

A  
Project Report On  
**“Phishing URL Detection using Machine Learning”**

Submitted in fulfillment of the requirements for the Degree of  
**BACHELOR OF TECHNOLOGY**  
**Computer Science and Engineering**

SUBMITTED BY

- 1) Miss. Anamika Dhananjay Gulumkar (T2065451242003)
- 2) Miss. Harshada Dattatray Jadhav (T2065451242017)
- 3) Miss. Sakshi Chandrashekhar Shinde (T2065451242035)
- 4) Mr. Shridhar Prakash Aware (T2065451242042)

Under the Guidance of  
**Mr. Pathak P. A.**



Department of Computer Science and Engineering  
**ARVIND GAVALI COLLEGE OF ENGINEERING, SATARA**  
DBATU, Lonere  
2023-24

Department of Computer Science Engineering  
**Arvind Gavali College of Engineering**  
**Panmalewadi, Satara - 415015**



**CERTIFICATE**

This is to certify that

Miss. Anamika Dhananjay Gulumkar  
Miss. Harshada Dattatray Jadhav  
Miss. Sakshi Chandrashekhar Shinde  
Mr. Shridhar Prakash Aware

has completed the Project entitled

**“Phishing URL Detection Using Machine Learning”**

Satisfactorily for the partial fulfillment of the requirements for the Degree  
in Computer Science Engineering from **DBATU, Lonere** during Academic  
Year 2023 – 2024

**Mr. Pranav Pathak**  
**Guide**

**Mr. Pranav Pathak**  
**Project Co-ordinator**

**Dr. Varsha Bhosale**  
**HOD**

**Dr. Vilas Pharande**  
**Principal**

# Certificate

This is to certify that the project report entitled “**Phishing URL Detection using Machine Learning**” is a bonafide work carried out by-

Miss. Anamika Dhananjay Gulumkar

Miss. Harshada Dattatray Jadhav

Miss. Sakshi Chandrashekhar Shinde

Mr. Shridhar Prakash Aware

under our supervision, during the year 2023-24 and submitted to the Faculty of Computer Science and Engineering, AGCE, Satara in fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering.

Mr. Shinde A.C.  
(Alumni Mentor)

Mr. Pathak P. A.  
(Project Guide)

Mr. Gujar V.B.  
(Project Co-Ordinator)

Prof. Dr. Bhosale V. K.  
(HOD)

Dr. Mirajkar G.S.  
(Dean R&D)

Dr. Pharande V. A.  
(Principal)

(Internal Examiner)

(External Examiner)

GST No.:- 27AACCV4474E1Z9

Date – 11<sup>th</sup> Feb 2024

To,  
The Principal,  
Arvind Gavali College of Engineering Satara

Subject: - Sponsorship Letter

Respected Sir/ Madam,

We would like to confirm that Miss. Anamika Dhananjay Gulumkar, Miss. Harshada Dattatray Jadhav, Miss. Sakshi Chandrashekar Shinde, Mr. Shridhar Prakash Aware, are the students from your college who have been selected for carrying out the final year project "**Phishing URL Detection**". They will work with our organization Vritti Solutions Ltd. from August 2023 for exactly 6 months. This Project has been sponsored by **Vritti Solutions Ltd.**

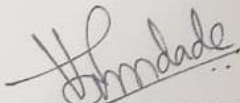
The working schedule of their project will be 3 days a week from 11.30 am to 6.00 pm. The details and scope of this project will be provided to them from the beginning of their tenure at the company facility. Upon successful completion of the project, we will be required to submit a copy of a detailed report before completion of the project with a sample model which is required to be submitted at the time of external examination held at the college.

The approximate cost sanctioned for the project is RS.30, 000/- All rights regarding the project application are reserved for the company. It is the joint responsibility of both company and the Institute to maintain confidentiality.

Thanks & Regards

Yours Truly,

**Vritti Solutions Limited**



**Veerendra Jamdade**

**CEO**



## **UNDERTAKING**

We hereby declare that the details furnished above are true and correct to the best of our knowledge and belief and we undertake to inform authorities about any changes therein immediately. In case of any of the above information is found to be false or untrue, our misleading or misrepresenting. We are aware that we may be held liable for it.

Sr. No.	Name of student	Roll No.	Sign
01	Miss. Anamika Dhananjay Gulumkar	2165451242003	
02	Miss. Harshada Dattatray Jadhav	2165451242017	
03	Miss. Sakshi Chandrashekhar Shine	2165451242035	
04	Mr. Shridhar Prakash Aware	2165451242042	

Place: Satara

Date:

## **ACKNOWLEDGEMENT**

It is our privilege to acknowledge my deep sense of gratitude to my guide Mr. Pathak P. V. in Computer Science and Engineering at Arvind Gavali College of Engineering, Satara for his valuable suggestions and guidance throughout our degree course and the timely help given to us in completion of our project work.

We are thankful to Dr. V. A. Pharande, Arvind Gavali Collage of Engineering, Satara and Prof. Mrs. Varsha Bhosale Head of Computer Science and Engineering Department for their kind cooperation and moral support.

Finally, we wish to express our sincere thanks to all the staff members of Arvind Gavali College of Engineering Satara for their direct and indirect help during the course of our project.

Place: Satara

Date:

## **ABSTRACT**

Cybersecurity threats, specifically those stemming from phishing attacks, have undergone a remarkable escalation in complexity, posing an acute threat to individuals and organizations globally. This project endeavors to address this critical concern by pioneering an innovative solution for the efficient detection and mitigation of phishing URLs. Through a synergistic integration of advanced technologies, intricate web scraping techniques, and cutting-edge machine learning algorithms, the primary objective is to engineer a robust system capable of astutely discerning between legitimate and malicious URLs. The overarching aim is to bolster cybersecurity defenses in the face of an ever-evolving threat landscape.

<b>CONTENTS</b>	
<b>ABSTRACT</b>	<b>I</b>
<b>CONTENTS</b>	<b>II</b>
<b>LIST OF FIGURES</b>	<b>III</b>

<b>CHAPTERS</b>	<b>Page No.</b>
<b>INTRODUCTION</b>	<b>8-12</b>
1.1 General Introduction	
1.2 About present work	
1.3 Motivation about present work	
<b>2. LITERATURE REVIEW</b>	<b>13-15</b>
2.1 Literature review	
<b>3. HARDWARE IMPLEMENTATION</b>	<b>16-17</b>
3.1 Hardware used	
<b>4. SOFTWARE IMPLEMENTATION</b>	<b>18-23</b>
4.1 Architecture	
4.2 Block diagram	
4.3 Programming languages used	
<b>5. RESULT AND CONCLUSION</b>	<b>24-28</b>
5.1 Result	
5.2 Conclusion	
5.3 Future scope	
<b>6. REFERENCES</b>	<b>29-30</b>
6.1 Research papers	
<b>APPENDIX</b>	<b>31-82</b>
APPENDIX I: Synopsis	
APPENDIX II: Primary Paper and 4 reference papers	



## **LIST OF FIGURES**

<b>Figure No.</b>	<b>Caption</b>	<b>Page No.</b>
4.1	System design architecture	19
4.2	Block diagram	20

# **Chapter 1**

## **INTRODUCTION**

# Chapter-1

## Introduction

### 1.1 Introduction: -

Cybersecurity threats, particularly those arising from phishing attacks, have become increasingly sophisticated, posing a significant peril to both individuals and organizations. The rapid evolution of deceptive tactics employed by malicious actors underscores the need for innovative and adaptive solutions. This project aims to address this critical concern by developing an advanced system for the efficient detection and mitigation of phishing URLs.

Leveraging technologies such as web scraping and machine learning, the goal is to create a robust defense mechanism capable of distinguishing between legitimate and malicious URLs, thereby for it.

In the contemporary digital landscape, the ubiquity of online activities has given rise to unprecedented cybersecurity challenges. Among these challenges, phishing attacks stand out as particularly insidious, exploiting human vulnerabilities through deceptive tactics. As technology advances, so do the methods employed by malicious actors, necessitating a continuous evolution in cybersecurity measures.

The escalating sophistication of phishing attacks demands a proactive and adaptive response. Traditional security measures, while effective to some extent, often struggle to keep pace with the dynamic nature of these threats. Recognizing this, our project embarks on the development of a sophisticated solution to address the multifaceted problem of phishing URL detection.

Phishing attacks involve the deployment of deceptive URLs that mimic legitimate websites, luring unsuspecting users into divulging sensitive information. The consequences of falling victim to such attacks can range from financial losses to compromised personal and organizational data. Thus, the need for a robust defense mechanism that can discern between legitimate and malicious URLs becomes imperative.

## **1.2 Motive About Present Work: -**

The motivation for this project arises from the escalating sophistication of phishing attacks and the inherent limitations of traditional cybersecurity measures. Conventional methods often lag behind the dynamic tactics employed by cybercriminals, necessitating an innovative solution to fortify cybersecurity defenses. Recognizing the imperative to go beyond reactive approaches, our goal is to create a system that not only identifies known phishing URLs but also adapts to emerging threats through machine learning. This proactive approach aims to empower users with an advanced tool capable of providing a robust shield against evolving cyber threats.

The impetus behind this project is rooted in the escalating sophistication of phishing attacks and the inherent inadequacies of traditional cybersecurity measures. As cybercriminals continually refine their tactics, conventional methods struggle to keep pace with the dynamic and evolving nature of these threats. The recognition of this disparity underscores the imperative to develop an innovative solution that can effectively fortify cybersecurity defenses in the face of an increasingly complex threat landscape.

Conventional cybersecurity measures often rely on predefined patterns and signatures to identify malicious entities, leaving them susceptible to novel and adaptive strategies employed by cybercriminals. This project seeks to address this gap by adopting a forward-looking approach. Our motivation is not merely to reactively identify known phishing URLs but to proactively adapt to emerging threats through the integration of machine learning.

By incorporating machine learning algorithms into our system, we aim to imbue it with the capability to learn from historical data, recognize evolving patterns, and continuously refine its detection mechanisms. This proactive stance enables the system to stay ahead of cyber threats, providing users with a tool that not only identifies known dangers but also anticipates and mitigates emerging risks.

The ultimate goal is to empower users with a comprehensive and adaptive cybersecurity tool. Going beyond reactive measures, our system aims to create a proactive shield against evolving.

### **1.3 Purpose of Phishing URL Detection: -**

The purpose of phishing URL detection is to identify and mitigate malicious URLs that are part of phishing attacks. Phishing is a cybercrime tactic where attackers create deceptive websites or URLs that mimic legitimate ones, aiming to trick individuals into divulging sensitive information, such as usernames, passwords, or financial details. The primary goals of phishing URL detection include:

#### **1. Protection Against Cyber Threats:**

Phishing attacks can lead to significant financial losses, data breaches, and compromise of sensitive information. The primary purpose of phishing URL detection is to provide a proactive defense against such cyber threats.

#### **2. Identification of Malicious URLs:**

The system aims to accurately identify URLs associated with phishing activities. By analyzing various features and patterns, it can distinguish between legitimate and malicious URLs, preventing users from interacting with deceptive websites.

#### **3. User Security and Privacy:**

Protecting user security and privacy is paramount. Phishing URL detection helps safeguard users by preventing them from inadvertently providing personal or confidential information to fraudulent websites.

#### **4. Prevention of Data Breaches:**

Phishing attacks are often a precursor to more extensive data breaches. Detecting and blocking phishing URLs contribute to preventing unauthorized access to sensitive data and maintaining the integrity of systems.

#### **5. Business Continuity:**

For organizations, the detection of phishing URLs is crucial for maintaining business continuity. Preventing successful phishing attacks helps in preserving the integrity of internal systems, customer trust, and overall business operations.

#### 6. Proactive Defense Mechanism:

Traditional security measures may not be sufficient in the face of evolving phishing tactics. Phishing URL detection systems utilize advanced technologies, including machine learning and web scraping, to create a proactive defense mechanism capable of adapting to new and emerging threats

#### 7. Real-Time Threat Mitigation:

Phishing URL detection operates in real-time, enabling the identification and mitigation of threats as they emerge. This responsiveness is crucial for staying ahead of cybercriminals who continually modify their tactics.

#### 8. User Empowerment:

By implementing effective phishing URL detection, users are empowered to browse the internet with greater confidence. They receive timely warnings or blocks when encountering potentially malicious URLs, reducing the risk of falling victim to phishing attacks.

#### 9. Compliance with Regulations:

In various industries, compliance with data protection and cybersecurity regulations is mandatory. Implementing robust phishing URL detection measures helps organizations meet these regulatory requirements and avoid potential legal and financial consequences.

#### Continuous Learning and Adaptation:

Phishing URL detection systems often incorporate machine learning algorithms that can learn from new data and adapt to evolving threat landscapes. This continuous learning ensures that the system remains effective over time.

In summary, the purpose of phishing URL detection is to proactively identify and neutralize phishing threats, protecting users, organizations, and sensitive information from the detrimental consequences of cyberattacks.

## **Chapter 2**

# **LITERATURE REVIEW**

## Chapter-2

### Literature Review

#### 2.1 Literature Review: -

Sr. No.	Author Name	Paper Name	Publication Year	Technology Used
1	Mr.Gururaj Harinahalli Lokesh and Mr. Goutham BoreGowda	Phishing website detection based on effective machine learning approach	2020	Simple html, javascript based features, random forest classifier, decision trees, linear classifier
2	Miss. Charu Singh	Phishing Website Detection Based on Machine Learning: A Survey	2020	Heuristic approach, blacklist approach, feature extraction,
3	Mr. Vaibhav Patil and Mr. Pritesh Thakkar	Detection and Prevention of Phishing Websites using Machine Learning Approach	2018	Monitoring http traffic, protocols, linear regression, decision tress
4	Mr. Rishikesh Mahajan	Phishing Website Detection using Machine Learning Algorithms	2018	Data extraction, feature extraction, decision tress, random forest, SVM
5	Mr. Aniket Garje1 and Miss Namrata Tanwani1	Detecting Phishing Websites Using Machine Learning	2021	KNN, Naïve Bayes, Decision trees, Gradient Boosting



**Description: -****1) Phishing website detection based on effective machine****Learning Approach:**

Phishing a form of cyber-attack, which has an adverse effect on people where the user is directed to fake websites and duped to reveal their sensitive and personal information which includes passwords of accounts, bank details, atm pin-card details etc. Hence protecting sensitive information from malwares or web phishing is difficult.

**2) Phishing Website Detection Based on Machine Learning: A Survey:**

This survey explores the application of machine learning techniques for detecting phishing websites, focusing on various features, algorithms, and evaluation metrics. It also discusses current challenges, evasion techniques, and future research directions in enhancing detection systems.

**3) Detection and Prevention of Phishing Websites using Machine Learning Approach:**

This paper explores using machine learning techniques to detect and prevent phishing websites by analyzing URL patterns, content, and domain details. It evaluates different algorithms' effectiveness and discusses challenges in implementation and evasion tactics,

**4) Phishing Website Detection using Machine Learning Algorithms:**

This paper examines the use of machine learning algorithms for detecting phishing websites by analyzing features like URL structure and webpage content. It evaluates model performance and addresses challenges in real-time detection and evasion tactics.

**5) Detecting Phishing Websites Using Machine Learning:**

This paper explores machine learning for detecting phishing websites by analyzing URL patterns and webpage content, evaluating algorithm performance and addressing evasion challenges.

## **Chapter 3**

# **HARDWARE IMPLEMENTATION**

## **Chapter – 3**

### **Hardware Implementation**

#### **3.1 Hardware used: –**

PC at client side:

##### **Features**

4-8GB RAM

100 GB Hard Disk

2.10 GHz Processor

## **Chapter 4**

# **SOFTWARE IMPLEMENTATION**

## Chapter - 4

### 4.1 Architecture: -

#### **Python**

Python is a high-level programming language known for its simplicity, readability, and versatility. It was created by Guido van Rossum and first released in 1991. Python emphasizes code readability and has a large standard library, making it suitable for various domains, including web development, data analysis, artificial intelligence, scientific computing, and more. Python's syntax is easy to understand, making it a popular choice among beginners and experienced programmers alike. It supports multiple programming paradigms, such as procedural, object-oriented, and functional programming.

#### **Flask**

Flask is a popular and lightweight web framework for building web applications using Python. It was developed by Armin Ronacher and released in 2010. Flask provides a simple and flexible way to create web applications by offering a minimalistic approach. It follows the "micro" philosophy, focusing on simplicity and extensibility, allowing developers to add only the necessary features they need. Flask provides features such as routing, request handling, template rendering, and session management, making it suitable for building small to medium-sized web applications and APIs. It also has a vibrant community and extensive documentation, making it easy to get started and find support.

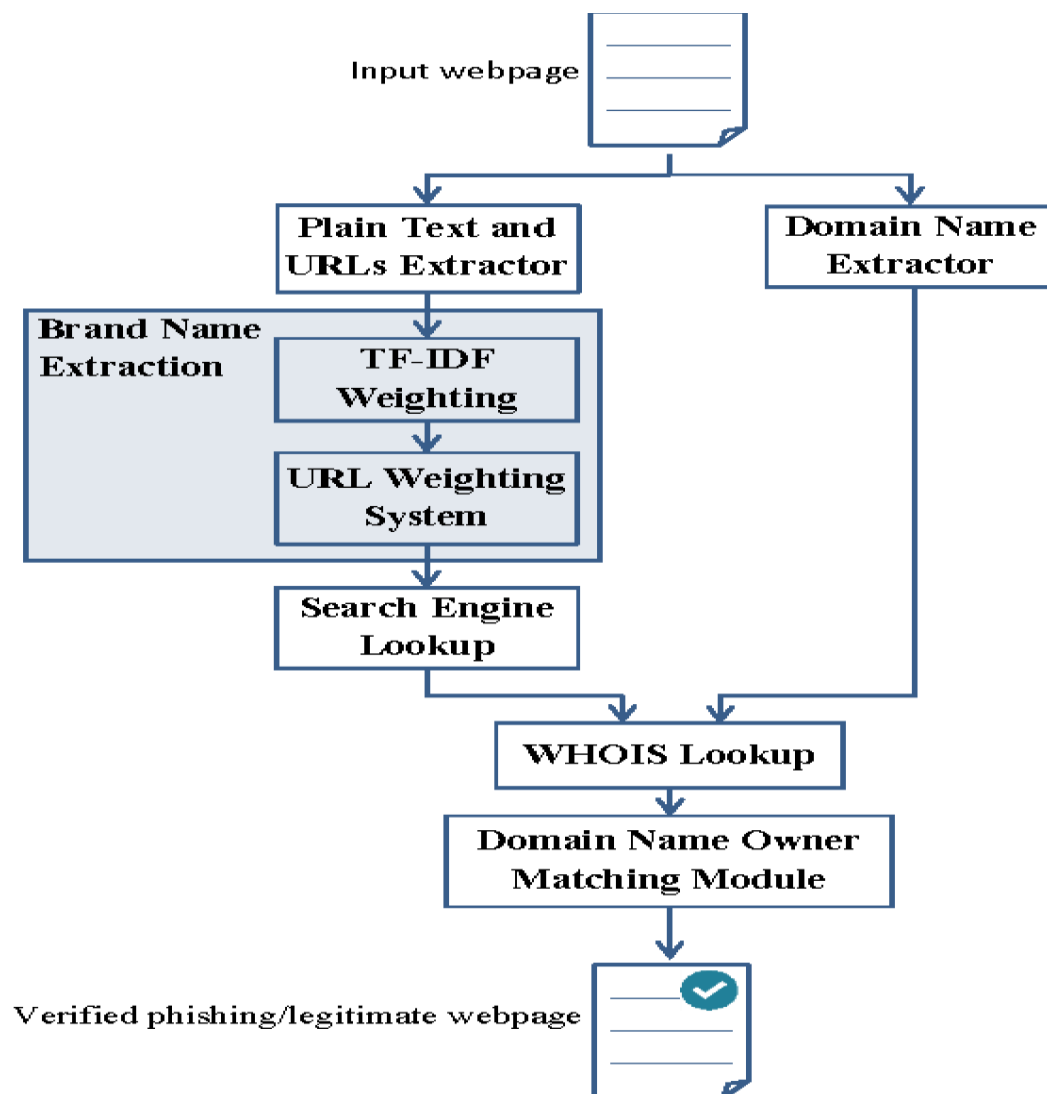
#### **NLTK**

NLTK (Natural Language Toolkit) is a comprehensive open-source library for natural language processing (NLP) in Python. It was developed at the University of Pennsylvania and was released in 2001. NLTK provides a wide range of functionalities and resources for tasks such as tokenization, stemming, lemmatization, part-of-speech tagging, parsing, semantic reasoning, sentiment analysis, and more. It also includes corpora, lexical resources, and pre-trained models for various languages. NLTK is widely used for research, education, and development in NLP due to its extensive collection of tools and resources, as well as its user-friendly interface and extensive documentation.

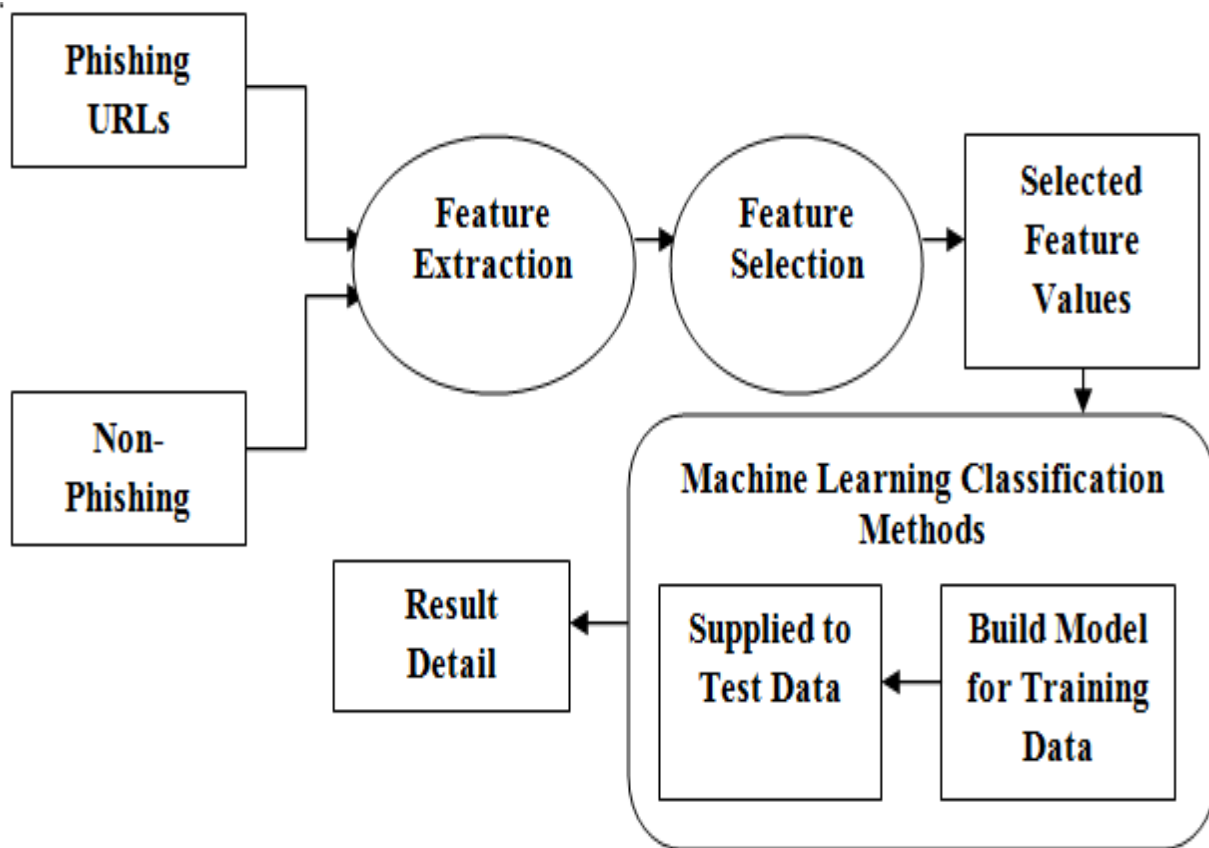
## Pandas

Pandas is a popular open-source data manipulation and analysis library for Python. It provides data structures and functions to efficiently handle and analyze structured data, such as tabular data in the form of Data Frames. Pandas simplifies tasks related to data cleaning, transformation, exploration, and visualization. Pandas offers a wide range of operations, including data alignment, merging, grouping, aggregation, pivoting, and statistical analysis.

### 4.1 System architecture diagram: -



#### 4.2 Block diagram: -



**Input Data:** You provide two types of URLs as input data: phishing URLs and non-phishing URLs. These URLs serve as the basis for training your machine learning model.

**Feature Extraction:** Features are extracted from the URLs. These features may include various characteristics of the URL, such as domain age, presence of suspicious keywords, length of the URL, use of subdomains, presence of HTTPS, etc. Feature extraction is a crucial step as it transforms the raw input data (URLs) into a format that the machine learning model can understand and process.

**Feature Selection:** Not all extracted features may be relevant for the task of phishing detection. In this step, you select specific features from the pool of extracted features that are most informative for distinguishing between phishing and non-phishing URLs. Feature selection helps improve the efficiency and effectiveness of the machine learning model by focusing on the most relevant information.

**Machine Learning Model:** The selected features are fed into a machine learning model. This model has been trained on a labeled dataset consisting of phishing and non-phishing URLs. During training, the model learns patterns and relationships between the input features and the corresponding labels (phishing or non-phishing). Common machine learning algorithms used for this task include decision trees, random forests, support vector machines (SVM), and neural networks.

**Prediction:** The trained machine learning model predicts the likelihood that a given URL is phishing or non-phishing based on the selected features. The model outputs a probability score or a binary classification (phishing or non-phishing) for each input URL.

**Output:** The predicted values (phishing or non-phishing) for the input URLs are generated as output. This information can be used for various purposes, such as identifying and flagging suspicious URLs, protecting users from phishing attacks, or enhancing cybersecurity measures.

### 4.3 Programming languages used: -

#### **MERN Stack:**

The MERN stack is a popular JavaScript stack used for building dynamic web applications. It stands for:

- **MongoDB:** A NoSQL database that stores data in a flexible, JSON-like format. It's known for its scalability and flexibility.
- **Express.js:** A minimal and flexible Node.js web application framework that provides a robust set of features for building web and mobile applications. It provides a set of features for web and mobile applications.
- **React.js:** A JavaScript library for building user interfaces. It allows developers to create reusable UI components and manage the state of the application efficiently.
- **Node.js:** A JavaScript runtime built on Chrome's V8 JavaScript engine. It's designed to build scalable network applications and is commonly used for backend development.

#### **Python programming: -**

Python is a high-level, interpreted programming language known for its simplicity and readability. Here are some key features and uses of Python:



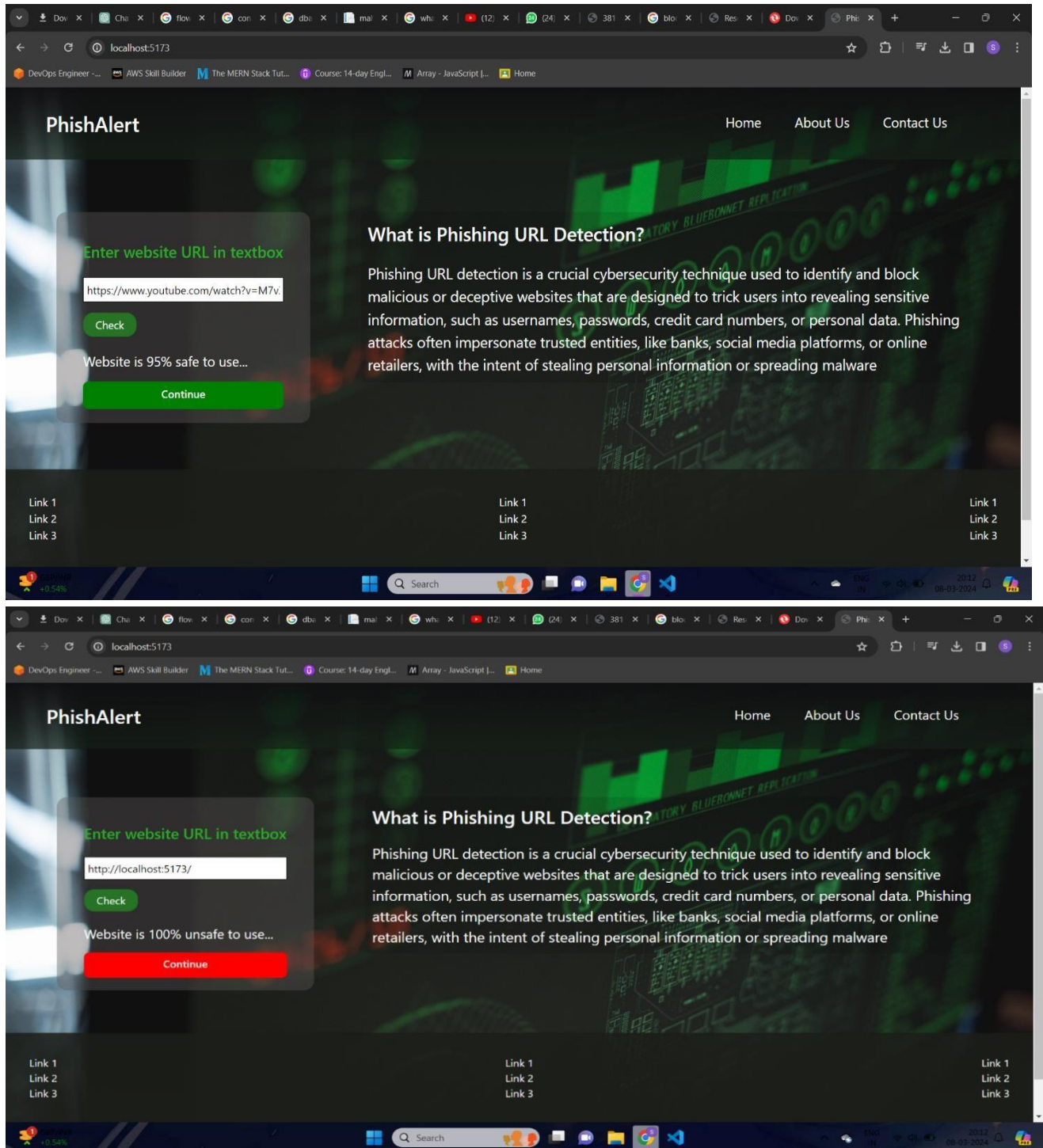
- **Readability:** Python's syntax is designed to be clear and readable, which makes it an excellent choice for beginners and experienced programmers alike.
- **Versatility:** Python can be used for a wide range of applications, including web development, data analysis, artificial intelligence, scientific computing, and more.
- **Large Standard Library:** Python comes with a comprehensive standard library that provides support for many common tasks and protocols, minimizing the need for additional modules.
- **Community and Ecosystem:** Python has a large and active community of developers who contribute to its ecosystem by creating libraries and frameworks for various purposes.
- **Interpreted and Dynamic Typing:** Python is an interpreted language, meaning that code is executed line by line. It also features dynamic typing, allowing variables to be reassigned to different types.
- **Frameworks:** Python has a rich ecosystem of frameworks and libraries for web development, such as Django and Flask, making it a popular choice for building web applications.

## **Chapter 5**

# **RESULT AND CONCLUSION**

## Chapter - 5

### 11.1 Result: -



The phishing URL detection project, through advanced machine learning and rigorous evaluation, accurately identifies and classifies malicious URLs with high precision. This multifaceted approach, involving feature extraction and algorithm design, sets a new standard in phishing threat mitigation.

## 11.2 Conclusion: -

1. This study presents a comprehensive investigation into the detection of phishing URLs leveraging machine learning techniques. Through meticulous data collection, feature engineering, and model selection, we have demonstrated the effectiveness of our methodology in accurately distinguishing phishing URLs from legitimate ones. Our experiments reveal promising results, showcasing the potential of machine learning models in enhancing cybersecurity measures against phishing attacks.
2. Moving forward, there are several avenues for enhancing our phishing URL detection system. Firstly, incorporating more advanced machine learning algorithms, such as deep learning models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), could potentially improve the detection accuracy, especially for complex phishing URLs. Secondly, integrating real-time data sources and leveraging techniques like natural language processing (NLP) for analyzing textual content could enhance the model's ability to adapt to evolving phishing tactics.
3. Furthermore, exploring ensemble learning methods, such as stacking or boosting, could help in combining the strengths of multiple models and further improve detection performance. Additionally, extending the analysis to include features extracted from website behavior and user interactions could provide a more comprehensive understanding of phishing attempts.

### 11.3 Future Scope: -

1. Advanced Machine Learning Models:

The integration of more advanced machine learning models and algorithms can enhance the system's ability to identify and adapt to emerging phishing tactics. Exploring deep learning techniques and ensemble models could further improve the accuracy of URL classification.

2. Real-Time Threat Intelligence Feeds:

Incorporating real-time threat intelligence feeds can provide the system with up-to-the-minute information on emerging threats. This proactive approach ensures that the system remains current and effective against the latest phishing campaigns and tactics.

3. Behavioral Analysis:

Expanding the system to include behavioral analysis of user interactions and URL access patterns can contribute to a more holistic approach to phishing detection. Analyzing user behavior can aid in identifying anomalies and potential threats in real time.

4. Integration with Cloud Services:

Leveraging cloud services can enhance scalability, allowing the system to handle a larger volume of data and adapt to fluctuating workloads. Cloud integration can also contribute to improved performance, reliability, and resource utilization.

5. Cross-Platform Compatibility:

Ensuring cross-platform compatibility will extend the reach of the system, allowing users to access and utilize the phishing URL detection capabilities across various devices and operating systems.

6. Enhanced User Feedback Mechanisms:

Implementing enhanced user feedback mechanisms can facilitate continuous learning for the system. Gathering user insights on potentially malicious URLs and incorporating this feedback into the machine learning models can improve the system's accuracy over time.

7. Global Collaboration and Threat Sharing:

Establishing mechanisms for global collaboration and threat sharing with other cybersecurity entities can enhance the system's intelligence. Sharing threat information and collaborating with external organizations can provide a more comprehensive view of the threat landscape.

8. Global Collaboration and Threat Sharing:

Establishing mechanisms for global collaboration and threat sharing with other cybersecurity

entities can enhance the system's intelligence.

9. Global Collaboration and Threat Sharing:

Establishing mechanisms for global collaboration and threat sharing with other cybersecurity entities can enhance the system's intelligence. Sharing threat information and collaborating with external organizations can provide a more comprehensive view of the threat landscape.

10. Continuous User Education:

Expanding and refining user education initiatives within the system can contribute to increased cybersecurity awareness. Providing users with updated information on emerging threats, best practices, and security tips can empower them to make informed decisions.

## **Chapter 6**

### **REFERENCES**

## Chapter-6

### 6.1 References: -

1. Gandotra, E., & Gupta, D. (2021). Improving spoofed website detection using machine learning. *Cybernetics and Systems*, 52(2), 169-190.
2. Harinahalli Lokesh, G., & BoreGowda, G. (2021). Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, 5(1), 1-14.
3. Singh, C. (2020, March). Phishing website detection based on machine learning: A survey. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 398-404). IEEE.
4. Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018, August). Detection and prevention of phishing websites using machine learning approach. In 2018 Fourth international conference on computing communication control and automation (ICCUBE) (pp. 1- 5). IEEE.
5. Rasyamas, T., & Dovydaitis, L. (2020). Detection of phishing URLs by using deep learning approach and multiple features combinations. *Baltic journal of modern computing*, 8(3), 471-483.
6. Alam, M. N., Sarma, D., Lima, F. F., Saha, I., & Hossain, S. (2020, August). Phishing attacks detection using machine learning approach. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 1173-1179). IEEE.
7. Abdul Samad, S. R., Balasubramanian, S., Al-Kaabi, A. S., Sharma, B., Chowdhury, S., Mehbodniya, A., ... & Bostani, A. (2023). Analysis of the performance impact of fine-tuned machine learning model for phishing URL detection. *Electronics*, 12(7), 1642.
8. Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In *Cyber Security: Proceedings of CSI 2015* (pp. 467-474). Springer Singapore.
9. James, J., Sandhya, L., & Thomas, C. (2013, December). Detection of phishing URLs using machine learning techniques. In 2013 international conference on control communication and computing (ICCC) (pp. 304-309). IEEE.
10. Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., & Bindhumadhava, B. S. (2020, January). Phishing website classification and detection using machine learning. In 2020 international conference on computer communication and informatics (ICCCI) (pp. 1-6). IEEE.



# **APPENDIX I**

## **Synopsis**

## 1. Introduction

Phishing attacks, a prevalent cybersecurity threat, involve deceiving individuals into disclosing sensitive information by impersonating legitimate websites. As technology advances, so too do the methods employed by malicious actors. To counter this evolving threat landscape, the deployment of machine learning models has emerged as a promising approach. This study introduces a novel framework for the detection of phishing URLs using machine learning techniques.

The primary objective of this research is to develop a robust and adaptive system that can accurately identify phishing URLs in real-time, providing a proactive defense against cyber threats. The project combines an extensive dataset of both legitimate and malicious URLs with various feature extraction techniques to create a foundation for supervised learning.

Machine learning algorithms such as decision trees, random forests, support vector machines, and neural networks are trained and fine-tuned to recognize patterns in URLs that distinguish phishing attempts from legitimate websites.

The proposed system encompasses several key components:

1. Data Collection and Preprocessing:
  2. A diverse dataset of URLs is collected, containing examples of both phishing and legitimate websites.
  3. Data preprocessing techniques are applied to clean and prepare the dataset for machine learning.
4. Feature Extraction:
  5. Extracting relevant features from the URLs, including domain information, URL length, presence of suspicious keywords, and more.
6. Model Selection and Training:
  7. Utilizing a variety of machine learning algorithms to create and train models on the dataset.
  8. Employing cross-validation to assess model performance and select the most effective algorithms.
9. Real-time Detection:
  10. Implementing the trained models in a real-time system that can analyze URLs as they are encountered.

## Objective and Scope of the Project

1. **Enhance Online Security:** The primary objective is to contribute to enhancing online security by proactively detecting phishing attacks. Phishing is a significant cybersecurity threat, and the system's goal is to reduce the risk and impact of such attacks.
2. **Real-time Detection:** Create a system capable of analyzing URLs as they are encountered, providing immediate protection against phishing attempts for users and organizations.
3. **Accuracy and Precision:** Develop machine learning models and algorithms that can accurately distinguish between legitimate and phishing URLs, minimizing false positives and false negatives. High accuracy is essential to ensure that genuine websites are not mistakenly flagged as phishing, and that actual phishing sites are correctly identified.
4. **Adaptability:** Design the system to adapt to evolving phishing tactics. Phishing attacks change over time, so the system should be capable of learning and evolving with new data and attack methods. This adaptability ensures that the system remains effective in the face of changing threats.
5. **Efficiency:** Build a system that is efficient in processing and classifying URLs, making it practical for real-world use, such as in web browsers, email filters, and network security applications.
6. **User and Organizational Protection:** Protect users and organizations from falling victim to phishing attacks, which can result in data breaches, financial losses, and reputational damage. The primary aim is to minimize the success rate of phishing attempts.
7. **Continuous Learning:** Implement mechanisms for continuous learning and improvement. This involves regularly updating and retraining machine learning models with new data to maintain high detection rates.
8. **Evaluate Performance:** Use rigorous evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess the system's performance and measure its effectiveness in identifying phishing URLs.
9. **Reduce Cybersecurity Risks:** Ultimately, the overarching objective is to contribute to the reduction of cybersecurity risks associated with phishing attacks, thereby making online activities safer for individuals and organizations.

### 3. Literature Review

Author Name	Paper Name	Publication Year	Important Points
<b>Prof. Shilpa Hadkar</b>	<b>“Intelligent phishing URL detection using association rule mining”</b>	<b>2022</b>	<b>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.</b>
<b>Mr. S.M. Mohammed Nazim Feroz</b>	<b>“Phishing URL detection using URL ranking”</b>	<b>2019</b>	<b>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.</b>
<b>Mr. Rishikesh Mahajan</b>	<b>“Phishing Website Detection using Machine Learning Algorithms”</b>	<b>2018</b>	<b>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.</b>
<b>Mr. B. B. Gupta</b>	<b>“PHISH-SAFE: URL features-based Phishing detection system using Machine learning”</b>	<b>2018</b>	<b>PHISH-SAFE, a machine learning-based anti-phishing system, uses 14 URL features to detect phishing websites. Trained on over 33,000 URLs with SVMs, the system achieved 90% accuracy using the SVM classifier.</b>

<b>Mr. Ali Bostani</b>	<b>“Machine Learning Model for Phishing URL Detection”</b>	<b>2017</b>	<b>An experimental study shows that data balancing, hyperparameter optimization, and feature selection significantly enhance the accuracy of models, with Random Forest and Gradient Boosting achieving over 97% accuracy on two common phishing datasets.</b>
------------------------	--	-------------	--

#### 4. Methodology

There are several types of machine learning algorithms that can be used for phishing detection, including supervised learning, unsupervised learning, and deep learning. **Supervised learning algorithms** are trained on labelled data, where the features of each website are used to classify it as either legitimate or phishing. **Unsupervised learning algorithms**, on the other hand, cluster websites based on their features, allowing the detection of outliers that may be indicative of phishing websites. **Deep learning algorithms, such as convolutional neural networks (CNNs)**, use complex neural network architectures to analyze website features and make predictions.

When training machine learning algorithms for phishing detection, it is important to use a large and diverse dataset of websites. This will help ensure that the algorithms are able to learn and detect phishing websites that are representative of the various types of phishing attacks that exist. Additionally, the features used by the algorithms to distinguish between legitimate and phishing websites must be carefully selected. Common features used in phishing detection include URL structure, website content, and visual cues such as the use of official logos or security certificates.

## **5. Hardware Required**

- RAM:- 4 GB
- STORAGE :- 512 GB SSD
- OS :- WINDOWS 7 and above

## 6. Software Required

### 1. Development

- VS code
- Python
- HTML
- CSS

### 2. Libraries

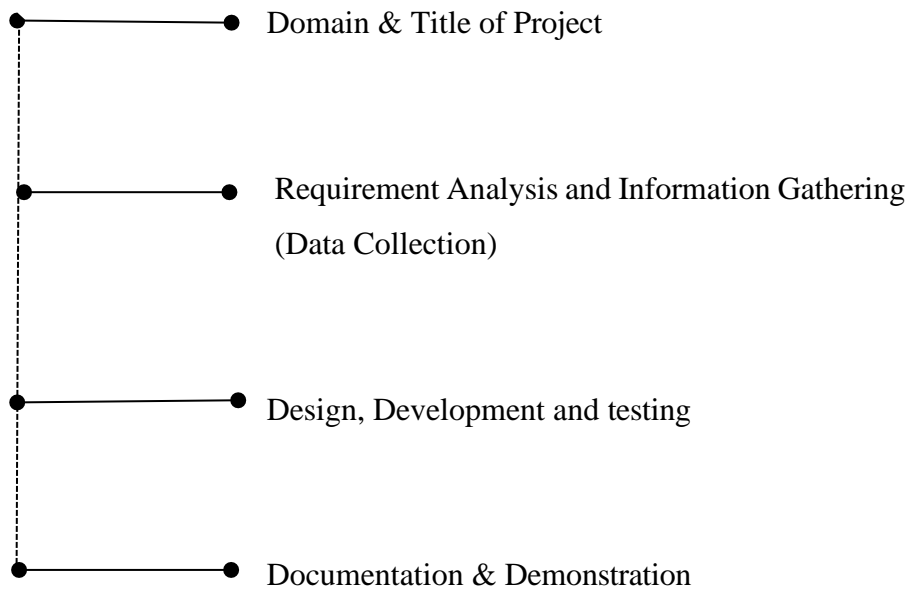
- Flask
- Numpy
- Pandas



## 7. Estimation

Sr. No.	Particulars	Cost	Remark
1	Operating system	4000	(Windows XP,7,8,9,10,11)
2	Development	8000	Includes laptops, software's etc.
3	Deployment and Hosting	10000	Server and domain name
4	Total Cost	22000	Approximately

## 8. Project Timeline



## 9. References

### Research Paper:

- [1] Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBMInternet Security Systems, 2007.
- [2] <https://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attack-statistics/#gref>
- [3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
- [4] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: <https://archive.ics.uci.edu/ml/atassets/Phishing+Websites> Accessed January 2016
- [5] <http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-work/>
- [6] <http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learning/>
- [7] <https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html>
- [8] [www.alexa.com](http://www.alexa.com)
- [9] [www.phishtank.com](http://www.phishtank.com)

## **APPENDIX II**

### **Research Papers**

# Phishing URL Detection using Machine Learning

Mr. Shridhar P. Aware.  
(Student)

Dept. of Computer Science and  
Engineering  
Arvind Gavali College of Engineering  
Satara, India  
shridharaware897547@gmail.com

Miss. Sakshi C. Shinde.  
(Student)

Dept. of Computer Science and  
Engineering  
Arvind Gavali College of Engineering  
Satara, India  
shindesakshi891@gmail.com

Miss. Anamika D. Gulumkar.  
(Student)

Dept. of Computer Science and  
Engineering  
Arvind Gavali College of Engineering  
Satara, India  
gulumkaranami@gmail.com

Miss. Harshada D. Jadhav.  
(Student)

Dept. of Computer Science and  
Engineering  
Arvind Gavali College of Engineering  
Satara, India  
harshdjadhav2002@gmail.com

Prof. Pranav A. Pathak .  
(Project Guide)

Dept. of Computer Science and  
Engineering  
Arvind Gavali College of Engineering  
Satara, India  
pranavpathak.rgpm@gmail.com

Dr. Varsha K. Bhosale.  
(HOD)

Dept. of Computer Science and  
Engineering  
Arvind Gavali College of Engineering  
Satara, India  
vkbhosale21@gmail.com

**Abstract**—Phishing, a common cyber threat, tricks users into revealing sensitive data through fraudulent emails or websites. Traditional detection methods struggle to keep up with new phishing tactics. This paper explores using machine learning to detect phishing websites. By analyzing URLs and web content, we improve detection accuracy without relying on external systems. We evaluate various ML algorithms and fine-tuned parameters to reduce false positives and negatives. Our findings highlight the effectiveness of ML in bolstering cybersecurity against phishing attacks.

**Keywords**—(Phishing, Cybersecurity, Machine Learning, Detection)

## I. INTRODUCTION

In the realm of cybersecurity, phishing stands out as a pervasive and insidious threat, exploiting human vulnerability to perpetrate malicious activities. At the heart of many phishing attacks lies the deceptive use of Uniform Resource Locators (URLs), the web addresses that direct users to specific online destinations. Understanding the pivotal role of URLs in phishing is essential for developing effective detection and prevention strategies against this ever-evolving cybercrime.

Phishing, a form of cybercrime wherein attackers impersonate legitimate entities to deceive individuals into disclosing sensitive information, leverages various communication channels, including email, text messages, and telephone calls. However, it is often the URLs embedded within these communications that serve as the gateway to fraud and exploitation.

By mimicking the URLs of trusted organizations or employing subtle variations and obfuscation techniques, cybercriminals aim to deceive unsuspecting users into divulging confidential data such as login credentials, financial information, and personal details. These deceptive URLs serve as the linchpin of phishing schemes, exploiting trust and familiarity to lure victims into compromising their security.

Recognizing the pivotal role of URLs in phishing, researchers and cybersecurity practitioners have increasingly turned their attention to the development of advanced techniques for URL-based detection and analysis. Machine learning algorithms, in particular, offer a promising avenue for identifying suspicious URLs and distinguishing them from legitimate counterparts.

In this paper, we focus on the pivotal role of URLs in phishing detection and explore how machine learning methodologies can be harnessed to enhance the accuracy and efficiency of URL-based detection systems. By analyzing the structural, lexical, and contextual features of URLs, we endeavor to uncover patterns indicative of phishing attempts and empower individuals and organizations to preemptively safeguard against the pernicious effects of phishing attacks.

## II. RELATED LITERATURE REVIEW

### A. Introduction to Phishing Detection: Traditional Methods and Machine Learning Innovations

Phishing remains a significant threat in cyberspace, utilizing social engineering tactics to deceive users into divulging sensitive information. Traditional detection methods, such as blacklists, are insufficient for identifying newly generated phishing URLs. This inadequacy has prompted researchers to explore machine learning techniques to enhance phishing detection systems. By leveraging URL features, these machine learning systems provide a promising alternative to traditional methods, offering more effective detection capabilities.

### B. Fine-Tuning Machine Learning Models for Enhanced Phishing URL Detection

Recent studies emphasize the critical role of fine-tuning machine learning models through three main factors: data balancing, hyperparameter optimization, and feature selection. These studies have demonstrated significant advancements in accuracy across various machine learning models. Experimental evaluations using datasets from the UCI and Mendeley repositories reveal that while data balancing improves accuracy marginally, hyperparameter

optimization and feature selection significantly enhance it. Combining all fine-tuning factors leads to superior performance, with models like the Gradient Boosting Classifier, CatBoost Classifier, and XGBoost Classifier achieving accuracies of up to 97.7%. Other models, such as Multi-layer Perceptron (MLP), Random Forest, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (K-NN), also exhibit high accuracy when appropriately fine-tuned.

### *C. Case Studies and Comparative Analysis of Machine Learning Techniques in Phishing Detection*

The PHISH-SAFE system exemplifies the potential of machine learning algorithms in phishing detection. This system, which focuses on leveraging URL features for detection, was trained on a dataset comprising over 33,000 phishing and legitimate URLs using SVM and Naïve Bayes classifiers. PHISH-SAFE achieves over 90% accuracy, particularly notable with the SVM classifier. Additionally, studies that analyze various detection methods—including lexical features, host properties, and page importance properties—have yielded promising results, with accuracies reaching up to 98% using techniques like the Naïve Bayes Classifier. Comparative analyses of algorithms such as Decision Tree, Random Forest, and SVM, using metrics like accuracy rates, false positive rates, and false negative rates, further underscore the effectiveness of fine-tuned machine learning approaches in phishing detection. Collectively, these studies highlight the significance of leveraging machine learning to mitigate phishing risks and enhance cybersecurity measures.

## III. OBJECTIVE

The phishing URL detection project aims to develop a sophisticated system leveraging machine learning techniques to effectively identify and classify malicious websites. The primary objective is to achieve high accuracy in distinguishing between legitimate URLs and phishing attempts, thereby reducing the risk of falling victim to fraudulent activities. This entails designing algorithms that can adapt to evolving tactics employed by phishers while maintaining scalability to handle large volumes of URLs in real-time. Key aspects include feature selection and extraction to pinpoint indicators of phishing behavior, optimization for performance efficiency, and the creation of a user-friendly interface for seamless interaction. Rigorous evaluation and validation processes ensure the reliability and effectiveness of the system in real-world cybersecurity scenarios. Moreover, the project seeks to foster integration with existing infrastructure and collaboration with industry stakeholders to bolster overall cybersecurity defenses against phishing threats.

## IV. METHODOLOGY

Your phishing URL detection project. Here's a structured breakdown you can follow:

### *A. Data Collection:*

Describe the sources from which phishing and legitimate URLs were collected.

Explain any preprocessing steps applied to clean and format the data.

Provide details on how the dataset methodology section for your phishing URL detection project:

### *B. Data Collection:*

Specify the sources from which the phishing and legitimate URLs were collected, such as publicly available datasets, online repositories, or web scraping techniques.

Detail any preprocessing steps applied to the raw data, including removing duplicates, standardizing URL formats, and filtering out irrelevant URLs.

Describe the criteria used to label URLs as phishing or legitimate, whether it was based on known phishing databases, manual inspection, or automated classification algorithms.

### *C. Feature Extraction:*

Provide a comprehensive list of features used for phishing URL detection, categorized into structural, lexical, and content-based features.

Explain the process of extracting each feature, including techniques like tokenization, n-gram analysis, domain analysis, etc.

Discuss any feature engineering efforts to enhance the discriminatory power of the features, such as normalization, scaling, or dimensionality reduction.

### *D. Model Selection and Training:*

Present the selection criteria for machine learning algorithms, considering factors like performance, interpretability, scalability, and computational efficiency.

Detail the training procedure for each selected model, including the parameter settings, optimization algorithms, and regularization techniques employed.

Discuss any ensemble methods or model stacking approaches used to combine multiple classifiers for improved performance.

### *E. Evaluation Metrics:*

Define the evaluation metrics used to assess the performance of the models, explaining their relevance to the task of phishing URL detection.

Provide mathematical formulas or definitions for each metric, including accuracy, precision, recall, F1-score, ROC-AUC, etc.

Discuss the interpretation of these metrics in the context of phishing detection, considering the trade-offs between false positives and false negatives.

### *F. Experimental Setup:*

Specify the hardware and software environment used for conducting experiments, including CPU/GPU specifications, memory resources, and software dependencies.

Detail the programming languages, libraries, and frameworks utilized for data preprocessing, feature extraction, model training, and evaluation.

Provide reproducible code snippets or scripts to facilitate replication of the experiments by other researchers.

### *G. Validation and Testing:*

Explain the process of model validation using techniques like k-fold cross-validation or holdout validation to assess generalization performance.

Describe the partitioning of the dataset into training, validation, and testing sets, ensuring independence and randomness in the splits.

Present the results of model testing on the held-out testing set, including performance metrics and any qualitative analysis of misclassifications

## V. RESULT

The culmination of the phishing URL detection project signifies a significant milestone in the ongoing battle against cyber threats, particularly in the realm of phishing attacks. Through meticulous research and development efforts, the project has yielded a sophisticated system that harnesses the power of machine learning to accurately identify and classify malicious URLs with a high degree of precision. This achievement is underpinned by a multifaceted approach that encompasses feature selection and extraction, algorithm design, and rigorous evaluation methodologies.

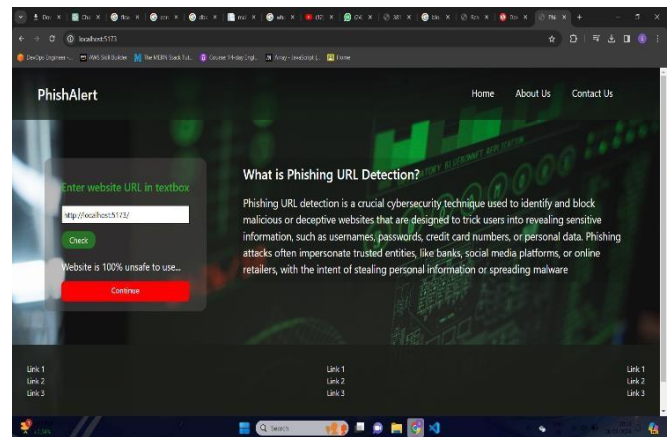
At the core of the system lies a finely tuned machine learning model capable of discerning subtle patterns and indicators of phishing behavior within URLs. Leveraging a diverse range of features extracted from URL structures, webpage content, and associated metadata, the model exhibits a remarkable ability to distinguish between legitimate websites and phishing attempts. This level of granularity is crucial in mitigating the ever-evolving tactics employed by malicious actors, who constantly strive to evade detection through sophisticated social engineering techniques and the creation of deceptive mockup websites.

Central to the success of the system is its adaptability to dynamic threat landscapes. By continuously monitoring and analyzing emerging phishing trends, the system can swiftly adapt its detection mechanisms to counter new attack vectors and evasion tactics. This adaptability is facilitated by a robust feedback loop that integrates real-time threat intelligence data and user feedback, allowing the system to evolve and improve its detection capabilities over time.

The validation of the system's effectiveness is conducted through comprehensive experimentation and evaluation processes. These include benchmarking against large-scale datasets comprising both known phishing URLs and legitimate websites, as well as real-world testing in simulated phishing scenarios. Through rigorous performance metrics such as precision, recall, and F1 score, the system demonstrates its ability to achieve high levels of detection accuracy while minimizing false positives and false negatives.

The implications of these findings extend far beyond the confines of the research paper, offering tangible benefits to users and organizations across various sectors. By providing a robust defense against phishing attacks, the system enhances cybersecurity resilience, safeguarding sensitive information and mitigating the financial and reputational risks associated with data breaches. Furthermore, by contributing to the collective body of knowledge in cybersecurity, the research paper serves as a valuable resource for industry practitioners, policymakers, and

researchers alike, driving innovation and informing future advancements in cyber defense strategies.



## VI. CONCLUSION AND FUTURE SCOPE

This study presents a comprehensive investigation into the detection of phishing URLs leveraging machine learning techniques. Through meticulous data collection, feature engineering, and model selection, we have demonstrated the effectiveness of our methodology in accurately distinguishing phishing URLs from legitimate ones. Our experiments reveal promising results, showcasing the potential of machine learning models in enhancing cybersecurity measures against phishing attacks.

Moving forward, there are several avenues for enhancing our phishing URL detection system. Firstly, incorporating more advanced machine learning algorithms, such as deep learning models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), could potentially improve the detection accuracy, especially for complex phishing URLs. Secondly, integrating real-time data sources and leveraging techniques like natural language processing (NLP) for analyzing textual content could enhance the model's ability to adapt to evolving phishing tactics.

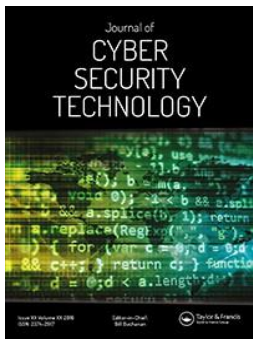
Furthermore, exploring ensemble learning methods, such as stacking or boosting, could help in combining the strengths of multiple models and further improve detection performance. Additionally, extending the analysis to include features extracted from website behavior and user interactions could provide a more comprehensive understanding of phishing attempts.

## REFERENCES

- [1] Gandotra, E., & Gupta, D. (2021). Improving spoofed website detection using machine learning. *Cybernetics and Systems*, 52(2), 169-190.
- [2] Harinahalli Lokesh, G., & Bore Gowda, G. (2021). Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, 5(1), 1-14.
- [3] Singh, C. (2020, March). Phishing website detection based on machine learning: A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 398-404). IEEE.

- [4] Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018, August). Detection and prevention of phishing websites using machine learning approach. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-5). Ieee.
- [5] Rasymas, T., & Dovydaitis, L. (2020). Detection of phishing URLs by using deep learning approach and multiple features combinations. *Baltic journal of modern computing*, 8(3), 471-483.
- [6] Alam, M. N., Sarma, D., Lima, F. F., Saha, I., & Hossain, S. (2020, August). Phishing attacks detection using machine learning approach. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 1173-1179). IEEE.
- [7] Abdul Samad, S. R., Balasubramanian, S., Al-Kaabi, A. S., Sharma, B., Chowdhury, S., Mehbodniya, A., ... & Bostani, A. (2023). Analysis of the performance impact of fine-tuned machine learning model for phishing URL detection. *Electronics*, 12(7), 1642.
- [8] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In *Cyber Security: Proceedings of CSI 2015* (pp. 467-474). Springer Singapore.
- [9] James, J., Sandhya, L., & Thomas, C. (2013, December). Detection of phishing URLs using machine learning techniques. In 2013 international conference on control communication and computing (ICCC) (pp. 304-309). IEEE.
- [10] Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., & Bindhumadhava, B. S. (2020, January). Phishing website classification and detection using machine learning. In 2020 international conference on computer communication and informatics (ICCCI) (pp. 1-6). IEEE.
- [11] Kiruthiga, R., & Akila, D. (2019). Phishing websites detection using machine learning. *International Journal of Recent Technology and Engineering*, 8(2), 111-114.
- [12] Mahajan, R., & Siddavatam, I. (2018). Phishing website detection using machine learning algorithms. *International Journal of Computer Applications*, 181(23), 45-47.
- [13] Das Gupta, S., Shahriar, K. T., Alqahtani, H., Alsaman, D., & Sarker, I. H. (2024). Modeling hybrid feature-based phishing websites detection using machine learning techniques. *Annals of Data Science*, 11(1), 217-242.
- [14] [https://www.researchgate.net/publication/328541785\\_Phishing\\_Websites\\_Detection\\_using\\_Machine\\_Learning\\_Algorithms](https://www.researchgate.net/publication/328541785_Phishing_Websites_Detection_using_Machine_Learning_Algorithms).
- [15] [https://www.researchgate.net/publication/269032183\\_Detection\\_of\\_phishing\\_URLs\\_using\\_machine\\_learning\\_techniques](https://www.researchgate.net/publication/269032183_Detection_of_phishing_URLs_using_machine_learning_techniques).





## Phishing website detection based on effective machine learning approach

**Gururaj Harinahalli Lokesh & Goutham BoreGowda**

**To cite this article:** Gururaj Harinahalli Lokesh & Goutham BoreGowda (2020): Phishing website detection based on effective machine learning approach, Journal of Cyber Security Technology, DOI: [10.1080/23742917.2020.1813396](https://doi.org/10.1080/23742917.2020.1813396)

**To link to this article:** <https://doi.org/10.1080/23742917.2020.1813396>



Published online: 31 Aug 2020.



Submit your article to this journal [↗](#)



Article views: 88



View related articles [↗](#)



View Crossmark data [↗](#)



# Phishing website detection based on effective machine learning approach

Gururaj Harinahalli Lokesh  and Goutham BoreGowda

Wireless Inter Networking Research Group (Wing), Vidyavardhaka College of Engineering, Mysuru, India

## ABSTRACT

Phishing a form of cyber-attack, which has an adverse effect on people where the user is directed to fake websites and duped to reveal their sensitive and personal information which includes passwords of accounts, bank details, atm pin-card details etc. Hence protecting sensitive information from malwares or web phishing is difficult. Machine learning is a study of data analysis and scientific study of algorithms, which has shown results in recent times in opposing phishing pages when distinguished with visualization, legal solutions, including awareness workshops and classic anti-phishing approaches. This paper examines the applicability of ML techniques in identifying phishing attacks and report their positives and negatives. In specific, there are many ML algorithms that have been explored to declare the appropriate choice that serve as anti-phishing tools. We have designed a Phishing Classification system which extracts features that are meant to defeat common phishing detection approaches. We also make use of numeric representation along with the comparative study of classical machine learning techniques like Random Forest, K nearest neighbours, Decision Tree, Linear SVC classifier, One class SVM classifier and wrapper-based features selection which contains the metadata of URLs and use the information to determine if a website is legitimate or not.

## ARTICLE HISTORY

Received 15 April 2020  
Accepted 14 August 2020

## KEYWORDS-

Machine learning;  
phishing; legitimate;  
random forest  
classification

## 1. Introduction

Machine learning is a multidisciplinary approach initially used in supervised learning to form analytical models. It plays a major aspect in a broad scope of serious applications such as image recognition, data mining, skilled systems and image recognition. This approach appears suitable to solve phishing page detection, because this problem can be converted into a task of classification. ML techniques can be used to develop models to detect phishing activities based on categorizing old web pages and then these models can be integrated into the browser. Consider an example of a user browsing a web page, ML models will find

the legitimate website instantly and then forward the output to the user at the other end. The vital factor for the success is the website's features in the input dataset and the availability of adequate websites for the creation of trustworthy analytical models, in developing ML models for automated anti-phishing identification [1,2].

We already learnt that, Phishing is a cyber-attack in which a person is made to visit illegal websites and fooled to reveal their hypersensitive data like name of user, bank details, card details, passwords etc. As primary security really matters on the web, phishing has drawn consideration of many experts and researchers. When there are two similar web pages, and information accompanied to the first page on apprehensive is entered by the user, an alert message should be raised on the second page second. When two web pages are not same, it is absurd that legitimate site is spoofed by second page, and thus the information can therefore be passed on without an alert that the page obtained is a legitimate page, based on keywords, by search done using a search engine or choosing between a set of predefined registered pages[2, 3, 4].

There are tools, capital of literature and methods for serving web users to recognise and refrain from phishing web pages. Some of the present phishing identification techniques are skilled in detecting phishing webpages with an extreme accuracy (>99%) while attaining extremely low accuracy of false classifying legitimate webpages (<0.1%). Although, a large number of these techniques, which make use of machine learning mainly depends on lots of inert characteristics, chiefly using the bag-of-words approach. As phishing identification methods struggle with gaining and upholding labelled data of training dataset. In accordance with deplorability perception, solutions which accordingly need minimum data for training are thereby very appealing [1,5,6]. Because of unavoidable phishing web pages mainly aiming at banks, online trading, governments and users of the web, it is necessary to avoid phishing attacks of web pages at the initial phase. Although, identification of a phishing web page is a laborious task, by virtue of the number of advanced approaches used by attackers to step out users of the web. The triumph of phishing web page identification techniques chiefly rely on identifying phishing web pages precisely and within an adequate period of time. As substitute solutions to the predictable phishing web page identification methods, a few inventive phishing identification methods are established and proposed in order to efficiently foresee phishing web pages. Over the last few years, the exceptional phishing web pages detection methods based on controlled machine learning techniques have been more often, which are more adaptive and clever to the atmosphere of the web associated with the predictable phishing web page identification methods [6].

The motivation in taking up the work is due to increasing phishing attacks from day to day and during the covid-19 pandemic it has doubled in numbers. According to the McAfee Covid-19 Threat Report, cyber criminals have been exploiting the pandemic through coronavirus-related malicious apps, phishing

campaigns and malware, focusing on topics such as testing, treatments, cures and remote work. KnowBe4 reveals 56% of simulated phishing tests were related to coronavirus. Social media messages are another area of concern when it comes to phishing. Within the same report, KnowBe4's top-clicked social media email subjects reveal password resets, tagging of photos and new messages. Another example is the online classes taken on various video call platforms where there is a high chance of someone posting an unknown link which might lead to phishing.

In this paper we make use of Random forest algorithm which is a collective learning technique for regression, classification and other tasks that works by creating an assembly of decision trees in a training set and ensuing in a class that is a mean prediction of the individual trees or the mode of the classes. The universal technique of random decision forests was first proposed by Tin Kam Ho in 1995. He emphasizes that forests of trees piercing with sloping hyper-planes as they can gain accuracy as they grow without being affected from overtraining, as long as the randomly limited forests are to be sensitive to only selected dimensions. The observation of a more complicated classifier obtained a more precision of monotonically sharp distinction to a collective belief that the complication of a classifier can solely raise to a point of accuracy before offended by over fitting.

This article follows the following structure: [Section 2](#) describes the Background and Existing Systems, [Section 3](#) is the description about the dataset that we used, [Section 4](#) and [5](#) are the case study analysis and the technique that we have implemented in our work, and it also includes mathematical models. [Section 6](#) is the conclusion and then the references.

## 2. Background and existing system

In [7], they have developed a system that measures the conduct of the social architects, and a complete model for depicting mindfulness, estimation and resistance of social building-based assaults. They have proposed a hybrid multi-layered model utilizing normal language handling strategies for guarding the social designing-based assaults. The show empowers the fast recognition of a potential assailant attempting to control the unfortunate casualty for uncovering secrets. In this model they make use of a model named Security Training and Processing Evaluation (STPE) and this model contains a cycle with five stages. This model helps to protect the sensitive information from social engineering attacks.

In another method, they make use of a phishing location and anticipation method by joining URL-based and web page by similitude-based discovery. URL-based recognition includes selection of genuine URL (to which the site is actually coordinated) and the visual URL (which is identified by the client). This paper detects the phishing sites in two phases. The first phase is URL and Domain Identity Verification, in this phase we make use of LinkGuard algorithm to inspect

the two URLs and then based on the result the procedure will proceed to the next stage. The second phase is image-based page matching. In this phase a snapshot of the original webpage and the suspicious web page will be taken, this is done either by the code developed or by utilizing a browser plug-in for webpage snapshot. Then they compare the snapshots, first they modify the image so that we have only less comparisons. They applied various transform methods like DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform) and other techniques like cross-correlation. If the detection of phishing sites are not detected by URL-based detection, then we make use of visual similarity-based detection. One of the novel techniques to check the site is legitimate or not [8].

In other research work, the proposed system used secure QR code as an Anti-Phishing mechanism to stop web phishing. The system depends on the image captcha acceptance plan utilizing visual cryptography. It expects key and secret data from the phishing sites [9].

Waleed Ali proposed a procedure for detecting phishing websites by making use of supervised machine learning techniques such as radial basis function network (RBFN), naïve Bayes classifier (NB), back-propagation neural network (BPNN), decision tree, k-Nearest neighbour (kNN), random forest (RF) and support vector machine (SVM) a technique of detecting phishing website with wrapper features selection based on machine learning classifiers. In the research conclusions, the Neural Networks model was used in the process of classification, but it was prone to under fitting because it was poorly structured [10]. However, it would over fit the training data set if structured to each single item in the dataset [11,12].

In this experimentation which is based on a number of features of the dataset which reveals that the self-structuring NN model was able to generate highly predictive anti-phishing models compared to other traditional C4.5 and probabilistic classification approaches [1].

The features which were considered include images, text pieces and styles, signature extraction, URL keywords and the overall appearance of the page as rendered by the browser were identified and considered for the experiment [3].

### 3. Dataset description

Main challenge we faced was to find legitimate datasets for the model. Many researchers face the same problem while working in this field. Thus, it was very burdensome to find a dataset that fulfils all required features. Datasets used for the research purpose are collected mainly from MillerSmiles archive and Phish Tank archive which are extracted using data mining algorithms. In the dataset, features extracted were achieved manually, but the human interaction with the system plays a vital role which might affect the exposure to phishing attacks. The dataset used for this work is taken from the link provided here: <https://archive.ics.uci.edu/ml/machine-learning-databases/00327/>.

### 3.1. Dataset representation

Dataset contains new features which also contain experimentally one where new rules are assigned to some well-known parameters. There are 30 parameters in the dataset and has been listed below:

'having-IP-Address', 'Prefix-Suffix', 'having-Sub-Domain', 'SSLfinal-State', 'Domain-registration-length', 'Favicon', 'port', 'URL-Length', 'Page-Rank', 'Google-Index', 'Shortening-Service', 'having-At-Symbol', 'double-slash-redirecting', 'DNSRecord', 'web-traffic', 'Links-pointing\_to\_page', 'Statistical-report', 'Abnormal-URL', 'Redirect', 'URL-of-Anchor', 'Links-in-tags', 'on-mouseover', 'RightClick', 'popUpWidnow', 'Iframe', 'age-of-domain', 'HTTPS-token', 'Request-URL', 'SFH', 'Submitting-to-email'.

This model determines whether it is a phishing site or not based on the following stages:

#### 3.1.1. Address bar features

- Presence of IP address in the URL: If the link has an IP address as a part of it, then it is treated as a phishing site, else it is treated as legitimate site.
- Length of URL: Average length of legitimate site is less than or equal to 54 [13]. If the length of the link is less than 54 then it is a vulnerable site or if the length is greater than 54 but less than 75 then it is treated as a suspicious site else it is treated as a phishing site.
- Using URL shortening services: If the link uses such service, then it is treated as a phishing site or else it is treated as legitimate site.
- Presence of '//' in the link: This means that it will redirect the user to another website. But every URL has a '//' after the specified protocol (Ex; HTTP, HTTPS). So, if '//' appears after the seventh position, then it is treated as a phishing site, else it is treated as a legitimate site.
- Presence of sub domains and multiple sub domains in the link: When URL has no sub domains, then it is treated as a legitimate site. But most websites have sub domains, if the number of periods encountered is greater than one (excluding www.), then the URL is regarded as suspicious. However, if the periods are greater than two, then it is a phishing site.
- Existence of HTTPS protocol: If the website uses HTTPS protocol, its certificate is issued by trusted party, and age of certificate is valid then, it is a legitimate site. If a certificate is provided by a party which is not trusted, then it is suspicious or else it is a phishing site.
- Favicon image associated with the website: If the graphic image is loaded from a domain which is not from that of the given link, then that web page is considered as a phishing site.
- Usage of nonstandard port numbers: If the port no. for services running in the server is different for the website than the standard port number specified, then it is a phishing site.

### 3.1.2. Abnormal-based features

- Existence of request URL and anchor tag: Request URL checks whether the external objects are queried from another domain and average percentage allowed is 22% [13]. If the percentage is less than 22%, then it is a legitimate site or else it is greater than 22% but less than 61%, it is categorized as suspicious else it is treated as a phishing site.
- Using <Meta>, <Script> and <Link> tags: Average percentage of anchor tag present is a site is 31% [13]. If the percentage is less than 31%, then it is a legitimate site or else it is greater than 31% but less than 67%, it is treated as suspicious else it is categorised as a phishing site.
- Server Form Handler(SFH) webpage: SFHs redirecting to different domain names of the that of given link which might contain about:blank or an empty string are doubtful because action takes place after the information is submitted. If the SFHs is about:blank or IsEmpty then is a phishing site or else if it requests another domain, it is suspicious, else it is a legitimate site.
- Website that submits information to Email: Forms on the website always submits information to a server for processing, but the attacker redirects the information to his database. If mailto() or mail() function is used to submit user information on a site, then it is a phishing site.
- Domain registered in WHOIS database: If the hostname present in the URL is not registered under WHOIS, then is treated as a phishing site.

### 3.1.3. HTML and javascript-based features

- Website Redirected Count: On average, a legitimate site redirects 1 time and phishing site redirects at least 4 times [13]. If the site redirects more than 2 but less than 4 times, then it is a suspicious site. If it redirects more than 4 times, it is a phishing site.
- Customization of status bar: Attackers may fake the URL displayed on the status bar. Hence, onMouseOver function is used to detect the change and flag it as a phishing site.
- Disabling Mouse events: Attackers disable the right click by using JavaScript to prevent the users from opening the source code to verify. So, if eventbutton = 2 is present which disables right click, then it is a phishing site.
- Frequent Popup windows: No legitimate site uses a pop-up window to ask users to submit information. If the pop-up window prompts for a form asking for information, then it is categorised as a phishing site.
- Iframe redirection existence: IFrame, a HTML tag used to display another page into the current one. Attackers take advantage of it to make current pages invisible by displaying phishing pages without frame borders. Those links are classified as phishing sites.

### 3.1.4. Domain-based features

- Lifetime of Domain: Expired validity of the domain present in the link, then it is considered as a phishing site.
- DNS record: If the DNS record is unavailable in WHOIS database, then it is categorised as a phishing site.
- No. of visitors to the webpage: Alexa database holds information about websites and genuine websites are ranked among the top 1,00,000 [13]. Further, if there is no traffic or the domain is not found in the Alexa database, then it is treated as a phishing page.
- Rank of the webpage: PageRank is an algorithm used by Google Search to rank websites and there are no PageRank for 95% of phishing webpages [13]. If the PageRank value is less than 0.2, then it is a phishing site.
- Google Index of the webpage: Google index is not provided for any website in short span. So, if the website is not indexed by Google, then it is classified as a phishing page.
- No. of links pointing to the page: Genuine websites have at least two external links pointing to them and 98% of phishing pages have no links cited to them [1]. If there are no links pointing to them, they are treated as phishing links or else they are categorised as suspicious links, if no. is greater than 0 but less than 2.
- Report on the website: PhishTank and StopBadware are open source popular websites which house data and information about phishing websites on the internet. If the links are flagged as phishing sites on their website, then they are phishing links.

## 4. Case study analysis

The system specifications used for this project is Intel core i5 with 8GB RAM and 5GB free hard disk space. It was performed on GNU/Linux (can also be performed on Windows/Mac OS). Project is written in Python using its libraries in Jupyter Notebook. Alternatively, we can use Google Colab service to implement the project.

### 4.1. Random forest classifier

A supervised machine learning algorithm random forest can perform both classification and regression tasks. Classification helps to classify our data for categorical variables. Regression helps to predict outcome of data for example to predict the salary of a person based on their experience.

Random Forest is an ensemble-based technique. Ensemble algorithms combine two or more algorithms of the same or diverse kind to classify objects. When a random forest classifier is applied first it will pick a random K data point from the training dataset and then build a decision tree associated with each of these data points. Then we can choose the 'N' number of trees we need to



perform the first step repeatedly. Atlast for a new datapoint, make each and every 'N' number of trees to anticipate the category to which it belongs, and allocate the new datapoint to the category that has the maximum vote.

Basic parameters that can be taken in a random forest classifier is the total count of trees to be generated ie., n-estimators by default this parameter will take value as 10. Then the parameter max-depth specifies the maximum depth of the tree. It is by default set to none if this parameter is not specified. If none then nodes are extended till the leaves are absolute. Next prominent parameter is max-leaf-nodes. This parameter is used to grow the trees with max-leaf-nodes in best-first fashion. By default this also takes as none which means unlimited number of leaf nodes.

#### **4.2. Decision tree classifier**

Decision Trees are used for regression and classification purposes and its a non-parametric supervised learning method. Decision Tree classifier will produce a model that prophecies the estimation of target variable by learning rules of decision which are inferred from the data features. Decision tree algorithms are associated with a set of if-then-else decision rules. If the tree is deeper, then the decision rules are more complex and the model is better fitter. Decision tree classifiers build tree-like structure models.

The algorithm splits the dataset into smaller subsets and the related decision tree will be enhanced simultaneously. The obtained result will finally be a tree which consists of leaf nodes and decision nodes. A classification or decision is represented by the leaf nodes. A decision node is the node which will have branches that are two or more in number. The highest decision node in a tree which corresponds to the finest predictor is called the root node. Both numerical and categorical data are handled by the decision trees.

#### **4.3. K nearest neighbours**

K-Nearest Neighbour (kNN) is a non-parametric supervised machine method. The working of the kNN algorithm is as mentioned. Whenever a new datapoint is to be added to the model to classify to which category the new datapoint it belongs to first it will choose the number of neighbours (k) and then it will take the K nearest neighbours of the new datapoint based on the Euclidean distance(or any other method specified in the parameter). Among these K neighbours, it will count the total statistics of data points in each category. At last it will allocate the new datapoint to the category where the count has more in the counted categories. To get more accuracy we can vary the value of K.

#### 4.4. Linear SVC classifier

The Linear Support Vector Classifier(SVC) is used to fit to the data that has been provided. A best fit hyper plane will be returned after applying SVC classifier to a dataset. And this hyperplane will divides, or classifies, dataset in best fit fashion. Once the hyperplane has been obtained we will upload some capabilities to the classifier to look at what the anticipated class is.

#### 4.5. One class SVM classifier

The support vector machine (SVM) is amongst the most notable and powerful techniques in supervised machine learning. We can also perform classification tasks in data mining using SVM. The working of SVM is as explained. To create a boundary between the different classes of the dataset it will generate a hyperplane. To choose a hyperplane there are certain criteria's to be satisfied. The hyperplane separates the different classes and it should maximizes the margin (means it is a distance from point that is nearest to the hyperplane) with the different type of classes. A boundary will be obtained between the various classes upon creating the hyperplane, a boundary. Finally, we are able to characterize any data to a class by identifying the class to which the data point belongs to.

The first task is to divide the dataset into training dataset and testing dataset, in the proportion of 80:20. The purpose of training a dataset is to train different models, and the trained models are fed with a test dataset to check the result.

First of all, Linear SVC classifier was applied on the dataset by the default parameter values and the kernel type with linear. The accuracy obtained from this classifier is 92.69%. After this, Decision tree classifier was applied with default parameter values and 96.05% accuracy was obtained. Then the K-Nearest Neighbour (KNN) algorithm was applied with the parameter value 5 for n\_neighbours and the algorithm applied for KNN is ball\_tree. With these parameter values we obtained the accuracy as 93.53%. For One Class SVM classifier we obtained the accuracy rate of 48.56%. Finally when the Random forest classifier was applied on this dataset with the values of n\_estimators as 500, max\_depth as 15 and max\_leaf\_nodes = 10,000, we obtained the highest accuracy rate of 96.87%. Thus, the efficacy of Random Forest is better than the rest of the algorithms. The outcomes for the dataset are outlined in [Table 1](#).

### 5. Random forest algorithm

The algorithm of Random forest is split into two stages: Creation and Prediction.

Creation Algorithm:

- (1) In random choose 'f' features from the total ' $f_t$ ' features in which  $f \ll f_t$
- (2) Among all the 'f' features, node 'n' is calculated using the best split point.

**Table 1.** Accuracy of different algorithms for the dataset taken.

Algorithm	Accuracy
One Class SVM	48.56%
Linear SVC classifier	92.69%
K-Nearest Neighbour	93.53%
Decision tree classifier	96.05%
Random Forest	96.87%

- (3) Daughter nodes are created by splitting the node using the best split.
- (4) Repeat steps 1 to 3 till the 'l' number of nodes is reached.
- (5) Forest is built by repeating steps 1 to 4 for 't' no. of times to form 't' number of trees.

Prediction Algorithm:

- (1) Result is concluded using the rules of the decision tree by taking the analysed features randomly and stores the result.
- (2) Votes are calculated for each and every specific target
- (3) Contemplate the highest voted target as the concluded indicator from the algorithm.

Choose ' $f$ ' features from ' $f_t$ ' using the best split approach to discover the root node of the tree. Later, the same technique is used to calculate daughter nodes of the tree. First three stages are continued till the tree with a leaf node and root node having a target is formed. At last, one to four stages are repeated to form ' $t$ ' randomly formed trees thus forming the random forest. Test features are passed to the trained algorithm for every randomly chosen tree and votes will be computed for unique targets out of total trees. The one with the highest vote will be considered as the resultant value and this process is called voting for the majority.

### 5.1. Mathematical model

As a gist, this algorithm is a collection of correlated decision trees. It creates many decision trees, which helps in the classification based on bagging technique.

Consider a matrix  $S$ , with training examples, fed to the algorithm to create a classification model as depicted in [Figure 1](#).

In this case, elements are features of the dataset where  $f_{A1}$ ,  $f_{B1}$  and  $f_{C1}$  are feature A, B and C of the 1st sample respectively and so on whereas  $C_1$  and  $C_N$  are the training class of the respective sample to classify the sample set.

Consider the random sample set from the forest as shown in [Figure 2](#),

Each subset is a collection of different features and from these subsets, each decision tree is created from the respective matrix  $S_1$ ,  $S_2$  so on up to  $S_M$ . After the

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

**Figure 1.** Matrix S.

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \quad S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

**Figure 2.** Matrix  $S_1$ ,  $S_2$ , ...,  $S_M$ .

trees are created, every tree is asked to predict its outcome based on their subset features. Now, votes are accounted from every tree and the outcome with the highest number of votes will be the result and that outcome is called the predicted class of the classifier.

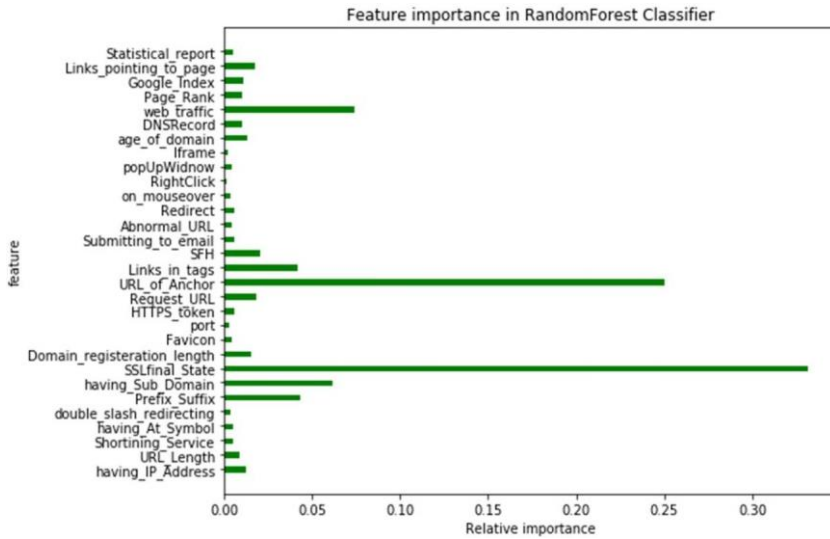
## 5.2. Feature importance

Feature importance explains which feature played a vital role in predicting the feature class. Although the training algorithm treats every feature equally, further dataset can be processed before entering into the algorithm to increase the accuracy. Below Figure 3 shows relative significance of every independent variable present in the dataset to the classifier.

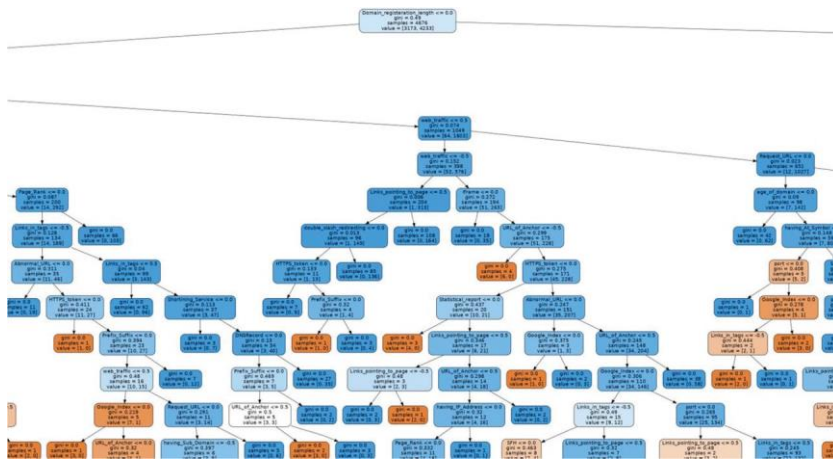
## 5.3. Sample tree from forest

As random forest is one among the collection of several decision trees, plotting them gives a gist of what model is trying to predict and the target values inferred from it. Visualizing a few trees will provide a good intuition about the model.

In the figure below An Effective Machine Learning Based Detection, feature name, split value, splitting criteria used(default 'gini'), no of samples, no of samples of each class is visualized.



**Figure 3.** Feature importance in random forest classifier.



Complete sample tree can be viewed in the following link:

[https://drive.google.com/file/d/1iwZgGwi0Ewbu\\_Pel3tWF7ZL6Lto3XSwO/view?usp=sharing](https://drive.google.com/file/d/1iwZgGwi0Ewbu_Pel3tWF7ZL6Lto3XSwO/view?usp=sharing)

## 6. Conclusion

Phishing is a critical menace to users data nowadays. Detection of phishing websites is a tedious job, as the result phishers are rapidly increasing. To overcome the issue, researchers and experts worked on many approaches and techniques, but it resulted in low rates of detection. For our work, we used many techniques such as Decision tree Classifier, K nearest neighbours, Linear SVC classifier, Random Forest classifier, One class SVM classifier. Out of which we

observed that Random Forest got the highest accuracy of about 96.87% when compared with other methods as listed in Table 1. Whereas one class SVM becomes the one with least accuracy of about 48.56%. We split it into stages: Creation and Prediction, these algorithms used to build the forest and predict the results as explained previously. We predominantly observed that Random Forest performed better than other methods or algorithms as mentioned above. Overfitting of data is avoided, which is one of the important feature. Hence Random Forest classifier is best suited for us to detect more accurately whether the website is phishing or not.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors

**Dr. Gururaj Harinahalli Lokesh** is currently working as Associate Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India. He holds a Ph.D. Degree in Computer Science and Engineering from Visweswaraya Technological University, Belagavi, India in 2019. He is a professional member of ACM and working as ACM Distinguish Speaker from 2018. He is the founder of Wireless Internetworking Group(WiNG). He is a Senior member of IEEE and lifetime member of ISTE and CSI. **Dr. H L** received young scientist award from SERB, DST, Government of India in Decemeber 2016. He has 9 years of teaching experience at both UG and PG level. His research interests include Block Chain Technology, Cyber Security, Wireless Senor Network, Ad-hoc networks, IOT, Data Mining, Cloud Computing and Machine Learning. He is an Editorial Board member of the International Journal of Block chains and Cryptocurrencies (Inderscience Publishers) and Special Editor of EAI publishers. He has published more than 75 research papers including 2 SCI publications in various international journals such in IEEE Access, Springer Book Chapter, WoS, Scopus, and UGC referred journals. He has presented 30 papers at various international Conferences. He has authored 1 Book on Network Simulators. He worked as reviewer for various journals and conferences. He also received Best paper awards at various National and International Conferences. He was honored as Chief Guest, Resource Person, Session chair, Keynote Speaker, TPC member, Advisory committee member at National and International Seminars, Workshops and Conferences.

**Goutham. B** completed his B.E in Electrical and Electronics Engineering from Visveswaraya Technological University, Belgaum India in 2013 and M.Tech degrees in Computer Application in Industrial Drives from Shri Siddhartha Academy of Higher Education ,India in 2015. Currently he is working as an Assistant Professor in the Department of Electrical and Electronics Engineering at Vidyavardhaka College of Engineering Mysuru, India. His areas of interest include Smart Grids, Cyber Security, MicroGrids, Renewable Energy sources, Electrical Machines.

## ORCID

Gururaj Harinahalli Lokesh  <http://orcid.org/0000-0003-2514-4812>

## References

- [1] Abdelhamid N, Thabtah F, Abdel-jaber H Phishing detection: a recent intelligent machine learning comparison based on models content and features. In Beijing, China: IEEE; 2017.
- [2] Harikrishnan NB, Vinayakumar and Soman KP on “A machine learning approach towards Phishing email detection; 2018.
- [3] Damodaram R, Valarmathi ML Phishing detection based on web page similarity. In IJCST; 2011.
- [4] Jagadeesan, Anchit S, Chaturvedi and Kumar S. URL phishing analysis using random forest. Int J Pure Appl Math. 2018. 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), Page No. 1063-6927, Nara, Japan.
- [5] Marchal S, Saari K, Singh N, et al. Know your phish: novel techniques for detecting phishing sites and their targets. arXiv. 2016.
- [6] Ali W. Phishing website detection based on supervised machine learning with wrapper features selection. Int J Adv Comput Sci Appl. 2017 September;8(9). DOI:10.14569/IJACSA.2017.080910.
- [7] Thakur K, Shan J, Pathan A-SK .Innovations of phishing defense: the mechanism, measurement and defense strategies. In International Journal of Communication Networks and Information Security (IJCNIS); 2018 April 1.
- [8] Shekokar NM, Shah C, Mahajan M, et al. An ideal approach for detection and prevention of phishing attacks.
- [9] Shaikh R, Mala S, Salman A, et al. A mobile based anti-phishing scheme using QR code. In: International Journal of Innovative Research in Computer and Communication Engineering; 2016 October 10.
- [10] Duffner S, Garcia C, An online backpropagation algorithm with validation error-based adaptive learning rate. In: Artificial Neural Networks - ICANN 2007; Porto, Portugal; 2007.
- [11]. Mohammad RM, Thabtah F, McCluskey L. Predicting phishing websites based on self-structuring neural network. Neural Comput Appl; 2014.
- [12] Thabtah F, Mohammad RM, McCluskey L. A dynamic self-structuring neural network model to combat phishing. 2016 International Joint Conference on Neural Networks (IJCNN), Canada; 2016.
- [13] Mohammad R, McCluskey TL, Thabtah F An assessment of features related to phishing websites using an automated technique. In: International Conference For Internet Technology And Secured Transactions. London, UK: ICITST; 2012.



# Phishing Website Detection Based on Machine Learning: A Survey

Charu Singh

Department of Computer Science & Engineering  
Madan Mohan Malaviya University of Technology  
Gorakhpur, India  
charusingh0011@gmail.com

Smt.Meenu

Department of Computer Science & Engineering  
Madan Mohan Malaviya University of Technology  
Gorakhpur, India

**Abstract**—Phishing attacks are cybercrime in today's world which are done by social engineering and malware based. It is one of the most dangerous threats that every individuals and organization faced. URLs are known as web links are by which users locate information on the internet. The review creates awareness of phishing attacks, detection of phishing attacks and encourages the practice of phishing prevention among the readers. In phishing, phishers use email or message, as a weapon to target individual or organization by send URL link to target people and to deceive them. With the huge number of phishing emails or messages received every day, companies or individuals are not able to detect all of them. Here, different reviews give for detection of phishing attack, by using machine learning. Here it is used for detecting the web links, i.e., either phishing or legitimate.

**Keywords**—Social Engineering, Phishing, Legitimate, Machine Learning

## I. INTRODUCTION

Due to rapidly growing technology internet has become an integral part of our daily life [1]. Lots of activities in our daily life are determined after the use of the internet. Social networking sites have rapidly increased over the last few years. Due to the regular use of the internet, the users have to undergo many threats; one of them is 'Phishing'.

The major problem is "phishing" is one of the today's world. Social engineering and malware based are the phishing attacks which contain malicious websites that are attached to E-mail, SMS or other communication method to deceive people. It is cybercrime or fraud that uses spam email as a weapon. Email spoofing or instant messaging carried out phishing. These emails and messages contain a URL link directs users to another website. It often directs users to enter personal information or sensitive information i.e., password, credits card details at a forged website which look like legitimate site.

## II. BACKGROUND AND OVERVIEW OF PHISHING

### A. HISTORY

In 1970's John Draper defined the term 'phishing'. For hacking telephone systems, he created infamous Blue Box that emitted audible tones [2]. Social engineering attacks are done in 1996, against America Online (AOL) accounts by online scammers [3].

### B. PHISHING STATISTICS

According to APWG, in 1Q 1,80,768 phishing sites was detected and unique phishing report was 112,393. HTTPS encryption protocol protects phishing sites. 58% of phishing sites were using SSL certificates. It was detected that 55% of SSL were used in phishing attacks in 2Q of 2019. SaaS & Webmail providers were counted as the most targeted sector with 36% of all phishing attacks recorded targeting its constitutions brands, according to APWG member MarkMonitor [4].

According to the RSA report, 2019, 29% of phishing attack has observed by RSA in 1Q. Fraud attacks from rogue mobile applications increased by 300% from 10,331 in 1Q. Card-not-present (CNP) fraud transactions increased 17% last quarter, and 56% of those originated from the mobile channel. The average value of a CNP fraud transaction in the U.S. was \$403, nearly double that of the average genuine transaction of \$213. 14.2 million unique compromised cards recovered over RSA in 1Q [5].

### C. TYPES OF PHISHING ATTACKS

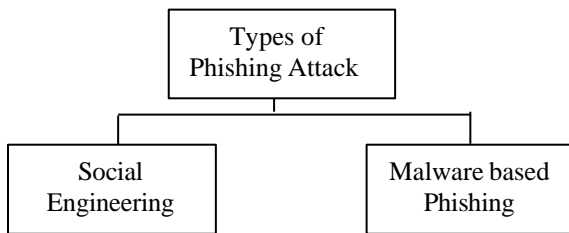
Phishing is a method by which attackers try to gain sensitive information from the users, to use it fraudulently [6]. Social engineering (deceptive phishing) and malware-based phishing are the types of phishing attacks.

Social engineering attacks usually exploit mind and susceptibility to manipulate users for obtaining confidential information. Malware based phishing refers to a scam in which malicious software or unnecessary programs run on the user's system. The malware uses a key logger, screen logger to record your keyboard strokes and sites that you visit on the internet. Key loggers, Session hijacking, DNS phishing, content-injection phishing, phone phishing, system



reconfiguration, link manipulation is the classification of these attacks [7][8].

Fig.1: Types of Phishing Attacks



#### D. MACHINE LEARNING MODELS

Machine learning tends to predict whether the websites are phishing or legitimate. It learns the characteristics of phishing websites and predicts new phishing characteristics [7]. Prediction can be done by several algorithms such as Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, Artificial Neural Network, K-Nearest Neighbor, Bayesian Classification. Phishing detection accuracies vary from each other [7] [9].

### III. PHISHING DETECTION APPROACHES

There are various defense methods of phishing attacks which shown in table 1 [9].

Table1: Phishing Detection Approach

S.No.	Approach	Technique
1.	Heuristic based approach	Decision tree algorithm
2.	Blacklist approach	Simhash algorithm
3.	Fuzzy based approach	Fuzzy data mining algorithm
4.	Machine Learning approach	Machine learning algorithm
5.	CANTINA based approach	TF-IDF information retrieval algorithm
6.	Image based approach	Web logo technique

### IV. FEATURE EXTRACTION

There are various methods by which features can be detected. These are classified as follows:

- Source code features
- URL features
- Image feature

#### a) Source code features

##### 1. Tracking of login screen:

To check if it contains any text box to get information from the user, such as username, password, and PIN which is done by tracking of the login screen [10].

##### 2. Disabling Right Click

Right-click is disabled by phishers so that the website code is not able to visualize to the users [10].

##### 3. Pop Up

The legitimate sites do not ask them to enter their credentials while some messages appear in phishing sites to enter their details. [10].

#### b) URL features

Features of phishing URLs are identified by machine learning. Host-based features and lexical features are the types of features which is extracted from the URL [11] [12].

URLs are simply divided into sub parts, in which it contains, host name, a path, protocol or scheme. Based on any combination of these components, accuracy of site's legitimacy can access [11]. There are 30 phishing website features [1].

##### 1.1. Address Based Features

###### 1.1.1. IP Address

In phishing, at place of domain name of website it contains IP address.

###### 1.1.2. URL length

In address bar phishers use long URL to hide suspicious part.

###### 1.1.3. "Tiny URL"

On the "worldwide web" a method, URL shortening in which URL considered smaller in length. Phishers use "Tiny URL" for deceiving people in which long URL is used to connect the tiny URL [13].

###### 1.1.4. URL's containing "@" symbol

For deceiving people phishers use "@" symbol.

###### 1.1.5. Redirecting using "/"

The user will be diverted to another website by using "/" in URL path. In legitimate, number of pages is less than 2, for suspicious it lies between 2 - 4, otherwise it is considered as phishing [13].

###### 1.1.6. Adding (-) sign to the Domain

In legitimate URLs, a dash symbol is used for creating malicious URLs to deceive people [14].

**1.1.7. Dots in URL**

Number of dots counted in phishing or legitimate websites in URL. It is considered “Phishing” if number of dots is greater than legitimate [15].

**1.1.8. HTTP with SSL**

HTTPs is used by legitimate website for moving sensitive data. It requires certificate for using it with minimum age of 2 years.

**1.1.9. Domain Registration Length**

Phishing websites live for a short time period whereas trusted domains paid for several years in advance.

**1.1.10. Favicon**

In web page a graphic image is created which is known as favicon. If there is inconsistency between favicon of domain and URL, then it is considered as Phishing [14] [13].

**1.1.11. Using Non-standard Port**

On a specific server certain service is up and down. User data is in danger, if all the ports are opened. For controlling intrusion, it is advice to open only those ports which are necessary.

**1.1.12. “HTTPS” Token in the Domain Part**

In URL, HTTPs symbol is added to deceiving people.

**1.2. Abnormal Based Features****1.2.1. Request URL**

Request URL examines whether the contains on the website are loaded from the same website or from another website. URL with greater than 61% is considered as phishing, and if it lies between 22%-61%, then suspicious, otherwise legitimate [15] [13]

**1.2.2. URL of Anchor**

Tag <a> is defined as Anchor element which is treated as “Request URL”.

**1.2.3. Links in <Meta>, <Script> and <Link> tags**

Legitimate website uses <Meta> which display about HTML, <Script> create a client-side script and <Links>, and <Links> tag accept other web sources. % of <Meta>, <Script> and <Links> is less than 17 % then it is considered as legitimate, if it lies between 17%-81% then it is considered as suspicious, otherwise phishing [14][13].

**1.2.4. Server form Handler**

An empty string or blank is considered as phishing. If webpage is differed from the SFHs, then it is considered as suspicious, otherwise legitimate [14] [13].

**1.2.5. Submitting Information to Email**

Submission of personal information is typically carried out by web services in which it redirects to the phisher mail. “mail()”, “mailto” function are used for server-side as a scripting language in PHP.

**1.2.6. Abnormal URL**

WHOIS database extract the feature. In URL, legitimate website is typically an identity part.

**1.3. HTML and JavaScript-based Features****1.3.1. Website Forwarding**

Distinguishes phishing websites from legitimate is based on redirection. Legitimate websites redirected one-time max whereas phishing websites redirected at least 4 times [13].

**1.3.2. Status Bar Customization**

Deceptive URLs are shown by phishers to deceive people with the help of Java Script. If source code of webpage is known, then any changes in the status bar can be done by “onMouseOver” event.

**1.3.3. Disabling Right Click**

Right-click function is disabled by phishers so that users are not able to see and save the web page source code [13].

**1.3.4. Using Pop-up window**

In pop-up window personal information is to ask by the user in phishing websites whereas legitimate website does not ask to submit [14].

**1.3.5. IFrame Redirection**

Phishers hide the webpage tag i.e., without frame border by ‘iframe’ and make it invisible to the users. So, for visual delineation phishers use frame border [14][13].

**1.4. Domain Based Features****1.4.1 Age Domain**

Age Domain checks the age of webpage. Phishing webpage remains for a shorter period of time whereas legitimate having minimum 6 months age. All these features are extracted from WHOIS database.[13]

#### 1.4.2. DNS Record

In phishing website WHOIS database does not identified host name. In DNS record if it found empty, then it is considered as phishing otherwise legitimate. [13].

#### 1.4.3. Page Rank

Weighted of a page rank lies between “0” to “1” where “0” indicates low page rank and “1” indicates higher page rank. Highest page rank is the most important for the webpage. Phishing webpage remains for a shorter period of time in which page rank does not exist. [14]

#### 1.4.3. Website Traffic

A website is defined as phishing or legitimate is determined by its popularity i.e., page rank. Legitimate website ranked among top 100,000 whereas greater than 100,000 is considered as phishing.[14].

#### 1.4.5. Google Index

Website is Google Index or not is based on indexing. Phishing webpage remains for shorter time period. When Google’s indexed a site, it is displayed on search results.[14]

#### 1.4.6. Links Pointing to Page

Website security depends on greater number of links. Website is phishing if number of links is 0, if it lies between 0-2 then considered as suspicious, otherwise legitimate [14]

#### 1.4.7. Statistical Reports based Features

Statistical reports on phishing websites have been defined by Phish Tank and Stop Badware over a period of time [15].

#### c) Image Features

##### 1. Grayscale:

0 or 1 value is contained by image. 0 is considered for black and 1 is for white, in which strength of the information is transmitted [10].

##### 2. Color Histogram:

According to intensity of the colored image, pixels are categorized [10].

## V.RELATED WORK

Studied on various phishing detection methods have been studied. They are classified as Blacklist, Heuristic, Content Analysis, Machine Learning techniques.

Sonmez et. al [1] studied classification technique on 30 phishing websites features using Extreme Learning Machine. Whole problem is divided into a certain number of classes in classification. Different classification methods (Artificial Neural Network, Support Vector Machine, Naïve Bayes) have been applied. In this ELM achieved higher accuracy i.e., 95.34% by using 6 different activation functions.

Ram Basnet et. al. [16] proposed the detection of phishing attacks by using machine learning models. In this, 16 features of phishing are used on 6 different machine learning models for detecting, i.e., either phishing or legitimate. Support Vector Machine (SVM) gives the best results, whereas Biased SVM and Artificial Neural Network give the same accuracy i.e., 97.98%.

Kamal et. al. [17] proposed the use of machine learning for phishing detection with features extracted from the URL only. In 2014, according to APWG, increasing phishing attacks due to cheapness & freeness of domain name. Naïve Bayes algorithm is used for the classification of phishing websites on Weka Platform. Using the ensemble methodology an accuracy of 97.08% can be achieved using Stacking, Bagging and Boosting along with the Naive Bayes, Decision Tree & Random Forest algorithm.

Baykara et. al. [18] proposed a software “Anti Phishing Simulator” for detecting phishing websites which contains malicious software and links and spam email by examining the mail contents. Also prevent serious threats like catches malicious email arriving at email addresses integrated into the system, providers a URL based control. As a result, Bayesian classification provides the weights of the words are calculated & spam word count are made.

Priya et. al. [10] proposed the ant colony optimization algorithm. Detection of phishing websites can be easily done by machine learning. In this proposed system, phishing websites features are extracted and to reduce features it is given to the ant colony optimization algorithm. Again, Naïve Bayes is used to reduce the features and classifies webpage.

Priyanka et. al. [15] proposed phishing detection using feature extraction based on machine learning. They used the Adaline algorithm and Backpropagation algorithm along with SVM to enhance the detection rate and classification of web pages. For better result Adaline is compared with SVM with 99.14%. Minimal time taken by Adaline network when compared with the Backpropagation network with SVM.

Mustafa Kaytan et. al. [19] proposed an Extreme Learning Machine classification algorithm to detected phishing webpages. In this paper, the classification of phishing website features based on “Request URL” and “Website Forwarding”. For evaluating performance 10-cross fold validation is used. The average classification accuracy was 95.05% and the best classification accuracy was 95.93%.

Amani et. al. [6] proposed the Random Forest algorithm to detect phishing websites. Random Forest technique is used for better performance as it gives high accuracy of 98.8% with the combination of 26 phishing websites features.

Xiang et. al. [20] proposed CANTINA+ which is the upgrade of CANTINA. False Positive (FP) and achieve runtime speedup reduces features. At CANTINA more features are applied and machine learning techniques are also applied in which 92% True Positive (TP) and 0.4% (FP).

Weina Niu et. al. [21] proposed a model Cuckoo- Search SVM(CS-SVM) for the detection of phishing of email with high accuracy. It improves the classification accuracy. To construct hybrid classifier, 23 features are extracted by using CS-SVM. In hybrid classifier, Cuckoo-Search (CS) combined with Support Vector Machine to enhance the parameter selection of Radial Basis Function (RBF). In this, it calculates higher accuracy than the SVM classifier by using RBF. Using CS-SVM classifier accuracy of 99.52% is obtained.

Ishant et. al. [14] proposed various machine learning techniques that are applied to the URL to check whether the website is phishing or legitimate. 30 attributes of phishing websites are considered for detection using Python. For accuracy calculation, the Generalized Linear Model (GLM) and Generalised Additive Model (GAM) are used. To gain more accuracy it uses a Decision tree, Random Forest. An accuracy of 98.4% is obtained by using Random Forest.

Taware et al. [22] proposed an MCAC which differentiate phishing websites from legitimate websites in which result of MACC algorithm is better than other data mining algorithm.

Amirreza et.al [23] proposed the PhishMon framework for detecting phishing webpages. It distinguishes legitimate and phishing webpages with high accuracy. In this paper, PhishMon detects 95.4%.

Yadollahi et al [24] proposed a real-time anti-phishing system. Online and feature-rich machine learning technique distinguish webpages. Approaches are extracted from discriminative features. The solution is based on the client-side, there is no service from the third party.

Gutierrez et. al. [25] proposed SAFE-PC (Semi-Automated Feature generation for Phish Classification) for detecting whether the webpage is phishing or legitimate.

Yang et. al [26] proposed multidimensional feature phishing learning (MFPD). This threshold is fixed for reducing the time. The highest accuracy is 98.61% by using CNN-LSTM.

## VI. PERFORMANCE EVALUATION FOR PHISHING DETECTION

Detection of phishing websites is a binary classification problem. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are categories in which webpage falls [27].

Table2: Classification Confusion Matrix

	Classified as Phishing	Classified as Legitimate
Phishing	NP→P	NP→L
Legitimate	NL→P	NL→L

NP → total number of phishing websites

NL → total number of legitimate websites [28]

❖ Performance evolution parameters are as following:

- **True Positive (TP):** Number of correctly classification of phishing website [28]:  

$$TP = (NP \rightarrow P) / NP$$

- **True Negative (TN):** Number of correctly classification of ham websites [28]:

$$TN = (NL \rightarrow L) / NL$$

- **False Positive (FP):** The number of ham websites wrongly classified [28]:

$$FP = (NL \rightarrow P) / NL$$

- **False Negative (FN):** Number of wrongly classification of phishing websites [28]:  

$$FN = (NP \rightarrow L) / NP$$

❖ Measures used for classification of webpages:

- **Precision:** Percentage of correct positive predictions is defined in precision.

$$\text{Precision} = TP / (TP + FP)$$

- **Recall:** Percentage of positive prediction of positive labeled instances [29].

$$\text{Recall} = TP / (TP + FN)$$

- **Accuracy:** Percentage of correct prediction is defined in accuracy [29].

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- **F-Score:** It measure weighted average of true positive rate/recall and precision [29].  

$$F = 2 * (\text{precision} * \text{recall} / (\text{precision} + \text{recall}))$$

## VII. CLASSIFICATION OF PROTECTION APPROACHES AGAINST PHISHING ATTACKS

### 1. Network level protection

Blacklist features which are also known as Network-level protection. Set of IP addresses or domain name to enter in a network does not allowed by network layer protection for implementation. It blocks the communication from those systems, which are identified as phishers. Example: Anti-Spam filters, DNS-Based filter [28].

### 2. Authentication

Authentication, in which it ensures or to verify the identification. It applies on both the user level and server/domain level to check whether the message is sent by trusted domain or not. Security communication is increased at both levels i.e., user and the domain level. Password is authenticated at the user level, but it can be easily broken by the phishers. Service providers ensured domain level. Domain level authentication examples are: Microsoft sender ID, Yahoo-based domain key [28].

### 3. Client-side tools

Client-side tools studied the phishing and attack initiate from the detecting phishing “Web browsers” directly. Another technique is domain check, URL examination, page content, etc. Blacklisting and whitelisting depend on these tools. Detection is failed for zero-day attack is the limitation of these methods.

### 4. Server-side filters and classifiers

It is considered to fighting zero-day attacks. By approval of statistical classifier for identification of weblinks, i.e., either phishing or legitimate is also trained on machine learning algorithm.

### 5. User Education

In today's generation phishing can be easily done by phishers due to lack of awareness about phishing. [30].

## VIII. ISSUES AND CHALLENGES

In previous work, for phishing attacks there are many solutions have been designed. No result is a “bullet of silver” for phishing [31]. Whenever researchers give solution for detecting and recovering phishing sites, then phishers breaks their solutions for fraudulent attempt. So, it is a rigid race between researcher and phisher.

Phishing attack is more successful due to lack of awareness about phishing. Therefore, one of the main challenges is the security, i.e., how to encourage users to protect themselves against phishing.

## IX. CONCLUSION

Internet is one of the most targeted phishing attacks, so the anti-phishing is used for protection. There is various defense technique for phishing. Better defense mechanism has adequate to identify phishing attack with low false positive (FP) [30]. It is a survey in which the machine learning algorithm was able to detect with approximate 99% accuracy by including a combination of 30 features.

Phishing detection techniques inform the users whether it is phishing, suspicious or legitimate websites.

## REFERENCES

- [1] Y. Sonmez, Turker Tuncer, Huseyin Gokal & Engin Avci (2018). “Phishing web Sites Features Classification Based on Extreme Machine Learning”. 6<sup>th</sup> International Symposium on Digital Forensic and Security (ISDFS).
- [2] Kay, R (2004) Sidebar: The Origins of Phishing. [Online]. Available: [http://www.computerworld.com/s/article/89097/Sidebar\\_The\\_Origins\\_of\\_Phishing](http://www.computerworld.com/s/article/89097/Sidebar_The_Origins_of_Phishing).
- [3] Mohmoud khonji, Youssef Iraqi and Andrew Jones(2013). “Phishing detection: A Literature Survey”. IEEE communications Systems and Tutorials, pp (99):1-31.
- [4] Anti-Phishing Working Group, “Phishing Activity Trends Report,” 2019.
- [5] RSA Online Fraud Report 2019.
- [6] Eduardo Benavides, Walter Fuertes, Sandra Sanchez & Manuel Sanchez(2019). “Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review”. Smart Innovation, Systems and Technologies, vol 152. Springer, Singapore.
- [7] Alswailem, A., Alabdullah, B., Alrumayh, N., & Alsedrani, A. (2019) “Detecting Phishing Websites Using Machine Learning”. 2nd International Conference on Computer Applications & Information Security (ICCAIS)
- [8] Himani Thakur & Supreet kaur(2016). “A Survey Paper On Phishing Detection”. International Journal of Advanced Research in Computer Science(IJARCS). ISSN: 0976-5697.
- [9] Kathrine, G. J. W., Praise, P. M., Rose, A. A., & Kalaivani, E. C. (2019). “Variants of phishing attacks and their detection techniques”. 3rd International Conference on Trends in Electronics and Informatics.
- [10] R.Priya (2016), “An Ideal Approach for Detection of Phishing Attacks using Naive Bayes Classifier”. International Journal of Computer Trends and Technology(IJCTI). ISSN: 2231-2803.
- [11] Arun Kulkarni, Leonard L. “Phishing Websites Detection using Machine Learning”, International Journal of Advanced Computer Science and Applications, 2019
- [12] Aron Blam, Brad Wardman, Tamar Solorio and Gary Warner (2010), “Lexical feature based phishing URL detection using online learning”, 3rd ACM Workshop on Security and Artificial Intelligence.
- [13] Rami M. Mohammad, Fadi Thabtah & Lee McCluskey “Phishing Websites Features”, (2014)
- [14] Tyagi, I., Shad, J., Sharma, S., Gaur, S., & Kaur, G. (2018). “A Novel Machine Learning Approach to Detect Phishing Websites”. 5th International Conference on Signal Processing and Integrated Networks (SPIN).

- [15] Singh, P., Maravi, Y. P. S., & Sharma, S. (2015). "Phishing websites detection through supervised learning networks". 2015 International Conference on Computing and Communications Technologies (ICCCCT)
- [16] Basnet, R., Mukkamala, S., & Sung, A. H. (n.d.). Detection of Phishing Attacks: A Machine Learning Approach. *Studies in Fuzziness and Soft Computing*, 373–383.
- [17] Gyan Kamal and Monotosh Manna, Detection of Phishing Websites Using Naive Bayes Algorithm, *Proceeding of International Journal of Recent Research and Review*, Vol. XI, Issue 4 December 2018, ISSN 2277-8322.
- [18] Baykara, M., & Gurel, Z. Z. mm(2018). Detection of phishing attacks. 2018 6th International Symposium on Digital Forensic and Security 355389(ISDFS).
- [19] M. Kaytan and D. Hanbay "Effective classification of Phishing Webpages Based on New Rules by Using Extreme Machine Learning" *Anatolian Journal of Computer Sciences*, AJCS 17, pp: 15-36, ISSN: 2548-1304, 2017.
- [20] Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). CANTINA+A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Transactions on Information and System Security*, 14(2), 1–28.
- [21] Niu, W., Zhang, X., Yang, G., Ma, Z., & Zhuo, Z. (2017). Phishing Emails Detection Using CS-SVM. 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC).
- [22] Wa'el Hadi, Faisal Aburub, Samer Alhawari. "A new fast associative classification algorithm for detecting phishing websites", *Applied Soft Computing*, 2016.
- [23] Niakanlahiji, A., Chu, B.-T., & Al-Shaer, E. (2018). PhishMon: A Machine Learning Framework for Detecting Phishing Webpages. 2018 IEEE International Conference on Intelligence and Security Informatics (ISI).
- [24] Yadollahi, M. M., Shoeleh, F., Serkani, E., Madani, A., & Gharaee, H. (2019). An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features. 2019 5th International Conference on Web Research (ICWR).
- [25] Gutierrez, C., Kim, T., Della Corte, R., Avery, J., Cinque, M., Goldwasser, D., & Bagchi, S. (2018). Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks. *IEEE Transactions on Dependable and Secure Computing*.
- [26] Yang, P., Zhao, G., & Zeng, P. (2019). Phishing Website Detection based on Multidimensional Features driven by Deep Learning. *IEEE Access*.
- [27] Khonji, M., Iraqi, Y., & Jones, A. (2013). "Phishing Detection: A Literature Survey". *IEEE Communications Surveys & Tutorials*, 15(4), 2091–2121.
- [28] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). "A Survey of Phishing Email Filtering Techniques". *IEEE Communications Surveys & Tutorials*, 15(4), 2070–2090.
- [29] Asha S Manek, D K Shamini, Veena H Bhat, P Deepa Shenoy, M. Chandra Mohan, K R Venugopal, L M Patnaik. "ReP-ETD: A Repetitive Preprocessing technique for Embedded Text Detection from images in spam emails", 2014 IEEE International Advance Computing Conference (IACC), 2014
- [30] A.MahaLakshmi, N.Swapna Goud, G.Vishnu Murthy (2018)."A Survey on Phishing And It's detection technique Based on support Vector Machine (SVM) and Software Defined Networking (SDN)". *International Journal of Engineering and Advanced Technology (IJEAT)*. ISSN: 2249-8958.
- [31] Gupta, B. B., Aakanksha Tewari, Ankit Kumar Jain, and Dharma P. Agrawal. "Fighting against phishing attacks: state of the art and future challenges", *Neural Computing and Applications*, 2016.

# Phishing Website Detection Based on Machine Learning: A Survey

Charu Singh

Department of Computer Science & Engineering  
Madan Mohan Malaviya University of Technology  
Gorakhpur, India  
charusingh0011@gmail.com

Smt.Meenu

Department of Computer Science & Engineering  
Madan Mohan Malaviya University of Technology  
Gorakhpur, India

**Abstract**—Phishing attacks are cybercrime in today's world which are done by social engineering and malware based. It is one of the most dangerous threats that every individuals and organization faced. URLs are known as web links are by which users locate information on the internet. The review creates awareness of phishing attacks, detection of phishing attacks and encourages the practice of phishing prevention among the readers. In phishing, phishers use email or message, as a weapon to target individual or organization by send URL link to target people and to deceive them. With the huge number of phishing emails or messages received every day, companies or individuals are not able to detect all of them. Here, different reviews give for detection of phishing attack, by using machine learning. Here it is used for detecting the web links, i.e., either phishing or legitimate.

**Keywords**—Social Engineering, Phishing, Legitimate, Machine Learning

## I. INTRODUCTION

Due to rapidly growing technology internet has become an integral part of our daily life [1]. Lots of activities in our daily life are determined after the use of the internet. Social networking sites have rapidly increased over the last few years. Due to the regular use of the internet, the users have to undergo many threats; one of them is 'Phishing'.

The major problem is "phishing" is one of the today's world. Social engineering and malware based are the phishing attacks which contain malicious websites that are attached to E-mail, SMS or other communication method to deceive people. It is cybercrime or fraud that uses spam email as a weapon. Email spoofing or instant messaging carried out phishing. These emails and messages contain a URL link directs users to another website. It often directs users to enter personal information or sensitive information i.e., password, credits card details at a forged website which look like legitimate site.

## II. BACKGROUND AND OVERVIEW OF PHISHING

### A. HISTORY

In 1970's John Draper defined the term 'phishing'. For hacking telephone systems, he created infamous Blue Box that emitted audible tones [2]. Social engineering attacks are done in 1996, against America Online (AOL) accounts by online scammers [3].

### B. PHISHING STATISTICS

According to APWG, in 1Q 1,80,768 phishing sites was detected and unique phishing report was 112,393. HTTPS encryption protocol protects phishing sites. 58% of phishing sites were using SSL certificates. It was detected that 55% of SSL were used in phishing attacks in 2Q of 2019. SaaS & Webmail providers were counted as the most targeted sector with 36% of all phishing attacks recorded targeting its constitutions brands, according to APWG member MarkMonitor [4].

According to the RSA report, 2019, 29% of phishing attack has observed by RSA in 1Q. Fraud attacks from rogue mobile applications increased by 300% from 10,331 in 1Q. Card-not-present (CNP) fraud transactions increased 17% last quarter, and 56% of those originated from the mobile channel. The average value of a CNP fraud transaction in the U.S. was \$403, nearly double that of the average genuine transaction of \$213. 14.2 million unique compromised cards recovered over RSA in 1Q [5].

### C. TYPES OF PHISHING ATTACKS

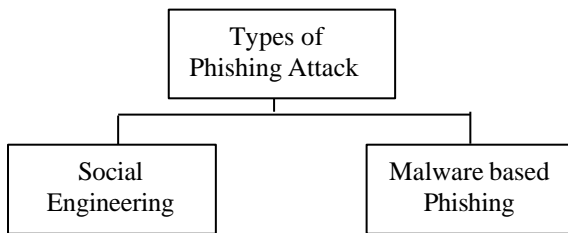
Phishing is a method by which attackers try to gain sensitive information from the users, to use it fraudulently [6]. Social engineering (deceptive phishing) and malware-based phishing are the types of phishing attacks.

Social engineering attacks usually exploit mind and susceptibility to manipulate users for obtaining confidential information. Malware based phishing refers to a scam in which malicious software or unnecessary programs run on the user's system. The malware uses a key logger, screen logger to record your keyboard strokes and sites that you visit on the internet. Key loggers, Session hijacking, DNS phishing, content-injection phishing, phone phishing, system



reconfiguration, link manipulation is the classification of these attacks [7][8].

Fig.1: Types of Phishing Attacks



#### D. MACHINE LEARNING MODELS

Machine learning tends to predict whether the websites are phishing or legitimate. It learns the characteristics of phishing websites and predicts new phishing characteristics [7]. Prediction can be done by several algorithms such as Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, Artificial Neural Network, K-Nearest Neighbor, Bayesian Classification. Phishing detection accuracies vary from each other [7] [9].

### III. PHISHING DETECTION APPROACHES

There are various defense methods of phishing attacks which shown in table 1 [9].

Table1: Phishing Detection Approach

S.No.	Approach	Technique
1.	Heuristic based approach	Decision tree algorithm
2.	Blacklist approach	Simhash algorithm
3.	Fuzzy based approach	Fuzzy data mining algorithm
4.	Machine Learning approach	Machine learning algorithm
5.	CANTINA based approach	TF-IDF information retrieval algorithm
6.	Image based approach	Web logo technique

### IV. FEATURE EXTRACTION

There are various methods by which features can be detected. These are classified as follows:

- a) Source code features
- b) URL features
- c) Image feature

#### a) Source code features

##### 1. Tracking of login screen:

To check if it contains any text box to get information from the user, such as username, password, and PIN which is done by tracking of the login screen [10].

##### 2. Disabling Right Click

Right-click is disabled by phishers so that the website code is not able to visualize to the users [10].

##### 3. Pop Up

The legitimate sites do not ask them to enter their credentials while some messages appear in phishing sites to enter their details. [10].

#### b) URL features

Features of phishing URLs are identified by machine learning. Host-based features and lexical features are the types of features which is extracted from the URL [11] [12].

URLs are simply divided into sub parts, in which it contains, host name, a path, protocol or scheme. Based on any combination of these components, accuracy of site's legitimacy can access [11]. There are 30 phishing website features [1].

##### 1.1. Address Based Features

###### 1.1.1. IP Address

In phishing, at place of domain name of website it contains IP address.

###### 1.1.2. URL length

In address bar phishers use long URL to hide suspicious part.

###### 1.1.3. "Tiny URL"

On the "worldwide web" a method, URL shortening in which URL considered smaller in length. Phishers use "Tiny URL" for deceiving people in which long URL is used to connect the tiny URL [13].

###### 1.1.4. URL's containing "@" symbol

For deceiving people phishers use "@" symbol.

###### 1.1.5. Redirecting using "/"

The user will be diverted to another website by using "/" in URL path. In legitimate, number of pages is less than 2, for suspicious it lies between 2 - 4, otherwise it is considered as phishing [13].

###### 1.1.6. Adding (-) sign to the Domain

In legitimate URLs, a dash symbol is used for creating malicious URLs to deceive people [14].



**1.1.7. Dots in URL**

Number of dots counted in phishing or legitimate websites in URL. It is considered “Phishing” if number of dots is greater than legitimate [15].

**1.1.8. HTTP with SSL**

HTTPs is used by legitimate website for moving sensitive data. It requires certificate for using it with minimum age of 2 years.

**1.1.9. Domain Registration Length**

Phishing websites live for a short time period whereas trusted domains paid for several years in advance.

**1.1.10. Favicon**

In web page a graphic image is created which is known as favicon. If there is inconsistency between favicon of domain and URL, then it is considered as Phishing [14] [13].

**1.1.11. Using Non-standard Port**

On a specific server certain service is up and down. User data is in danger, if all the ports are opened. For controlling intrusion, it is advice to open only those ports which are necessary.

**1.1.12. “HTTPS” Token in the Domain Part**

In URL, HTTPs symbol is added to deceiving people.

**1.2. Abnormal Based Features****1.2.1. Request URL**

Request URL examines whether the contains on the website are loaded from the same website or from another website. URL with greater than 61% is considered as phishing, and if it lies between 22%-61%, then suspicious, otherwise legitimate [15] [13]

**1.2.2. URL of Anchor**

Tag <a> is defined as Anchor element which is treated as “Request URL”.

**1.2.3. Links in <Meta>, <Script> and <Link> tags**

Legitimate website uses <Meta> which display about HTML, <Script> create a client-side script and <Links>, and <Links> tag accept other web sources. % of <Meta>, <Script> and <Links> is less than 17 % then it is considered as legitimate, if it lies between 17%-81% then it is considered as suspicious, otherwise phishing [14][13].

**1.2.4. Server form Handler**

An empty string or blank is considered as phishing. If webpage is differed from the SFHs, then it is considered as suspicious, otherwise legitimate [14] [13].

**1.2.5. Submitting Information to Email**

Submission of personal information is typically carried out by web services in which it redirects to the phisher mail. “mail()”, “mailto” function are used for server-side as a scripting language in PHP.

**1.2.6. Abnormal URL**

WHOIS database extract the feature. In URL, legitimate website is typically an identity part.

**1.3. HTML and JavaScript-based Features****1.3.1. Website Forwarding**

Distinguishes phishing websites from legitimate is based on redirection. Legitimate websites redirected one-time max whereas phishing websites redirected at least 4 times [13].

**1.3.2. Status Bar Customization**

Deceptive URLs are shown by phishers to deceive people with the help of Java Script. If source code of webpage is known, then any changes in the status bar can be done by “onMouseOver” event.

**1.3.3. Disabling Right Click**

Right-click function is disabled by phishers so that users are not able to see and save the web page source code [13].

**1.3.4. Using Pop-up window**

In pop-up window personal information is to ask by the user in phishing websites whereas legitimate website does not ask to submit [14].

**1.3.5. IFrame Redirection**

Phishers hide the webpage tag i.e., without frame border by ‘iframe’ and make it invisible to the users. So, for visual delineation phishers use frame border [14][13].

**1.4. Domain Based Features****1.4.1 Age Domain**

Age Domain checks the age of webpage. Phishing webpage remains for a shorter period of time whereas legitimate having minimum 6 months age. All these features are extracted from WHOIS database.[13]

#### 1.4.2. DNS Record

In phishing website WHOIS database does not identified host name. In DNS record if it found empty, then it is considered as phishing otherwise legitimate. [13].

#### 1.4.3. Page Rank

Weighted of a page rank lies between “0” to “1” where “0” indicates low page rank and “1” indicates higher page rank. Highest page rank is the most important for the webpage. Phishing webpage remains for a shorter period of time in which page rank does not exist. [14]

#### 1.4.3. Website Traffic

A website is defined as phishing or legitimate is determined by its popularity i.e., page rank. Legitimate website ranked among top 100,000 whereas greater than 100,000 is considered as phishing.[14].

#### 1.4.5. Google Index

Website is Google Index or not is based on indexing. Phishing webpage remains for shorter time period. When Google’s indexed a site, it is displayed on search results.[14]

#### 1.4.6. Links Pointing to Page

Website security depends on greater number of links. Website is phishing if number of links is 0, if it lies between 0-2 then considered as suspicious, otherwise legitimate [14]

#### 1.4.7. Statistical Reports based Features

Statistical reports on phishing websites have been defined by Phish Tank and Stop Badware over a period of time [15].

#### c) Image Features

##### 1. Grayscale:

0 or 1 value is contained by image. 0 is considered for black and 1 is for white, in which strength of the information is transmitted [10].

##### 2. Color Histogram:

According to intensity of the colored image, pixels are categorized [10].

## V. RELATED WORK

Studied on various phishing detection methods have been studied. They are classified as Blacklist, Heuristic, Content Analysis, Machine Learning techniques.

Sonmez et. al [1] studied classification technique on 30 phishing websites features using Extreme Learning Machine. Whole problem is divided into a certain number of classes in classification. Different classification methods (Artificial Neural Network, Support Vector Machine, Naïve Bayes) have been applied. In this ELM achieved higher accuracy i.e., 95.34% by using 6 different activation functions.

Ram Basnet et. al. [16] proposed the detection of phishing attacks by using machine learning models. In this, 16 features of phishing are used on 6 different machine learning models for detecting, i.e., either phishing or legitimate. Support Vector Machine (SVM) gives the best results, whereas Biased SVM and Artificial Neural Network give the same accuracy i.e., 97.98%.

Kamal et. al. [17] proposed the use of machine learning for phishing detection with features extracted from the URL only. In 2014, according to APWG, increasing phishing attacks due to cheapness & freeness of domain name. Naïve Bayes algorithm is used for the classification of phishing websites on Weka Platform. Using the ensemble methodology an accuracy of 97.08% can be achieved using Stacking, Bagging and Boosting along with the Naive Bayes, Decision Tree & Random Forest algorithm.

Baykara et. al. [18] proposed a software “Anti Phishing Simulator” for detecting phishing websites which contains malicious software and links and spam email by examining the mail contents. Also prevent serious threats like catches malicious email arriving at email addresses integrated into the system, providers a URL based control. As a result, Bayesian classification provides the weights of the words are calculated & spam word count are made.

Priya et. al. [10] proposed the ant colony optimization algorithm. Detection of phishing websites can be easily done by machine learning. In this proposed system, phishing websites features are extracted and to reduce features it is given to the ant colony optimization algorithm. Again, Naïve Bayes is used to reduce the features and classifies webpage.

Priyanka et. al. [15] proposed phishing detection using feature extraction based on machine learning. They used the Adaline algorithm and Backpropagation algorithm along with SVM to enhance the detection rate and classification of web pages. For better result Adaline is compared with SVM with 99.14%. Minimal time taken by Adaline network when compared with the Backpropagation network with SVM.

Mustafa Kaytan et. al. [19] proposed an Extreme Learning Machine classification algorithm to detected phishing webpages. In this paper, the classification of phishing website features based on “Request URL” and “Website Forwarding”. For evaluating performance 10-cross fold validation is used. The average classification accuracy was 95.05% and the best classification accuracy was 95.93%.

Amani et. al. [6] proposed the Random Forest algorithm to detect phishing websites. Random Forest technique is used for better performance as it gives high accuracy of 98.8% with the combination of 26 phishing websites features.

Xiang et. al. [20] proposed CANTINA+ which is the upgrade of CANTINA. False Positive (FP) and achieve runtime speedup reduces features. At CANTINA more features are applied and machine learning techniques are also applied in which 92% True Positive (TP) and 0.4% (FP).

Weina Niu et. al. [21] proposed a model Cuckoo- Search SVM(CS-SVM) for the detection of phishing of email with high accuracy. It improves the classification accuracy. To construct hybrid classifier, 23 features are extracted by using CS-SVM. In hybrid classifier, Cuckoo-Search (CS) combined with Support Vector Machine to enhance the parameter selection of Radial Basis Function (RBF). In this, it calculates higher accuracy than the SVM classifier by using RBF. Using CS-SVM classifier accuracy of 99.52% is obtained.

Ishant et. al. [14] proposed various machine learning techniques that are applied to the URL to check whether the website is phishing or legitimate. 30 attributes of phishing websites are considered for detection using Python. For accuracy calculation, the Generalized Linear Model (GLM) and Generalised Additive Model (GAM) are used. To gain more accuracy it uses a Decision tree, Random Forest. An accuracy of 98.4% is obtained by using Random Forest.

Taware et al. [22] proposed an MCAC which differentiate phishing websites from legitimate websites in which result of MACC algorithm is better than other data mining algorithm.

Amirreza et.al [23] proposed the PhishMon framework for detecting phishing webpages. It distinguishes legitimate and phishing webpages with high accuracy. In this paper, PhishMon detects 95.4%.

Yadollahi et al [24] proposed a real-time anti-phishing system. Online and feature-rich machine learning technique distinguish webpages. Approaches are extracted from discriminative features. The solution is based on the client-side, there is no service from the third party.

Gutierrez et. al. [25] proposed SAFE-PC (Semi-Automated Feature generation for Phish Classification) for detecting whether the webpage is phishing or legitimate.

Yang et. al [26] proposed multidimensional feature phishing learning (MFPD). This threshold is fixed for reducing the time. The highest accuracy is 98.61% by using CNN-LSTM.

## VI. PERFORMANCE EVALUATION FOR PHISHING DETECTION

Detection of phishing websites is a binary classification problem. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are categories in which webpage falls [27].

Table2: Classification Confusion Matrix

	Classified as Phishing	Classified as Legitimate
Phishing	NP→P	NP→L
Legitimate	NL→P	NL→L

NP → total number of phishing websites

NL → total number of legitimate websites [28]

❖ Performance evolution parameters are as following:

- **True Positive (TP):** Number of correctly classification of phishing website [28]:  

$$TP = (NP \rightarrow P) / NP$$

- **True Negative (TN):** Number of correctly classification of ham websites [28]:

$$TN = (NL \rightarrow L) / NL$$

- **False Positive (FP):** The number of ham websites wrongly classified [28]:

$$FP = (NL \rightarrow P) / NL$$

- **False Negative (FN):** Number of wrongly classification of phishing websites [28]:  

$$FN = (NP \rightarrow L) / NP$$

❖ Measures used for classification of webpages:

- **Precision:** Percentage of correct positive predictions is defined in precision.

$$\text{Precision} = TP / (TP + FP)$$

- **Recall:** Percentage of positive prediction of positive labeled instances [29].

$$\text{Recall} = TP / (TP + FN)$$

- **Accuracy:** Percentage of correct prediction is defined in accuracy [29].

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- **F-Score:** It measure weighted average of true positive rate/recall and precision [29].  

$$F = 2 * (\text{precision} * \text{recall} / (\text{precision} + \text{recall}))$$

## VII. CLASSIFICATION OF PROTECTION APPROACHES AGAINST PHISHING ATTACKS

### 1. Network level protection

Blacklist features which are also known as Network-level protection. Set of IP addresses or domain name to enter in a network does not allowed by network layer protection for implementation. It blocks the communication from those systems, which are identified as phishers. Example: Anti-Spam filters, DNS-Based filter [28].

### 2. Authentication

Authentication, in which it ensures or to verify the identification. It applies on both the user level and server/domain level to check whether the message is sent by trusted domain or not. Security communication is increased at both levels i.e., user and the domain level. Password is authenticated at the user level, but it can be easily broken by the phishers. Service providers ensured domain level. Domain level authentication examples are: Microsoft sender ID, Yahoo-based domain key [28].

### 3. Client-side tools

Client-side tools studied the phishing and attack initiate from the detecting phishing “Web browsers” directly. Another technique is domain check, URL examination, page content, etc. Blacklisting and whitelisting depend on these tools. Detection is failed for zero-day attack is the limitation of these methods.

### 4. Server-side filters and classifiers

It is considered to fighting zero-day attacks. By approval of statistical classifier for identification of weblinks, i.e., either phishing or legitimate is also trained on machine learning algorithm.

### 5. User Education

In today's generation phishing can be easily done by phishers due to lack of awareness about phishing. [30].

## VIII. ISSUES AND CHALLENGES

In previous work, for phishing attacks there are many solutions have been designed. No result is a “bullet of silver” for phishing [31]. Whenever researchers give solution for detecting and recovering phishing sites, then phishers breaks their solutions for fraudulent attempt. So, it is a rigid race between researcher and phisher.

Phishing attack is more successful due to lack of awareness about phishing. Therefore, one of the main challenges is the security, i.e., how to encourage users to protect themselves against phishing.

## IX. CONCLUSION

Internet is one of the most targeted phishing attacks, so the anti-phishing is used for protection. There is various defense technique for phishing. Better defense mechanism has adequate to identify phishing attack with low false positive (FP) [30]. It is a survey in which the machine learning algorithm was able to detect with approximate 99% accuracy by including a combination of 30 features.

Phishing detection techniques inform the users whether it is phishing, suspicious or legitimate websites.

## REFERENCES

- [1] Y. Sonmez, Turker Tuncer, Huseyin Gokal & Engin Avci (2018). “Phishing web Sites Features Classification Based on Extreme Machine Learning”. 6<sup>th</sup> International Symposium on Digital Forensic and Security (ISDFS).
- [2] Kay, R (2004) Sidebar: The Origins of Phishing. [Online]. Available: [http://www.computerworld.com/s/article/89097/Sidebar\\_The\\_Origins\\_of\\_Phishing](http://www.computerworld.com/s/article/89097/Sidebar_The_Origins_of_Phishing).
- [3] Mohmoud khonji, Youssef Iraqi and Andrew Jones(2013). “Phishing detection: A Literature Survey”. IEEE communications Systems and Tutorials, pp (99):1-31.
- [4] Anti-Phishing Working Group, “Phishing Activity Trends Report,” 2019.
- [5] RSA Online Fraud Report 2019.
- [6] Eduardo Benavides, Walter Fuertes, Sandra Sanchez & Manuel Sanchez(2019). “Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review”. Smart Innovation, Systems and Technologies, vol 152. Springer, Singapore.
- [7] Alswailem, A., Alabdullah, B., Alrumayh, N., & Alsedrani, A. (2019) “Detecting Phishing Websites Using Machine Learning”. 2nd International Conference on Computer Applications & Information Security (ICCAIS)
- [8] Himani Thakur & Supreet kaur(2016). “A Survey Paper On Phishing Detection”. International Journal of Advanced Research in Computer Science(IJARCS). ISSN: 0976-5697.
- [9] Kathrine, G. J. W., Praise, P. M., Rose, A. A., & Kalaivani, E. C. (2019). “Variants of phishing attacks and their detection techniques”. 3rd International Conference on Trends in Electronics and Informatics.
- [10] R.Priya (2016), “An Ideal Approach for Detection of Phishing Attacks using Naive Bayes Classifier”. International Journal of Computer Trends and Technology(IJCTI). ISSN: 2231-2803.
- [11] Arun Kulkarni, Leonard L. “Phishing Websites Detection using Machine Learning”, International Journal of Advanced Computer Science and Applications, 2019
- [12] Aron Blam, Brad Wardman, Tamar Solorio and Gary Warner (2010), “Lexical feature based phishing URL detection using online learning”, 3rd ACM Workshop on Security and Artificial Intelligence.
- [13] Rami M. Mohammad, Fadi Thabtah & Lee McCluskey “Phishing Websites Features”, (2014)
- [14] Tyagi, I., Shad, J., Sharma, S., Gaur, S., & Kaur, G. (2018). “A Novel Machine Learning Approach to Detect Phishing Websites”. 5th International Conference on Signal Processing and Integrated Networks (SPIN).

- [15] Singh, P., Maravi, Y. P. S., & Sharma, S. (2015). "Phishing websites detection through supervised learning networks". 2015 International Conference on Computing and Communications Technologies (ICCCCT)
- [16] Basnet, R., Mukkamala, S., & Sung, A. H. (n.d.). Detection of Phishing Attacks: A Machine Learning Approach. *Studies in Fuzziness and Soft Computing*, 373–383.
- [17] Gyan Kamal and Monotosh Manna, Detection of Phishing Websites Using Naive Bayes Algorithm, *Proceeding of International Journal of Recent Research and Review*, Vol. XI, Issue 4 December 2018, ISSN 2277-8322.
- [18] Baykara, M., & Gurel, Z. Z. mm(2018). Detection of phishing attacks. 2018 6th International Symposium on Digital Forensic and Security 355389(ISDFS).
- [19] M. Kaytan and D. Hanbay "Effective classification of Phishing Webpages Based on New Rules by Using Extreme Machine Learning" *Anatolian Journal of Computer Sciences*, AJCS 17, pp: 15-36, ISSN: 2548-1304, 2017.
- [20] Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). CANTINA+A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Transactions on Information and System Security*, 14(2), 1–28.
- [21] Niu, W., Zhang, X., Yang, G., Ma, Z., & Zhuo, Z. (2017). Phishing Emails Detection Using CS-SVM. 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC).
- [22] Wa'el Hadi, Faisal Aburub, Samer Alhawari. "A new fast associative classification algorithm for detecting phishing websites", *Applied Soft Computing*, 2016.
- [23] Niakanlahiji, A., Chu, B.-T., & Al-Shaer, E. (2018). PhishMon: A Machine Learning Framework for Detecting Phishing Webpages. 2018 IEEE International Conference on Intelligence and Security Informatics (ISI).
- [24] Yadollahi, M. M., Shoeleh, F., Serkani, E., Madani, A., & Gharaee, H. (2019). An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features. 2019 5th International Conference on Web Research (ICWR).
- [25] Gutierrez, C., Kim, T., Della Corte, R., Avery, J., Cinque, M., Goldwasser, D., & Bagchi, S. (2018). Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks. *IEEE Transactions on Dependable and Secure Computing*.
- [26] Yang, P., Zhao, G., & Zeng, P. (2019). Phishing Website Detection based on Multidimensional Features driven by Deep Learning. *IEEE Access*.
- [27] Khonji, M., Iraqi, Y., & Jones, A. (2013). "Phishing Detection: A Literature Survey". *IEEE Communications Surveys & Tutorials*, 15(4), 2091–2121.
- [28] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). "A Survey of Phishing Email Filtering Techniques". *IEEE Communications Surveys & Tutorials*, 15(4), 2070–2090.
- [29] Asha S Manek, D K Shamini, Veena H Bhat, P Deepa Shenoy, M. Chandra Mohan, K R Venugopal, L M Patnaik. "ReP-ETD: A Repetitive Preprocessing technique for Embedded Text Detection from images in spam emails", 2014 IEEE International Advance Computing Conference (IACC), 2014
- [30] A.MahaLakshmi, N.Swapna Goud, G.Vishnu Murthy (2018)."A Survey on Phishing And It's detection technique Based on support Vector Machine (SVM) and Software Defined Networking (SDN)". *International Journal of Engineering and Advanced Technology (IJEAT)*. ISSN: 2249-8958.
- [31] Gupta, B. B., Aakanksha Tewari, Ankit Kumar Jain, and Dharma P. Agrawal. "Fighting against phishing attacks: state of the art and future challenges", *Neural Computing and Applications*, 2016.

# Detection and Prevention of Phishing Websites using Machine Learning Approach

Vaibhav Patil  
Dept. of Computer Engineering  
Sinhgad Academy of Engineering  
Pune, India  
vaibhav95.patil@gmail.com

Tushar Bhat  
Dept. of Computer Engineering  
Sinhgad Academy of Engineering  
Pune, India  
tusharbhat002@gmail.com

Pritesh Thakkar  
Dept. of Computer Engineering  
Sinhgad Academy of Engineering  
Pune, India  
priteshtakkar53@gmail.com

Prof. S. P. Godse  
Dept. of Computer Engineering  
Sinhgad Academy of Engineering  
Pune, India  
sachin.gds@gmail.com

Chirag Shah  
Dept. of Computer Engineering  
Sinhgad Academy of Engineering  
Pune, India  
chirag041@gmail.com

**Abstract**—Phishing costs Internet user's lots of dollars per year. It refers to exploiting weakness on the user side, which is vulnerable to such attacks. The phishing problem is huge and there does not exist only one solution to minimize all vulnerabilities effectively, thus multiple techniques are implemented. In this paper, we discuss three approaches for detecting phishing websites. First is by analyzing various features of URL, second is by checking legitimacy of website by knowing where the website is being hosted and who are managing it, the third approach uses visual appearance based analysis for checking genuineness of website. We make use of Machine Learning techniques and algorithms for evaluation of these different features of URL and websites. In this paper, an overview about these approaches is presented.

**Keywords**—phishing, security, blacklist, whitelist, URL, anti-phishing, web-page

## I. INTRODUCTION

Phishing is a social engineering attack that aims at exploiting the weakness found in the system at the user's end. For example, a system may be technically secure enough for password theft but the unaware user may leak his/her password when the attacker sends a false update password request through forged (phished) website. For addressing this issue, a layer of protection must be added on the user side to address this problem.

A phishing attack is when a criminal sends an email or the url pretending to be someone or something he's not, in order to get sensitive information out of the victim. The victim in regard to his/her curiosity or a sense of urgency, they enter the details, like a username, password, or credit card number, they are likely to acquiesce. The recent example of a Gmail phishing scam that targeted around 1 billion Gmail users worldwide.

The Fig. 1 looks exactly like a Gmail sign-in form, the URL is slightly changed, but it's not the . Filling in this form would give the attacker full access to the victim's Gmail account. The kind of theft and fraud that could take place by just acquiring the details of someone's or some organizations' account couldn't really be imagined. All other account are controlled by the Gmail account. That could be a huge threat. Microsoft Outlook fraud is the second-most targeted and Google drive being the third. Other targets are facebook, bank logins and paytm, paypal etc.

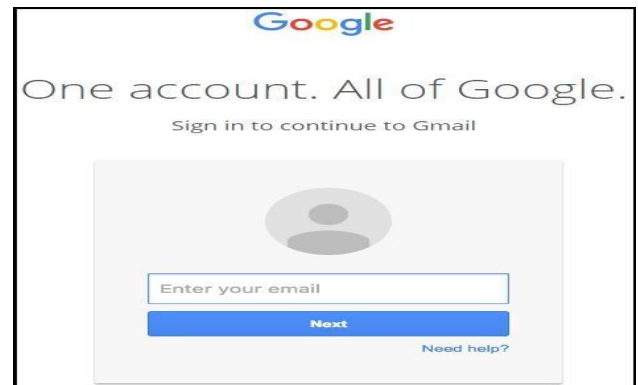


Fig. 1. Gmail Phishing Scam Url

## II. RELATED WORK

Many researchers have previously been carried out in this field of phishing detection. We have gathered the information from various such works and have profoundly reviewed them which has helped us in motivating our own methodologies in the process of making a more secure and accurate system.

### A. Blacklist Approach and Whitelist Approach

In [13], Pawan Prakash, Manish Kumar, Ramana Rao Kompella, Minaxi Gupta (2010) proposed a predictive blacklist approach to detect phishing websites. It identified new phishing URL using heuristics and by using an appropriate matching algorithm. Heuristics created new URL's by combining parts of the known phished websites from the available blacklist. The matching algorithm then calculates the score of URL. If this score is more than a given threshold value it flags this website as phishing website. The score was evaluated by matching various parts of the URL against the URL available in the blacklist.

In [14], Jung Min Kang and DoHoon Lee described approach which detected phishing based on users online activities. This method maintained a white list as a part of users' profile. This profile was dynamically updated whenever a user visited any website. An engine used here identified a website by evaluating a score and then comparing it with a threshold score. The score was calculated from the entries available in the user profile and details of the current website.

### B. Heuristic Approach

In [7], Aaron Blum, Brad Wardman, Thamar Solorio proposed a work which focused on the exploration of surface level features from URLs to train a confidence-

weighted learning algorithm. The idea is to restrict the source of possible features to the character string of the URL and avoid having the vulnerability of extracting host-based information. Every URL is displayed as a vector of binary feature. These vectors are fed to the online algorithm where at time of testing, previously unseen URLs in the binary feature vector is then mapped to it. The learner continues this new vector and output into the final result, either phish or non phish.

In [15], Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor proposed CANTINA+, a comprehensive feature-based approach in the literature including eight novel features, which exploits the HTML Document Object Model (DOM), search engines and third party services with machine learning techniques to detect phish. Also two other filters are designed in it to help reduce FP and achieve good runtime speedup. The first is a near-duplicate phish detector that uses hashing to catch highly similar phish. The second is a login form filter, which directly classifies webpages with no identified login form as legitimate.

In [8], Joby James, Sandhya L, Ciza Thomas proposed a work which with the combined help of blacklisting approach and the Host based Analysis applied certain classifiers which can be used to help detect and take down various phishing sites. The host based, popularity based and lexical based feature extractions are applied to form a database of feature values. The database is knowledge mined using different machine learning methods. After evaluating the classifiers, a particular classifier was selected and was implemented in MATLAB.

In [9], APWGM published a case study citing the importance of the WHOIS tool and how invaluable it has been for the rapid phishing site shutdown over the past few years all around the globe.

### C. Visual Similarity Approach

In [2], A. Mishra and B. B. Gupta presented a hybrid solution based on URL and CSS matching. In this approach it can detect embedded noise contents like an image in a web page which is used to sustain the visual similarity in the webpage. They used the technique used in [3] by Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang to compare the CSS similarity and used it in their technique. The different types of visual features are - text content and text features. Text features are like font colour, font size, background colour, font family and so forth. This approach matches the visual features of different websites because the attacker copies the page content from the actual website.

In [5] Matthew Dunlop, Stephen Groat, and David Shelly proposed a browser based plug in called goldphish to identify phishing websites. It uses the website logos to identify the fake website. The attacker can use the real logo of the target website to trap the internet users. Three stages to it is:

- *Logo Extraction* : Goldphish is used to extract the website logo from the suspicious website. Then it converts it into text using optical character recognition (OCR) software.
- *Legitimate website extraction* : The text obtained is used as a query for the search engine. Generally, search engine "google" is used because it always return genuine websites in their top results.

- *Comparisons* : Suspicious website is compared with the top result obtained from the search engine based on different features. If any domain is matched with the current website then it is declared legitimate or else make it phishing site.

## III. PROPOSED WORK

### A. Overview of our approach

Out of all the previous work, only the blacklist and whitelist are implemented which has a drawback of not being updated in long time. The basic idea of our proposed solution is the hybrid solution which uses all the three approaches – blacklist and whitelist, heuristics and visual similarity. Our proposed system has the following algorithm.

1. Monitor all "http" traffic of end-user system by creating a browser extension. The benefit of an extension over an application or software is that the system will be based purely in real time and at the same time will also be quite agile in delivering the outputs.
2. Compare domain of each URL with the white-list of trusted domains and also the black-list of illegitimate domains. The data required for both the lists would be extracted dynamically by web scraping and stored on the server. If domain of the URL is found under the white-list, mark the URL as innocent (Exact Matching), else go further and use the other approaches.
3. Furthermore, the whole website analysis would now be done by considering various details (features). The set of features we took are : website protocol (secure or unsecure), length of the URL, number of hyphen (-) in URL, number of @ symbol in URL, number of dots in the URL, using direct IP address or not, alexa rank, bounce rate, daily page view, whois availability, registration and expiration date of website, alexa.com availability, rank2traffic.com availability, siterankdata.com availability, daily unique visitor, favicon icon similarity and google indexing.  
Example :  
If hyphen in URL > 1 – Phished website  
If hyphen in URL = 1 – Suspicious website  
If hyphen in URL < 1 – Legitimate website  
All the feature take into consideration at the same time increases the accuracy of the system.
4. Intuitively, the higher similarity between the phishing page and the target page indicates a greater chance of the users being deceived. This is the reason, attackers always try their best to clone the target pages.
5. To counter such antics, our next approach would be to extract and compare CSS of suspicious URL and compare it with the CSS of each of the legitimate domains in queue. This method will look into visual based features of the phished websites.
6. The machine learning classifiers such as decision tree, logistic regression and random forest will be applied to the collected data and a score is generated.
7. The match score and similarity score is calculated. If the score is greater than threshold then we mark the URL as phishing and block it.

8. This approach basically provides a three level security block and hence can prove to be more effective and accurate than any of the other existing systems

### B. Requirement Analysis

The System which deals with providing security concern using new and effective technology like Machine Learning with the help of user's personal computer and the browser extension.

#### (i) Software Requirements

- Python 3.6
- BeautifulSoup (Package in Python)
- Scikit-learn (Package in Python)
- JavaScript
- Browser (Chrome)

#### (ii) Hardware Requirements

- Windows 7 above
- Hard disk of at least 64 GB

### C. Design Phase

The flow of the proposed system is shown in Fig. 2.

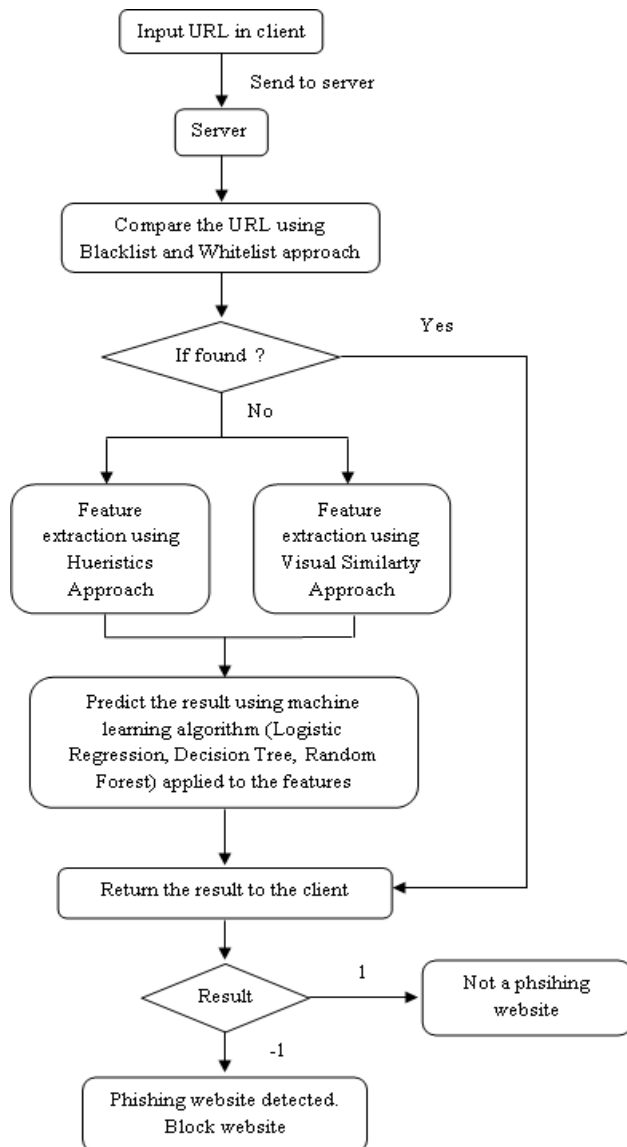


Fig. 2. Flow chart of the proposed system

### D. Analysis Phase

For different features we put different rules based on the analysis of phished and non-phished website scraped over from internet.

For example Fig. 3 and Fig. 4 shows the hyphen count of the phished and legitimate websites respectively. Y axis denotes count of websites and X axis denotes count of hyphens in the website. Based on this analysis, we concluded that phished websites do consist of hyphen in the domain part of the URL and legitimate websites don't.

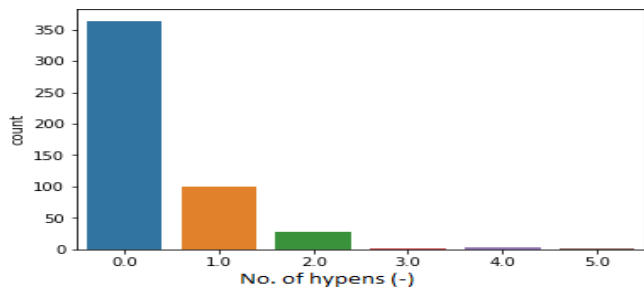


Fig. 3. Hyphen count of phished websites

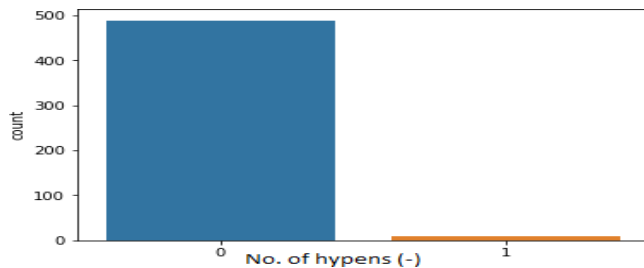


Fig. 4. Hyphen count of legitimate websites

Following are the detailed rules that we created based on analysis :

1. Has protocol ?  
If yes then legitimate, else suspicious
2. Length of domain in url  
If length in between 3 to 20, then legitimate  
Else if length in between 20 to 24, then suspicious  
Else if length greater than 24, then phished.
3. Number of hyphen in domain  
If number of hyphen is 0, then legitimate  
Else phished
4. @ symbol in domain  
If number of @ symbol is 0, then legitimate  
Else phished
5. In between domain the keyword 'http'  
If 'http' found in domain, then phished  
Else legitimate
6. Direct IP Address  
If url is a numeric IP address, then suspicious  
Else legitimate
7. Alexa.com, rank2traffic.com, and siterankdata.com availability



If the website is available in the database of any of these website, then legitimate  
Else suspicious

8. Time difference of date of expiration and data of registration of the website

If time difference is greater than 90 days then legitimate  
Else suspicious

9. Daily unique visitors

If daily unique visitor details are available on internet, then legitimate  
Else suspicious

10. Google indexing using title

If the title of the website queried on google search engine shows the exact same url of website in the top results, then legitimate  
Else suspicious

11. Google indexing using url

If the url of the website queried of google search engine shows the exact same url of website in top result, then legitimate  
Else phished

12. Favicon similarity using google indexing

If the favicon of the two websites are similar and domain of url is different, then phished  
Else legitimate

All these features combined together will lead to accurate results.

#### IV. RESULT

The linear regression plot of expected output versus predicted output is show in Fig. 5. This was predicted by the random forest algorithm. It has a slight deviation from the expected output for the phished websites.

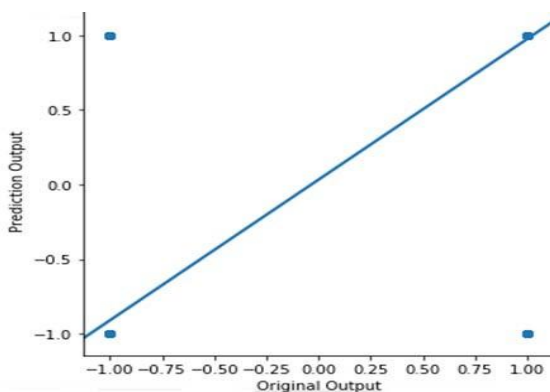


Fig. 5. Linear regression plot of original output versus predicted output

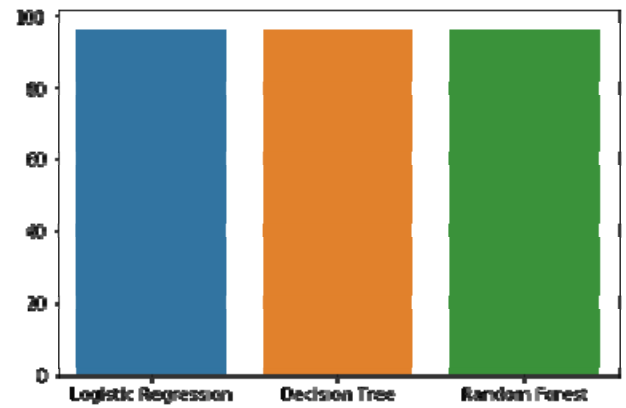


Fig. 6. Machine learning accuracy bar plot

The true positive, false positive, true negative, false negative count and accuracy results of 9076 test websites is as shown in Table I.

TABLE I. CONFUSION MATRIX RESULTS

Algorithm	TN	TP	FP	FN	Accuracy
Logistic Regression	6447	2287	325	17	96.23 %
Decision Tree	6393	2341	326	16	96.23 %
Random Forest	6392	2374	297	13	96.58 %

#### V. CONCLUSION

The proposed system enables the internet users to have a safe browsing and safe transactions. Its helps users to save their important priivate details that should not be leaked. Providing our proposed system to users in the form of extension makes the process of delevering our system much easier. The results points to the efficiency that can be achieved using the hybrid solution of hueristic features, visual features and blacklist and whitelist approach and feeding these features to machine learning algorithms. A particular challenge in this domains is that criminals are constantly making new strategies to counter our defense meausres. To succeed in this context, we need algorithms that continually adapt to new examples and features of phishing URL's. And thus we use online learning alorithms. This new system can be designed to avail maximum accuracy. Using different approaches altogether will enhance the accuracy of the system, providing an efficient protection system. The drawback of this system is detecting of some minimal false positive and false negative results. These drawbacks can be eliminated by introducing much richer feature to feed to the machine learning

algorithm that would result in much higher accuracy.

#### REFERENCES

- [1] Ankit Kumar Jain and B. B. Gupta, "Phishing Detection Analysis of Visual Similarity Based Approaches", Hindawi 2017.
- [2] A. Mishra and B. B. Gupta, "Hybrid Solution to Detect and Filter Zero-day Phishing Attacks", ERCICA 2014.
- [3] Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, "Bait Alarm Detecting Phishing Sites Using Similarity in Fundamental Visual Features", INCS 2013.
- [4] Eric Medvet, Engin Kirda and Christopher Kruegel, "Visual-Similarity-Based Phishing Detection", ACM 2015.
- [5] Matthew Dunlop, Stephen Groat, and David Shelly, "GoldPhish Using Images for Content-Based Phishing analysis", IEEE 2010.
- [6] Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, "Textual and Visual Content-Based Anti-Phishing A Bayesian Approach", IEEE 2011
- [7] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner; "Lexical Feature Based Phishing URL Detection Using Online Learning", Department of Computer and Information Sciences The University of Alabama at Birmingham, Alabama, 2016
- [8] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, MinaxiGupta,Purdue University, Indiana University "PhishNet: Predictive Blacklisting to Detect Phishing Attacks".
- [9] The Anti-Phishing Working Group, DNS Policy Committee;" Issues in Using DNS Whois Data for Phishing Site Take Down",The Anti-Phishing Working Group Memorandum, 2011.
- [10] Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor ,;"CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites", School of Computer Science Carnegie Mellon University, ACM Society of computing Journal, 2015.
- [11] Joby James,SandhyaL,Ciza Thomas "Detection of phishing websites using Machine learning techniques", 2013 International Conference on Control Communication and Computing (ICCC).
- [12] Mohsen Sharifi and Seyed Hossein Siadati "A Phishing Sites Blacklist Generator".
- [13] JungMin Kang and DoHoon Lee "Advanced White List Approach for Preventing Access to Phishing Sites".
- [14] Y. Zhang, J. I. Hong, and L. F. Cranor. Cantina: a content-based approach to detecting phishing web sites. In WWW '07: Proceedings of the 16<sup>th</sup> international conference on World Wide Web, pages 639–648, New York, NY, USA, 2007. ACM.