Phishing Detection Using Machine Learning

A PROJECT REPORT
Submitted by

Adithyan M S	**PRP19CS006**
Aswani N K	**LPRP19CS056**
Amal Soman	**PRP19CS012**
Harikrishnan K B	**PRP19CS029**

to

the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree

of

Bachelor of Technology
In
Computer Science and Engineering

Department of Computer Science and Engineering
COLLEGE OF ENGINEERING & MANAGEMENT PUNNAPRA
PUNNAPRA, ALAPPUZHA

JANUARY 2023

DECLARATION

We hereby declare that the project report "Phishing Detection Using Machine Learning " , submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by us under supervision of Mrs. Krishnapriya V J. This submission represents our ideas in our own words and where ideas or words of others have been included, We have adequately and accurately cited and referenced the original sources. We also declare that We have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Punnapra
Date: 09-01-2023

ADITHYAN M S
ASWANI N K
AMAL SOMAN
HARIKRISHNAN K B

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

COLLEGE OF ENGINEERING AND MANAGEMENT PUNNAPRA

CERTIFICATE

This is to certify that the project report entitled " **Phishing Detection Using Machine Learning** " submitted by : **ADITHYAN M S (KTU ID : PRP19CS006), ASWANI N K (KTU ID : LPRP19CS056), AMAL SOMAN (KTU ID : PRP19CS012), HARIKRISHNAN K B (KTU ID : PRP19CS029)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering is a bonafide record of the project work carried out by them under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Mrs. KRISHNA PRIYA V J	Mrs. SMITHA M JASMINE	Mrs. NEETHU SATHYAN M
Asst. Professor	Asst. Professor	Asst. Professor
Dept. of IT	Dept. of CSE	Dept. of CSE
(Project Guide)	(Project Co-Ordinator)	(HOD)

CONTENTS

ACKNOWLEDGEMENT

First and foremost We thank God Almighty for his blessings for this project. We take this opportunity to express our gratitude to all those who have guided in the successful completion of this Endeavor. It has been said that gratitude is the memory of the heart. We wish to express our sincere gratitude to our Principal Dr. ROOBIN V VARGHESE for providing infrastructural facilities and for providing good faculty for guidance.

We owe a great depth of gratitude towards our Head of the department, CSE, Mrs. NEETHU SATHYAN M, Assistant Professor for her full-fledged support.We owe Mrs. KRISHNA PRIYA V J, Associate Professor, Department of IT, a deep sense of gratitude for his unwavering support and guidelines. We are also deeply indebted to our project co-ordinator Ms. SONY M S , Asst. Professors , Dept. of CSE for their keen interest and ample guidance throughout the project.

We are indebted to our beloved teachers whose cooperation and suggestions throughout the project which helped me a lot. We also thank all our friends and classmates for their interest, dedication and encouragement shown towards the project. We convey our hearty thanks to our parents for the moral support, suggestions and encouragement to make this venture a success.

ADITHYAN M S
ASWANI N K
AMAL SOMAN
HARIKRISHNAN K B

Chapter 1

INTRODUCTION

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

Chapter 2

LITERATURE SURVEY

PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System [1] 2020 : Design a PhishHaven which detects and classifies a URL using three subcomponents. First subcomponent, URL Hit The second subcomponent is Features Extractor. The third subcomponent is Modelics.In this new paradigm executes ensemble-based machine learning models in parallel using multi-threading technique, and results in real-time detection by significant speed-up in the classification process.

Phishing Happens Beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of a Phishing Process [2] 2021 : Participants play a risk-taking game and answer questions related to two psychological scales to measure their behaviours, and then conducted a simulated phishing campaign to assess their phishability throughout the three phishing steps selected. Analysed the effect of some personal behaviours and demographic factors in each of the three phishing steps described.

A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques [3] 2022 : The use of Natural Language Processing (NLP) techniques for detection of phishing except one that shed light on the use of NLP techniques for classification and training purposes, while exploring a few alternatives. In this research, journal, conference, and workshop papers were carefully analysed, published between 2006 and 2022, with different techniques to investigate the trend of phishing email detection.

AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites [4] 2020 : Data Source and Preparation Algorithm Implementation Model Development Model Evaluation. This paper implemented and presented four different AI-based meta-learner models using Extra-tree algorithm base learner for detecting phishing websites.

Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection [5] 2021 : Phishing instances are usually derived from PhishTank Other legitimate instances are from Alexa, DMOZ, and Common Crawl. Features used in phishing detection are usually extracted from URLs (protocol, domain, path, parameters). This feature selection framework achieves a remarkable 87.6% reduction in feature quantity with suffering from only a 0.1% deterioration in detecting accuracy, making it possible for up-date training and real-time detecting in a production environment.

PDGAN: Phishing Detection With Generative Adversarial Networks [6] 2022 : The proposed PDGAN model consists of a generator and a discriminator trained in adversarial processes. The generator is an LSTM model which generates synthetic phishing URLs, and the discriminator is a CNN model which decides whether a URL is phishing or legitimate.PDGAN achieved 97.58% accuracy and 98.02% precision without depending on third-party services and greater accuracy than other compared models.

A Comprehensive Survey for Intelligent Spam Email Detection [7] 2019 : Looked into several papers selected based on the listed index terms and thoroughly analyzed the presented method, whether it has effectively used machine learning principles; how robust and impactful the proposed solution really. High adoption of supervised approaches is quite obvious SVM and Naïve Bayes are in high demand. Single-algorithm anti-spam systems are quite common.

Eth-PSD: A Machine Learning-Based Phishing Scam Detection Approach in Ethereum [8] 2021 : Detect phishing scam-related transactions using a novel machine learning-based approach. Eth-PSD tackles some of the limitations in the existing works, such as the use of imbalanced datasets, complex feature engineering, and lower detection accuracy. Proposed Eth-PSD to detect the phishing scam in Ethereum. Started with derived requirements based on the limitations of related works and other effective IDSs from previous related works.

Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning [9] 2019 : Character sequence features of the given URL are extracted and used for quick classification by deep learning,we combine URL statistical features, webpage code features, webpage text features, and the quick classification result of deep learning into multidimensional features. Found that the MFPD approach is effective with high accuracy, low false positive rate and high detection speed.

OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network [10] 2019 : In the proposed OFS-NN, a new

index, feature validity value (FVV), is first introduced to evaluate the impact of sensitive features on the phishing websites detection. Then, based on the new FVV index, an algorithm is designed to select the optimal features from the phishing websites. This algorithm could properly deal with problems of big number of phishing sensitive features and the continuous changes of features. Consequently, it can mitigate the over-fitting problem of the neural network classifier.

Detecting Phishing Web Pages with Visual Similarity Assessment Based on EMD [11] 2006 : Convert the involved Web pages into low resolution images Use EMD to calculate the signature distances of the images Train an EMD threshold vector for classifying a Web page as a phishing or a normal.10,281 suspected Web pages are carried out to show high classification precision, phishing recall, and applicable time performance for online enterprise solution.

Counteracting Phishing Page Polymorphism:An Image Layout Analysis Approach [12] 2009 : Analyze the layout of webpages rather than the HTML codes, colors, or content. Specifically, compute the similarity degree of a suspect page and an authentic page through image processing techniques.This mechanism is more robust than the HTML-based approachbecause it is more adaptable to phishing page polymorphism.

A Computer Vision Technique to Detect Phishing Attacks [13] 2015 : The proposed approach is a combination of white list and visual similarity based techniques. Use computer vision technique called SURF detector to extract discriminative key point features from both suspicious and targeted websites. This proposed solution is efficient, covers a wide range of websites phishing attacks and results in less false positive rate.

Fighting Phishing with Discriminative Keypoint Features [14] 2009 : An effective image-based antiphishing scheme based on discriminative keypoint features in Web pages. Their invariant content descriptor, the Contrast Context Histogram (CCH), computes the similarity degree between suspicious and authentic pages.The results show that the proposed scheme achieves high accuracy and low error rates.

Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions [15] 2018 : Discuss the history of phishing attacks and the attackers' motivation in details Provide taxonomy of various solutions proposed in literature to protect users from phishing based on the attacks identified in our taxonomy.Conclude paper discussing various issues and challenges that still exist in the literature, which are important to fight against with phishing threats.

A Layout-Similarity-Based Approach for Detection [16] 2007 : In this paper, an extension of our system (called DOMAntiPhish) that mitigates the shortcomings of previous system. In particular, novel approach leverages layout similarity information to distinguish between malicious and benign web pages.This makes it possible to reduce the involvement of the user and significantly reduces the false alarm rate, experimental evaluation demonstrates that our solution is feasible in practice.

School of phish: a real-world evaluation of anti-phishing training [17] 2009 : Teaches users to avoid falling for phishing attacks by delivering a training message when the user clicks on the URL in a simulated phishing email.Adding a second training message to reinforce the original training Training does not decrease users' willingness to click on links in legitimate messages.

Phishing for user security awareness [18] 2007 : Taken the concept of using an exercise and modified it in application to evaluate a users propensity to respond to email phishing attacks in an unannounced test. This paper describes the considerations in establishing and the process used to create and implement an evaluation of one aspect of our user information assurance education program.

Chapter 3

PROPOSITIONAL METHOD

3.1 Problem Formulation

Phishing has a list of negative effects on a Business, including loss of money, loss of intellectual property, damage to reputation, and disruption of operational activities. An attack is disguised as a message from a legitimate company. It s facilitated by communicating a sense of urgency in the message, which could threaten account suspension, money loss or loss of the targeted user's job. Users tricked into an attacker's demands don't take the time to stop and think if demands seem reasonable.

Therefore, we suggest a phishing detection model based on machine learning that compares the features of the target websites mainly the URLs.

3.2 Objectives

- Automatically Storing URLs of different available websites in our Country which are detected as phishing URLs
- To analyse the accuracy level for different machine learning algorithms and implementing the best among them
- To design and Implement a software to search and detect whether it is phishing or not.

3.3 Schedule of works

Table 3.1: **Schedule Of Works**

Sl No	Module Name	Tentative Date
1	Identification of accurate ML Algorithm	18/02/2023
2	Implementation of ML Algorithm	25/02/2023
3	Training pf Existing Data	04/03/2023
4	Testing threaten websites	11/03/2023
5	Implementation of UI &	18/03/2023
6	System Integration	25/03/2023

Chapter 4

CONCLUSION

Phishing detection is now an area of great interest among the researchers due to its significance in protecting privacy and providing security. There are many methods that perform phishing detection by classification of websites using trained machine learning models. URL based analysis increases the speed of detection. Furthermore, by applying feature selection algorithms and dimensionality reduction techniques, we can reduce the number of features and remove irrelevant data. There are many machine learning algorithms that perform classification with good performance measures and we choose the most accurate among them.