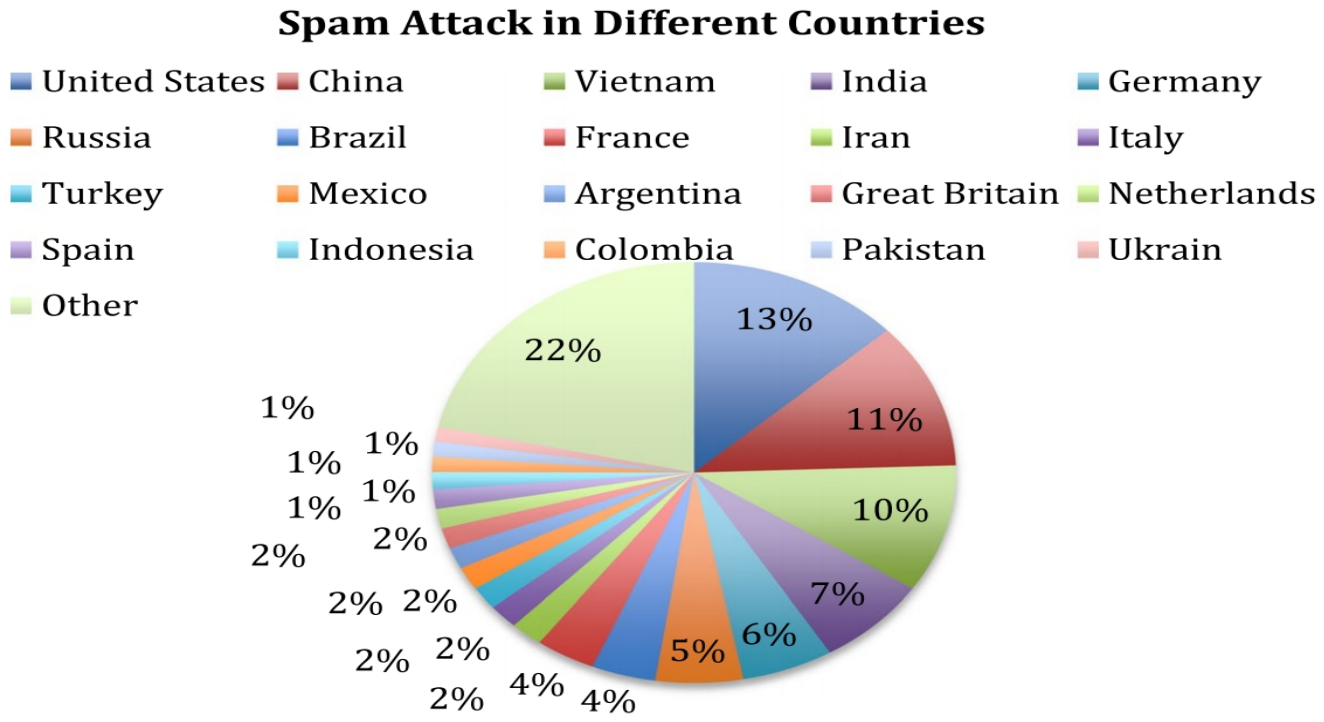


Fig-Steps of Web phishing process.

The term ‘phishing’ is derived from the analogy of ‘‘fishing’’ for victims’ passwords and credentials in the web. The phrase ‘‘ph’’ comes from ‘‘phone phreaking’’, which was very common technique that attacked telephone systems during 1970s. The word ‘phishing’was used for the first time over the Internet by a group of hackers in 1996, who stole America Online (AOL) accounts by tricking unaware AOL users into disclosing their passwords.[10].

2.Motivation

Attackers is always take mis-advantage of human nature that generally ignores critical warning message. Lack of awareness about the phishing attacks in the sociaty is also the main reason why phishing attacks have been so much succesful. Whenever any researcher come with some many technique to prevent these attacks, phishers try to find out associated loophole to commit successful attacks.



Phishing mainly used for financial gains, there are other factors that also motivate phishers to commit the crime. Motivation behind these activity are as below:

Theft of login credentials: Phishers steals login credential of online services like- flipkart, Gmail, Amazon and Phonepe from the user using spoofed email as warning message to chnage there password and provided hyperlinks.

Theft of banking credential: Online login credentials and credit card details such as card num. , expiry date, card holder name, CCV number and several other popular online banking organization like paypal, online SBI, HDFC and citibank.

3.Literature Survey

In [1] AK JAIN & BB GUPTA has presented a novel approach for filtering phishing websites at client side where URL, hyperlinks, CSS, login form and identity features is used. In this Paper author proposed various machine learning approaches like Random forest, SVM, Neural Networks, Logistic Regression and Naive Bayes with 99.39% true positive rate and only 1.25% False positive rate.

AK JAIN et al. [2] proposed survey to understand the history, current trends of attacks and Failure of various available solution. The defense techniques proposed in this survey are black-List, Data mining and heuristic, Machine Learning and Soft computing Algorithms.Machine Learning techniques gives the best results as compared to other techniques. It able to detect TP Upto 99%.

In [3]AK JAIN AND BB GUPTA has designed a phishing detection system is implemented in Java platform Standard edition 7(JDK 1.7) .In this paper author proposed a novel approach to protect against phishing attacks using auto-updated white-list of legitimate sites accessed by the individual user. Moreover, author proposed system is efficient to detect various other types of phishing attacks (i.e DNS poisoning, embedded objects, zero-hour attacks) and suitable for a real-time enviroment.

In [4] Waleed Ali has designed the wrapper-based features selection method which is used for Selecting the most significant fetures to be utilize in predicting the phishing website accurately. Author implemented many supervised machine learning algorithm like Back-Propagation Neural Network(BPNN), Support Vector Machine (SVM), Naive Bayes Classifier (NB), Decision Tree(C4.5) and Random Forest(RF) & K-Nearest Neighbour (KNN). Furthermore, The wrapper-based features selection can be used with ensemble learning to improve the performance of the intelligent phishing website detection techniques.

In [5]AK JAIN AND BB GUPTA, have proposed a novel anti-phishing method based on Search engine and source code based heuristic. Proposed phishing webpages detection method depends on source code and the URL of the website that is able to detect the webpages designed using HTML code only. Moreover, the proposed approach cannot detect the phishing attack if the phishing webpages are hosted on the hijacked or compromised legitimate website. Proposed method achieves 98.15% TPR and 0.05% FPR.

In [15]. S.Neelamegam & Dr.E.Ramaraj, proposed various Classification Algorithm used in data mining. Data Classification is a data mining technique used to predict group membership for data instances Various Classification Algorithm discussed are decision tree, Bayesian networks, k nearest neighbor classifier, Neural Network, Support vector machine.

Kulkarni et al[13] proposed a system to detect a phishing website using Novel Algorithm This detection algorithm can find out the maximum number of phishings URLs because it executes multiple tests such as Blacklist search Test, Alexa ranking test, and different URL features test. But this solution is effective only for HTTP URLs.

4.Proposed work

4.1 Overview of phishing detection approaches

Overview of phishing detection approaches proposed in the literature

Phishing detection approaches split into two classes:

1. User education based approaches

   User education approach aims to improve the capacity of Internet users in the detection of phishing attacks Internet users can be educated to distinguish the characteristics of phishing and legitimate emails and websites.

2. Software-based approaches

   1. Phishing blacklist: A blacklist contains the list of malicious domains, URLs and IP addresses. The blacklist needs to be the regular update from their source because thousands of fake websites launch every day.

   2.Visual similarity based Techniques:These techniques utilize various features to compute the similarity between websites like page source code, images,textual content, text formatting, HTML tags, CSS, website logo, etc. Most of the visual similarity based approaches compare the new website with previously visited or storedwebsites.Therefore, these techniques cannot detect the new phishing websites and produce high false negative rate.

   3.Search Engine based techniques: The search Engine (SE) based techniques extract identity features (e.g., title, copyright, logo, domain name, etc.) from the webpage and make use of the search engine to check the

legitimacy of webpage.

4.Machine Learning based techniques: These methods train a machine learning algorithm with some features that can distinguish a genuine website from the phishing one. In this, a website is declared as phishing, if the design of the websites matches with the predefined feature set. The performance of these solutions depends on features set, training data and classification algorithm. These features are extracted from various sources like URL, page source, website traffic, search engine, DNS, etc.

4.2 Design Objectives

Detection Accuracy: This technique aims to develop higher detection accuracy. The correct classification of legitimate websites must be higher and incorrect labeling of phishing webpages should be minimum.

Content Language: The detection method must identify the fake webpages written in any textual language.

Low prediction time: The technique must predict phishing webpage before disclosing the credential of the user.

Zero-hour detection: Filtering the malicious websitesshould not be independent from a particular sector (i.e., banking, e-commerce). An efficient method must identify all kind of fake websites.

Lightweight: The detection method must require minimum resources, low computation and memory consumption

4.3 Proposed Approach

There some below mentioned are the steps involved in the completion of this project-

- Collect dataset containing phishing and legitimate websites from the open source platforms.
- Write a code to extract the required features from the URL database.
- Analyze and preprocess the dataset.
- Divide the dataset into training and testing sets.
- Run selected machine learning and deep neural network algorithms like SVM, RandomForest, on the dataset.
- Write a code for displaying the evaluation result considering accuracy metrics.
- Compare the obtained results for trained models and specify which is better.

4.3.1 Data Collection

Legitimate URLs are collected from the dataset provided by University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html

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.

https://www.phishtank.com

4.3.2 Feature Selection

The following category of features are selected:

- Address Bar based Features
- Domain based Features

Address Bar based Features considered are:

- Domian of URL
- IP Address in URL
- '@' Symbol in URL
- Length of URL
- Redirection '//' in URL
- 'http/https' in Domain name
- Using URL Shortening Service
- Prefix or Suffix "-" in Domain

Domain based Features considered are:

- •DNS Record
- •Age of Domain
- •Website Traffic
- •End Period of Domain

- **All together 12 features are extracted from the dataset.**

4.3.3 MACHINE LEARNING MODELS

1. Decision Tree: The Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

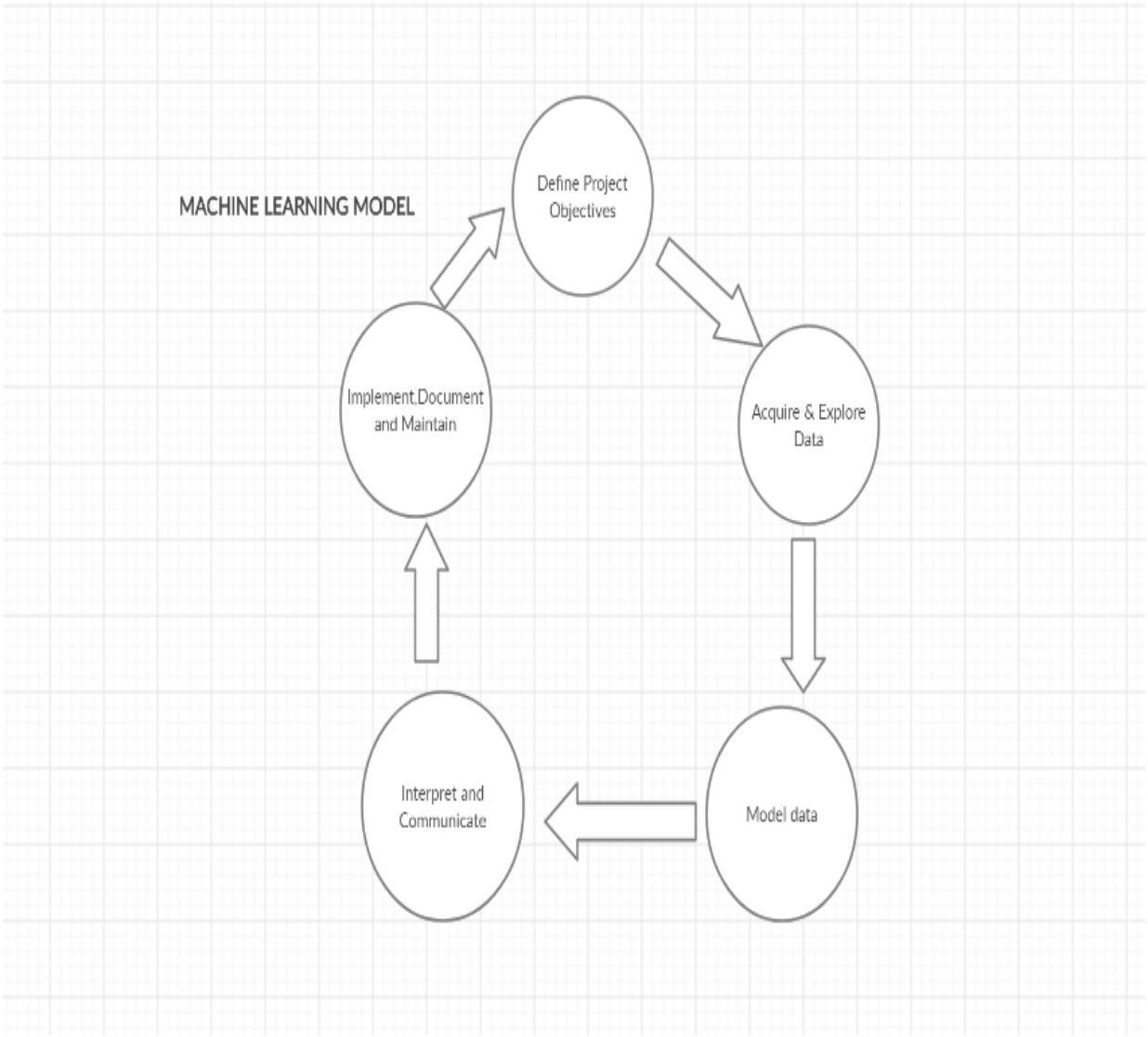
2.Random Forest: The Random Forest (RF) is another popular decision tree, which can be used for both classification and regression. RF is an ensemble of a number of decision trees independently trained on selected training datasets. The classification information is then determined by voting among all the trained decision trees. Therefore, Random Forest usually achieves a better classification accuracy compared to a single tree.

3.Multilayer Perceptrons:The Multilayer perceptrons (MLPs) are also known as (vanilla) feed-forward neural networks, or sometimes just neural networks. Multilayer perceptrons can be applied for both classification and regression problems. MLPs can be viewed as generalizations of linear models that perform multiple stages of processing to come to a decision.

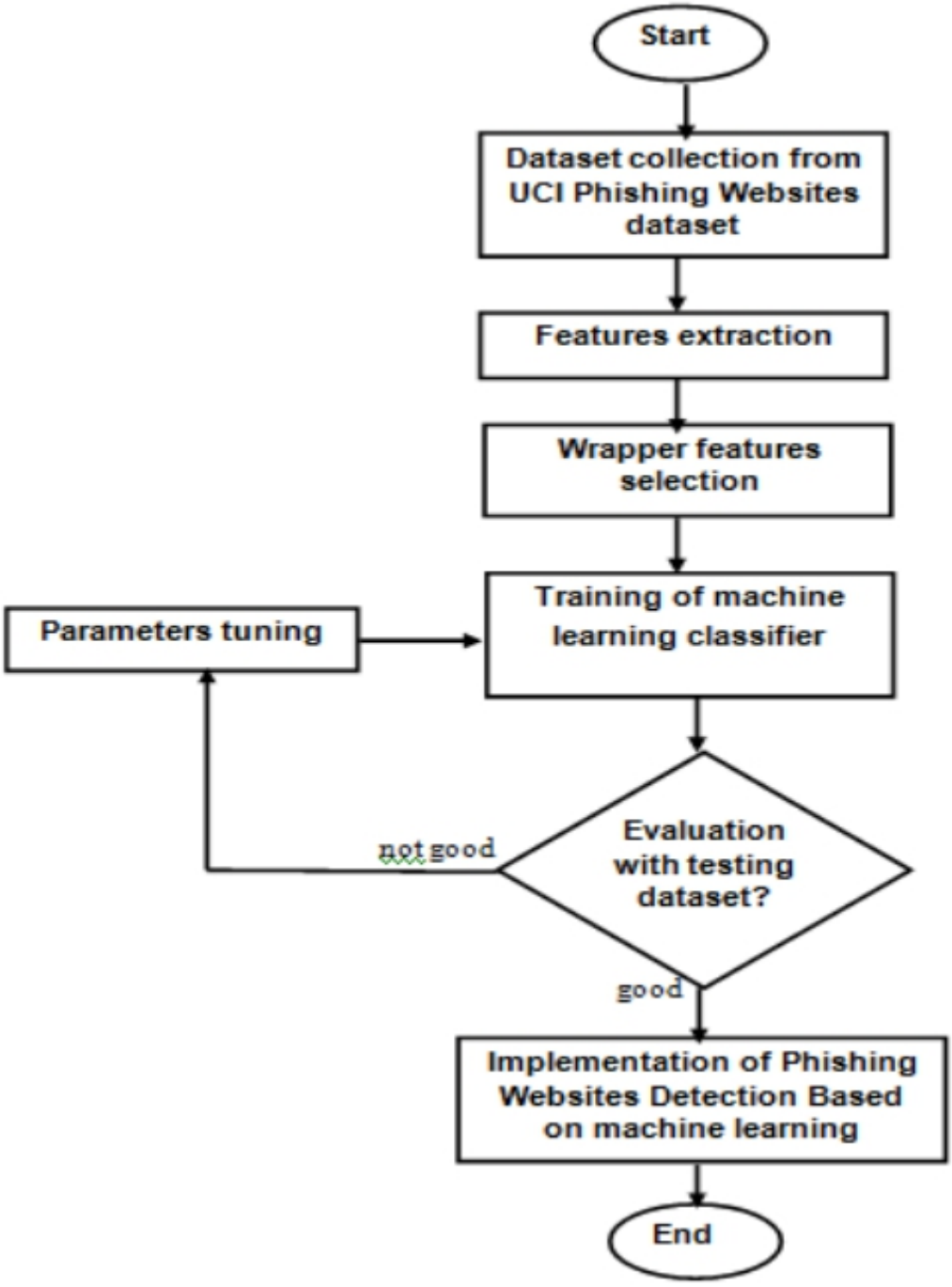
4.Support Vector Machines: Support vector machine (SVM) is one of the most well known and robust supervised machine learning techniques, which has been utilized effectively in many science and engineering applications. SVM is based on maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances to reduce an upper bound on the expected generalization error.

5.Data Flow Diagram

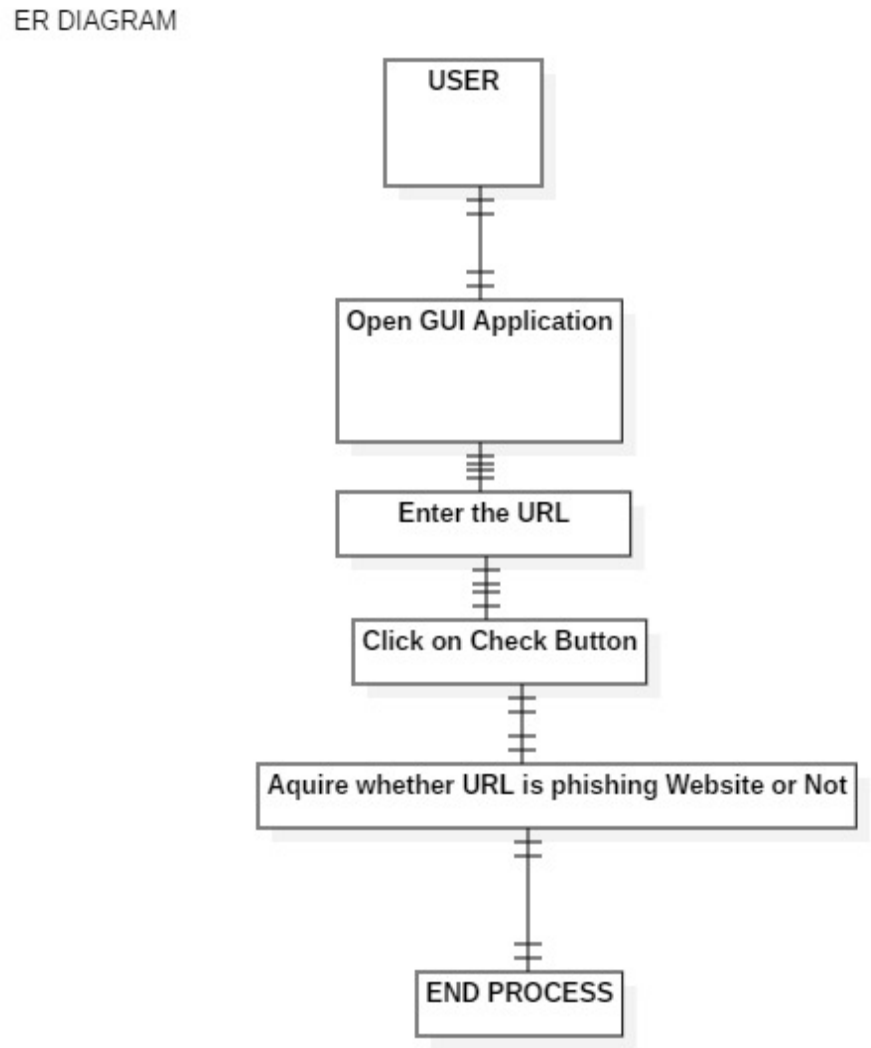
5.1 Level 0 DFD



5.2 Level 1 DFD



5.3 Level 2 DFD



6.Implementation Details

A laptop machine having intel core i5 9th Generation with 8 GB RAM is to used to implement the proposed approach is implemented using the python programming language.

Python offers vast support of its libraries, and it has a reasonable compile time.We have created the seperate function for each feature. These libararies can be installed individually using either the pip installer for python downloading and extracting them from official website. Following libraries that are used during execution of the code are-

Beutifulsoup: This library is used for pulling data from HTML and XML files.

urllib: This library is used to get response object from the URL, which extract all the resource from webpage.

whois: python-whois is a python model, use to see who is the registered owner of the domain name.

Pandas: Pandas is a library written for python for data manipulation and analysis.

NumPy: NumPy is a library in python for multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Matplotlib: Matplotlib is an amazing visualization library in python for 2D plots of arrays.

Seaborn: Seaborn is data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

7.Results & Observations

Below Figure shows the training and test dataset accuracy by the respective models:

	ML Model	Train Accuracy	Test Accuracy
2	Multilayer Perceptrons	0.864	0.850
1	Random Forest	0.821	0.814
3	SVM	0.802	0.802
0	Decision Tree	0.814	0.801

For the above it is clear that the Multilayer Perceptrons model gives better performance.

8. Conclusion & Future Plan

In this project we have proposed a novel machine machine learning approach to detect the legitimate website and phshing website. The proposed machine learning algorithm to detect the phishing website are - 1. Decision Tree 2. Multilayer Perceptron 3. Random Forest 4. Support Vector Machine(SVM). Our proposed methods achieves above 80% Train and Test accuracies. The most important way to protect to users from the phishing attacks is the education awareness. Internet users must be aware of all security which is given by the experts. Every users should also be trained they do not blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the websitetes.Working on this project is very knowledgeable and worth the effort. Through this project, one can know a lot about the phishing websites and how they are differentiated from legitimate ones.

This project can be taken further by creating a browser extensions of developing a GUI.The limitation of this project is we considered a small data set that contains only 10,000 URLs and there 12 features for each it can be extended further.

9.REFERENCES

[1]. A.K. Jain & B.B. Gupta, Towards detection of phishing websites on client-side using Machine learning based approach , Springer Science+Business, LLC, part of Springer Nature 2017.

[2]. A.K. Jain, B.B. Gupta, A Tewari & DP Agrawal, Fighting against phising attacks: Stae of the art and future challenges. The Neural Computing & Applications Forum 2016

[3]. A.K. Jain & B.B.Gupta,A novel approach to protect against phishing attack at client side Using auto-updated white-list, EURASIP Journal on Information Security (2016) 2016:9

[4]. Waleed Ali, Phishing website detection based on Supervised machine learning with Wrapper features selection, International Journal of Advanced Computer Science and Application - January 2017

[5]. A.K. Jain & B.B.Gupta, Phishing Attack using a Search Engine and Heuristics-based Technique, Journal of Information Technology Research Volume 13:Issue 2:Apr-June2020.

[6]. A.K. Jain & N. Choudhary, Comparative Analysis of Mobile Phishing Detection and Prevention Approaches, Springer International Publishing AG2018, ICTIS-2017 Volume1

[7]. A.K. Jain & D. Goel, Mobile phishing attacks and defence mechanisms: state of art and open research challenges, *Computers & Security* (2017), https://doi.org/10.1016/j.cose.2017.12.006.

[8]. A.K. Jain, S.K. Yadav & N. Choudhary, A Novel Approach to Detect Spam and Smishing SMS using Machine Learning Techniques, International Journal of E-Services and Mobile Applications Volume 12 • Issue 1 • January-March 2020

[9]. A.K Jain, D. Goel, S. Agarwal, Y. Singh, G. Bajaj, Predicting Spam Messages Using Back Propagation Neural Network, © Springer Science+Business Media, LLC, part of Springer Nature 2019 Wireless Personal Communications https://doi.org/10.1007/s11277-019-06734-y

[10]. The Phishing Guide Understanding & Preventing Phishing Attacks By: Gunter Ollmann, Director of Security Strategy, IBM Internet Security Systems, 2007

[11]. Legitimate URLs are collected from the dataset provided by University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html

[12]. Phishing URLs are collected from opensource service called PhishTank. https://www.phishtank.com

[13]. H. Sampat, M. Saharkar, A. Pandey & H. Lopes, Detection of Phishing Website Using Machine Learning, International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 03 | Mar-2018.

[14]. A. Kulkarni & L.L.Brown, Phishing Websites Detection using Machine Learning, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019.

[15]. S.Neelamegam & Dr.E.Ramaraj, Classification algorithm in Data mining: An Overview International Journal of P2P Network Trends and Technology (IJPTT) - Volume 3 Issue 5 September to October 2013

[16]. A.K. Jain & B.B.Gupta, Two-level authentication approach to protect from phishing attacks in real time, © Springer-Verlag GmbH Germany 2017, J Ambient Intell Human Comput. DOI 10.1007/s12652-017-0616-z

[17]. Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi. 2019. Malicious URL Detection using Machine Learning: A Survey. 1, 1 (August 2019), 37 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn