Summary of changes

#	Change Requested by the supervisor	**How** you addressed the supervisor request
	Reduce the plagiarism	Reduced to below 20%

Phishing Website Detection

J.W.I.A. Botejue

2021

Phishing Website Detection

A Thesis Submitted for the Degree of Master of Business Analytics

J.W.I.A. Botejue

University of Colombo School of Computing

2021

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: J.W.I.A. Botejue

Registration Number: 18880056

Index Number: 2018/BA/005

J.W.I.A. Botejue
_____

Signature: Date: 2021/10/13

This is to certify that this thesis is based on the work of

Mr. J.W.I.A. Botejue

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name:

_____

Signature: Date:

I would like to dedicate this thesis to world wide web users.

Acknowledgements

First and foremost, my sincere thanks go to the University of Colombo School of Computing for selecting me to follow this degree of master of Business Analytics.

I would like to express my sincere gratitude to my thesis supervisor, Dr D. A. S. Atukorale for the tremendous support and guidance I received from the very beginning of this research and throughout the research sparing his valuable time and sharing his knowledge with me to get solved whatever the problems whenever I had to come across regarding the research. His continuous inspiration and motivation always encouraged me to continue this research. His guidance helped me for the research without any failures and to analyse and reach the expected goals.

If it is not for my beloved parents, I would never made this far in my life. I am what I am today is all thanks to them. Would like to make this an opportunity to show my sincere gratitude for them for never giving up on me and encouraging me whenever I need it the most with love and affection.

I want to thank all my close friends and colleagues who have helped me. Always glad to have them in my life as they have become the pillars of encouragement to achieve my goals in life.

Finally, I would love to thank all who helped me to make this thesis a success.

Abstract

Online phishing is certainly considered one among the largest and maximum famous net crime programs. Phishing Attackers are sorting out new strategies like phishing to trick sufferers into the use of faux web sites to acquire touchy facts like on-line account credentials, usernames, passwords, etc. Phishing is the maximum typically used social engineering method that goals to make the weakest point discovered in device techniques as as a result of device users. a preferred step is to evaluate laptop addresses to a blacklist of rated phishing web sites, which generally help guide scanning and are ineffective. As the internet evolves, computerized URL detection has step by step come to be crucial to imparting speedy safety to quit users.

Phishing is taken into consideration a shape of internet chance described because the artwork of impersonating valid company web sites designed to gain touchy records. Currently there is no one-size-fits-all strategy to catching each phishing attack.

In this article, a real-time anti-phishing machine, which makes use of seven exclusive type algorithms and functions primarily based totally on cope with bar, functions primarily based totally on Html and JavaScript, and functions primarily based totally on domain, is offered. The machine reveals the subsequent figuring out residences from opportunity research withinside the literature: linguistic independence, use of a massive length of phishing and valid information, execution period, detection of latest websites, independence from vis-à-vis third-celebration offerings and use of feature-wealthy functions. classifiers. To degree the overall performance of the machine, a alternative information set is constructed, and the experimental outcomes also are examined on it. constant with the experimental and comparative outcomes of the algorithmic compelled type program.

In this thesis, an green and bendy phishing detection system comes with diverse capabilities that reflect the distinct capabilities of a phishing internet site and a chrome extension to test whether the internet site is a valid internet site or not..

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1 – Introduction

1.1 Project Overview

People are victimization digital trade searching over conventional purchasing with the evolution of the Internet. Nowadays, criminals use this platform to searching for out their sufferers in the computer online community with a few unique tricks. Attackers confirm new strategies, comparable to phishing, to mislead sufferers with the usage of fake web sites to acquire touchy records along with on line account IDs, usernames, passwords, etc. Phishing is that the most normally used social engineering approach that objectives at exploiting the weak point discovered in modern techniques as because of online customers. The phisher objectives on-line customers via way of means of tricking them into revealing personal information, with the intention of victimization it fraudulently. More than one strategies are usually enforced to mitigate unique attacks (Sahingoz *et al.*, 2019).

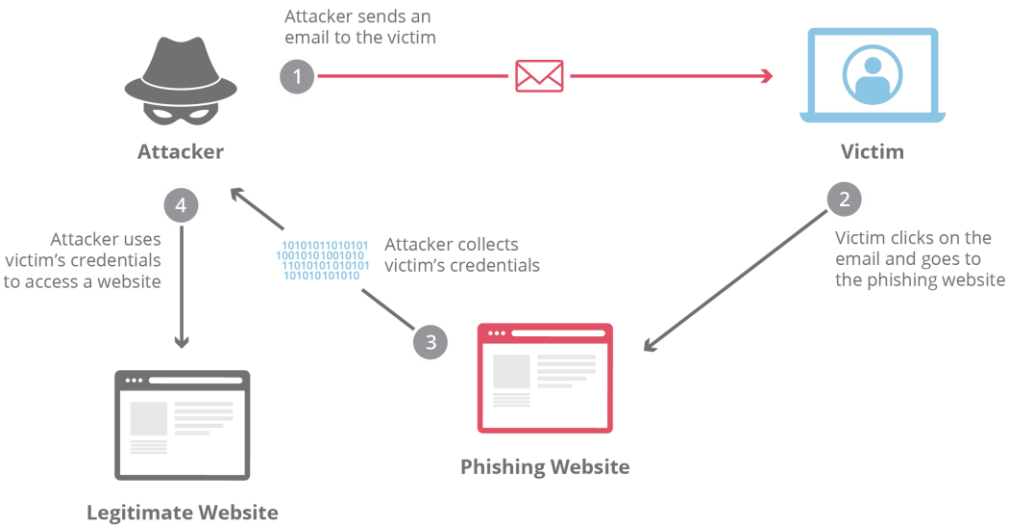

Figure 1: Deceptive Phishing Attack Diagram

1.2 Motivation

Phishing web sites are predicted to be extra elegant withinside the future. Therefore, a promising answer that need to be progressed continuously had to preserve tempo with this non-stop evolution. As net customers experience secure of being phished, they make use of anti-phishing gear. This throws an excellent responsibility at the anti-phishing gear to be correct in predicting phishing web sites. Predicting and preventing fraudulent web sites is an essential step in the direction of protective on-line transactions. Several tactics had been proposed to find out and save online users. Those attacks and anti-phishing measures can also additionally take numerous bureaucracies consisting of legal, schooling and technical answers. The technical answers are the concern of our interest, particularly, heuristic-primarily based totally phishing detection approach. The accuracy of the heuristic-primarily based totally answer especially relies upon on a fixed of discriminative standards extracted from the website. Hence, the manner wherein the ones functions are processed performs an intensive position in classifying web sites correctly. Therefore, a powerful and speedy information retrieval technique is vital for making a great decision. Data mining is one of the strategies that may employ the functions extracted from the web sites to discover styles in addition to members of the features amongst them. (Witten, Frank and Geller, 2002).

1.3 Objectives

A phishing internet site is a famous form of social engineering that mimics net pages and sincere uniform Resource locators (URLs). This task specifically goals to educate gadget gaining knowledge of fashions and deep neural networks at the dataset accumulated from unique reasserts to forecast web sites for phishing. Website phishing and URLs are accumulated to shape a dataset and URL and internet site content-primarily based totally functions are required from them.

1.4 Background Study

Phishing has been around since 1995, but became more popular in July 2003 whilst it began out focused on huge economic institutions. Before 2003, phishing changed

into nearly unknown. This is usually only used to steal AOL user credentials. Over time, phishing became popular in 2004. Nearly 2 million US citizens have checking accounts compromised by cybercriminals. With an estimated median loss reported per incident of $ 1,200, the total loss was nearly $ 2 billion. According to a study by research firm Gartner Inc. based in Stamford Cohn, consumers in the United States were scammed by phishing scams which amounted to approximately US $ 3.2 billion in 2007. This is a significant increase over the previous year (Mei, 2008).

In recent years, attacking online banking systems has become increasingly popular. One of the reasons is the increasing interest of financial institutions in offering online services.Most people have never heard of Taobao, Alibaba's online trading site in China. However, in latest years the organisation has grown rapidly. Reorts indicates that Taobao's marketplace percentage expanded from 9% to 40% in 2004..

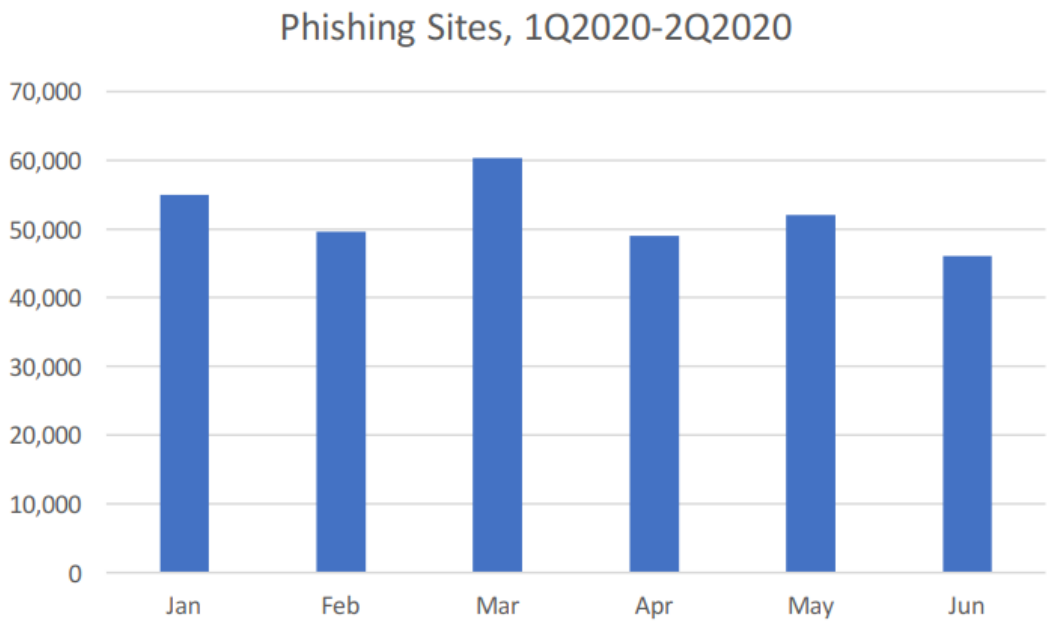

Figure 2: Phishing report from Anti-Phishing Work Group

Figure 3 demostrates the Cosmic Lynx targets. Cosmic Lynx is a Russia-primarily based totally BEC cybercriminal business enterprise that has notably impacted the e-mail risk panorama with sophisticated, high-greenback phishing attacks. In

contrast,Cosmic Lynx has a clean goal preference: large, multinational businesses. Nearly all the goal businesses we've identifed have a international presence and lots of them are Fortune 500 or Global 2000 companies. These goal businesses are situated in 20 international locations; however, due to the fact maximum of the businesses are international, personnel centered with the aid of using Cosmic Lynx BEC campaigns are placed in forty six international locations on six continents. Even personnel in international locations now no longer generally visible in phishing marketing campaign focused on sets, like Namibia and Mongolia, had been centered with the aid of using Cosmic Lynx (Griffin *et al.*, 2020).

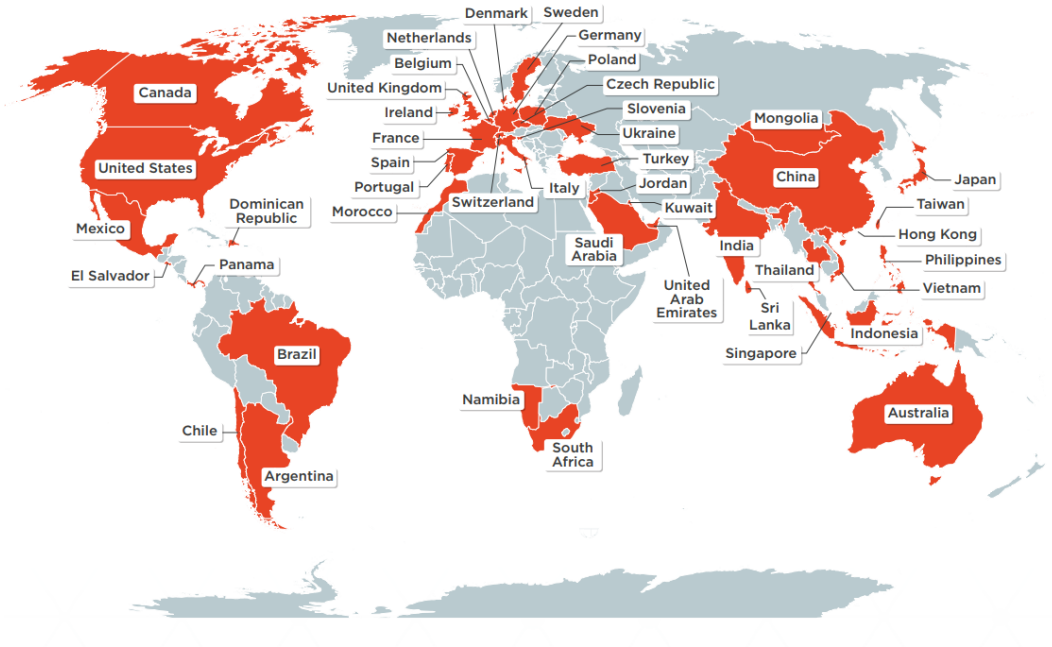

Figure 3: Map of Cosmic Lynx targets.

Currently, phishing assaults do now no longer handiest goal gadget end-users, however additionally technical personnel at carrier providers, and can set up state-of-the-art strategies which includes MITB assaults.

1.5 Scope of The Study

Scope of the thesis is to seek out a far better prediction using available free dataset and extract a greater number of features and trained them with several model and find out the highest accurate model. Then trained the model based on feature types and compare the accuracies.

1.6 Limitations

All the collected dataset should be up to dated and live websites to retrieve the features which are expected to be extracted to have a good result.

Dataset should not be class imbalanced. In order to archive that enough data should be collected from various recent data sources.

1.7 Feasibility Study

The normal overall performance and overall performance of the tool are right away related to them. In order to archive better cease end result dataset, ought to comprise greater data and applicable capabilities set.

1.8 Data Collection

- Legitimate URLs are collected from the dataset published by **Aalto University.** URLs dataset with features built and used for evaluation in the paper "Phish Storm: Detecting Phishing with Streaming Analytics" published in IEEE TNSM.
  https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset

- Phishing and Legitimate URLs Dataset is collected from Kaggle.com
  https://www.kaggle.com/akashkr/phishing-website-dataset

1.9 Data Description

There are three types of features Categories that can be extracted from the URL. (Total features: 16)

- Address bar features
- Html and JavaScript features
- Domain features

1.10 Structure of The Dissertation

The thesis is prepared as follows:

- Chapter 2 survey Background and Related paintings that describe withinside the shape of an up to date and complete overview of applicable literature; the overall theories and of phishing, anti-phishing and associated knowledge.
- Section 3 shows the suggested Approach & Methodology
- Section 4 demonstrate Research / Solution Design. The architecture of the phishing detecting system and the structure of program are included.
- Section 5 shows the Evaluation of the paper which give the result of study.
- Section 6 has conclusions of model trained and future work need to be done.

Chapter 2 – Related Projects/ Research

This chapter reviews the literature on phishing detection. There are some human factors which caused phishing attacks to target existing vulnerabilities in systems. Many cyber-assaults unfold thru mechanisms that make the most weaknesses located in quit customers, making customers the weakest a part of the safety chain. The trouble of phishing is vast, and there's no one-size-fits-all way to efficiently mitigating all vulnerabilities. Therefore, numerous strategies are frequently applied to mitigate precise attacks. This assessment pursuits to observe many phishing mitigation techniques proposed today. An excessive stage evaluation of numerous training of phishing mitigation techniques is likewise presented, such as detection, offensive defense, remediation, and prevention, which we agree that it's miles critical to offer in which detection techniques phishing are a part of the overall mitigation process (Khonji, Iraqi and Jones, 2013).

People are using online shopping as opposed to traditional shopping with the evolution of the internet. Nowadays, criminals use this platform to discover their sufferers in our on-line world with unique tricks. Attackers find out new techniques, like phishing, to trick sufferers into the usage of faux web sites to accumulate touchy statistics like on line account credentials, usernames, passwords, etc. Phishing is the maximum generally used social engineering method that targets to take advantage of weak point in machine procedures resulting from machine users (Indrajeet kumar, Shipra shalvi, no date). Phisher targets online users by tricking them into disclosing confidential information, with the aim of fraudulently victimizing.

2.1 Background

Phishing has been around since 1995, but became more popular in July 2003 whilst it began out focused on huge economic institutions. Before 2003, phishing changed into nearly unknown (Mei, 2008). This is usually only used to steal AOL user credentials. Over time, phishing became popular in 2004. Nearly 2 million US citizens have checking accounts compromised by cybercriminals. With an estimated

median loss reported per incident of $ 1,200, the total loss was nearly $ 2 billion. According to a study by research firm Gartner Inc. based in Stamford Cohn, consumers in the United States were scammed by phishing scams which amounted to approximately US $ 3.2 billion in 2007. This is a significant increase over the previous year.

In recent years, attacking online banking systems has become increasingly popular. One of the reasons is the increasing interest of financial institutions in offering online services.

2.2 Literature Review

While there are numerous techniques suggested for detecting phishing websites, this phase describes and analyzes simplest the maximum latest recommendations for detecting phishing. Its goal is to offer a whole image of the country of the artwork withinside the subject of phishing detection that could assist industry, researchers, and scientists to study the latest systems with their advantages and disadvantages and the most suitable schemes for detection. (Varshney, Misra and Atrey, 2016b), Past proposals can be categorized as follows:

- Search engine-based techniques
- Heuristics and machine learning-based techniques
- Phishing blacklist and whitelist-based techniques
- Visual similarity-based techniques
- DNS-based techniques
- Proactive phishing URL detection-based techniques

2.2.1 Search Engine-Based Techniques

Search engine strategies extract and use a website's Document Object Model (DOM), snap shots or URLs as a seek string to decide the popularity of an internet site and use engines like Google to locate phishing.

Varshney *et al.* (Varshney, Misra and Atrey, 2016a) centered at the want for a light-weight method to phishing detection the use of engines like google and diagnosed the lightest feasible features (web page identify and area name) in an effort to be pulled from an internet web page without loading a whole internet web page. Not simplest acknowledges the person's real site, however additionally gives it if the person reaches a fraudulent or phishing web page withinside the browser. However, LPD can supply fake positives on lately released benign web sites which are light-weight and addicted to simply features. The writer indicates incorporating different capabilities in destiny paintings to adapt the RPF even as retaining useful resource efficiency. that is frequently the principal concept of the LPD proposal.

2.2.2 Heuristics and Machine Learning-Based Techniques

All strategies on this class extract some of traits including net web page content, URL and / or community traits, and a chain of category or device gaining knowledge of strategies used to create a model.

Moghimi *et al.* (Moghimi and Varjani, 2016) counseled the use of Levenshtein distance for string matching to locate the connection among content material and URL of an internet web page, and used SVM for category. Like (Singh, Maravi & Sharma, 2015), the authors used SVM as a category of extra than 17 characteristics. They proposed 8 new functions that pick out the connection among the content material and the URL of an internet web page the use of a rough string-matching algorithm (Levenshtein Distance). The writer used 3066 pages from the Phishtank database and 686 valid Yahoo websites. Directory as a records record. If internet site has been cautiously designed with the aid of using Phisher, the extracted functions won't offer sufficient facts to come across phishing. Likewise, phishers use photographs and flash media in preference to text.

Singh *et al.* (Singh, Maravi and Sharma, 2015) recommended education via the Adaline community as extra green and correct for neural networks for phishing detection. The authors examined Adaline and Neural Network Backward Propagation Training further to SVM to charge phishing web sites on 15 features.

Adaline claimed it become extra green and correct for type. Multi-stage type thru SVM and neural community calls for extra assets and is the barrier.

2.2.3 Phishing Blacklist and Whitelist-Based Techniques

The techniques on this class use a whitelist of ordinary web sites and a blacklist of atypical phishing web sites. Blacklists are received both through consumer enter or through 1/3 celebration reviews that stumble on phishing URLs the usage of one of the different phishing detection schemes. The strategies are one of a kind in phrases of,

- How the blacklist or whitelist is created, saved and accessed
- If whitelist, blacklist, or each are used for detection

In the anti-phishing method utilized by the Google Chrome browser (‘Developers G. Safe browsing API-developer guide V3’, 2014), every URL opened with the aid of using the consumer is in comparison towards Google's blacklist of phishing web sites the use of the Google Safe Browsing API. For every of those URLs, an HTTP API request is dispatched and the reaction to that request is used to come across phishing. It is a light-weight solution; however, it can't come across phishing assaults in actual time if the blacklist isn't up to date with the brand-new phishing URLs.

The Firefox web browser (‘Firefox M. How does built-in phishing and malware protection work?’, 2014) additionally makes use of the Google Safe Browsing API for phishing detection. It signals the consumer if the API outcomes imply phishing. The primary disadvantage is that it is predicated at the Google Safe Browsing API for phishing detection and cannot stumble on phishing assaults in actual time if the blacklist isn't always as much as date.

In the approach of Li *et al.* (Li *et al.*, 2014), An evaluation of anti-phishing equipment primarily based totally on blacklists and whitelists embedded in a browser turned into presented. The paintings is new because it research the usage of blacklist and whitelist in phrases of accuracy and applicability and offers a hard and fast of tips to implementers on their use for phishing detection, for instance example, blacklist and

whitelist may be used for more precision. The authors recognized that there has been no distinction withinside the detection accuracy of the toolbar, whether or not a blacklist or a whitelist turned into used for detection. He cautioned that safety pop-ups must simply comprise constrained records for green users.

The scheme of Krishnamurthy *et al.* (Krishnamurthy B, Spatscheck O and A, 2009) Classifies goal net domain names / domain names as valid or phished and creates URL whitelist and URL blacklist. URLs are first as compared to whitelisted URLs and if there's no healthy then they're as compared to blacklisted URLs. Finally, the URLs withinside the filtered set are those which might be closest to the phishing URLs. Instead of the usage of present blacklists and whitelists like the ones used by (‘Developers G. Safe browsing API-developer guide V3’, 2014; ‘Firefox M. How does built-in phishing and malware protection work?’, 2014), The authors created their personal blacklist and whitelist the usage of a fixed of ordinary expressions to become aware of unacceptable and applicable Internet domains from a goal organization. Multiple listing detection will increase computational and garage complexity.

2.2.4 Visual Similarity-Based Techniques

All strategies on this class use the visible similarity among actual and phishing internet pages and their visible traits to stumble on phishing. They are but exceptional in phrases of

i. visible capabilities extracted to discover similarity;
ii. The visible matching set of rules used
iii. the use of the common sense of different anti-phishing techniques.

Most strategies primarily based totally on visible similarity are capable of stumble on a particular form of phishing assault known as tab nabbing. (Raskin A. Tabnabbing, 2014).

Chiew *et al.* (Chiew *et al.*, 2015) Proposed a brand extraction thru device mastering strategies as opposed to matching the whole web site or a fixed of pix from the web site and photograph seek-primarily based totally phishing detection scheme. Logo is

searched thru Google photograph seek and back area is matched to discover phishing site. TNR may be very low for this scheme. Currently, it desires to down load pix and do brand identity that may be accomplished greater correctly thru taking the whole display shot of the web site.

Sarika and Paul (Sarika and Paul, 2017) Proposed a framework that includes 3 ranges of agents. Level 1 includes the URL agent and the nab Tab agent. The URL agent tests for URL obfuscations and the nab Tab agent tests for format changes. The degree 1 agent forwards the message to the extent 2 agent, who makes a selection and forwards it to the extent three agent. The degree three agent notifies and indicators the consumer in case of phishing. Singh and Tripathy (Singh and Tripathy, 2014) came up with a scheme that compares the URL of the internet web page with a whitelist of URLs. If no suit is found, the SHA-1 digest of the web page is calculated. When the web page is re-centered, this digest is as compared to the brand-new calculated digest. If the summaries do now no longer suit, the web page is said valid simplest if no login shape is present. Although this schema calculates the SHA hash of the webpage's source, we taken into consideration this schema on this class as it solves the trouble of locating tabs, which changed into solved more often than not through VSB schemas. However, the authors do now no longer point out how regularly the alternate of content material is monitored; many web sites inclusive of information websites always replace their content material; overhead prices boom dramatically if assessment is needed at common intervals.

2.2.5 DNS-Based Techniques

All strategies use DNS facts to pinpoint the credibility of domains and associated laptop addresses for on-sight phishing. The strategies are tricky withinside the technique they use DNS to reap the required data to become aware of phishing. this may variety from acquiring IP addresses of domains to acquiring logs of area inquiries to become aware of often visited hosts. Phishing internet site detection.

Chen *et al.* (Chen CS, Su SA, 2011) Proposed a scheme wherein a consumer aspect module to extract web page signatures obtains the signature of the modern internet web page. This is despatched as a DNS request to a far-flung server that allows you

to be as compared with the signatures of phishing sites. Upon receipt of the reaction, the consumer-aspect coverage enforcement module takes reaction actions (Sun, Wen and Liang, 2010; Gastellier-Prevost, Granadillo and Laurent, 2011), DNS protocol is used to ship a question containing the internet site signature that fits the phishing internet site signature on a far off server. Test and validation consequences had been now no longer to be had from the author. Communication-extensive solution, given that communique with the far-off server and DNS.

In the approach of Bo *et al.* (Hong *et al.*, 2011), Recursive DNS question logs are used to discover all residing hosts visited through a person with a view to discover suspicious phishing hosts. Known phishing URLs also are used to discover common phishing paths and, in all likelihood, an energetic phishing webpage. He added the idea of studying logs of DNS queries that don't require energetic use of DNS servers, as visible in (Sun, Wen and Liang, 2010; Gastellier-Prevost, Granadillo and Laurent, 2011). Computationally complex as it relies on recursively querying a user's DNS query records, which requires large resources from the client, resulting in poor performance.

2.2.6 Proactive Phishing URL Detection-Based Techniques

All of the strategies on this class use a couple of mechanisms to generate or discover some of feasible phishing URLs. These URLs are retrieved or flagged throughout the internet for lively detection earlier than customers or different phishing detection structures locate or document them. Diets fluctuate in phrases of,

i. The unique technique or approach used to generate feasible phishing URLs.
ii. The strategies used to move slowly the internet or discover feasible URLs at the internet.

He *et al.* (He *et al.*, 2010) Proposed an AN technique to proactively decide new registered malicious domains. this could be supported with the aid of using the commentary that domains are constructed from pregnant English phrases are legitimate. It often videos display units newly registered domains and detects malicious domains as quickly as they're registered. The second-order Markov

version identifies beneficial functions and a random type of forests is implemented for detection. It may be bypassed if the attacker makes use of significant URLs for phishing; Computing in depth as it calls for each day extraction and type of the traits of all newly registered domains.

Wu *et al.* (Wu, Du and Wu, 2016)(Hou J, 2012) released the MobiFish phishing device for Android. It includes modules, WebPhish and AppPhish. WebFish suits the area call of the URL to the whitelist. If no suit is found, the internet web page is looked for the login shape, and if the shape is found, OCR is used to extract the photograph textual content of the web page, and if it includes the area, the second one degree of the URL is the valid web page. AppFish turned into additionally advanced for phishing detection thru cell applications. Android to take screenshots and preserve a listing of suspicious apps. AS calls for preservation of respectable software servers and the listing desires to be updated. It calls for a valid area whitelist.

2.3 Research Gap

Fully seek engine-primarily based totally techniques are new to phishing detection and are gaining recognition for ease of implementation and client-facet implementations. For an extended time, in addition, they seem withinside the nice seek consequences for phishing. For top high-satisfactory domain names that run at the community with a totally brief lifespan, the range of fake positives ought to be decreased so they do now no longer seem withinside the nice seek consequences.

Because machine learning is computationally excessive and calls for a training record, now no longer realistic as a client-side deployment answer like a light-weight browser plug-in or IDS. Find simple, opportunity device mastering strategies that require much less education and consequently fewer sources, at the same time as supplying similar accuracy. In the age of cloud computing, the call for for sources is growing isn't a limitation, however hiring devoted sources to set up phishing detection answers will increase costs; However, if the client can enforce the identical answer, those extra sources may be saved; increase scalable and dynamic answers

which can adapt to converting environments. Design an answer that acknowledges that phishers regularly skip textual content and visible functions which are taken into consideration phishing detection schemes, and dynamically find out and down load new features based totally on self-control.

Phishing strategies and blacklist whitelisting have the equal drawbacks as they each take time to replace the listing and locate new phishing website URLs. There is a sure time period and then the listing is up to date both with the aid of using the purchaser or on the important server end. Until the listing is up to date, the brand-new URLs will stay undetected with the aid of using this answer. Therefore, it's miles vital to have studies and improvement on this identification. An efficient, quicker and extra lively listing updating mechanism in order that new phishing URLs may be detected immediately. Regardless of whether or not the answer runs and saves URLs on the purchaser end, all blacklisted URLs absorb a massive quantity of space. If the URL is whitelisted, it takes up extra garage space. An answer is wanted to successfully control URLs saved in browser-primarily based totally purchaser answers consisting of browser add-ons. A conversation value discount scheme is needed whilst lists are saved on faraway important servers. Methods ought to be explored to lessen conversation fees with every test the usage of cache or different transient storage.

Due to the want for extra reaffirmations with OCR, Pixel Matching API, and others, VSB responses are complicated and platform precise. For example, the Google Pixel Match API will now no longer be to be had in unmarried browsers with Firefox. OCR readers to be had on Windows will now no longer be to be had on structures jogging Linux. The required CSS capabilities and layouts will now no longer be to be had on all websites. Websites can use undeniable HTML without or with precise photographs and functions. In order to make this software an anti-phishing response that may be extensively utilized by precise clients and organizations, it's far critical to:

- A quantity of worldwide familiar visible capabilities that may be anticipated on maximum web sites need to be identified.

- A quantity of platform-impartial visible matching strategies wants to be evolved.
- Because visible similarity strategies are extra complicated than textual content customization strategies, extra powerful visible customization strategies want to be evolved to make the answer a possible and aggressive proposition.

Using DNS statistics to pick out phishing is a feasible idea; however, it provides weight to DNS servers as they should procedure requests from customers or a valuable server. Research and improvement are had to lessen verbal exchange costs. Using caching and different clever garage strategies can lessen community verbal exchange costs, latency, and strain on DNS and networks. Solutions in Proactive Phishing URL Detection-Based Techniques are commonly averted as they require quite a few internet mining assets to pick out phishing URLs. They additionally use computationally in-depth algorithms to generate viable phish URLs. Hence, those questions need to be responded in destiny research.

Chapter 3 – Proposed Approach & Methodology

3.1 Data Pre-Processing

Non legitimate and valid website URLs from numerous data sources were accrued. These website URLs encompass a few pinnacles banking sites, on line purchasing sites, reservation sites, etc. The dataset contains a big range of real-global internet site URLs.

The valid website URLs are acquired from the public datasets of the Aalto University, URLs dataset with functions constructed and used for assessment withinside the paper "Phish Storm: Detecting Phishing with Streaming Analytics" posted in IEEE TNSM. https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset. This dataset has a set of benign and phishing URLs. From this dataset, random legitimate URLs are accumulated to teach the Machine learning models.

The set of phishing URLs are accumulated from an open-supply carrier referred to as PhishTank. This carrier affords a fixed of phishing URLs in more than one codecs like CSV, JSON, etc. that receives up to date hourly.

Another big wide variety of Phishing and Legitimate URLs Dataset is accrued from Kaggle.com

3.2 Data Cleaning

Processing the information training and filtering steps may be time-consuming. The URLs want to be eliminated. In this step, we put off any redundant URLs, in addition to the ones which have been blocked or have expired. Therefore, blocked website URLs from the list were eliminated.

3.3 Feature Extraction

There are numerous capabilities that outline the traits of phishing sites. Based on those traits, Many capabilities were defined that may be very powerful in rating a

website. very beneficial in detecting phishing sites. In those capabilities are identified as Address Bar Features, HTML and JavaScript Features and Domain Features. There are three forms of capabilities Categories that we are able to extract from the URL.

- Address bar features
- Html and JavaScript features
- Domain features

3.3.1 Address Bar Features

Many capabilities may be extracted that may be take into account as deal with bar base capabilities. Out of them, beneath noted have been taken into consideration for this thesis.

- IP Address in URL

  This exams for the presence of IP deal with withinside the URL. URLs may have IP deal with rather than area call. If an IP deal with is used as an opportunity of the area call withinside the URL, that may be positive that a person is attempting to scouse borrow private facts with this URL.

  If the area a part of URL has IP deal with, the value assigned to this selection is 1 (phishing) otherwise 0 (legitimate).

- "@" Symbol in URL

  Here, its exams for the presence of '@' image withinside the URL. Using “@” image withinside the website URL commands the browser to disregard the whole thing previous the “@” image and the actual cope with frequently follows the “@” image.

  If the URL has '@' image, the cost assigned to this option is 1 (phishing) in any other case 0 (legitimate).

- Length of URL

Computes the duration of the URL. Most of the time phishers use long website URL to cover the dubious element withinside the deal with bar. In this thesis, if the duration of the URL is more than or same fifty-four characters then the URL categorized as phishing in any other case legitimate.

If the duration of URL >= fifty-four, the cost assigned to this selection is 1 (phishing) otherwise 0 (legitimate).

- Depth of URL
  Computes the intensity of the URL. This characteristic calculates the range of subpages withinside the given URL primarily based totally at the '/'.

  The value of characteristic is numerical primarily based totally at the URL.

- Redirection
  If the "//" is everywhere withinside the URL other than after the protocol, thee price assigned to this selection is 1 (phishing) in any other case 0 (legitimate).

- "http/https" in Domain name
  This tests for the presence of "http/https" withinside the area a part of the URL.

  If the URL has "http/https" withinside the area part, the cost assigned to this selection is 1 (phishing) otherwise 0 (legitimate).

- URL Shortening Services
  If the website URL is the usage of Shortening Services, the output value assigned to this selection is 1 (phishing) in any other case 0 (legitimate).

- Prefix or Suffix "-" in Domain

Checking the existence of '-' withinside the area a part of the URL. The sprint image is hardly ever utilized invalid URLs. Phishers have a tendency to feature prefixes or suffixes separated by (-) to the area call in order that customers sense that they may be coping with a valid webpage.

If the URL of Domain has '-' symbol withinside the area a part of the URL, the fee assigned to this option is 1 (phishing) in any other case 0 (valid).

3.3.2 Domain based Features

- DNS Record

  For phishing websites, both the claimed identification isn't always identified with the aid of using the WHOIS or no statistics based for the hostname. If the DNS report is empty or now no longer observed then, the fee assigned to this option is 1 (phishing) in any other case 0 (legitimate).

- Website Traffic

  This function measures the recognition of the internet site through figuring out the range of traffic and the range of pages they visit. However, considering the fact that phishing web sites stay for a quick duration of time, they will now no longer be diagnosed through the Alexa database. If the area has no visitors or isn't always diagnosed through the Alexa database, it's far labeled as “Phishing”.

  If the rank of the domain < 100000, the output of this option is 1 (phishing) else 0 (legitimate).

- Domain Age

  This function may be determined from WHOIS. Most non legitimate web sites stay for a quick length of time. The minimal age of the valid area is taken into consideration to be one year for this thesis. Age right here is not anything however special among advent and expiration time.

If domain age > one year, the output of this option is 1 (phishing) else 0 (valid).

- End Period of Domain
  For this characteristic, the closing area time is calculated with the aid of using locating the distinctive among expiration time & modern time. The stop duration taken into consideration for the valid area is 6 months or much less for this thesis.

  If the end duration of the domain is > 6 months, the cost of this selection is 1 (phishing) else 0 (legitimate).

3.3.3 HTML & JavaScript Features

- IFrame Redirection
  If the iframe is empty or response isn't determined then, the fee assigned to this option is 1 (phishing) otherwise 0 (legitimate).

- Status Bar Customization
  Phishers can also additionally use JavaScript to expose a fake URL withinside the reputation bar to users. To extract this feature, we need to dig out the web site supply code, especially the “onMouseOver” event, and test if it makes any modifications at the reputation bar.

  If the response is empty or onmouseover is located then, the price assigned to this selection is 1 (phishing) otherwise 0 (legitimate).

- Disabling Right Click
  Phishers use JavaScript to disable the right-click on function, in order that customers can not view and store the website supply code. This function is handled precisely as “Using onMouseOver to cover the Link”. Nonetheless, for this function, we can look for event “event.button==2” withinside the web site supply code and take a look at if the proper click on is disabled.

If the reaction is empty or onmouseover isn't always observed then, the price assigned to this selection is 1 (phishing) in any other case 0 (legitimate).

- Website Forwarding

  The quality line that distinguishes phishing web sites from valid ones is how typically an internet site has been redirected. In our dataset, we discover that valid web sites were redirected one time max. On the alternative hand, phishing web sites containing this option were redirected as a minimum four times.

3.4 Normalization

In this section all capabilities that have been captured in preceding step are normalized. Class facts statistics is as follows- zero (0) is assigned to valid URLs and one (1) are assigned to phishing URLs.

3.4.1 Data collection and Preparation

Extracted capabilities had been used to symbolize the enter neurons. A dataset includes big wide variety of phishing, and valid web sites had been used to extract the sixteen capabilities the usage of python. The dataset composed of valid web sites amassed from specific directories, and phishing web sites amassed the open datasets of the Aalto University, Phishtank archive and Millersmiles archive. The amassed dataset holds specific values the ones are "valid" and "phishy", and those values have to be transformed to numerical values, just so the neural community can do its calculations, and thus, the values zero and 1 will get replaced with "valid" and "phishy" respectively for training the community. And also, those models could be used to degree the accuracy.

3.5 Training with Models

An artificial neural network, or neural network, is a mathematical model stimulated with the aid of using organic neural networks (Ningxia Zhang, no date). In maximum instances it's far an adaptive device that adjustments its shape throughout learning.

In this section normalized dataset that is withinside the shape of 0s and 1s, exceeded to the multiple models after trained with this dataset. Here we'll train the models generally for purchasing better accuracy and after the training, the ones dataset values that are misclassified with the aid of using the training models, eliminated from the training dataset. Models are used to degree the accuracy;

- Decision Tree Classifier
- Logistic Regression
- Random Forest Classifiers
- XG Boost Classifier
- K Neighbors Classifier
- Multilayer Perceptron (MLPs): Deep Learning
- Autoencoder Neural Network
- Support Vector Machines

Chapter 4 – Research / Solution Design

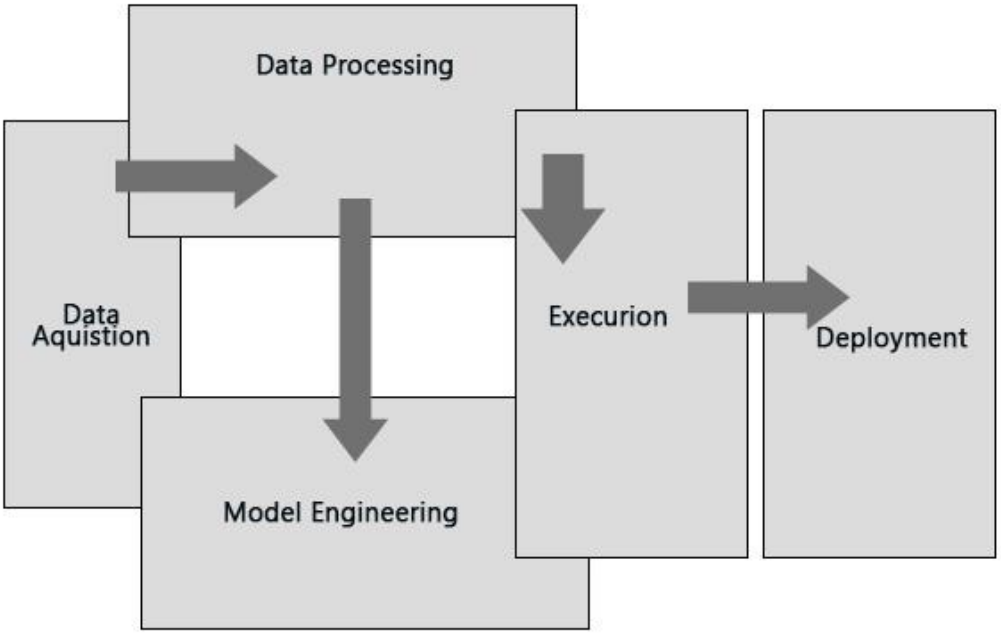

Figure 4: Block diagram of decision flow architecture for machine learning systems

4.1 Data Acquisition

Data availability is a key issue for a system gaining knowledge of mission device to make decisions. This includes gathering records, making ready and isolating the state of affairs from the case consistent with positive traits worried withinside the choice cycle, and passing the records to the processing unit for in addition categorization. This step is every so often called the records preprocessing step. The records version predicts reliable, fast, and elastic records that may be discrete or durable. The records is then transmitted to a non-stop processing device and for perennial records saved in a batch records warehouse (for discrete records) earlier than being handed to the modeling or records processing stages..

4.2 Data Processing

The facts obtained withinside the facts acquisition layer is then exceeded to the facts processing layer wherein it undergoes superior integration and processing and involves,

- Feature Extraction
- Normalization
- Training and Test Sets Preparation

4.3 Data Modeling

This degree of structure includes the selection of diverse algorithms which could adapt the machine to satisfy the hassle that the education is being processed, those algorithms evolve or are inherited via way of means of a set of libraries. Algorithms are used to version the statistics accordingly, making the machine equipped for execution.

4.4 Execution

This level of gadget studying is in which experimentation is done, checking out is involved, and upgrades are made. The universal purpose at the back of optimizing the set of rules to extract the specified output from the gadget and maximize machine performance, step output is an elegant answer that could offer the statistics had to make decisions.

4.5 Deployment

Machine studying outputs ought to paintings or be dispatched for in addition exploration processing. The end result may be visible as a non-deterministic session that should be applied withinside the decision-making system. It is suggested that the output of system studying be transferred seamlessly at once to production, wherein it'll permit the system to make selections at once primarily based totally at the output and decrease the want for extra exploratory steps..

Chapter 5 – Evaluation

This section gives the experimental info of the skilled version type algorithms and the kinds of function extraction used are detailed. Then the outcomes of the comparative checks among those algorithms with associated traits are presented.

Jupyter Notebook and Google Colab program was used in this phase. All the experiments were carried out in a system with CPU Pentium Intel(R) CoreTM i7-4720HQ CPU @ 2.60 GHz, RAM 16.00 GB. The environment is Windows 10 64-bit Operating System.

First part of this phase is to extract the features from the original dataset where it contains 0s (legitimate) and 1s (phishing).

The extracted features are categorized into,

- Address Bar based Features
- Domain based Features
- HTML & JavaScript based Features

The status of the website is checked before the extraction process starts. Most of the websites were taken down by relevant authorities. Therefore, the dataset was filtered out with python functionality to check whether the website is still live or not. The dataset contains of 5000 legit websites and 5000 phishing websites was extracted from Original Dataset after performing the feature extraction. Dataset consist of 16 features.

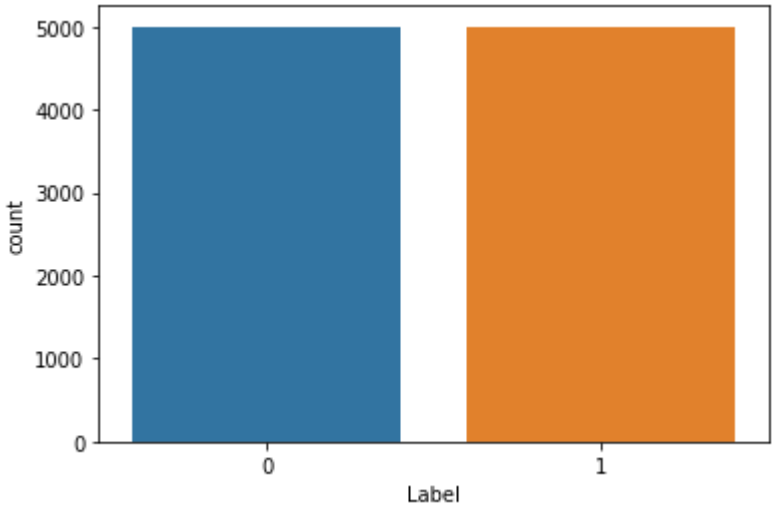

Figure 5: Classes count of dataset

Since we took 5000 legitimate websites and 5000 phishing websites the class are well balanced and there are no null values in the dataset. Therefore, further preprocessing is not required.

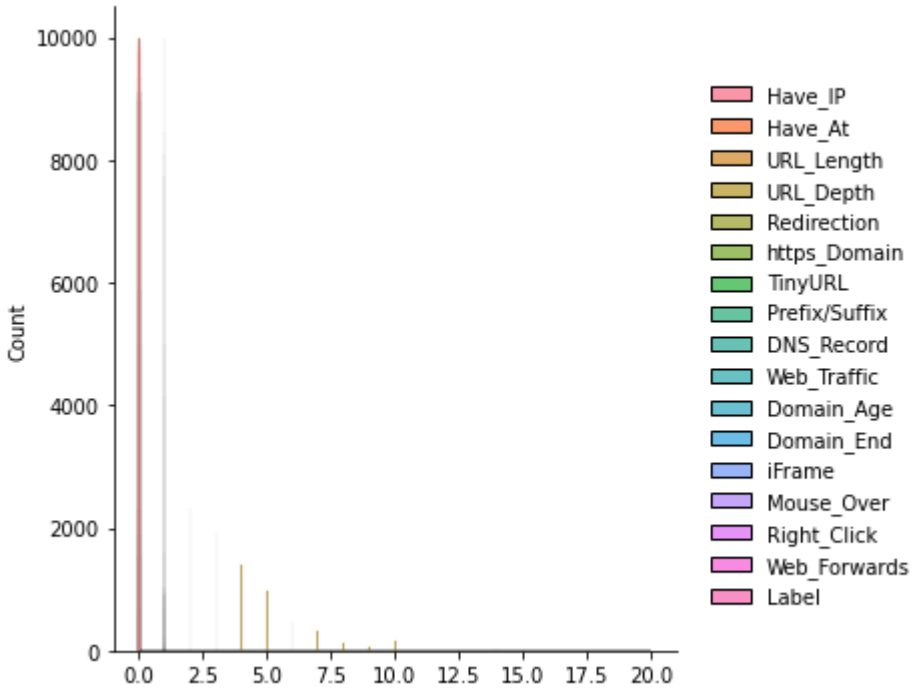

Figure 6: Overall distribution of continuous data variables

Finally, The, dataset become handed to the Models after which Models are trained with this dataset. Dataset has been shuffled to get a very good distribution. Here we train numerous Models for purchasing better accuracy and after the training, we ranked them maximum to lowest.

Before pointing out the Machine Learning model training, the record set are split into 80. From the dataset, this problem comes under supervised machine learning there are fundamental kinds of the supervised machine learning problems, referred to as classification and regression.

Classification models that can be trained using this dataset are:

- Decision Tree Classifier
- Logistic Regression
- Random Forest Classifiers
- XG Boost Classifier
- K Neighbors Classifier
- Multilayer Perceptron (MLPs): Deep Learning
- Autoencoder Neural Network
- Support Vector Machines

Four specific facts as sensitivity, f-measure, precision and accuracy are calculated to degree the usefulness and performance of the algorithms by the usage of the values in the confusion matrix. These facts, whose system is depicted in Eqs. (1–4), also are essential for creating a contrast among the examined system getting to know approaches.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{2}$$

$$F - Measure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where TN means true negative, TP means the true positive, FP describes false positive, and FN implies the false negative rate of the classification algorithms.

- Decision Tree Classifier

  The TP rate indicates the true positive rate for each class, which defines the numeric value of positive classifications by the model. In Decision Tree Classifier, Class 0 (legitimate) has the highest TP Rate. FN rate is indicating the number of false negatives classified by the model. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.70	0.99	0.82	932
1	0.98	0.64	0.77	1068

Table 1: Classification Report of Decision Tree Classifier

  Precision is the proportion of true positive instances among the instances that the model classified as positive, and sensitivity (Recall) is the rate of the total number of positive instances classified as positive. In Decision Tree Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

- Logistic Regression

In Logistic Regression Classifier, Class 0 (legitimate) has the highest TP Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.71	0.97	0.82	932
1	0.96	0.65	0.78	1068

Table 2: Classification Report of Logistic Regression

Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

- Random Forest Classifiers

  In Random Forest Classifier, Class 0 (legitimate) has the highest TP Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.71	0.99	0.83	932
1	0.99	0.64	0.78	1068

Table 3: Classification Report of Random Forest Classifiers

In Random Forest Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate). This indicates a good performance.

- XG Boost Classifier

In XG Boost Classifier, Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.78	0.95	0.86	932
1	0.95	0.77	0.77	1068

Table 4: Classification Report of XG Boost Classifier

Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate) in XG Boost Classifier. Comparatively Sensitivity and Precision values for all the classes are higher than other models in this. Therefore, it delivers a good performance.

- K Neighbors Classifier

  Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.77	0.77	0.77	932
1	0. 80	0.80	0.80	1068

Table 5: Classification Report of K Neighbours Classifier

Class 1 (phishing) has higher Sensitivity and Precision than Class 0 (legitimate) in K Neighbors Classifier.

- Multilayer Perceptron (MLPs): Deep Learning

Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.79	0.93	0.86	932
1	0.93	0.78	0.85	1068

Table 6: Classification Report of Multilayer Perceptron (MLPs): Deep Learning

In Multilayer Perceptron (MLPs): Deep Learning Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

- Support Vector Machines

  Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.69	0.98	0.80	932
1	0.97	0.61	0.75	1068

Table 7: Classification Report of Support Vector Machines

In Support Vector Machines Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

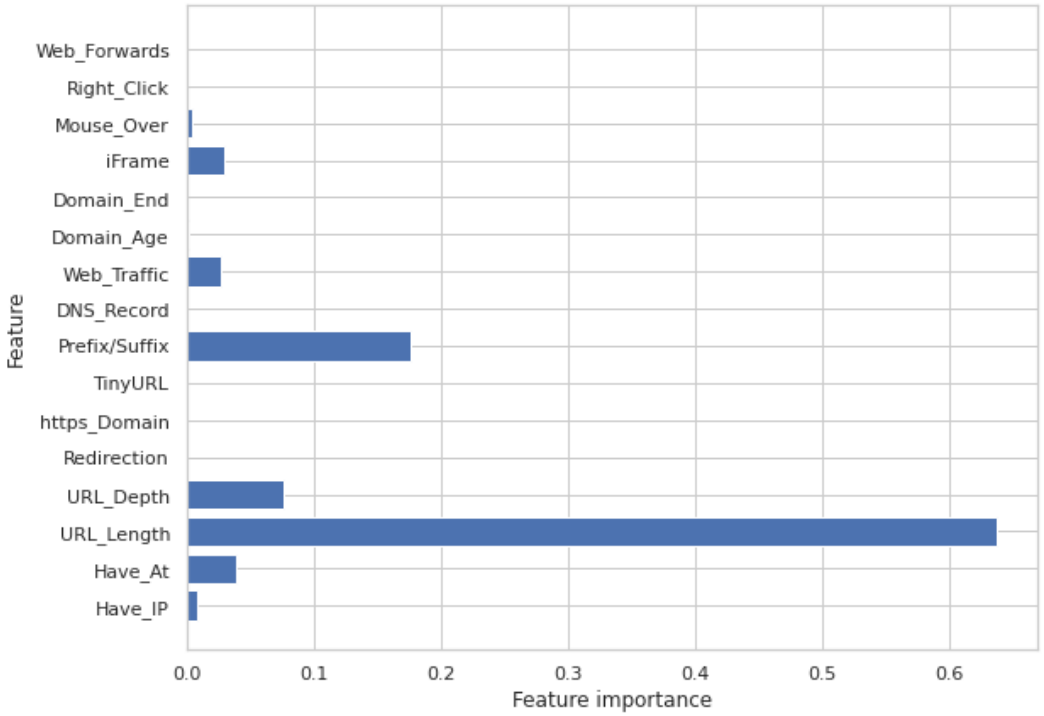

Figure 7: Feature Importance

URL Length has a highest rate of feature importance in the dataset. It has broader range of values comparing to other features. Secondly Prefix/Suffix feature has the highest rate feature importance in the dataset while other features have same rate of feature importance.

Here are the result obtained after training the model,

	ML Model	**Accuracy**
1	Logistic Regression	0.812
2	Decision Tree	0.820
3	Random forest	0.822
4	K Neighbors Classifier	0.826
5	XGBoost	0.872

	ML Model	Accuracy
6	SVM	0.811
7	Auto Encoder	0.858
8	Multilayer Perceptron	0.867

Table 8: Model Accuracy before splitting on feature type

XGBoost algorithm has the best accuracy classification performance of 87.2% (table 8). Multilayer Perceptron algorithm has second highest accuracy of 86.7%.

	ML Model	Address Bar Based Features Accuracy	Html and JavaScript based Features Accuracy	Domain based Features Accuracy
1	Logistic Regression	0.789	0.554	0.554
2	Decision Tree	0.803	0.533	0.559
3	Random forest	0.798	0.530	0.559
4	K Neighbors Classifier	0.808	0.509	0.509
5	XGBoost	0.815	0.533	0.559
6	SVM	0.797	0.532	0.542
7	Auto Encoder	0.848	0.508	0.503
8	Multilayer Perceptron	0.810	0.533	0.558

Table 9: Model Accuracy after splitting on feature type

Then the dataset was divided into three sub datasets based on the feature categorization. Each dataset was trained separately with above machine learning algorithms again. Domain Features and Html and JavaScript features have lower values comparing to the previous results which was trained before spiling the dataset into three sub datasets. In here, Auto Encoder algorithm has the best classification performance with 84.8% accuracy. Based on the experimental results, it is clear that the features based on the address bar do better than other features.

This accuracy rating is interpreted as a great and appropriate end result for fraud detection. An exact ratio of 100% is impossible. Because while sysadmins are trying to use some new methods, attackers are trying to improve their attack methods in order to avoid mistakes. In a typical phishing attack, the website is designed as if it was a legitimate website, So attackers attempt to disguise it the use of a protracted URL the use of a few unique phrases to misinform customers as URLs can clutch extra quick clips made by users who has basic knowledge of phishing attacks.

Finally, the tested model saved for API development and chrome extension was developed using that API. Once a user browses a website chrome extension automatically detect the website and send it to trained model via the API to check whether the website is legitimate or not. If the website is legitimate chrome extension will notify the user with green check mark if not danger mark.

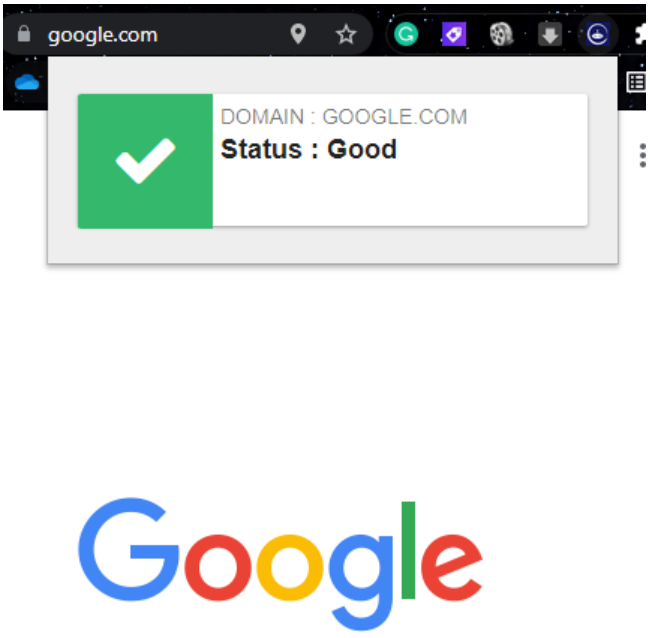

Figure 8: Chrome Extension for Phishing Prediction

Chapter 6 – Conclusion

6.1 Lessons Learnt

It has been documented that a reliable anti-phishing systems predicts phishing attacks over a period of time. Providing an honest anti-phishing remedy with an honest expiration date is also important for increasing the predicted number of phishing sites.

Due to rapid growth of phishing scenarios, different techniques were used to fool the user who browse the internet. Sine it is very difficult to find a up to date dataset to perform the machine learning algorithms dataset was created by extracting features based on techniques which is used by phishers recently. However, Feature extraction is a process that is very difficult and time consuming for this dataset. Not only for feature extraction but also for uptime of the websites were tested before doing the extraction. All these processes are time taking and good processing power is required.

In this thesis, Phishing detection system has been implemented by using eight different machine learning algorithms, as XG Boost Classifier, Logistic Regression, Random Forest Classifiers, Decision Tree Classifier, K Neighbours Classifier, Multilayer Perceptron (MLPs): Deep Learning, Autoencoder Neural Network and Support Vector Machines. It is important to create an efficient list of functions to enhance the accuracy of the detection system. Therefore, feature list has been grouped in three different classes.

After training with above mentioned machine learning algorithms, XGBoost algorithm delivered the best classification performance (87.2% accuracy). Multilayer Perceptron algorithm has the second highest accuracy of 86.7%. Then the dataset was divided into three sub datasets based on the feature categorization. Each dataset was trained separately with above machine learning algorithms again. By analyzing the result of machine learning algorithms' accuracy, it showed that there are some features performing low accuracy values and it might affect the entire

dataset training accuracy. Therefore, finding the right combination of features is difficult task in this context. Attackers find out new techniques to perform the phishing attacks. Therefore, it became challenging to find out the latest techniques and have a up to date feature set. And also, the lifetime of the phishing website is short (less than a year). Websites should be live to perform all the feature extraction tasks.

6.2 Future Modifications

All dataset may be used for building the statistics base with the employment of this deep gaining knowledge of technology, having huge collection of datasets will be good for training and accuracy. Therefore, a few multiprocessing strategies may be personalized to the system.

Further, the system can be developed to get higher accuracy by trying the vectors of word that rely on using the words in the website URL without doing alternate operations. Even though finding and having a large dataset is challenging it will give more accuracy to the final output. One of the future improvements to the model is to add a methodology for evaluating the importance of features. The system can be integrated with a website where user can check a website is legitimate or not rather than having a chrome extension where it detects the website automatically.

References

Abu-Nimeh, S. *et al.* (2007) 'A comparison of machine learning techniques for phishing detection', in *ACM International Conference Proceeding Series*. doi: 10.1145/1299015.1299021.

Chang, E. H. *et al.* (2013) 'Phishing detection via identification of website identity', in *2013 International Conference on IT Convergence and Security, ICITCS 2013*. doi: 10.1109/ICITCS.2013.6717870.

Chen CS, Su SA, H. Y. (2011) 'Protecting computer users from online frauds,to Google Patents'.

Chiew, K. L. *et al.* (2015) 'Utilisation of website logo for phishing detection', *Computers and Security*. doi: 10.1016/j.cose.2015.07.006.

'Developers G. Safe browsing API-developer guide V3' (2014). Available at: https://developers.google.com/safe-browsing/developers_guide_v3.

Dunlop, M., Groat, S. and Shelly, D. (2010) 'GoldPhish: Using images for content-based phishing analysis', in *5th International Conference on Internet Monitoring and Protection, ICIMP 2010*. doi: 10.1109/ICIMP.2010.24.

'Firefox M. How does built-in phishing and malware protection work?' (2014). Available at: https://support.mozilla.org/%0Aen-US/kb/how-does-phishing-and-malware-protection-work.

Gastellier-Prevost, S., Granadillo, G. G. and Laurent, M. (2011) 'A dual approach to detect pharming attacks at the client-side', in *2011 4th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2011 - Proceedings*. doi: 10.1109/NTMS.2011.5721063.

Griffin, A. J. *et al.* (2020) 'AGNs at the cosmic dawn: Predictions for future surveys from a ΛCDM cosmological model', *Monthly Notices of the Royal Astronomical Society*. doi: 10.1093/mnras/staa024.

He, Y. *et al.* (2010) 'Mining DNS for malicious domain registrations', in *Proceedings of the 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2010*. doi: 10.4108/icst.collaboratecom.2010.5.

Hong, B. *et al.* (2011) 'A hybrid system to find&fight phishing attacks actively', in *Proceedings - 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011*. doi: 10.1109/WI-IAT.2011.94.

Hou J, Y. Q. (2012) 'Defense against mobile phishing attack. Computer Security Course Project'.

'How can you tell if your computer's security has been compromised?' (no date). Available at: https://leadingedgeprovider.com/how-can-you-tell-if-your-computers-security-has-been-compromised/.

Huh, J. H. and Kim, H. (2012) 'Phishing detection with popular search engines: Simple and effective', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-27901-0_15.

Indrajeet kumar, Shipra shalvi, R. soni (no date) 'A System to Detect Phishing Mail', (Department of Computer Engineering, Sinhgad Institute of Technology). Available at: https://www.ijedr.org/papers/IJEDR1502180.pdf.

Khonji, M., Iraqi, Y. and Jones, A. (2013) 'Phishing detection: A literature survey', *IEEE Communications Surveys and Tutorials*. doi: 10.1109/SURV.2013.032213.00009.

Krishnamurthy B, Spatscheck O, V. D. M. J. and A, R. (2009) 'Krishnamurthy B, Spatscheck O, Van Der Merwe J, Ramachandran A'.

Kumaraguru, P. *et al.* (2008) 'Lessons from a real world evaluation of anti-phishing training', in *eCrime Researchers Summit, eCrime 2008*. doi: 10.1109/ECRIME.2008.4696970.

Li, L. *et al.* (2014) 'Towards a contingency approach with whitelist- and blacklist-based anti-phishing applications: What do usability tests indicate?', *Behaviour and Information Technology*. doi: 10.1080/0144929X.2013.875221.

Marchal, S. *et al.* (2012) 'Proactive discovery of phishing related domain names', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-33338-5_10.

Mei, Y. (2008) 'Anti-phishing system Detecting phishing e-mail'. Available at: https://www.diva-portal.org/smash/get/diva2:205981/FULLTEXT01.pdf.

Moghimi, M. and Varjani, A. Y. (2016) 'New rule-based phishing detection method', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2016.01.028.

Mohammad, R. M., Thabtah, F. and Mccluskey, L. (2013) *Predicting Phishing Websites using Neural Network trained with Back-Propagation*, *World Congress in Computer Science, Computer Engineering, and Applied Computing*.

Mohammad, R. M., Thabtah, F. and McCluskey, L. (2014a) 'Intelligent rule-based phishing websites classification', *IET Information Security*. doi: 10.1049/iet-ifs.2013.0202.

Mohammad, R. M., Thabtah, F. and McCluskey, L. (2014b) 'Predicting phishing websites based on self-structuring neural network', *Neural Computing and Applications*. doi: 10.1007/s00521-013-1490-z.

Ng, A. (2000) 'CS229 Lecture notes', *CS229 Lecture notes*. doi: 10.1111/j.1466-8238.2009.00506.x.

Ningxia Zhang, Y. Y. (no date) 'Phishing Detection Using Neural Network'.

Phishtank (2015) 'Developer Information'. Available at: https://www.phishtank.com/developer_info.php.

Ramesh, G., Krishnamurthi, I. and Kumar, K. S. S. (2014) 'An efficacious method for detecting phishing webpages through target domain identification', *Decision Support Systems*. doi: 10.1016/j.dss.2014.01.002.

Raskin A. Tabnabbing (2014) 'A new type of phishing attack'. Available at: http://www.azarask.in/blog/post/a-new-type-ofphishing-attack/.

De Ryck, P. *et al.* (2013) 'TabShots: Client-side detection of tabnabbing attacks', in *ASIA CCS 2013 - Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and*

Communications Security. doi: 10.1145/2484313.2484371.

Sahingoz, O. K. *et al.* (2019) 'Machine learning based phishing detection from URLs', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2018.09.029.

Sarika, S. and Paul, V. (2017) 'Parallel phishing attack recognition using software agents', *Journal of Intelligent and Fuzzy Systems*. doi: 10.3233/JIFS-169270.

Singh, A. and Tripathy, S. (2014) 'TabSol: An efficient framework to defend tabnabbing', in *Proceedings - 2014 13th International Conference on Information Technology, ICIT 2014*. doi: 10.1109/ICIT.2014.56.

Singh, P., Maravi, Y. P. S. and Sharma, S. (2015) 'Phishing websites detection through supervised learning networks', in *Proceedings of the International Conference on Computing and Communications Technologies, ICCCT 2015*. doi: 10.1109/ICCCT2.2015.7292720.

Sun, B., Wen, Q. and Liang, X. (2010) 'A DNS based anti-phishing approach', in *NSWCTC 2010 - The 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing*. doi: 10.1109/NSWCTC.2010.196.

Unlu, S. A. and Bicakci, K. (2010) 'NoTabNab: Protection against the "tabnabbing attack"', in *General Members Meeting and eCrime Researchers Summit, eCrime 2010*. doi: 10.1109/ecrime.2010.5706695.

Varshney, G., Misra, M. and Atrey, P. K. (2016a) 'A phish detector using lightweight search features', *Computers and Security*. doi: 10.1016/j.cose.2016.08.003.

Varshney, G., Misra, M. and Atrey, P. K. (2016b) 'A survey and classification of web phishing detection schemes', *Security and Communication Networks*. doi: 10.1002/sec.1674.

Witten, I. H., Frank, E. and Geller, J. (2002) 'Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations', *SIGMOD Record*. doi: 10.1145/507338.507355.

Wu, L., Du, X. and Wu, J. (2016) 'Effective Defense Schemes for Phishing Attacks on

Mobile Computing Platforms', *IEEE Transactions on Vehicular Technology*. doi: 10.1109/TVT.2015.2472993.

Xiang, G. and Hong, J. I. (2009) 'A hybrid phish detection approach by identity discovery and keywords retrieval', in *WWW'09 - Proceedings of the 18th International World Wide Web Conference*. doi: 10.1145/1526709.1526786.