Mini Project Report on

# DETECTION OF PHISHING WEBSITE USING MACHINE LEARNING

A Synopsis Submitted
in Partial Fulfilment of the Requirements
for the Degree of
**Bachelor In Technology**
in
**Computer Science and Technology**

Submitted By:

**Group No - 15**


**46 Unnati Pimple**
**47 Tanisha Purohit**
**50 Surabhi Raut**



DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY
**USHA MITTAL INSTITUTE OF TECHNOLOGY**
S.N.D.T Women's University
Santacruz(W)


**January 2024**

# Abstract

In the ever-growing world of online transactions, the threat of phishing, a deceptive cyber-attack tricking users into revealing sensitive information through fake websites, has become more pronounced. Traditional methods like using lists of known bad websites and educated guesses have proven insufficient, leading to a rising number of victims. This study proposes a smarter approach using a "recurrent neural network" (RNN), a kind of smart detective that excels at distinguishing fake from real websites. Testing it with thousands of sites showed that the RNN outperformed existing methods. This technology is part of a family of smart tools capable of understanding website details, helping detect subtle differences between genuine and malicious sites. As the digital landscape evolves, the study emphasises the need for such intelligent tools to safeguard against the ever-tricky phishing scams and ensure a safer online experience for users.

## Problem Statement:

- The rise of online transactions has increased the risk of phishing attacks, where users are deceived by fake websites. Current methods, like using known bad website lists, are not effectively countering this threat, leading to a growing number of victims.
- The project proposes an advanced solution with a "recurrent neural network" (RNN) for better phishing detection.
- However, the fast-paced digital landscape and evolving phishing tactics demand more intelligent tools. This project addresses these gaps, aiming to enhance phishing detection methods and contribute to a safer online experience for users.

## Introduction:

Phishing stands as a malicious online attack with the intent to steal a user's private information, constituting a scam where unauthorised individuals attempt to acquire sensitive data. Given the prevalent use of online platforms for various activities like business transactions, money transfers, and bill payments, the detection of phishing websites becomes crucial in our daily lives. Identifying such deceptive sites proves challenging, prompting exploration into list-based anti-phishing approaches, such as blacklists or whitelists, which maintain databases of URLs. However, these methods have limitations, particularly in detecting newly created phishing URLs absent from the database.

A phishing attack unfolds when an unauthorised person sends deceptive emails or URLs to obtain users' sensitive information for misuse. Victims, unaware of the deception, unwittingly submit details like passwords, usernames, and credit card numbers, falling prey to the attacker's schemes. In light of these risks, it becomes imperative for users to stay vigilant, adopt advanced techniques like machine learning algorithms for detection, and prioritise security measures such as two-factor authentication and regular software updates to bolster overall protection against phishing threats.

## Objective of Project:

- **Awareness Building:** Increase awareness among users about the pervasive threat of phishing attacks and the importance of safeguarding private information.
- **Limitations of Conventional Approaches:** Highlight the drawbacks of list-based anti-phishing methods, such as blacklists and whitelists, in effectively identifying newly created phishing URLs.
- **Empowering Users:** Empower individuals with knowledge on common phishing tactics, enabling them to recognize and avoid falling prey to deceptive schemes.
- **Promotion of Advanced Technologies**: Advocate for the adoption of advanced technologies, such as machine learning algorithms, to enhance the dynamic detection of phishing threats beyond static databases.
- **Encourage Best Practices:** Promote the adoption of best practices, including user education initiatives, two-factor authentication, and regular software updates, to fortify overall online security and create a safer digital environment.

## Scope of Project:

The comprehensive scope of projects in phishing and anti-phishing measures involves developing advanced detection systems, incorporating machine learning, artificial intelligence, recurrent neural networks, and deep learning techniques to enhance dynamic identification of phishing threats. This multifaceted approach is coupled with educational campaigns, security audits, biometric integration, global collaboration, ongoing research, and advocacy for robust legislation, forming a holistic strategy to address the evolving challenges posed by online phishing attacks and continually advance cybersecurity measures.

# Literature Survey:

Study of papers

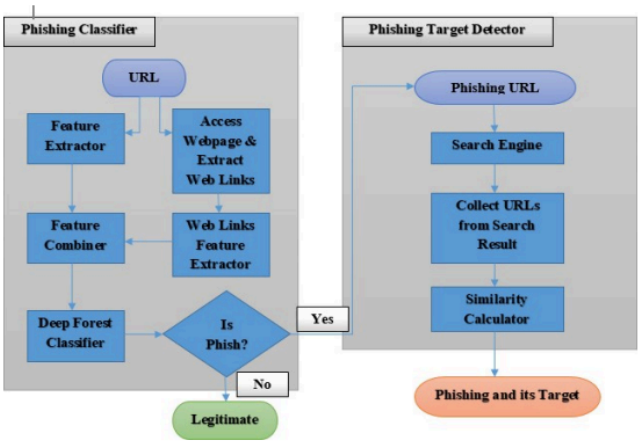
1. ”A Machine learning approach for phishing attack detection”.Phishing is the easiest method for gathering sensitive information from unwary people.In order to distinguish between legal and phishing URLs,machine learning technology is used.Extreme Gradient Boosting,Decision Tree,Logistic Regression and Support Vector Machine were used in this research work.The goal was to identify phishing URLs and determine most effective ML technique by comparing accuracy rates of each algorithm.
TarunChoudhary,Siddhesh Mhapankar,Rohit Buddha,Ashish Kharuk,Rohini patil.Journal of Artificial Intelligence and Technology(2023).

2.”Phishing Website Detection Using Machine Learning” :A Review”.The algorithms,including decision trees,support vector machines,and random forest,analyse multiple website features,such as URL structure,website content,and presence of specific keywords or patterns,to ascertain the likelihood of a website being a phishing site.Overall,machine learning algorithms serve as potent tool in identification of phishing websites,thereby safeguarding users against falling prey to such malicious attacks. Marwa Abd Al Hussein Qasim ,Dr Nahla Abbas Flayh. Waist Journal of Pure Sciences.

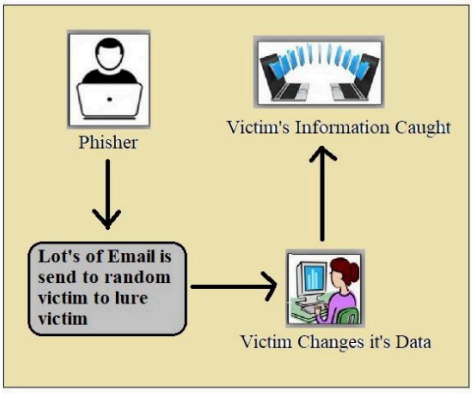
3.”Detection of Phishing Website Using Machine Learning Approach”.In this research work anti phishing solutions are used,it uses various approaches.Heuristic approach is used for classifying the URLs.There are some combined solution which uses three algorithm-whitelist and blacklist,heuristics and visual similarity.
Mahajan Mayuri Vilas ,Sawant Purva Jaypralash, Kakade Prachi Ghansham Pawar Shila.ICEECCOT 2019.

## Proposed Work:

The algorithm under consideration relies on machine learning processes and automated real-time phishing detection. It involves the extraction of phishing URLs by utilising specific features. These extracted features are then employed in a machine learning classification system to identify phishing websites in real-time. Following extensive analysis and surveys, which involved comparing different classification algorithms, the performance and accuracy of each algorithm were evaluated using the Waikato Environment for Knowledge Analysis (WEKA). To enhance efficiency, the Extreme Learning Machine (ELM) was chosen as the classification algorithm, and the RStudio tool was utilised for more comprehensive analysis.




Today the web has become a very popular platform where different activities are carried out by people or users like online transactions, entering id and password while login process etc. But while doing these activities people suffer from various security attacks. To avoid these types of security attacks different machine learning algorithms are used.



A. PHISHING:

Phishing is a form of Internet fraud that combines technology and social engineering to extract personal information during online activities such as shopping, email, or chatting. Phishing attacks can occur in four main ways:

- Fake Websites: Phishers create illegal websites that closely mimic legitimate ones to deceive users.
- Deceptive Links: Phishers trick users into visiting their fake websites by sending links disguised as legitimate ones from reputed organisations.
- Victim Interaction: Users, deceived by the link, visit the fake website and unknowingly provide personal information.
- Illegal Actions: Phishers exploit the obtained private information for unauthorised activities, such as unauthorised money transfers from the victim's account.

B. BLACKLIST AND WHITELIST APPROACH:

Utilising the Whitelist and Blacklist approach facilitates the identification of legal or unauthorised websites. The Blacklist contains categorised spam websites, often maintained by organisations like Google. On the other hand, the Whitelist approach involves comparing the current URL with a prebuilt list of authorised URLs to identify phishing sites. However, a significant drawback is its inability to distinguish recently created phishing websites from legitimate ones, limiting its effectiveness in real-time detection.

C. MACHINE LEARNING APPROACH:

Features are extracted and classified using ML techniques, a computational approach that generates rules and patterns to create general models. In supervised learning, labels are provided during training. Common ML methods include Random Forest (RF), Support Vector Machine (SVM), Back-Propagation Neural Network (BPNN), k-Nearest Neighbour (kNN), and Naive Bayes Classifier (NB), all essential for automated feature classification, particularly in phishing website detection.

D. HEURISTIC BASED APPROACH:

Heuristics are problem-solving techniques that provide alternative solutions within a limited time frame. In URL classification, heuristics play a vital role. They involve collecting and evaluating specific website features to identify influential characteristics crucial for phishing detection. Standardised features related to URL, Search Engine, Lookup, HTML DOM, and website traffic are assessed. If a website's heuristic structure aligns with predefined rules, it is categorised as a phishing site, showcasing the effectiveness of this method in rapid and constrained decision-making processes.

E. HYBRID APPROACH:

In a Hybrid Approach, various techniques like blacklisting and URL heuristics are combined to effectively determine the legitimacy of a website. This model incorporates 30 features to address phishing issues, recognizing that a single model may not suffice. The approach enhances efficiency, accuracy, and execution rates by combining multiple models. The process involves evaluating individual classifiers, selecting the best-performing one, and then combining it with others to create a more robust hybrid classification model.

F. ANTI PHISHING APPROACH:

This approach is a knowledge base service that helps to prevent illegitimate access to secure and sensitive information.Anti-phishing services protect a different type of data in other ways beyond the variety of stages. Anti Phishing software comprises computer programs that try to determine phishing content.

## Resources :

**Hardware Requirements:**

1.Multi-core Processor
2.Sufficient RAM (8GB and more)
3.Dedicated GPU
4.Storage(SSDS)
5.Internet connection
6.Keyboard -Standard Windows Keyboard
7.Mouse
8.Monitor-SVGA

**Software Requirements:**

1.Python
2.Machine Learning Libraries(Scikit,TensorFlow,Pytorch)
3.Feature Extraction Libraries
4.Data Manipulation and Analysis(Pandas)
5.Model Evaluation(Scikit -learn metrics)
6.Cloud Services
7.Web Framework
8.Virtual Environment(Virtualenv)
9. Deep Learning Neural Network

## Conclusion:

In conclusion, the project of detecting phishing websites using machine learning is a vital endeavour in enhancing online security. By leveraging Python-based machine learning libraries, web scraping tools, and feature extraction techniques, we can develop a robust model capable of identifying potential threats. The utilisation of frameworks like Scikit-learn, TensorFlow, or PyTorch, coupled with proper data manipulation using Pandas, ensures efficient model training and evaluation.This project not only addresses the immediate need for phishing detection but also lays the foundation for ongoing research and development in the field of cybersecurity, promoting a safer online environment for users.

## References:

1.S. Yadav, et al., "A novel approach for phishing website detection using machine learning techniques," in Procedia Computer Science, 2018.

2.R. Ramya, et al., "Phishing website detection using machine learning and social network analysis," in Materials Today: Proceedings, 2020.

3.K. Y. Wang, J. Beck, "Automated Detection of Phishing Targeted at E-commerce Websites," in IEEE Transactions on Dependable and Secure Computing, 2014.

4.F. Ahmed, et al., "A novel hybrid model for phishing detection using random forest, decision tree, and multilayer perceptron," in Security and Communication Networks, 2019.

5.H. Singh, A. Yadav, "Phishing Detection Using Machine Learning Algorithms: A Review," in Procedia Computer Science, 2018.

6.Comprehensive Study on Phishing Detection Using Machine Learning Techniques.Authors: D. Udhayakumar, S. Manogaran, Published in: Computers, Materials & Continua, 2020.

7.Phishing Website Detection using Machine Learning Techniques Authors: Sachin Sharma, Parvinder S. Sandhu .Published in: Procedia Computer Science, 2018.