**MALICIOUS URL DETECTION USING MACHINE LEARNING TECHNIQUES**

A PROJECT REPORT

Submitted by

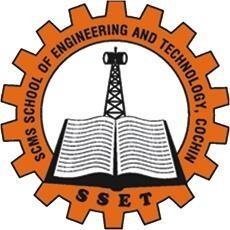
**REJITHA K P [SCM22MCA-2022]**

to

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

of

*Master of Computer Application*



### Department of Computer Science and Engineering

#### SCMS SCHOOL OF ENGINEERING AND TECHNOLOGY

**(*Affiliated to APJ Abdul Kalam Technological University*) VIDYA NAGAR, PALISSERY, KARUKUTTY**

#### ERNAKULAM - 683582

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SCMS SCHOOL OF ENGINEERING & TECHNOLOGY**

**(*Affiliated to APJ Abdul Kalam Technological University)* VIDYANAGAR, PALISSERY,KARUKUTTY ERNAKUALM - 683582**



**CERTIFICATE**

This is to certify that the report entitled **‘MALICIOUS URL DETECTION USING MACHINE LEARNING TECHNIQUES’** submitted by **REJITHA K.P** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Computer Application is a bonafide record of the project work carried out by her under my guidance and supervision.

**PROJECT GUIDE HEAD OF THE DEPARTMENT**

Ms.Aiswarya Menon Dr.Manish TI

**PROJECT COORDINATOR EXTERNAL EXAMINER**

Ms.Greshma P Sebastian

**ACKNOWLEDEMENT**

I am greatly indebted to Dr.Anitha G.Pillai, Principal, SCMS SCHOOL OF ENGINEERING AND TECHNOLOGY, Ernakulam and Dr. Manish TI, Head of department, Department of Computer Science and Engineering, SSET, who whole heartedly granted me the permission to do the project.I would like to thank my guide, Ms. Aiswarya Menon , and project coordinator, Ms. Greshma P Sebastian, Department of Computer Science and Engineering, SSET who has given me valuable guidance and support all the way.I would like to express my sincere gratitude to all the teachers of Computer Science Department who gave me moral and technical support. I would like to thank the supporting staff in the Computer lab whose dedicated work kept the lab working smoothly, thus enabling me to have access to various resources which helped me understand more about the project topic. I would also like to thank friends and family members for providing me with necessary resources and support. Last but not the least, I would like to thank god almighty for helping me to do my project hassle free.

**ABSTRACT**

The detection of malicious URLs poses a critical challenge in contemporary cyber security due to the proliferation of online threats. This paper presents a comprehensive approach to detecting malicious URLs using machine learning techniques. The study focuses on feature extraction from URLs, encompassing attributes such as URL length, presence of special characters, domain reputation, and HTTPS encryption. These features are utilized to train machine learning classifiers including Random Forest, Logistic Regression, and Decision Trees. Evaluation of the classifiers is conducted using metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate the efficiency of the proposed approach in accurately identifying malicious URLs, with significant improvements over traditional blacklist-based methods. The findings underscore the importance of feature-rich analysis in enhancing the detection capabilities of cyber security systems, thereby contributing to the ongoing efforts in safeguarding online environments against malicious activities. The goal of this project is to determine malicious URLs and evaluate the precision of each algorithm, with the aim of determining the most effective machine learning algorithm.

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** |  |
| **CONTENTS** | **PAGE NO.** |
| Acknowledgement | i |
| Abstract | ii |
| Chapter 1. Introduction | 1 |
| 1.1. General Background | 1 |
| 1.2. Objective | 2 |
| 1.3. Scope | 2 |
| 1.4. Organization of report | 2 |
| Chapter 2. Literature survey | 3 |

**CHAPTER-1** **INTRODUCTION**

* 1. **GENERAL BACKGROUND**

The project focusing on malicious URL detection using machine learning techniques falls within the realm of cyber security, which is a critical field aimed at protecting computer systems, networks, and data from various forms of cyber threats. Malicious URLs pose a significant risk to users' online security, as they can lead to phishing attacks, malware downloads, identity theft, and other harmful activities.

Traditionally, detecting malicious URLs has relied on methods such as blacklist-based systems, which maintain lists of known malicious URLs and block access to them. However, these systems have limitations, as they may not be able to detect newly created or previously unknown malicious URLs.

To address these challenges, researchers and practitioners have turned to machine learning techniques. By leveraging machine learning algorithms and feature-rich analysis of URLs, it becomes possible to automatically classify URLs as either malicious or benign based on their characteristics. This approach offers several advantages, including the ability to adapt to new and evolving threats, scalability, and efficiency.

The general background of this project involves leveraging machine learning algorithms to analyze features extracted from URLs and accurately classify them as malicious or benign. Key components of the project include feature extraction, model training and evaluation, and the deployment of a detection system capable of identifying and mitigating malicious URLs in real-time. Overall, the project aims to contribute to the advancement of cyber security measures by enhancing the ability to detect and respond to malicious online activities.

* 1. **OBJECTIVE**

Malicious attacks can be effectively detected using Machine Learning (ML) methods rather than static techniques. The objective of this project is to develop a comprehensive solution using a machine learning model that accurately detects phishing attacks, evaluating and verifying the results using various datasets.

Supervised machine learning algorithms are applied to categorize websites as either phishing or legitimate. The process of Feature Extraction is utilized to identify crucial features by applying new rules to well-known features and updating certain others.

* 1. **SCOPE**

Development of a system or tool for detecting Malicious URLs in real-time.This could include Machine learning models,rule-based systems,or a combination of both.Evaluation and testing of the developed system/tool using a dataset of known malicious and legitimate URLs.Implementation of the system/tool in a real-world scenario,such as integrating it into a company’s email or web filtering sysytems.Documentation of the project,including a user manual and technical report on the system/tool’s design and performance.Maintenance and updating of the system/tool to adapt to new phishing techniques and methodologies.

* 1. **ORGANIZATION OF REPORT**

The report is divided into five sections, including an introduction to the background, objective, and scope of the project in chapter one, a review of relevant literature in chapter two, a discussion on the system's architecture and operation in chapter three, experimental results and analysis in chapter four, and a summary of conclusions and potential future developments in the final chapter, with references listed at the end.

**CHAPTER-2**

**LITERATURE SURVEY**

1. Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions [Malak Aliabri 1,2, Hanan S.Altamimi 2, Shahd A.Albelali 2,Maimunah Al-Harbi 2,Haya T.Alhuraib 2]

The article reviews recent research studies on detecting malicious URLs using machine learning (ML) algorithms, focusing on both Arabic and non-Arabic content. It presents taxonomies and comparison results, emphasizing lexical features as the most commonly used for detection. While Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) are prevalent algorithms, Convolutional Neural Network (CNN) and XGBoost models demonstrate superior performance, achieving 99.98% accuracy. Dataset sources vary, with Arabic studies often creating their own datasets, while non-Arabic studies use open-source datasets like PhishTank and Alexa. The article also discusses challenges, including dataset size and feature selection, offering insights for future research in the field.

1. Malicious URL: Analysis and Detection using Machine Learning [Mansi Mehndiratta; Naman Jain; Ankit Malhotra; Ishika Gupta; Rachna Narula]

Malicious URLs pose significant risks, leading to scams and malware. Existing prevention strategies, like blacklisting, are limited by the short lifespan of phishing websites. To address this, a real-time detection method is crucial. This study proposes characterizing URLs based on connections between domain levels and paths, using attributes for estimations. Machine learning techniques, including the K-Nearest Neighbors Algorithm, achieve up to 90% accuracy. Graph visualization aids in distinguishing patterns between malicious and benign URLs, enhancing cyber threat detection and prevention.

1. Robust Detection of Malicious URLs With Self-Paced Wide & Deep Learning.[Yunji Liang , Qiushi Wang , Kang Xiong, Xiaolong Zheng ,]

In response to the growing scale of cybercrimes, This paper introduce CyberLen, a deep learning-based system for robust and effective detection of malicious URLs. By leveraging factorization machine (FM), position embedding, temporal convolution network (TCN), and a self-paced wide & deep learning strategy, CyberLen addresses the challenges posed by diverse cyber threats and the proliferation of obfuscated malicious URLs. Evaluation on a large-scale dataset demonstrates the effectiveness of CyberLen in reducing ambiguity in URL tokens and achieving superior performance in terms of F1 score and convergence speed.

1. A Markov Detection Tree-Based Centralized Scheme to Automatically Identify Malicious

Webpages on Cloud Platforms[Jianhualiu 1,2,Mengda XU2, Xin Wang3,Shign Shen 1, and Mingluli4]

This paper introduces a malicious webpage analysis method based on decision tree classification. Attributes of webpages are extracted using the JavaScript standard, and a decision tree model is established to quantify these attributes using information entropy and information gain. Additionally, a Markov detection tree model, integrating decision trees with Markov decision processes, is proposed. In this model, webpage attributes are organized into a URL relational tree based on URL links, enabling quick categorization of benign and malicious webpages. Two methods for processing missing values are designed, namely AMD and MVPD, which effectively adapt to webpage attribute distribution and achieve high accuracy and efficiency in webpage classification compared to the original C4.5 algorithm.

1. Deep learning methods for malicious URL detection using embedding techniques as Logistic Regression with Lasso penalty and Random Forest[ Isha Thakur; Kajal Panda; Sanjeev Kumar]

This paper proposes a deep learning-based approach for detecting malicious URLs, addressing limitations of traditional techniques like blacklisting and pattern analysis. It utilizes Natural Language Processing (NLP) with TF-IDF vectorizer and N-gram parameter for feature extraction, along with embedded feature selection methods like Logistic Regression with L1 penalty and Random Forest Selection. Neural network models including DNN, LSTM, and CNN are constructed for detection. Experimental results show the DNN model achieves the highest performance, with 96.95% accuracy, 99% precision, 100% recall, and 99% F1-score using Logistic Regression with L1 penalty. The study concludes by comparing various deep learning methods for malicious URL detection.

1. Identification and Classification of Malicious and Benign URL using Machine Learning Classifiers [Danish Kundra]

This study explores machine learning algorithms for identifying harmful websites using a dataset of over 640,000 URL records. Comparing various classifiers, the random forest classifier outperforms others with a 91.49% accuracy rate. The research extensively extracts and analyzes features to determine optimal characteristics for predicting hazardous URLs, contributing to improved cybersecurity measures in combating cybercriminal acts.

1. Malicious URL Detection Based on a Parallel Neural Joint Model Jianting Yuan1 , Guanxin Chen 2 , Shengwei Tian 1 , and Xinjun Pei1

In the paper [3], suggests a machine learning approach for detecting malicious URLs. It incorporates the utilization of both URL text and metadata to make predictions about the malicious nature of a URL. The model uses multiple parallel branches, each processing a different aspect of the URL, and then combines the results to make a final prediction. The authors claim that this approach results in improved accuracy compared to models that only consider text or meta-data alone. The paper includes an experimental evaluation of the proposed model on a dataset of URLs, showing that it outperforms existing models in detecting malicious URLs.

8. A malicious URLs detection system using optimization

and machine learning classifiers [Ong Vienna Lee¹, Ahmad Heryanto2, Mohd Faizal Ab Razak3, Anis Farihan Mat Raffei]

This paper has presented the performance of the proposed approach in detecting malicious URLs. The proposed approach that implements the optimization has optimized the selection of URL features and the machine learning classifier has correctly classified the relevant malicious features. In the experiments, this paper considers applied real URL malware and benign samples application dataset. The experiment results show that the proposed approach recorded high accuracy in classifying the URLs malware samples.

9. Phishing website detection using machine learning classifiers optimized by feature selection [Rishikesh Mahajan]

This study investigates the use of machine learning classification algorithms to combat phishing websites, a significant online threat. By applying feature selection methods from Weka and testing three classification algorithms (KNN, decision tree, and RF), we aimed to reduce harm caused by these malicious sites. Utilizing a less commonly used database, we achieved a remarkable accuracy of 100%, surpassing previous results. Our approach involved multiple feature selection filters to identify the most valuable features for classification. Interestingly, we observed that reducing the number of features maintained the same accuracy level. Additionally, we significantly reduced the time needed to build models, enhancing overall performance and contributing to the effectiveness of our approach.

10. Detection of Malicious URLs through an Ensemble of Machine Learning Techniques

[Shreya Venugopal; Shreya Yuvraj Panale; Manav Agarwal; Rishab Kashyap; U Ananthanagu]

This paper focuses on classifying URLs and web pages as either legitimate or malicious to enhance internet browsing safety. By analyzing various attributes related to domain registration, URL text, web page structure, and content, we identify characteristics of malicious sources to prevent potential harm. Leveraging models such as BERT, LSTM, and Decision Trees, combined into an ensemble, yields a pragmatic solution with 95.3% accuracy. Additionally, concepts like web page reputation, internal links, and external links are incorporated into the classification process. This novel approach, integrating Natural Language Processing techniques and Machine Learning models with diverse features, has not been previously explored. The paper concludes by suggesting methods for further improvement in solving this problem.