

# Malicious URL Detection using ML & AI

*Submitted in partial fulfilment of the  
requirements of the degree of*

## **BACHELOR OF ENGINEERING in INFORMATION TECHNOLOGY (A.Y. 2021-2022)**

by

**Aman Gupta (Roll No. 25)  
Dev Jindani (Roll No. 46)  
Manu Khandelwal (Roll No. 50)**

**Under the Guidance of  
Mr. Rahul Neve  
Assistant Professor, IT Department, TCET**



**Choice Based Credit Grading System with Holistic Student Development  
(CBCGS-H)**



*Thakur Singh Charitable Trust's (Regd.)*  
**THAKUR COLLEGE OF  
ENGINEERING & TECHNOLOGY**  
*Autonomous College Affiliated to University of Mumbai  
Approved by All India Council for Technical Education (AICTE) and Government of Maharashtra (GoM)*  
Conferred Autonomous Status by University Grants Commission (UGC) for 10 years w.e.f. A.Y. 2019-20  
Among Top 250 Colleges in the Country where NIRF India Ranking 2020 in Engineering College category  
• ISO 9001:2015 Certified • Programmes Accredited by National Board of Accreditation (NBA), New Delhi  
• Institute Accredited by National Assessment and Accreditation Council (NAAC), Bangalore

Website : [www.tcetmumbai.in](http://www.tcetmumbai.in)



# Malicious URL detection using ML & AI

*Submitted in partial fulfillment of the requirements*

*of the degree of*

## BACHELOR OF ENGINEERING *in* INFORMATION TECHNOLOGY (A.Y. 2021-2022)

by


Aman Gupta (Roll No.: 25)  
Dev Jindani (Roll No.: 46)  
Manu Khandelwal (Roll No.: 50)

Under the Guidance of  
**Mr. Rahul Neve**

Assistant Professor, I.T Department, TCET



### Choice Based Credit Grading System with Holistic Student Development (CBCGS-H)

 <p>Estd. in 2001</p>	<p><i>Zagdu Singh Charitable Trust's (Regd.)</i></p> <p><b>THAKUR COLLEGE OF ENGINEERING &amp; TECHNOLOGY</b></p> <p><i>Autonomous College Affiliated to University of Mumbai</i></p> <p><i>Approved by All India Council for Technical Education (AICTE) and Government of Maharashtra (GoM)</i></p> <p><i>Conferred Autonomous Status by University Grants Commission (UGC) for 10 years w.e.f. A.Y 2019-20</i></p> <p><i>Amongst Top 200 Colleges in the Country, Ranked 193<sup>rd</sup> in NIRF India Ranking 2019 in Engineering College category</i></p> <ul style="list-style-type: none"><li>• ISO 9001:2015 Certified • Programmes Accredited by National Board of Accreditation (NBA), New Delhi</li><li>• Institute Accredited by National Assessment and Accreditation Council (NAAC), Bangalore</li></ul>
--	---



## **Certificate**

This is to certify that Mr. Aman Gupta, Mr. Dev Jindani, and Mr. Manu Khandelwal are bonafide students of the Information Technology Department, Thakur College of Engineering and Technology, Mumbai. They have satisfactorily completed the requirements of PROJECT-II as prescribed by **Thakur College of Engineering and Technology (An Autonomous College affiliated to the University of Mumbai)** while working on “Malicious URL detection using ML & AI”.

Signature :-----

Name : Mr. Rahul Neve  
Assistant Professor

Signature :-----

Name : Dr. Bijith Markarkandy  
HOD-IT

Signature: -----

Name : Dr. B. K. Mishra  
Principal,  
Thakur College of Engineering and Technology.

**Internal Examiner:**

**External Examiner:**

Signature :-----

Signature :-----

Name :

Name :

Thakur College of Engineering and Technology, Kandivali(East) Mumbai.

Date:

Place:

## **Declaration**

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-----  
(Signature)

1.Aman Gupta (BE IT A-25)

2.Dev Jindani(BE IT A-46)

3.Manu Khandelwal(BE IT A-50)

Date:

# ACKNOWLEDGEMENT

We sincerely thank our guide **Mr. Rahul Neve** for his guidance and support in carrying out our project work. We also thank our guide for giving us new ideas to develop our project well. We also give our sincere thanks to our Principal Dr. B. K. Mishra, Vice-Principal Dr. Kamal Shah, HOD-IT, Dr. Bijith Marakarkandy, our project coordinators for arranging the necessary facilities to carry out the project work.

We thank Thakur College of Engineering and Technology for their support and for providing us a platform to explore new trending technologies in the world.

1.Aman Gupta(BE IT A-25)

2.Dev Jindani(BE IT A-46)

3.Manu Khandelwal(BE IT A-50)

# **ABSTRACT**

Since the beginning of the Internet, phishing has been a major issue. It's a classic Internet con. As a result, no antivirus or other technical security can entirely eliminate this threat. However, research is being undertaken by intellectuals all over the world to overcome this online threat. Different techniques to dealing with this challenge are being developed by researchers. Black Listing and Machine Learning are the two main concentrated tactics attempted by researchers to combat phishing. The use of machine learning to combat phishing is a novel and unique approach.

Malicious URLs are harmful to every aspect of computer users. Detecting of the malicious URL is very important. Currently, detection of malicious web pages techniques includes black-list and white-list methodology and machine learning classification algorithms are used. However, the black-list and white-list technology is useless if a particular URL is not in list. In this paper, we propose a comparative study between different algorithms of machine learning for detecting malicious URL and have also provided the outcomes. The algorithm used can directly determine if the URL provided is malicious or not. We also used an example to verify that the model can improve the accuracy of URL detection.

To overcome phishing, we have used Machine Learning and heuristic-based technique in this thesis. This thesis is a comparison of various Machine Learning algorithms such as Logistic Regression, SVM, Naïve Bayes, and Decision Tree. Each URL goes through a series of algorithms and is being detected whether it is malicious or not. If the URL is malicious our model will give an alert box while if it is safe to use it will redirect to the web page using the default browser.

# CONTENTS

- List of Figures i
- List of Tables ii

Chapter No.	Topic	Pg. No.
<b>Chapter 1</b>	<b>Overview</b>	<b>01</b>
	1.1 Introduction	
	1.2 Background	
	1.3 Importance of the Project	
	1.4 Perspective of stakeholders and customers	
	1.5 Objectives and Scope of the project	
	1.6 Summary	
<b>Chapter 2</b>	<b>Literature Survey &amp; Proposed Work</b>	<b>05</b>
	2.1 Introduction	
	2.2 Literature Survey Table	
	2.3 Problem definition (Phase wise)	
	2.4 Feasibility Study	
	2.5 Methodology used	
	2.5.1 Scrum/XP/Agile	
	2.5.2 Customer interaction details	
	2.6 Summary	
<b>Chapter 3</b>	<b>Analysis and Planning</b>	
	3.1 Introduction	
	3.2 Product Backlog or Sprint backlog	
	3.3 Project planning (Resources, Tools used, etc.)	
	3.4 Scheduling (Timeline chart or Gantt chart) according to sprint backlog	
	3.5 Summary	
<b>Chapter 4</b>	<b>Design and Implementation</b>	
	4.1 DFD (if applicable) or Kanban Chart	
	4.2 Block Diagram (if applicable)	
	4.3 Flow Chart (if applicable)	
	4.4 UML (if applicable)	
	4.5 GUI screenshot	
	4.6 Database screenshot	
<b>Chapter 5</b>	<b>Results &amp; Discussion</b>	
	5.1 Actual Results	
	a. Outputs (sprint wise)	
	b. Outcomes	
	c. Discussion of the results	
	5.2 Future Scope (further phases)	
	5.3 Testing	
	5.4 Deployment	
<b>Chapter 6</b>	<b>Conclusion</b>	
	6.1 Conclusion	



**References:**

Style of list in references of some standards are as below;

**[1] Text book references**

**[2] Journal references**

**[3] Web references**

**APPENDIX:**

**[A] Plagiarism check report:**

1 page plagiarism self- evaluation report. (Use Quetext or Plagscan for generating report)

**[B] Graduate Attributes and its mapping with the project**

**[C] Copy of your technical paper published in any journal or conference**

## **List of Figures**

Figure 1 : Timeline Chart.

Figure 2 : Process of proposed detection method.

Figure 3 : Features used in thesis.

Figure 4 : System flow.

Figure 5 : UML diagram .

Figure 6 : GUI screenshot.

Figure 7 : Training Dataset.

Figure 8 : Code Implementation.

Figure 9 : Logistic Regression Confusion matrix.

Figure 10 : Logistic regression ROC curve.

Figure 11 : Logistic regression classification report.

Figure 12 : Comparative study.

Figure 13 : Testing Output.

Figure 14 : Feature Extraction of testing data.

Figure 15 : Deployment.

## **List of Tables**

Table 1 : Key finding and Identification of research gap based on literature survey

Table 2 : Libraries used in our project

# **Chapter 1: Overview**

# Chapter 1: Overview

## 1.1 Introduction

Phishing websites are used to steal personal information like credit cards and passwords, as well as to do drive-by downloads. Muggers like phishing because it is easier to deceive someone. In most situations, such obnoxious behaviour diverts network resources intended for other purposes into visiting a malicious link that appears legitimate rather than attempting to breach a computer's defences. In most circumstances, such obnoxious behaviour interferes with network attributes intended for other purposes, and it almost always jeopardises the network's and/or data's security. Intruders can be deterred by properly developing and installing a Phishing URL. qualities of a phishing domain (or a fraudulent domain), the traits that distinguish them from acceptable domains, why it's crucial to recognise these domains, and how machine learning techniques may be used to detect them.[1]

## 1.2 Background

Internet security is a broad topic of computer science that encompasses a variety of approaches for dealing with online and offline dangers such as those posed by flash drives, hard drives, and other storage devices. Many other types of assaults against an Internet user's personal credentials may be confiscated over the Internet, including viruses, Trojans, and other malware. With the advancement of antivirus and anti-malware software, our security science has progressed significantly to address these issues. These risks have been much reduced thanks to these antiviruses. But, in addition to this, there is another type of Internet assault or danger that is still the least known and most vulnerable. Since the creation of the Internet, phishing has been a major issue. Phishing had a large influence on the world's economy at the period of the Internet's development, resulting in major business, personal, and government losses. It was a major danger to the Internet at the time, and it had an impact on the actual world. Phishing has been a major issue since the dawn of the Internet and continues to be so today. Antivirus software cannot detect or remove any virus, worm, or Trojan. It's a type of Internet fraud that uses social engineering to get individuals to divulge their credentials online (Bajaj & Hansen, 2008).[3] As a result, no antivirus or other technological security can entirely eliminate this threat. However, research is being undertaken by intellectuals all over the world to counter this internet fraud. Researchers are coming up with a variety of solutions to cope with this issue. Black Listing and Machine Learning are two of the most common approaches used by researchers to combat phishing. The use of machine learning to combat phishing is a novel and unique approach. Machine Learning may one day be able to entirely combat phishing. Machine Learning concepts are also used in our study.

### **1.3 Importance of the Project**

From the view from above, it is clear that phishing is not an issue to be treated lightly. It's significantly more hazardous than any virus or malware since no antivirus software can identify it; it's simply social engineering. To combat this problem, IT businesses and researchers must pay special attention. Phishing is a problem that many researchers overlook throughout the world. Different researchers have undoubtedly conducted studies on phishing. However, no solid strategy for combating phishing has yet been created. As a result, we chose to work with a theme.

### **1.4 Perspective of stakeholders and customers**

The creators of the project have to ensure the efficiency and the robustness of the project and the customers expect that they are well protected from such malicious content and are not getting scammed in any way possible.

### **1.5 Objectives and Scope of the project**

The objective of this project is to develop a software which can detect newly generated malicious URLs by using the database and their common features using ML and AI applications. The Aim is also to detect as many common features between the malicious url so as to make it easy for detecting in the future.

The proposed system can be installed in any PC or Laptop. With the help of this system users can be saved from malpractices that occur due to clicking on these malicious urls and will help the user to keep their data safe.

### **1.6 Summary**

The problem statement was written and extensively researched, and the project's scope was established to drive the project's flow and requirements.

## **Chapter 2: Literature Survey & Proposed Work**

# Chapter 2: Literature Survey & Proposed Work

## 2.1 Introduction

A literature review is a study – or, more precisely, a survey – that uses scholarly material to discuss previously published information on a particular topic or research question. As a result, in order to produce a literature review, you must be an expert in the subject of study. The discoveries and results will be published and made available to the general public, specifically scientists working in the same field.

## 2.2 Literature Survey Table

Table 1 : Key finding and Identification of research gap based on literature survey

Ref No.	TITLE OF PAPER	AUTHOR, YEAR	METHODOLOGY	KEY FINDINGS	RESEARCH GAP
1.	A system to detect malicious URLs by using lexical features..	Mohammed et al., 2009	The creators Due to speed and lightweight processing, they primarily exploited lexical features from URL.	<ol style="list-style-type: none"><li>1. For feature selection, they used correlation-based feature selection (CFS) and information gain. To identify the URLs, a few machine learning techniques were applied, including KNN, Random Forest, and C4.5.</li><li>2. Due to speed and lightweight processing, they primarily exploited lexical features from URL.</li></ol>	<p>They looked at two datasets and put generic features to the test.</p> <p>Despite the fact that these two datasets contain distinct variations of malicious URLs, they can still be used to further check our obtained generic features by comparing them to another dataset from a different source.</p>



2.	Detection of Malicious URLs Using Deep Learning Approach	Rajni Kushwaha, 2019	They devised a method that takes a URL string as input and applies CNNs to both the characters and words in the URL. Identify unique characters in the training corpus and represent each character as a vector for character-level CNNs.	<p>The full URL (a string of characters) is converted to a matrix representation, which may then be convolutioned. Character CNNs identify vital information from groupings of characters that appear together, which could indicate malice. Word-level CNNs start by identifying unique words in the training corpus, free of special characters. They get a matrix representation of the URL by using a wordembedding matrix (which in this context, is a sequence of words). After that, convolution can be used.</p>	<p>Their methodology is deficient in the analysis and detection of obfuscated JavaScripts in Webpages, which is the primary cause of attacks such as drive-by downloads, XSS, malware-deliver, and so on.</p> <ul style="list-style-type: none"> <li>• There is a need to investigate more discriminative spam URL features to distinguish them efficiently from benign URLs;</li> <li>• There is a need to investigate more features of short URLs for effective detection and attack type identification, because it is the most popular trend today on microblogging sites such as Twitter, Facebook, and others.</li> </ul>
3.	Detection and analysis of Malicious URLs	B Murali,2015	They gathered a significant number of long and short URLs from various SNS sources, which were then tested against malicious and non-malicious detectors, and their properties were analysed to classify the URLs.	<p>. It begins with a data collecting model in which data is gathered from Facebook and Twitter, followed by the extraction of URLs and their features, labelling, feature extraction, classification, and the outcome of the classification algorithm. They used the online detection method described above to identify our gathered URLs, which included both long and short URLs, as malicious or non-malicious. We</p>	<p>To determine some characteristics of URLs using web resources such as Virus Total, Phish Tank, and Bit.ly. They can't vouch for the legitimacy of these services. They may update their evaluation approach in the future to detect rogue URLs.</p>

				also examined the URLs against a continually updated list of phishing and malware on public blacklists such as McAfee, SiteAdvisor, URIBL, and SURBL.	
4.	A Malicious URL Detection Model Based on Convolutional Neural Network	Tao Yang, Shuhao Li, 2021	This work presents a DCNN-based malicious URL detection model. To extract features automatically and understand the URL's expression, it uses word embedding based on the character embedding method.	This work presents a DCNN-based malicious URL detection model. To the initial multilayer convolution structure, the dynamic convolution method adds a new folding layer. The k-max-pooling layer takes the role of the pooling layer. We also conduct three contrast experiments, demonstrating that using a network structure comprised of a DCNN and various fields derived from the URL can produce a superior result.	By designing these features, attackers can escape being identified by current detection approaches, making it extremely difficult to maintain a detection system based on typical machine learning. Furthermore, when detecting fraudulent URLs on a broad scale, a trained model may lose some essential information from the URL.
5.	Convolutional neural networks for sentence classification	K Yoon, 2014	It classifies URLs generated by DGA using a cyclic neural network model at the character level. Extreme machine learning is a method proposed in the literature for detecting malicious URLs.	A brand-new Literature uses character-level semantic information to detect whether DGA creates the URL by combining an n-gram model with deep learning. The literature lists a number of deep learning architectures for malicious URL identification, including single-layer long short term memory (LSTM), bidirectional LSTM, the combined structure of CNN and LSTM, and deep convolution structure contextual meaning.	One of its drawbacks is that they employ a preset CNN structure to recognise URLs. Because the model parameters cannot be modified according to the input vector's dimension, extracting in-depth features over a large range is challenging.

6.	A systematic review of insider threat detection	K. Aram and J. O. SoK, 2019	DNNs (Deep Neural Networks) are artificial neural networks with numerous hidden layers between the input and output layers. Through multiple hidden layers, DNN can simulate complex non-linear connections.	Deep learning allows models to extract features automatically. Deep learning is a type of complex function approximation that uses a linear combination of weights, neuron values, and a nonlinear activation function to approximate a complex function.	Although this method is simple and effective, it is limited in that it cannot detect newly produced malicious URLs. According to the literature, attackers can use a random seed to construct a variety of malicious domain names, thereby evading the typical blacklist detection method.
----	---	-----------------------------	--	---	--

## 2.3 Problem definition

Typically, malicious links are used to lure a victim into clicking through to a payload that is hosted on third-party sites rather than the malicious content being directly available from the social media platform. One-click attacks, such as those used for account takeover, could be easily propagated through social media and, once clicked, might abuse the victim in terms of profile takeover or misguiding people for fraudulent adverts.

## 2.4 Feasibility Study

**Operational Feasibility:** The only thing a user needs during visiting the website is a computer and a connection to our servers. There won't be the need for special equipment to ensure the working of the website, which makes the whole process hassle free.

**Economic Feasibility:** The program can be executed on a computer of normal configuration of at least 4 GB RAM. The project can be executed independently of other programs or modules and there is no need to install any extra dependent modules or programs to run this project without facing overhead or load on the system.

**Technical Feasibility:** The internal working and complexities of the project are hidden from the user and all input/outputs can be executed without any technical knowledge, which is a critical aim of our project. All the algorithms and techniques are optimized to yield best and fast possible results.

**Legal Feasibility:** All the programs are developed using either open source software or freely available framework without any licensing issues. All the frameworks used to develop our project allows us to use and distribute our application freely and can be used and integrated anywhere free of cost.

**Social Feasibility:** This project can be helpful for implementing the 'studies should not be affected due to the pandemic' concept in a small form as well as large scale.

## **2.5 Methodology used**

### **2.5.1 Agile Methodology**

The Agile software development technique is one of the most straightforward and efficient methods for transforming a vision for a business need into software solutions. Continuous planning, learning, improvement, team collaboration, evolutionary development, and early delivery are all terms used to describe agile software development methodologies. It increases adaptability in the face of change.[4]

The four essential values of agile software development are highlighted.

1. Interactions between individuals and teams over processes and tools
2. Working software trumps thorough documentation.
3. Collaboration with customers rather than contract negotiations
4. Adapting to change in accordance with a strategy

The agile methodology is a very easy-to-follow methodology where the project is broken down into several step-by-step modules to achieve results. We used agile methodology to carry out our tasks phase-wise: research, designing, development, testing, and additional improvements.

### **2.5.2 Customer interaction details**

Customers play an active role in applications designed for them. The visually impaired user should be familiar with the computer system that has been created for their comfort. The system will just facilitate the verification process. For our current application. A survey in the form of google forms was circulated among people of different age groups ranging from 17 to 51. After collecting the responses, we have observed that most people are aware of Malicious Urls and how it can be used to harm people. The questions which were asked were mostly about if the person has ever changing me accross any malicious URL and if yes then what were the circumstances after clicking on the website. We also asked people if they recognised any common feature on those malicious URL and also took feedback for the interface of the desktop application.

## **2.6 Summary**

With the help of literature survey, the research gaps were identified from the existing papers/system. The research gaps identified has led to the problem definition of our project. The problem definition was discussed in two phases.

**Phase 1:** Included planning, analysis, design and implementation where the project plan, requirement gathering, the layout of the system and the coding technique to be used was elaborated.

**Phase 2:** Included Coding, Testing and deployment where and how the system will be tested using various cases and how the system will be deployed was discussed.

## **Chapter 3: Analysis and Planning**

# Chapter 3: Analysis and Planning

## 3.1 Introduction

Traditional wisdom is that planning and analysis are very important and the more there is in a project, the more successful the project will be. Time spent on these activities will reduce risk and increase project success. For our project, A database will be embedded in the Malicious URL detector software which will contain all the previously registered malicious urls which will protect/warn the user's system to not enter into any threats. For detecting newly generated url we used different machine learning algorithms and by training our model using the available dataset. took the dataset from PhishTank to train our model.

## 3.2 Product Backlog or Sprint backlog

Product Backlog is the systematic prioritized task listing so that the work is performed efficiently. The project was carried out in the following manner:

1. Research and planning
2. First we have to develop our dataset for training our model using machine learning algorithms such as SVM, Logistic regression, etc.
3. Feature extraction from the database
4. Testing different algorithms on the current training dataset.
5. Final app development.
6. Maintenance and improvements

## 3.3 Project planning (Resources, Tools used, etc.)

Hardware Requirements:

4GB RAM.

10 GB HDD.

Intel 1.66 GHz Processor Pentium 4 processor

Software Requirements:

Windows 7 and above

Python 3.6.3

Programming Prerequisites:

Python( Version 3.6.3)

Visual Studio CODE

Python Environment setup:

The project was implemented using python language, so Visual Studio Code was used.

It comes with a pre-installed set of environments. So Visual Studio Code was used for the overall development of this research.

### Libraries Used:

Table 2. Libraries used in our project

pandas	import pandas as pd
numpy	import numpy as np
Confusion Matrix	from sklearn.metrics import confusion_matrix
sklearn	Import sklearn as sk
matplotlib	import matplotlib.pyplot as plt
wxPython	import wx
Support Vector Machine	from sklearn import svm
metrics	from sklearn import metrics
preprocessing	from sklearn import preprocessing
seaborn	import seaborn as sns

### 3.4 Scheduling (Timeline chart or Gantt chart) according to sprint backlog

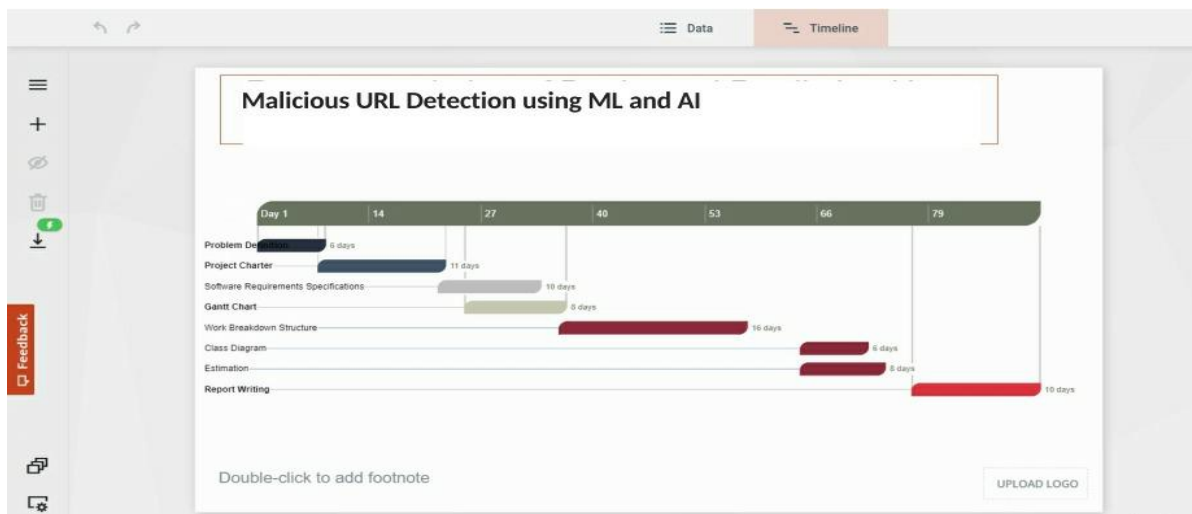


Figure 1. Timeline Chart

### **3.5 Summary**

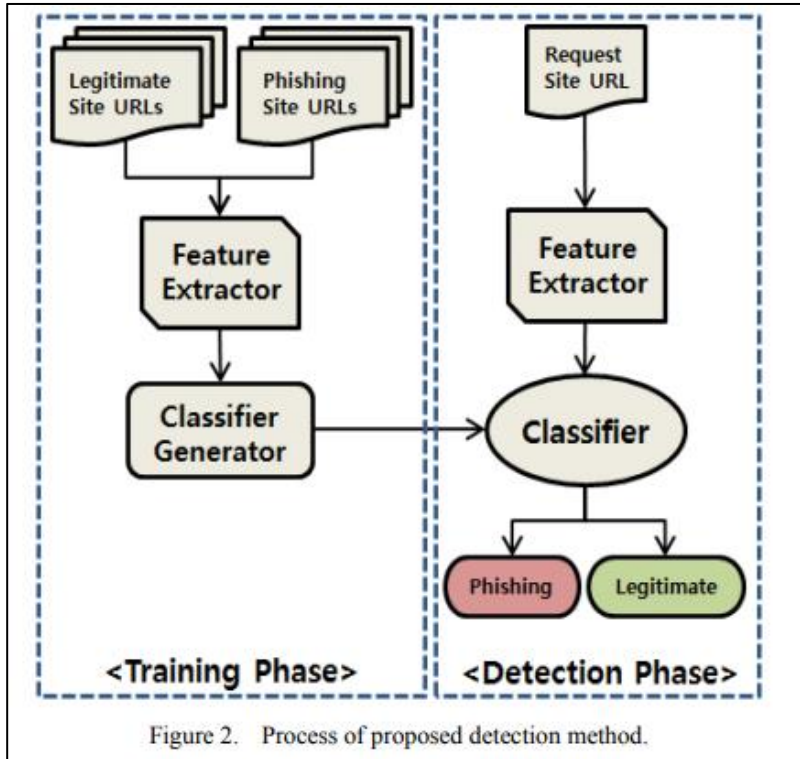
With analysis and planning, we are able to set time duration in which we would complete each module of our project. Analysis helps us take advantage and be informed of the gaps in previous works done by others. With proper planning we would be able to complete the project as per the given time period.



## **Chapter 4: Design and Implementation**

# Chapter 4: Design and Implementation

## 4.1 Block Diagram



No	Feature name	Description
1	IP address	Whether domain is in the form of an IP address
2	Length of URL	Length of URL
3	Suspicious character	Whether URL has _@', _//
4	Prefix and suffix	Whether URL has _'
5	Length of sub domain	Length of sub domain
6	Number of _/'	Number of _/' in URL
7	HTTPS protocol	Whether URL use https.

Figure 3. Set of features used for this thesis

## 4.2 Flow chart

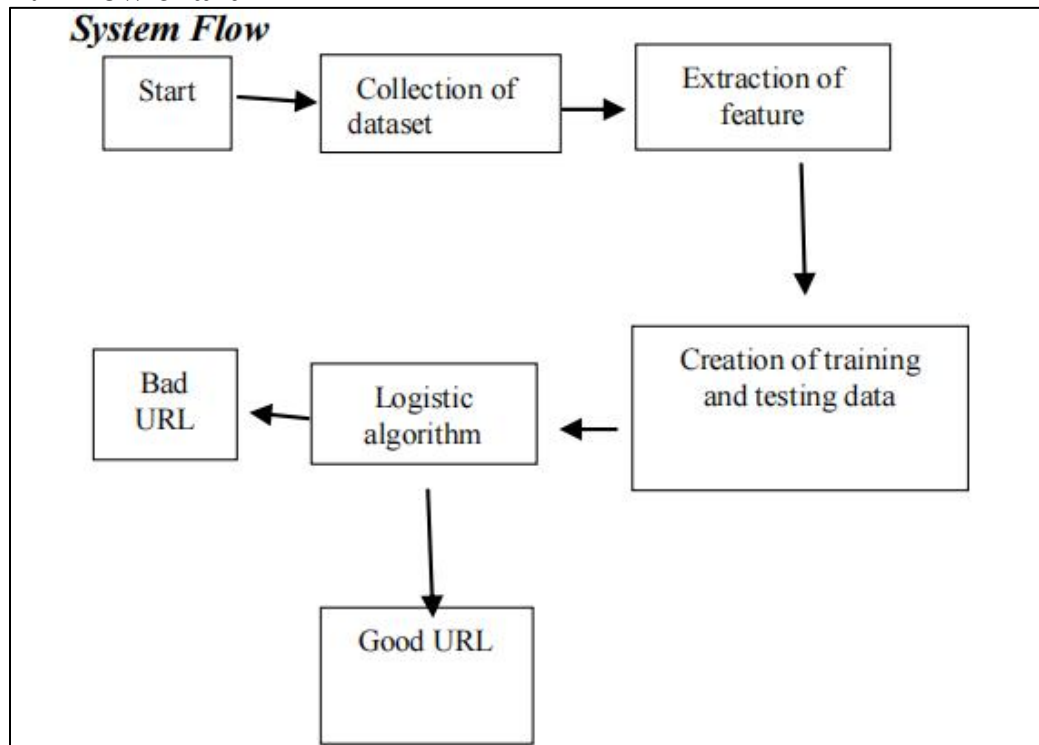


Fig 4. System Flow

## 4.3 UML

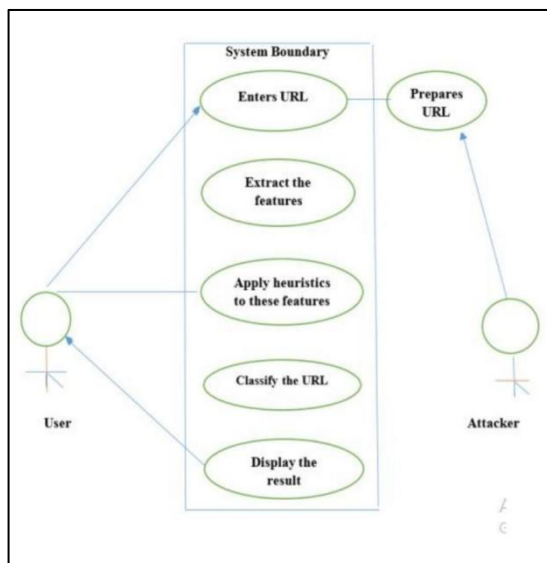


Figure 5. UML Diagram

## 4.4 GUI screenshot

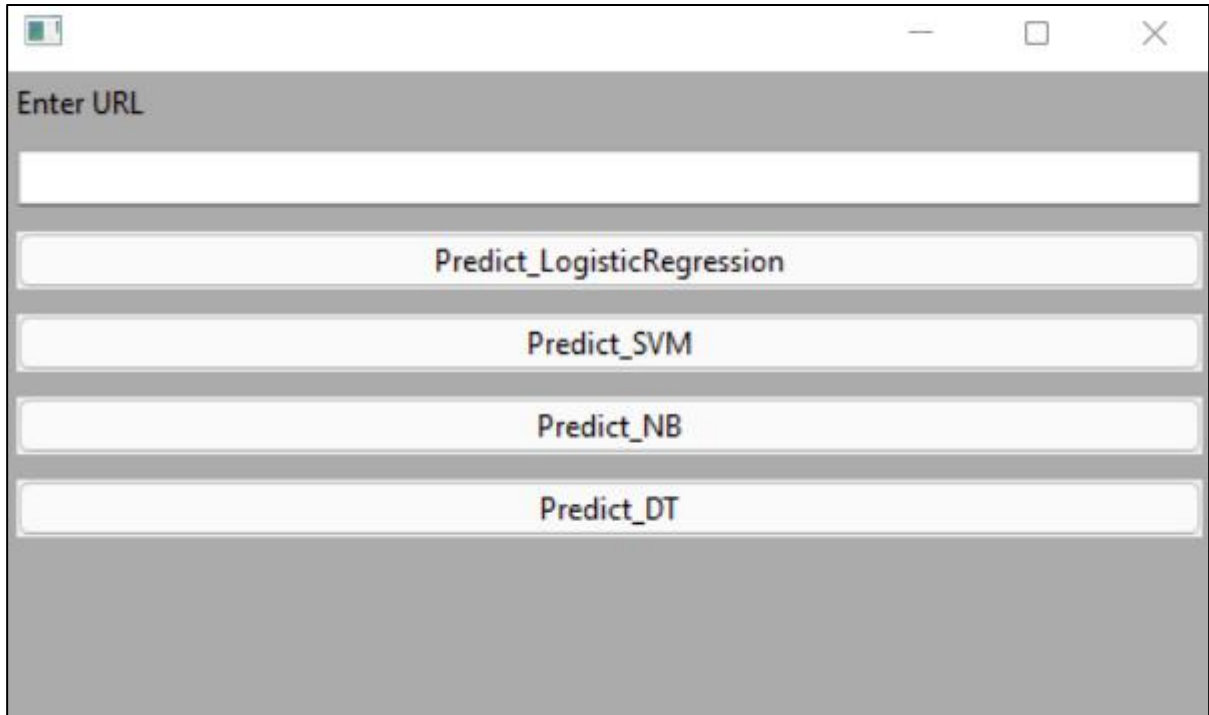


Figure 6. GUI Screenshot

## 4.5 Dataset screenshot

	A	B	C	D	E
1	phish_id	url	online	submissio	phishing
2	4912855	<a href="https://hotelsundram.com/EserviceMain/secure/irs/ir/index.html">https://hotelsundram.com/EserviceMain/secure/irs/ir/index.html</a>	yes	2017-03-2	yes
3	4912845	<a href="https://dice-profit.top/EserviceMain/irs/ir/index.html">https://dice-profit.top/EserviceMain/irs/ir/index.html</a>	yes	2017-03-2	yes
4	4912843	<a href="https://glprinters.com/EserviceMain/irs/ir/index.html">https://glprinters.com/EserviceMain/irs/ir/index.html</a>	yes	2017-03-2	yes
5	4912627	<a href="http://www.veflat.com/Jhoubert/bod.php/">http://www.veflat.com/Jhoubert/bod.php/</a>	yes	2017-03-2	yes
6	4912460	<a href="https://3mtoyoo.000webhostapp.com/">https://3mtoyoo.000webhostapp.com/</a>	yes	2017-03-2	yes
7	4912387	<a href="http://107-198-79-168.lightspeed.nsvltn.sbcglobal.net/login.html?assets.adobedtm.com/5165c8c319825f6ec3fb78d0a8dcc1054ab35a3d/satelliteLib-08b84ffc82250dd93a29554">http://107-198-79-168.lightspeed.nsvltn.sbcglobal.net/login.html?assets.adobedtm.com/5165c8c319825f6ec3fb78d0a8dcc1054ab35a3d/satelliteLib-08b84ffc82250dd93a29554</a>	yes	2017-03-2	yes
8	4912307	<a href="http://www.aol.anmolsecurity.com/secure/aol_update.php">http://www.aol.anmolsecurity.com/secure/aol_update.php</a>	yes	2017-03-2	yes
9	4912205	<a href="https://www.cadunico.com.br">https://www.cadunico.com.br</a>	yes	2017-03-2	yes
10	4912180	<a href="http://qmelskkjuw98-001-site1.1tempurl.com/Multi/">http://qmelskkjuw98-001-site1.1tempurl.com/Multi/</a>	yes	2017-03-2	yes
11	4912177	<a href="http://tinyurl.com/m9u7jny">http://tinyurl.com/m9u7jny</a>	yes	2017-03-2	yes
12	4912175	<a href="http://www.rollencenter.eu/wells/wells3/index.htm">http://www.rollencenter.eu/wells/wells3/index.htm</a>	yes	2017-03-2	yes
13	4912076	<a href="http://processa.ind.br/Pessoa.Fisica06/ativar.cadastr0s/chave.desatualizada/cliente.preferencial/pendente.atualizacao/comunicado/index1.php">http://processa.ind.br/Pessoa.Fisica06/ativar.cadastr0s/chave.desatualizada/cliente.preferencial/pendente.atualizacao/comunicado/index1.php</a>	yes	2017-03-2	yes
14	4912073	<a href="http://troycarstar.com/se@sess/">http://troycarstar.com/se@sess/</a>	yes	2017-03-2	yes
15	4912071	<a href="http://tkshipyard.com/images/program/index2.htm">http://tkshipyard.com/images/program/index2.htm</a>	yes	2017-03-2	yes
16	4912065	<a href="http://hairbeautyfurniture.com.au/image/2/update.htm">http://hairbeautyfurniture.com.au/image/2/update.htm</a>	yes	2017-03-2	yes
17	4911990	<a href="http://grupomomex.com.mx/mxenlinea.com/Home/portal/">http://grupomomex.com.mx/mxenlinea.com/Home/portal/</a>	yes	2017-03-2	yes
18	4911989	<a href="http://loogin-updates.esy.es/payment/recovery-checkpoint-login.html">http://loogin-updates.esy.es/payment/recovery-checkpoint-login.html</a>	yes	2017-03-2	yes
19	4911984	<a href="http://csfb2017kvfln.esy.es/secure/recovery-checkpoint-login.html">http://csfb2017kvfln.esy.es/secure/recovery-checkpoint-login.html</a>	yes	2017-03-2	yes
20	4911983	<a href="http://secure-account-information-update-your-payment-method.goldennimages.com/SignIn/">http://secure-account-information-update-your-payment-method.goldennimages.com/SignIn/</a>	yes	2017-03-2	yes
21	4911980	<a href="http://recovery-continue2017.esy.es/recovery-checkpoint-login.html">http://recovery-continue2017.esy.es/recovery-checkpoint-login.html</a>	yes	2017-03-2	yes
22	4911971	<a href="https://srv220.prodns.com.br/~contagemcro/includes/folterr/sp/index.php?passo=home">https://srv220.prodns.com.br/~contagemcro/includes/folterr/sp/index.php?passo=home</a>	yes	2017-03-2	yes
23	4911970	<a href="http://secure-validaton.com.sicconingenieros.com/red1/index.php">http://secure-validaton.com.sicconingenieros.com/red1/index.php</a>	yes	2017-03-2	yes
24	4911969	<a href="http://cekpoint-loogin900.esy.es/recovery-checkpoint-login.html">http://cekpoint-loogin900.esy.es/recovery-checkpoint-login.html</a>	yes	2017-03-2	yes
25	4911968	<a href="http://phoenixwebwiz.com/aaccv/aol.htm">http://phoenixwebwiz.com/aaccv/aol.htm</a>	yes	2017-03-2	yes
26	4911965	<a href="http://bridgesforwomen.org/aol/index.htm">http://bridgesforwomen.org/aol/index.htm</a>	yes	2017-03-2	yes
27	4911963	<a href="http://www.ccregypte.com/ar/modules/mod_breadcrumbs/validaaoaodia/">http://www.ccregypte.com/ar/modules/mod_breadcrumbs/validaaoaodia/</a>	yes	2017-03-2	yes
28	4911962	<a href="http://cekpoln-looginn65.esy.es/recovery-checkpoint-login.html">http://cekpoln-looginn65.esy.es/recovery-checkpoint-login.html</a>	yes	2017-03-2	yes
29	4911917	<a href="http://gsbzvxx.online/dd/A0L.htm">http://gsbzvxx.online/dd/A0L.htm</a>	yes	2017-03-2	yes
30	4911881	<a href="http://cookstruck.com/aol.htm">http://cookstruck.com/aol.htm</a>	yes	2017-03-2	yes
31	4911776	<a href="http://www.tudoazull.com/home/login.htm">http://www.tudoazull.com/home/login.htm</a>	yes	2017-03-2	yes
32	4911775	<a href="https://elight-photo.com/uta.edu/office/">https://elight-photo.com/uta.edu/office/</a>	yes	2017-03-2	yes
33	4911770	<a href="http://re-confirms-7.esy.es/recovery-checkpoint-login.html">http://re-confirms-7.esy.es/recovery-checkpoint-login.html</a>	yes	2017-03-2	yes
34	4911769	<a href="https://contasinativas.org/caixa.gov.br/pages/inter/index.php">https://contasinativas.org/caixa.gov.br/pages/inter/index.php</a>	yes	2017-03-2	yes
35	4911767	<a href="http://anades.com.br/Pessoa.Fisica05/ativar.cadastr0s1/chave.desatualizada/cliente.preferencial/pendente.atualizacao/comunicado/index1.php">http://anades.com.br/Pessoa.Fisica05/ativar.cadastr0s1/chave.desatualizada/cliente.preferencial/pendente.atualizacao/comunicado/index1.php</a>	yes	2017-03-2	yes
36	4911762	<a href="http://www.sicredativacao@online.com.br">http://www.sicredativacao@online.com.br</a>	yes	2017-03-2	yes

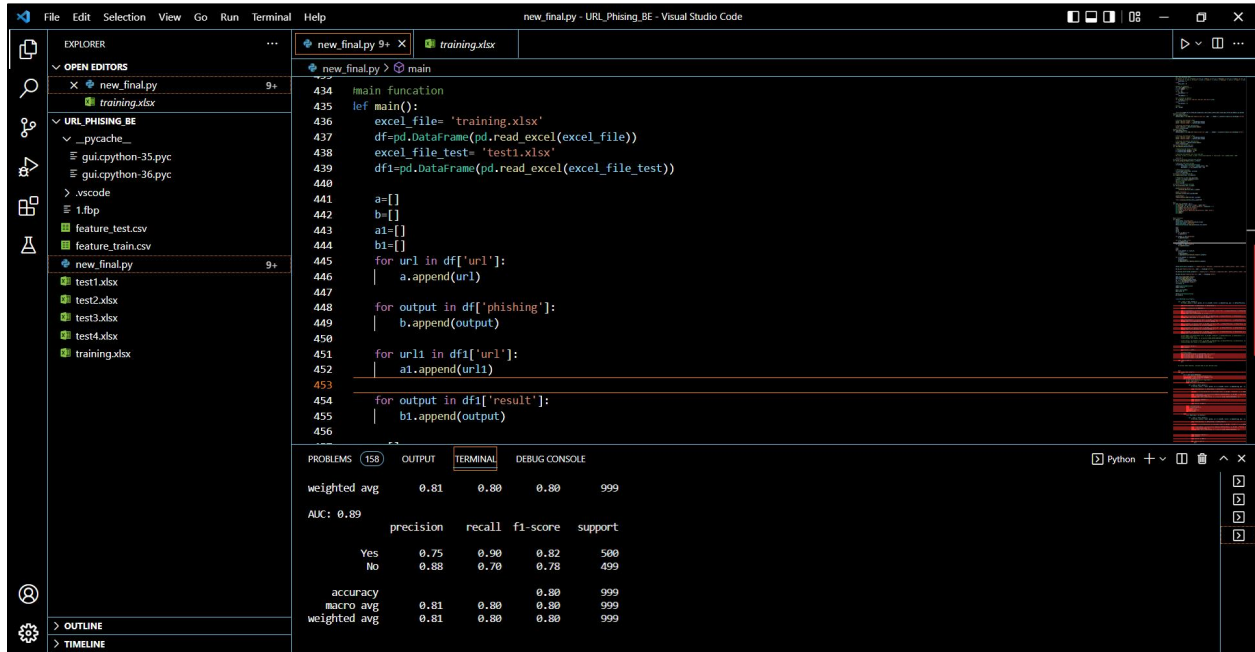
Figure 7. Training Dataset

## **Chapter 5: Results & Discussion**

# Chapter 5: Results & Discussion

## 5.1 Actual Results

### a. Outputs



The screenshot shows the Visual Studio Code interface with a Python file named `new_final.py` open. The code implements a main function that reads training and test data from Excel files, processes URLs, and calculates various metrics. The terminal output displays the following results:

```
weighted avg    0.81    0.80    0.80    999

AUC: 0.89
precision    recall    f1-score    support
Yes          0.75     0.90     0.82     500
No           0.88     0.70     0.78     499

accuracy          0.80    999
macro avg         0.81    999
weighted avg      0.81    999
```

Figure 8. Code Implementation

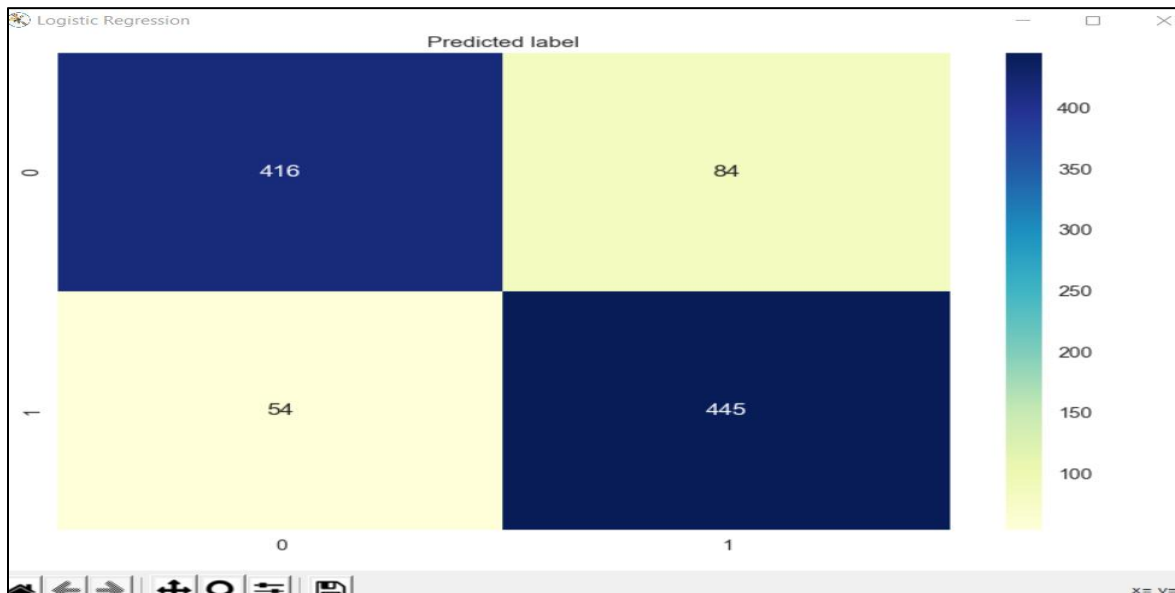


Figure 9. Logistic Regression Confusion Matrix

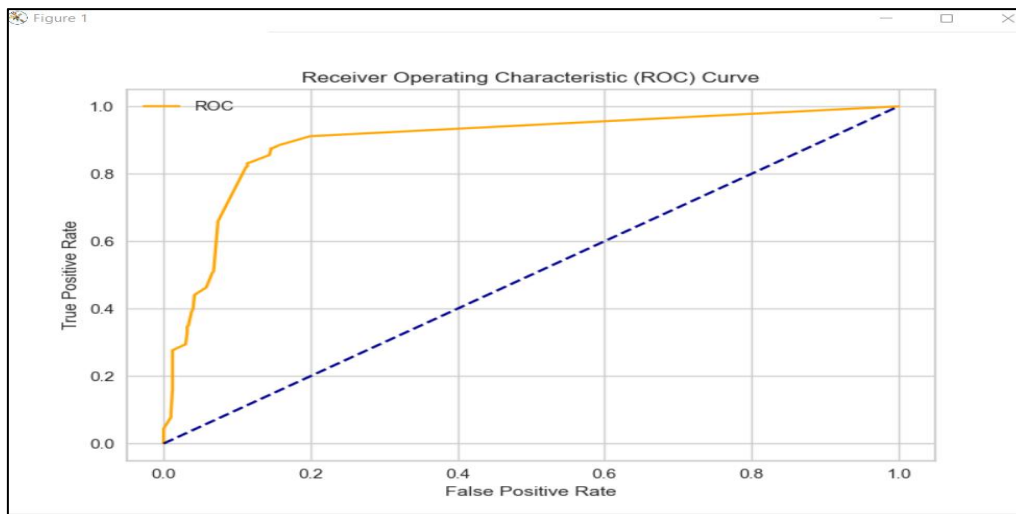


Figure 10. Logistic Regression ROC Curve

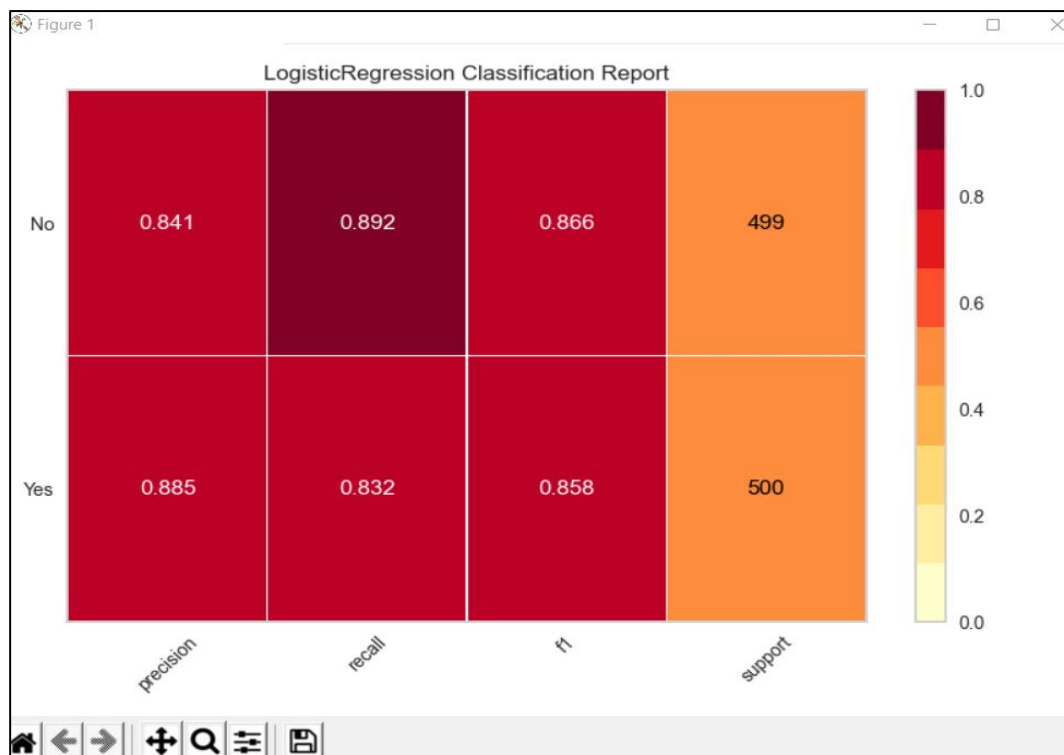


Figure 11. Logistic Regression Classification Report

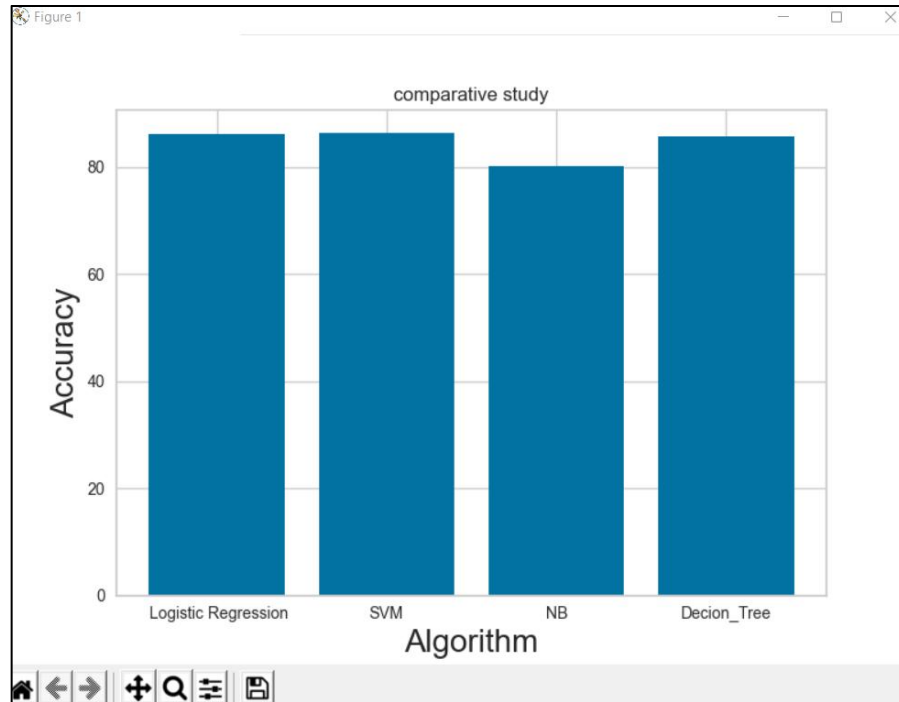


Figure 12. Comparative Study

## b. Outcomes

We have created our own model in which we have taken four machine algorithm namely Logistic Regression,SVM,Decision Tree and Naive Bayes.All the algorithm are trained and are giving high accuracy rate when tested with test dataset.

## c. Discussion of the results

The user will test URL by pasting it in the text box and selecting any of the machine learning algorithm.On click of the algorithm the model will test the URL and give the output as ‘Legitimate’ and redirect to the url page or alert the user if the URL is malicious by the message ‘Phishing’.

## 5.2 Future Scope (further phases)

Many cyber security and networking applications rely on malicious URL detection. The majority of computer assaults begin with a rogue webpage being visited. On a phishing page, a user might be misled into freely giving out personal information, or a drive-by download can result in malware infection. We demonstrated phishing URL detection using a machine learning method called logistic regression, which has the highest learning accuracy when compared to other algorithms like naive bayes and random forest. In the future, there is a plan to expand training and testing data and uncover different levels of accuracy, which may then be deployed as online content to all networked devices.Adding some host based features would make our model more accurate.



### 5.3 Testing

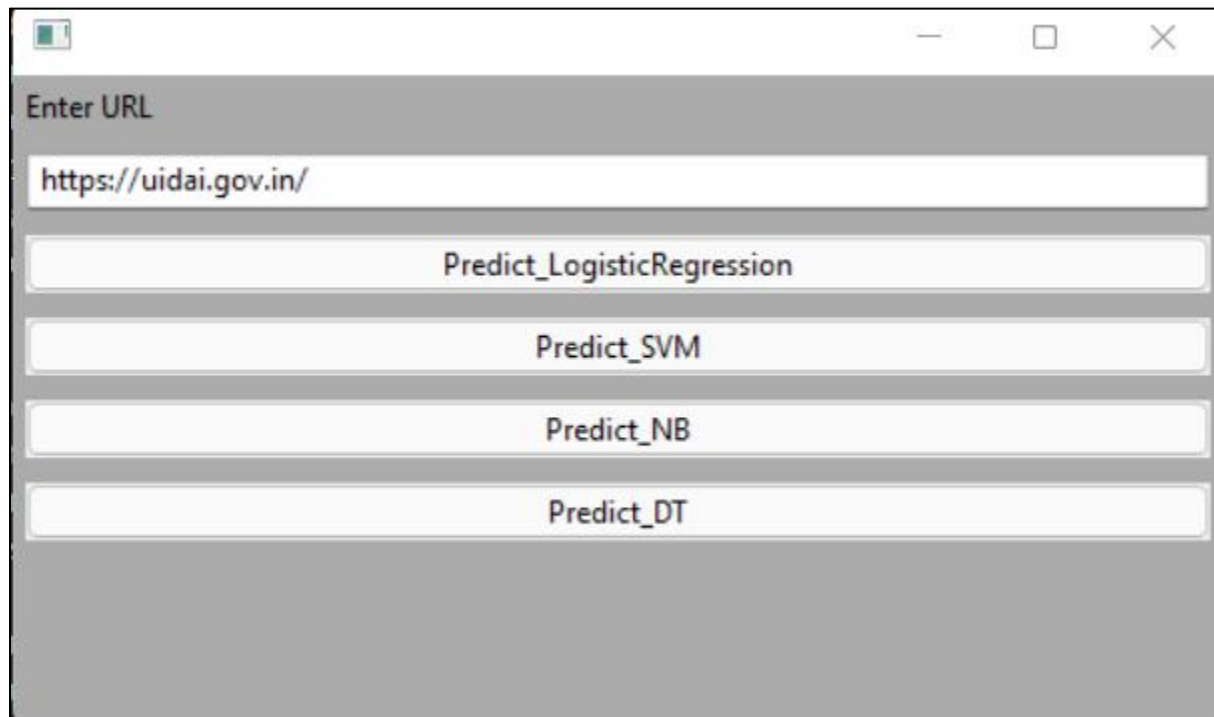
A	B
http://www.union.gr/components/com_poll/Avast/crypted/index.php?email=	yes
http://pavestonejeans.com/chases/home/	yes
http://www.tgm.cl/sitio/cgi-bin/drup%20box.html	yes
http://www.hrt.is/xxx.php	yes
http://www.anugrahabadi.co.id/newinfo/oldme/oods/oods/oods/oods/gdoc/filewords/index.php	yes
http://ip6.si#lzyZb0	yes
http://hvgoosechasers.com/LLC/detailsdocuments/login.php?action=online_login=true&_sessi	yes
http://sytinmobiliaria.com/templates/system/css/go.php	yes
http://pm1.kostrzyn.pl/Dropboxait/dropbox/dropbox/	yes
http://www.sicher.cl/css/includes/templates/Home/index.php	yes
https://www.nasa.gov	yes
https://www.rottentomatoes.com	no
https://home.mcafee.com/root/landingpage.aspx?lpname=get-it-now&affid=0&cid=170789	no
https://www.trustedsource.org/?p=mcafee	no
https://community.mcafee.com/docs/DOC-4932	no
https://www.sumhr.com/top-pages-follow-facebook-popular-list/	no
https://www.socialbakers.com/statistics/facebook/pages/total/india/	no
https://www.facebook.com/saLe.DoSt.kO.tHaNKs.bOITa.HaI	no

Figure 13. Testing Output

	A	B	C	D	E	F	G	H	I	J	K
			length_of_url	http_has	suspicious_char	prefix_suffix	dots	slash	phis_term	sub_domain	ip_contain
1		r									
2	0	yes	1	1	1	1	1	1	1	0	0
3	1	yes	1	1	1	1	1	1	1	0	0
4	2	yes	1	1	1	0	0	0	0	0	0
5	3	yes	1	1	1	1	1	1	1	0	0
6	4	yes	1	1	1	1	1	1	1	0	0
7	5	yes	1	1	1	1	1	1	1	0	0
8	6	yes	1	1	1	1	1	1	1	0	0
9	7	yes	1	1	1	1	1	1	1	0	0
10	8	yes	1	1	1	1	1	1	1	0	0
11	9	yes	0	1	1	0	1	1	0	0	0
12	10	yes	1	1	1	1	1	1	1	1	0
13	11	yes	1	1	1	0	1	0	0	1	0
14	12	yes	0	1	1	0	1	1	0	0	0
15	13	yes	1	1	1	0	1	0	0	1	0
16	14	yes	1	1	1	0	1	0	0	0	0
17	15	yes	1	1	1	0	1	1	0	0	0
18	16	yes	0	1	1	1	1	0	1	0	0
19	17	yes	1	1	1	1	1	0	0	1	0
20	18	yes	0	1	1	1	1	1	0	1	0
21	19	yes	1	1	1	0	1	0	0	1	0
22	20	yes	0	1	1	0	1	0	0	1	0
23	21	yes	0	1	1	0	1	1	0	0	0
24	22	yes	0	1	1	0	1	1	0	0	0
25	23	yes	0	1	1	1	1	0	1	0	0
26	24	yes	1	1	1	1	1	1	1	0	0
27	25	yes	1	1	1	0	1	0	0	0	0
28	26	yes	0	1	1	0	1	0	0	0	0
29	27	yes	0	1	1	0	1	1	0	1	0
30	28	yes	1	1	1	1	1	1	1	1	0

Figure 14. Feature Extraction of Testing Data

## 5.4 Deployment



The image shows a web application interface for deployment. It features a window with a title bar containing a small icon and standard minimize, maximize, and close buttons. The main content area has a dark gray header with the text "Enter URL". Below this is a white text input field containing the URL "https://uidai.gov.in/". Underneath the input field are four light gray buttons with dark gray text, stacked vertically. The buttons are labeled "Predict\_LogisticRegression", "Predict\_SVM", "Predict\_NB", and "Predict\_DT". The bottom portion of the window is a solid dark gray area.

Figure 15.Deployment

## **Chapter 6: Conclusion**

# Chapter 6: Conclusion

## 6.1 Conclusion

The research was performed using dataset from PhishTank and OpenDNS. We found that the Logistic Regression, SVM, Decision Tree and Naive Bayes algorithms were quite effective in predicting phishing websites using features extraction. Also the features we selected were quite accurate in giving maximum data to our algorithms to predict. Thus, the idea of using ensembles algorithms were quite effective. We received an accuracy of more than 80%. We can say from the above implementation that Machine Learning can be a very significant factor or may be the future solution to Phishing. As we can see that the accuracy we obtained is good but definitely not the best. We made a software that detects phishing websites. The system could have definitely performed better if trained better. After the research we felt that the data set that we used were not of a very high quality. The reason for this is it is very hard to find a high quality phishing URLs in bundle and to find legitimate URLs that falls on that shady line between Legitimate and illegitimate, so that our algorithms can have better understanding of thin line between legitimate and fake.

### References:

- [1] Yazhmozhi, V., 2019. Natural language processing and Machine learning based phishing website detection system. *IEEE*.
- [2] Zhu, E., Chen , Y. & Ye, C., 2019. OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. *IEEE*.
- [3] Abbasi, A., Zhang, . Z. & Chen, H., 2008. A statistical learning based system for fake website detection. *IEEE*.
- [4] Abunadi, A., Akanbi , O. & Zainal, A., 2013. Feature extraction process: A phishing detection approach. *IEEE*.
- [5] Aburrous, M., Hossain, M., Dahal, K. & Thabt, F., 2010. "Predicting phishing websites using classification mining techniques with experimental case studies. *IEEE*.
- [6] Alswailem, A., Alabdullah, B., Alrumayh, N. & Alsedrani, D., 2019. Detecting Phishing Websites Using Machine Learning. *IEEE*.

### Web references

#### List of Websites:

1. <https://www.phishing.org/what-is-phishing>
2. <https://www.phishtank.com/>
3. <https://cio.economictimes.indiatimes.com/news/digital-security/efforts-being-made-to-nab-frauds-creating-fake-accounts-to-siphon-off-donation-money-for-covid-19/74913053>
4. <https://wxpython.org/>
5. [https://www.learnpython.org/en/Pandas\\_Basics](https://www.learnpython.org/en/Pandas_Basics)
6. <https://www.geeksforgeeks.org/stacking-in-machine-learning/>

## APPENDIX:




### Plagiarism check report:



#### Document Information

Analyzed document	G-12_Malicious URL Deteaction_Blackbook(1).pdf (D135396211)
Submitted	2022-05-04T05:00:00.0000000
Submitted by	sangeeta
Submitter email	sangeeta.vhatkar@thakureducation.org
Similarity	5%
Analysis address	sangeeta.vhatkar.thakur@analysis.orkund.com

#### Sources included in the report

<b>W</b>	URL: <a href="http://ethesis.nitrklac.in/7569/1/2015_MT_Design_Murali.pdf">http://ethesis.nitrklac.in/7569/1/2015_MT_Design_Murali.pdf</a> Fetched: 2021-04-24T17:20:53.4270000	 3
<b>W</b>	URL: <a href="https://hammer.purdue.edu/articles/thesis/A_MACHINE_LEARNING_BASED_WEB_SERVICE_FOR_MALICIOUS_URL_DETECTION_IN_A_BROWSER/11359691/1/files/20163926.pdf">https://hammer.purdue.edu/articles/thesis/A_MACHINE_LEARNING_BASED_WEB_SERVICE_FOR_MALICIOUS_URL_DETECTION_IN_A_BROWSER/11359691/1/files/20163926.pdf</a> Fetched: 2022-05-04T05:00:17.7470000	 1
<b>W</b>	URL: <a href="https://www.semanticscholar.org/paper/Short-links-under-attack%3A-geographical-analysis-of-Klien-Strohmaier/226b50617f199f6ca50c041b3928a1412a940ef6">https://www.semanticscholar.org/paper/Short-links-under-attack%3A-geographical-analysis-of-Klien-Strohmaier/226b50617f199f6ca50c041b3928a1412a940ef6</a> Fetched: 2022-05-04T05:00:51.4800000	 1
<b>W</b>	URL: <a href="https://www.hindawi.com/journals/scn/2021/5518528/">https://www.hindawi.com/journals/scn/2021/5518528/</a> Fetched: 2021-07-04T07:41:19.6570000	 4
<b>W</b>	URL: <a href="https://www.mdpi.com/1424-8220/21/24/8281/pdf">https://www.mdpi.com/1424-8220/21/24/8281/pdf</a> Fetched: 2021-12-18T11:08:41.0430000	 1

# Malicious-URL Detection using Logistic Regression Technique

\*Note: Sub-titles are not captured in Xplore and should not be used

Aman Gupta  
dept. of Information Technology  
TCET  
avgupta3100@gmail.com

Dev Jindani  
dept. of Information Technology  
TCET  
devjindani68@gmail.com

Manu Khandelwal  
dept. of Information Technology  
TCET  
manukhandelwal20@gmail.com

**Abstract**—Over the last few years, the Web has seen a massive growth in the number and kinds of web services. Web facilities such as online banking, gaming, and social networking have promptly evolved as has the faith upon them by people to perform daily tasks. As a result, a large amount of information is uploaded on a daily to the Web. As these web services drive new opportunities for people to interact, they also create new opportunities for criminals. URLs are launch pads for any web attacks such that any malicious intention user can steal the identity of the legal person by sending the malicious URL. Malicious URLs are a keystone of Internet illegitimate activities. The dangers of these sites have created a mandates for defences that protect end-users from visiting them. The proposed approach is that classifies URLs automatically by using Machine-Learning algorithm called logistic regression that is used to binary classification. The classifiers achieves 97URLs.

## I. INTRODUCTION

Phishing websites are being employed to steal personal information, such as credit cards and passwords, and to implement drive-by downloads. Phishing is popular among muggers since it is easier to trick someone. In most cases, such annoying activity engages network resources intended for other use into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. In most cases, such annoying activity engages network properties intended for other uses, and nearly always threatens the security of the network and/or its data. Properly designing and deploying a Phishing URL will help block the intruders. phishing domain (or Fraudulent Domain) characteristics, the features that discriminate them from appropriate domains, why it is important to detect these domains, and how they can be detected using machine learning techniques.

### Background Study

This section discusses related methodologies used by researchers who have tried to solve the problem of phishing URL detection and classification. The authors Mohammed Nazim Feroz and, Susan Mengel[3] has describes an approach that classifies URLs automatically based on their lexical and host-based features. These methods are able to learn highly analytical models by extracting and automatically Mahout is

established for such scalable machine learning problems, and online learning is considered over batch learning. The classifier achieves 93-95number of phishing hosts, while maintaining a modest false positive rate. Justin Ma, Lawrence K. Saul, Stefan Savag and, Geoffrey M. Voelker[4] describes an approach to this problem based on automated URL classification, using statistical methods to discover the tell-tale lexical and host-based properties of malicious Web site URLs. These methods are able to learn highly analytical models by extracting and repeatedly examining tens of thousands of features potentially indicative of suspicious URLs. The resulting classifiers obtain 91-94large numbers of malicious Web sites from their URLs, with only modest false positives. Frank Vanhoenshoven, Gonzalo N apoles, Rafael Falcot, Koen Vanhoof and Mario K"oppent Universiteit Hasselt Campus Diepenbeek [1]determines online learning approaches for detecting malicious Web sites (those involved in criminal scams) using lexical and host-based features of the related URLs. We show that this application is mostly suitable for online algorithms as the size of the training data is larger which can be efficiently processed in batch also the distribution of features that typify malicious URLs is changing unceasingly.

## II. PROPOSED METHODOLOGY

To ripen a defined manners from the data-sets, the model is to be sketched out like obliquely identify the data from which it has to be practised. The pillar of this model is data-sets and hence it should be sufficient and perfect data for good as well as bad URLs existing in the data for the model to be trained upon. A list of URLs that have been classified as either malicious or benevolent and characterize each URL via a set of attributes such as number of dots presents in URL, distance of the URL, token-based diagrams such as google.com. To train a model, binary classification technique which is also called as binary regression technique is used in a model.

### A. Advantages of proposed method

The proposed method acquires maximum learning accuracy comparing to other machine learning algorithms.

### III. UNIFORM RESOURCE LOCATOR (URL)

A URL is a exclusive identifier used to locate a resource on the internet. It is also denoted to as a web address. URLs consist of multiple parts – including a protocol and domain name – that tell a web browser how and where to recover a resource. End operators use URLs by typing them directly into the address bar of a browser or by ticking a hyperlink found on a webpage, bookmark list, in an email or from additional application. A URL is the most collective type of Uniform Resource Identifier (URI). URIs are strings of typescripts used to identify a source over a network. URLs are vital to traversing the internet. URL Structure The URL encompasses the name of the protocol required to access a resource, as well as a resource name. The first portion of a URL identifies what protocol to use as the primary access medium. The second portion identifies the IP address or domain name – and possibly subdomain – where the resource is located. After the domain, a URL can also specify:

- A path to a exact page or file within a domain
- A network port to use to make the link.
- A request or search parameters used – commonly found in URLs for search results

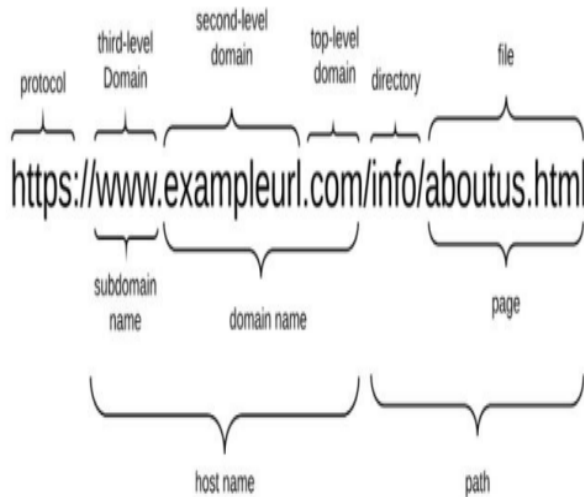


Fig. 1. URL Structure

#### A. Malicious URL

Within the gathering of cyber threats out there, mischievous websites play a critical role in today's attacks and scams. Malicious URLs can be carried to users via email, text message, pop-ups or sheltered advertisements. The end effect can often be downloaded malware, spyware, ransom ware, compromised accounts. It should be obvious that being aware of what a Malicious URL is, and how it can do harm. Launch phishing movements meant to bargain your private information, When we ticking a URL it directs us to phishing Sites and get you to install malware, viruses or Trojans, whether by transferring a file or as a drive-by-download that is provoked by something as simple as a mouse-over or other trick

#### B. First Chunk- Machine learning

Machine learning is a subsection of artificial intelligence (AI) that offers systems the skill to mechanically learn and improve from experience without being explicitly programmed. Machine learning concentrates on the development of computer programs that can access data and use it learn for themselves. The procedure of learning begins with data, such as examples, direct understanding, or instruction, in order to look for outlines in data and make better conclusions in the feature based on the examples that provide. The main aim is to permit the computers to learn automatically without human interference or assistance and regulate actions consequently. Supervised learning Supervised learning, in the background of artificial intelligence (AI) and machine learning, is a type of system in which both input and preferred output data are provided. Input and output data are labelled for classification to deliver a learning basis for future data processing. Supervised learning models have some benefits over the unsupervised approach, but they also have boundaries. The systems are more likely to make decisions that humans can relate to, for example, because humans have provided the basis for decisions. However, in the case of a retrieval-based method, supervised learning systems have distress dealing with new information.

1) *Regression*: Regression predictive modeling is the task of approaching a mapping function ( $f$ ) from input variables ( $X$ ) to a continuous output variable ( $y$ ). A constant output variable is a real-value, such as an integer or floating point value

#### C. Stepwise regression

It is used when there is doubt about which of a set of analyst variables should be included in a regression model. It works by adding and/or removing separate variables from the model and detecting the resulting effect on its accuracy. Stepwise regression is no longer stared as a valid tool for dimensionality reduction because it yields unstable results that heavily over fit the training data.

#### D. Multivariate Adaptive Regression Splines (MARS)

It is a form of regression analysis. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and communicate between variables.

#### E. Logistic Regression

The logistic regression technique includes dependent variable which can be signified in the binary (0 or 1, true or false, yes or no) values, means that the result could only be in either one form of two. For example, it can be applied when we need to find the probability of positive or fail event. Here, the same method is used with the additional sigmoid function, and the value of  $Y$  ranges from 0 to 1. Consider a model with two predictors,  $x_1$  and  $x_2$ ; these may be constant variables or indicator functions for binary variables (taking value 0 or 1). Fig 5 represents the comparison method [7].

#### F. Comparison of linear and Logistic Regression

Linear and Logistic regression are the furthestmost basic form of regression which are usually used. The crucial difference between these two is that Logistic regression is used when the dependent variable is binary in nature. In difference, Linear regression is used when the dependent variable is continuous and nature of the regression line is linear. Regression is a method is used to predict the value of a response (dependent) variables, from one or more predictor variables, where the variable is numeric. There are several forms of regression such as linear, multiple, logistic, polynomial, non-parametric, etc.

### IV. WORKING METHODOLOGY

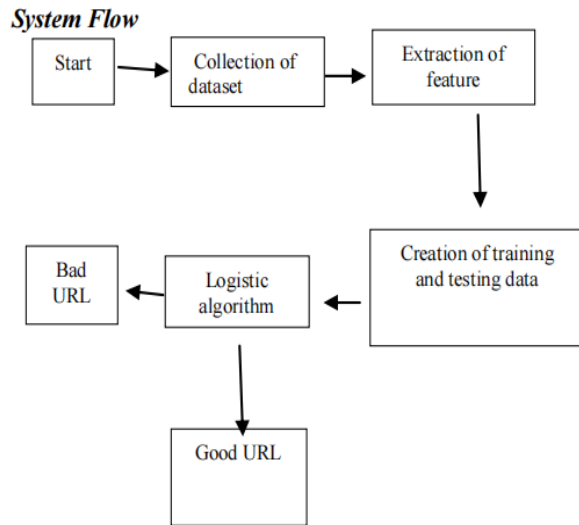


Fig. 2. System Flow

#### A. Second Chunk -Data Sets

The training data set in Machine Learning is the genuine dataset used to train the model for performing various actions. This is the actual data the current development process models learn with several API and algorithm to train the machine to work automatically.

#### B. Training Dataset

There are two types of data sets – Training, and Test that are used at several stage of development. Training dataset is the leading of two of them, while test data functions as closure of approval and you don't need to use till the end of the development.

#### C. Test Dataset

This is the data typically used to provide an balanced evaluation of the final that are completed and fiton the training dataset. Essentially, such data is used for testing the model whether it is responding or working properly or not. All of URL in our dataset are labelled Data sets are collected from [https://github.com/VAD3R-95/Malicious-URLDetection/blob/master/data\\_URL.csv](https://github.com/VAD3R-95/Malicious-URLDetection/blob/master/data_URL.csv) yahoo – phishtank

#### D. Extraction of Feature

In machine learning, a feature is an separate assessable property or characteristic of a phenomenon being detected. Picking informative, perceptive and independent features is a vital step for effective algorithms in pattern recognition, classification and regression. When the input data to an algorithm is too huge to be processed and it is suspected to be redundant. then it can be converted into a reduced set of features The selected features are expected to contain the appropriate information from the input data, so that the desired task can be performed by using this reduced demonstration instead of the complete initial data. Since the URLs are in our dataset are different from our normal text documents so we have to use text feature extraction method for construct a feature vector. Fig 7 shows the feature factorizing methods

#### E. Count Vectorizer

The most straightforward one, it counts the number of times a token shows up in the document and uses this value as its weight.

#### F. Hash Vectorizer

This one is measured to be as memory efficient as possible. In its place of storing the tokens as strings, the vectorizer applies the hashing trick to encode them as numerical indexes. The problem of this method is that once vectorized, the features' names can no longer be recovered.

#### G. TF-IDF Vectorizer

TF-IDF stands for “term frequency-inverse document frequency”, means the weight allocated to each token not only depends on its frequency in a document but also how persistent that term is in the entire corpora.

#### H. Preparing Data

Subsequently the URLs are dissimilar from our typical text documents, we need to engrave our own purification method to get the appropriate data from raw URLs. To contrivance our distillation function in python to filter the URLs with following code as shown in trial code. This will give us the desired URL data-set values to sequence the model and test it. The data-set will partake two pillar, one is for URLs and other is for labels. Here we have proceeded with the use of Tf-idf machine learning text feature extraction approach from the python module of sk-learn. Feature Vector Construction <http://www.Bfuduuioolfb.mobi/ws/ebayisapi.dll> Fig:8 Vector Construction

### V. FEATURES CONSIDERED

Blacklist Queries Lexical Features Blacklist List of known malicious sites from yahoo phish tank, google crawlers. List of malicious URLs from various domain Providers like SORBS, URIBL, SURBL, Spamhaus.



### A. Lexical Features

Tokens in URL hostname + path Length of URL Entropy of the domain name Reading Data It is essential to recite the data-sets into data frames and matrix, which can be presumed by the Vectorizer. After Vectorizer data are arranged and distributed onto the term-frequency and inverse document frequency, which is called as text extraction approach. Pandas component in python is used for the task to be implemented.

### B. Splitting Data

The data we use is typically split into training data and test data. The training set covers a known output and the model learns on this data in order to be universal to other data later on. The test dataset (or subset) is to test our model's prediction on this subset. In order to use the splitting method we have to import pandas library training set—a subset to train a model. (80 test set—a subset to test the trained model. (20

### C. Training Model

To train model call the logistic algorithm that is imported using sklearn model from python sci-kit library. (From sklearn.linear model import Logistic Regression). It uses train data set for learning. After learning it prints score of trained model.

### D. Conclusion and Future Enhancement

Malicious URL detection plays a serious role for many cyber security applications, and networking applications. The majority of computer attacks are launched by visiting a malicious webpage. A user can be tricked into voluntarily giving away private information on a phishing page or become target to a drive-by download resulting in a malware infection. In this approach we showed phishing URL detection by using machine learning algorithm called logistic regression, it obtains maximum learning accuracy comparing to other algorithms such as naïve bays, random forest. In future there is an idea to increase training and testing data and to find vary of accuracy, and can deploy as web content for all the network connected devices. In addition to that adding some more feature like host based (WHOIS) features makes our model more accurate.

### E. Figures and Tables

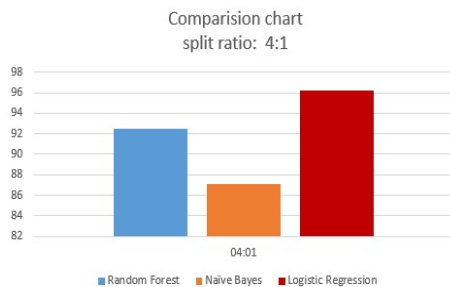


Fig. 3. comparison of Logistic Regression with Other methods with split ratio 1:1

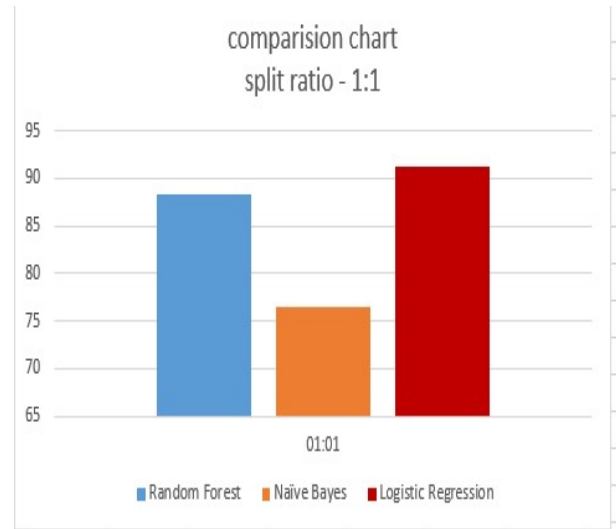


Fig. 4. Comparison of Logistic Regression with Other

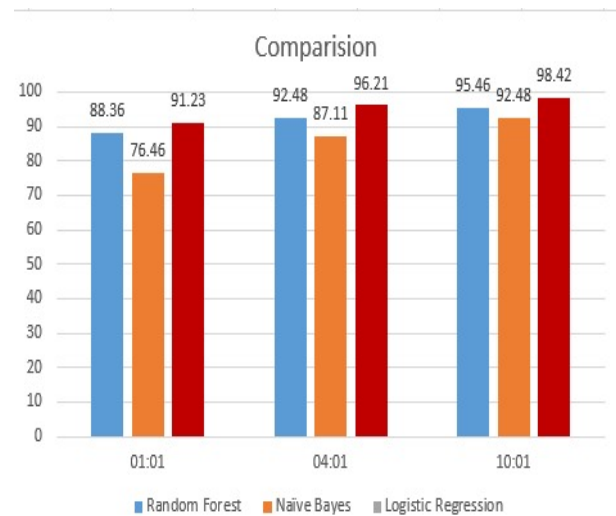


Fig. 5. Comparison

### REFERENCES

- [1] Yazhmozhi, V., 2019. Natural language processing and Machine learning based phishing website detection system. IEEE.
- [2] Zhu, E., Chen, Y., Ye, C., 2019. OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. IEEE.
- [3] Abbasi, A., Zhang, Z., Chen, H., 2008. A statistical learning based system for fake website detection. IEEE.
- [4] Abunadi, A., Akanbi, O., Zainal, A., 2013. Feature extraction process: A phishing detection approach. IEEE.
- [5] Aburrous, M., Hossain, M., Dahal, K., Thabt, F., 2010. "Predicting phishing websites using classification mining techniques with experimental case studies. IEEE.
- [6] Alswailem, A., Alabdullah, B., Alrumayh, N., Alsedrani, D., 2019. Detecting Phishing Websites Using Machine Learning. IEEE.