**AN ENGINEERING PROJECT REPORT**

**ON**

**“Phishing URL detection using ML”**

**Submitted By**

**Sushil Paudel – 200348**

**Bibek Pandeya – 200311**

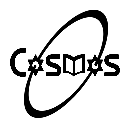
**Bhushan Bartaula – 200310**

**Prakash Lamichhane – 200323**

**Submitted To**

**Department of IT and Computer Engineering**

**In partial fulfillment of requirement for the degree of Bachelor of engineering in Computer Engineering.**



**Cosmos College of Management & Technology**

**(Affiliated with Pokhara University)**

**Tutepani, Lalitpur, Nepal**

**Date of Submission: - 2082/02/02**

# **ACKNOWLEDGEMENT**

We would like to express our special thanks of gratitude to all faculty members who supported and encouraged us, without their guidance this project would not have been possible.

We are deeply grateful to the Department of Information and Communications Technology, **Cosmos College of Management and Technology, Tutepani, Lalitpur** for providing us with all the consultation and knowledge to conduct this project. We show our sincere gratitude to our teachers for effective suggestions and strengthening in the completion of this proposal.

We would also like to acknowledge the immense contribution of our supervisor for the kind support and guidance during the project implementation phase as well as his inspiration to move forward for the completion of project on time.

Any consecutive criticism and suggestions for improvements are warmly welcomed.

**Sushil Paudel**

**Bibek Pandeya**

**Bhushan Bartaula**

**Prakash Lamichhane**

# **ABSTRACT**

Phishing attacks are a rapidly expanding threat in the cyber world, costing internet users billions of dollars each year. It is a criminal crime that involves the use of a variety of social engineering tactics to obtain sensitive information from users. Phishing techniques can be detected using a variety of types of communication, including email, instant chats, pop-up messages, and web pages. This study develops and creates a model that can predict whether a URL link is legitimate or phishing.

The data set used for the classification will be sourced from an open-source service called ‘Phish Tank’ which contains phishing URLs in multiple formats such as CSV, JSON, etc. and from the University of New Brunswick dataset bank which has a collection of benign, spam, phishing, malware & defacement URLs. Over six (6) machine learning models and deep neural network algorithms all together will be used to detect phishing URLs.

This study aims to develop a web application software that detects phishing URLs from the collection of over 5,000 URLs which are randomly picked respectively and are fragmented into 80,000 training samples & 20,000 testing samples, which are equally divided between phishing and legitimate URLs. The URL dataset will be trained and tested based on some feature selection such as address bar-based features, domain-based features, and HTML & JavaScript-based features to identify legitimate and phishing URLs.

Table of Contents

[ACKNOWLEDGEMENT ii](#_Toc198273356)

[ABSTRACT iv](#_Toc198273357)

[LISTS v](#_Toc198273358)

[a) List of Figures: v](#_Toc198273359)

[b) List of Tables: - vi](#_Toc198273360)

[**1. INTRODUCTION** 1](#_Toc198273361)

[1.1 Background 1](#_Toc198273362)

[1.2 Statements of the problems 1](#_Toc198273363)

[1.3 Objectives 1](#_Toc198273364)

[**2. LITERATURE REVIEW** 2](#_Toc198273365)

[**3. REQUIREMENT ANALYSIS** 3](#_Toc198273366)

[3.1 Technologies used 3](#_Toc198273367)

[3.2 Other Tools used 3](#_Toc198273368)

[**4. METHODOLOGY** 4](#_Toc198273369)

[4.1 Software Process Model 4](#_Toc198273370)

[**5. ALGORITHM: -** 6](#_Toc198273371)

[**6. FLOWCHART: -** 7](#_Toc198273372)

[**7. USE CASE DIAGRAM: -** 8](#_Toc198273373)

[**8. ER DIAGRAM: -** 9](#_Toc198273374)

[**9. TIME SCHEDULE: -** 10](#_Toc198273375)

[**10. CONCLUSION: -** 11](#_Toc198273376)

[**11. BIBLIOGRAPHY: -** 12](#_Toc198273377)

# **LISTS**

# List of Figures:

[Figure 1 :Phases of Iterative Increment Model 8](#_Toc198884436)

[Figure 2: Architectural Design of the system 9](#_Toc198884437)

[Figure 3: Flowchart of the system 12](#_Toc198884438)

[Figure 4: Flowchart of the web interface 13](#_Toc198884439)

[Figure 5: Use case diagram of system 14](#_Toc198884440)

## List of Tables: -

Gantt chart

## List of abbreviations: -

**URL** Uniform Resource Locators

**HTML** Hyper Text Markup Language

**EDA** Exploratory Data Analysis

**SVM** Support Vector Machine

**CSV** Comma Separated Values

**JSON** JavaScript Object Notation

**API** Application Programming Interface

**IDE** Integrated Development Environment

# **1. INTRODUCTION**

## **1.1 Background** **Information:**

Phishing is a social engineering (a type of attack used to steal user data, including login credentials and credit card numbers) as a type of attack that is one of the most common social engineering attacks. The attack happens when an attacker fools a victim into opening an email, instant message, or text message as if it were from a trusted source. Upon clicking the link, the recipient is fooled into believing that they've received a gift and unsuspectingly clicks a malicious link, resulting in the installation of malware, the freezing of the system as part of a ransomware attack, or the disclosure of sensitive information.

## **1.2 Statements of the problems:**

Phishing attacks have gotten increasingly complex; it is very difficult for an average person to determine if an email message link or website is legitimate. Cyber-attacks by criminals that employ phishing schemes are so prevalent and successful nowadays. Hence, this project seeks to address fake URLs and domain names by identifying phishing website links. Therefore, having a web application that provides the user with an interface to check if a URL is Phishing or legitimate will help decrease security risks to individuals and organizations.

## **1.3 Scope of the Study:**

This study explores data science and machine learning models that use datasets gotten from open-source platforms to analyze website links and distinguish between phishing and legitimate URL links.

The model will be integrated into a web application, allowing a user to predict if a URL link is legitimate or phishing. This online application is compatible with a variety of browsers.

## **1.4 Objectives:**

To accomplish the project's purpose, the following objectives have been established:

* dataset collection and pre-processing
* machine-learning model selection and development
* development of a web-based application for detection
* Integration of the developed model to a web application.

# **2. LITERATURE REVIEW**

## 2.1 Overview of the Study

This chapter offers an insight into various important studies conducted by excellent scholars from articles, books, and other sources relevant to the detection of phishing websites. It also provides the project with a theoretical review, conceptual review, and empirical review to demonstrate understanding of the project.

Ankit and Gupta (2017) mentioned that Statistics show that according to Internet world stats ("Internet world stats usage and population statistics", 2014), the total numbers of Internet users worldwide are 2.97 billion in 2014; that is, more than 38% of the world population uses the Internet. Hackers take advantage of the insecure Internet system and can fool unaware users to fall for phishing scams. Phishing e-mail is used to defraud both individuals and financial organizations on the Internet. (“RSA Anti-Fraud Command Center”, n.d.) Said the Anti-Phishing Working Group (APWG) is an international consortium that is dedicated to promoting research, education, and law enforcement to eliminate online fraud and cyber-crime. In 2012, total phishing attacks increased by 160% over 2011, signifying a record year in phishing volumes. The total phishing attacks detected in 2013 were approximately 450,000 and led to financial losses of more than 5.9 billion dollars (“RSA Anti-Fraud Command Center”, n.d.). Total attack increases by 1% in 2013 as compared to 2012. The total number of phishing attacks noticed in Q1 (first quarter) of 2014 was 125,215, a 10.7 percent increase over Q4 (fourth quarter) of 2013. More than 55% of phishing websites contain the name of the target site in some form to fool users and 99.4% of phishing websites use port 80 ("Anti-Phishing Working Group (APWG)Phishing activity trends report first quarter",2014). According to the APWG report in the first quarter of 2014, the second-highest number of phishing attacks ever recorded was between January and March 2014 ("Anti Phishing Working Group (APWG) Phishing activity trends report first quarter", 2014) and payment services are the most targeted industry. During the second half of 2014, 123,972 unique phishing attacks were observed ("APWG report", 2014). In the year 2011, total financial losses were 1.2 billion, and they rose to 5.9 billion dollars in 2013.

# **3. REQUIREMENT ANALYSIS**

## 3.1 Technologies used

* Jupyter notebook
* Pycharm
* Postman
* Chrome Browser

## 3.2 Other Tools used

• Visual Studio Code: Integrated Development Environment (IDE) for development

•Brave, Google Chrome: Web Browsers for testing source code

# **4. METHODOLOGY**

## 4.1 Software Process Model

A software process model is an abstraction of the software development process. The models specify the stages and order of a process. So, think of this as a representation of the order of activities of the process and the sequence in which they are performed. There are various types of Software Process Model but in this project, we are going to use Iterative and Increment Model.

**Iterative Incremental Model:**

Since, our project has multiple parts (Agent, API server, ML engine, Dashboard), which can be built and tested in small increments with the help of Iterative Increment Model. It has the following phases:

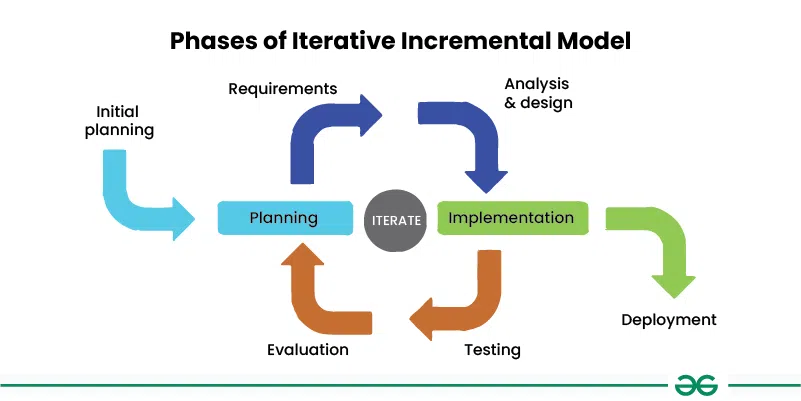


Figure 1 :Phases of Iterative Increment Model

**1. Planning Phase**

In this phase, the team identifies the goals and objectives of the project, along with the project scope, requirements, and constraints on them. The team then identifies different iterations that would be needed to complete the project successfully.

**2. Requirements Analysis and Design Phase**

In this phase, the requirements met are then analyzed and the according system is designed based on these requirements. The projected design should be modular, which would allow easy modification and testing in subsequent iterations.

**3. Implementation Phase**

In this phase, the system is implemented based on the design created in the previous phase. The implementation should be done in small, manageable pieces or increments, which can then be tested in the next phase of the cycle.

**4. Testing Phase**

In this phase, the system is tested against the requirements identified in the planning phase. Testing is done for each iteration, and any defects or issues are identified and resolved, and this helps in each iteration.

**5. Evaluation Phase**

In this phase, the team evaluates the performance of the system based on the results of testing. Feedback is gathered from users and stakeholders, and changes are made to the system as needed, which makes the system more scalable and flexible.

**6. Incremental Release**

In this phase, the completed iterations are released to users and stakeholders. Each release builds on the previous release, providing new functionality or largely improving existing functionality.

Overall, following a structured methodology ensures that the **Threat Detection System** is developed efficiently and effectively, meets the project requirements, and provides positive user experience for each generation of the user.

Model of the system

*A computer screen shot showing a diagram

AI-generated content may be incorrect.*

Figure 2: Architectural Design of the system

# **6. FLOWCHART: -**

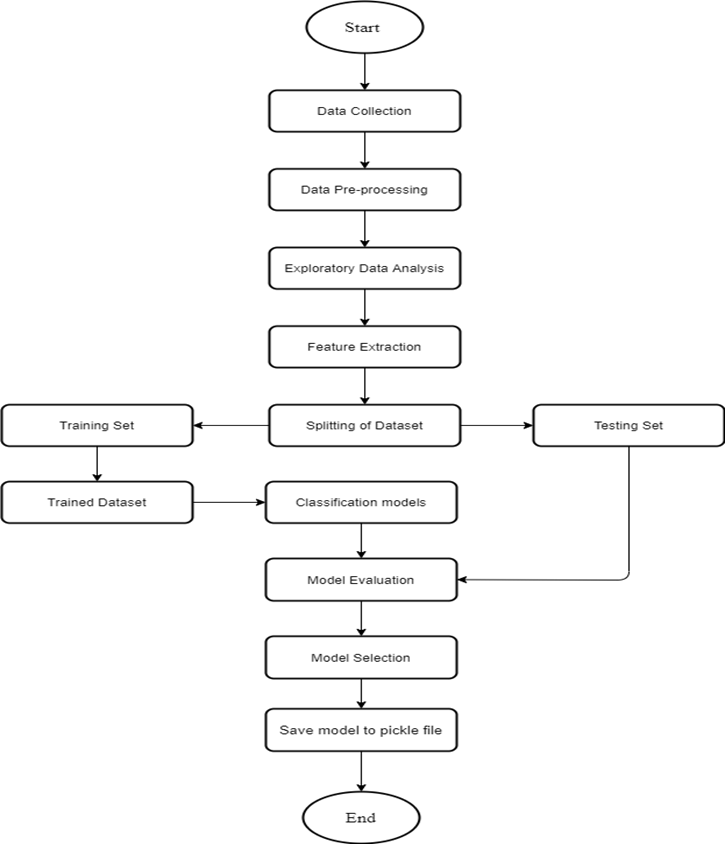


Figure : Flowchart of the system

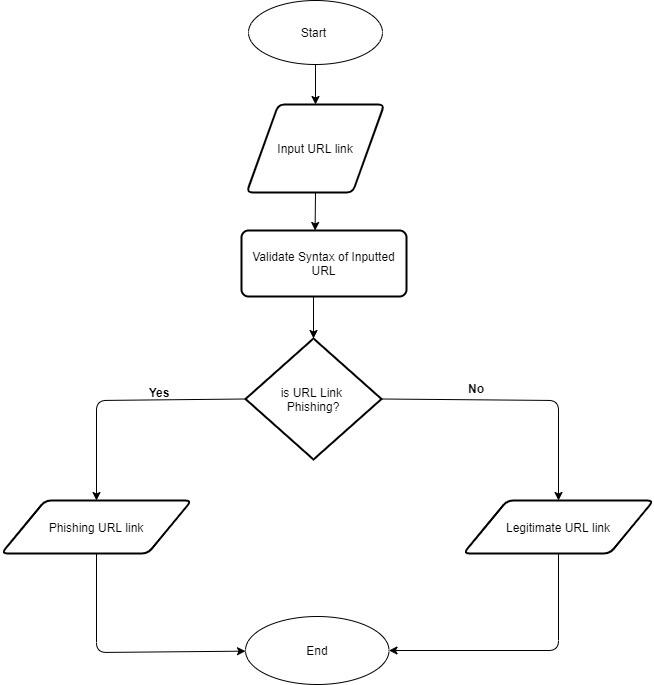


Figure : Flowchart of the web interface

# **7. USE CASE DIAGRAM: -**

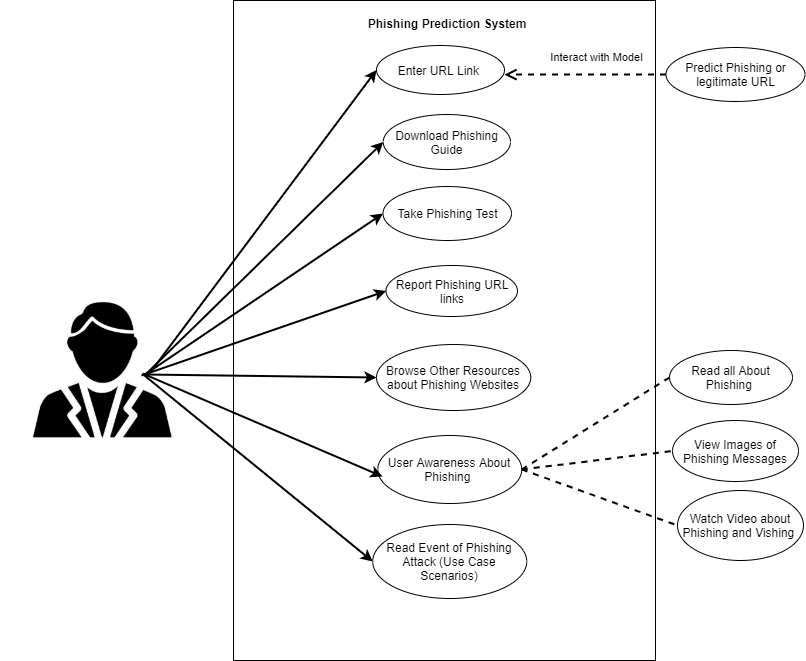


Figure : Use case diagram of system

# **9. TIME SCHEDULE: -**

Fig 9.1 Gantt chart

# **10. CONCLUSION: -**

Our project introduces a proactive and intelligent approach to cybersecurity through the development of a lightweight agent-based threat detection system. By continuously collecting system and user activity logs and analyzing them using machine learning techniques, the system can detect anomalies in real-time, allowing for quicker threat identification and response. This solution reduces reliance on manual monitoring and provides organizations with a scalable, cost-effective way to safeguard their digital infrastructure. With its modular design and automated alert system, our project demonstrates a significant step towards modernizing endpoint security and addressing evolving cyber threats.

The contribution of this research work is to help both individuals and organization identify and understand phishing techniques used by phisher as well as help them detect phishing URL attack effectively and efficiently by employing machine learning models instead of the previous method of detecting phishing website.

# 

# **11. BIBLIOGRAPHY: -**

[1] N. Abdelhamid, F. Thabtah, and H. Abdel-Jaber, "Phishing detection: A recent intelligent machine learning comparison based on models’ content and features," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, pp. 72–77, 2017, doi: 10.1109/ISI.2017.8004877.

[2] N. S. Anjum, M. S. Antesar, and M. A. Hossain, "A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps," 10th Int. Conf. on Software, Knowledge, Information Management & Applications (SKIMA), Chengdu, China, pp. 9–15, 2016, doi: 10.1109/SKIMA.2016.7916190.

[3] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2070–2090, 2013.

[4] R. M. Rishikesh and S. Irfan, "Phishing Website Detection using Machine Learning Algorithms," International Journal of Computer Applications, vol. 23, pp. 45–47, 2018, doi: 10.5120/ijca2018918026.

[5] R. Kiruthiga and D. Akila, "Phishing Websites Detection Using Machine Learning," ResearchGate, 2019. [Online]. Available: <https://www.researchgate.net/publication/337049054>

[6] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: State of the art and future challenges," Neural Computing and Applications, 2016.

[7] T. Peng, I. Harris, and I. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," 12th IEEE International Conference on Semantic Computing (ICSC), pp. 300–301, 2018.

[8] Anti-Phishing Working Group (APWG), Phishing Activity Trends Report – 1st Quarter 2014, [Online]. Available: http://docs.apwg.org/reports/apwg\_trends\_report\_q1\_2014.pdf

[9] Imperva, "Phishing Attacks," 2021. [Online]. Available: <https://www.imperva.com/learn/application-security/phishing-attack-scam/>