International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Detection Of Phishing Website Using Machine Learning

Karthikeyan K[1], Koshik Rishi M[2],Naveen Prasad K[3]

[1]*Assistant Professor & Department of Computer Science and Engineering*
[2]*Student & Department of Computer Science and Engineering*
[3]*Student & Department of Computer Science and Engineering*

Abstract - Phishing is a widespread tactic used to trick gullible people into disclosing their personal information by using bogus websites. Phishing website URLs are designed to steal personal data, including user names, passwords, and online financial activities. Phishers employ websites that resemble those genuine websites both aesthetically and linguistically. Utilizing anti-phishing methods to identify phishing is necessary to stop the rapid advancement of phishing techniques as a result of advancing technology. A strong tool for thwarting phishing assaults is machine learning. Attackers frequently use phishing because it is simpler to fool a victim into clicking a malicious link that looks authentic than to try to get past a computer's security measures. The malicious links within the message body are intended to appear to go to the spoofed company utilising that company's logos and other genuine information. In the method that is being presented, machine learning is used to create a revolutionary approach for detecting phishing websites. Gradient Boosting Classifier is the model we utilised in our suggested strategy to identify phishing websites based on aspects of URL significance. By extracting and comparing different characteristics between legitimate and phishing URLs, the suggested method uses gradient boosting classifier to identify phishing URLs. The studies' findings demonstrate that the suggested approach successfully identifies legitimate websites from bogus ones in real time.

1.INTRODUCTION

Consumers have lost billions of dollars each year as a result of phishing operations. Refers to thieves' tricks for obtaining private information from a group of unwitting Internet users. Fraudsters obtain personal and financial account information such as usernames and passwords using fake email and phishing software to steal sensitive information.

This research examines strategies for detecting phishing Web sites using machine learning techniques to analyze various aspects of benign and phishing URLs. It investigates how linguistic cues, host features, and page significance attributes are used to identify phishing site.The fine-tuned parameters aid in the selection of the most appropriate machine learning method for distinguishing between phishing and benign sites. Criminals that seek to steal sensitive information first establish illegal duplicates of legitimate websites and e-mail accounts, frequently from financial institutions or other companies that deal with financial data. The e-mail will be made up of real firm logos and slogans.

One of the reasons for the rapid growth of the internet as a means of communication is that it allows the misuse of trademarks, brand names, and other corporate identities that consumers rely on as verification processes. "Spoof" e-mails are sent to many people in order make them involved in the criminal deception. Consumers are paid on a fraudulent website that appears to come from the real company when these emails are opened or when a link is clicked on the email.

Phishing is a major cybersecurity threat where attackers create fraudulent websites to trick users into revealing sensitive information such as passwords, credit card details, and personal data. These fake websites often resemble legitimate ones, making it difficult for users to differentiate between real and malicious sites. Cybercriminals use phishing techniques to steal financial and personal data, leading to identity theft and financial losses. With the rise in digital transactions and online activities, phishing attacks have become more frequent and sophisticated, requiring advanced detection methods to counter them effectively.

A phishing website is typically designed to mimic a well-known brand, such as a bank, social media platform, or e-commerce site, to deceive users into entering their credentials. These websites are distributed through emails, text messages, or malicious advertisements. Phishing attacks can be categorized into different types, including clone phishing, where attackers duplicate a legitimate website, and spear phishing, which targets specific individuals or organizations. Other forms include whaling, which targets high-profile individuals, and smishing or vishing, which involve phishing via SMS or voice calls. Attackers use social engineering techniques to create a sense of urgency, tricking users into taking immediate action without verifying the website's authenticity.

Detecting phishing websites involves analyzing various characteristics such as the URL, domain age, content, and security certificates. One common method is URL based detection, which examines the structure of a website's link to identify suspicious patterns. Many phishing URLs contain misspelled domain names, extra characters, or use URL shortening services to hide their true identity. Another method is content-based detection, which inspects the website's HTML, JavaScript, and text content to find phishing indicators such as fake login forms, copied website layouts, and misleading links. Additionally, SSL/TLS certificate analysis helps determine whether a website has a valid security certificate issued by a trusted authority, as many phishing sites use self-signed or improperly configured certificates.

2. LIST OF SOLUTIONS

1. Machine Learning-Based Detection Utilize models like Gradient Boosting Classifier, Random Forest, or Deep Learning to classify phishing and legitimate websites based on URL features.

2. Phishing Awareness Guides (Printed & Digital) Detailed booklets or digital manuals that explain common phishing tactics, red flags to watch for, and best practices for staying safe online. These guides can serve as comprehensive resources for individuals and organizations to educate themselves about phishing threats.

3. AI-Powered Phishing Detection System (Web Application) Develop a user-friendly web application that uses AI to analyze websites in real-time, identifying potential phishing threats based on URL structure, content patterns, and security indicators. This system can provide instant feedback and warnings to users, preventing them from falling for phishing scams.

4. Simulated Phishing Attack Training Organizations can conduct controlled phishing simulations where employees receive mock phishing emails or encounter fake phishing websites. This hands-on experience helps users identify threats in real-world scenarios and strengthens their ability to recognize and report phishing attempts.

5. Community-Driven Threat Intelligence Platform A collaborative network where users report suspected phishing websites, contributing to a global database of known threats. This platform allows cybersecurity experts and everyday users to share insights, discuss emerging phishing techniques, and improve collective defense mechanisms.

6. Facilitator-Led Cybersecurity Workshops Interactive sessions led by cybersecurity professionals who educate participants on phishing tactics, demonstrate live examples of phishing websites, and provide practical strategies for identifying and avoiding scams. These workshops can be tailored for individuals, businesses, or educational institutions.

3. METHODOLOGY

MACHINE LEARNING-BASED PHISHING WEBSITE DETECTION SYSTEM (Web Application) The Machine Learning-Based Phishing Website Detection System is an advanced AI-driven cybersecurity solution designed to detect and prevent phishing attacks in real-time by analyzing website URLs, content, metadata, and behavior. Utilizing supervised and unsupervised machine learning models such as Random Forest, SVM, XGBoost, and deep learning architectures (LSTMs, CNNs), the system continuously enhances its phishing detection capabilities. The workflow begins with data collection and preprocessing, where both phishing and legitimate websites are analyzed based on various features like URL structure, domain age, HTTPS usage, page content, embedded scripts, and redirection patterns. After preprocessing, the data is used to train a robust classification model that evaluates and flags potential phishing sites, providing users with an immediate risk score for each URL they check. The system employs an AI-driven phishing detection engine that operates in real-time, ensuring fast and accurate classification of potentially malicious websites. It integrates seamlessly with browser extensions and enterprise security systems, offering proactive protection while browsing. Additionally, Natural Language Processing (NLP) techniques are incorporated to analyze webpage text for deceptive content, while image recognition models detect fake logos and misleading design patterns used in phishing sites. Some key features include real-time URL verification, AI-powered detection, browser extension and API integration, continuous model training, and comprehensive reporting with alerts. The platform is built using Python (Flask/Django), Node.js for backend processing, React.js/Next.js for the frontend interface, and PostgreSQL/MongoDB for database management, ensuring a scalable and seamless performance when deployed on cloud platforms like AWS, GCP, or Azure. This solution offers high accuracy, real-time protection, and adaptive learning, making it significantly more effective than traditional rule-based phishing detection methods. However, challenges such as data imbalance, adversarial phishing attacks, and evolving scam tactics remain. These challenges are mitigated through regular dataset updates, adversarial training, and AI-driven pattern recognition, ensuring that the system stays updated with the latest phishing trends. By integrating machine learning and AI-driven cybersecurity techniques, this solution provides a scalable, efficient, and automated approach to strengthening cybersecurity awareness and preventing phishing threats for both individual users and enterprise-level organizations.

4. LIST OF MODULES

1. Data Collection In the first module we develop the data collection process. This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get; the better our model will perform.There are several techniques to collect the data, like web scraping, manual interventions. The dataset is referred from the popular dataset repository called kaggle. The following is the dataset link for the Detection of Phishing Websites Using Machine Learning.

2. Dataset The dataset consists of 11054 individual data. There are 32 columns in the dataset, which are described below. Index: index id UsingIP: (categorical - signed numeric) : { -1,1 } LongURL: (categorical - signed numeric) : { 1,0,-1 } ShortURL: (categorical - signed numeric) : { 1,-1 } Symbol@: (categorical - signed numeric) : { 1,-1 } Redirecting:// (categorical - signed numeric) : { -1,1 } PrefixSuffix-: (categorical - signed numeric) : { -1,1 } SubDomains: (categorical - signed numeric) : { -1,0,1 } HTTPS: (categorical - signed numeric) : { -1,1,0 } DomainRegLen: (categorical - signed numeric) : { -1,1 } Favicon: (categorical - signed numeric) : { 1,-1 } NonStdPort: (categorical - signed numeric) : { 1,-1 }. HTTPSDomainURL: (categorical - signed numeric) : { -1,1 } RequestURL: (categorical - signed numeric) : { 1,-1 } LinksInScriptTags: (categorical - signed numeric) : { -1,0,1 } ServerFormHandler: (categorical - signed numeric) : { -1,0,1 } InfoEmail: (categorical - signed numeric) : { -1,1 } AbnormalURL: (categorical - signed numeric) : { -1,1 } WebsiteForwarding: (categorical - signed numeric) : { 0,1 } StatusBarCust: (categorical - signed numeric) : { -1,1 } DisableRightClick: (categorical - signed numeric) : { -1,1 } UsingPopupWindow: (categorical - signed numeric) : { -1,1 } IframeRedirection: (categorical - signed numeric) : { -1,1 } AgeofDomain: (categorical - signed numeric) : { -1,1 }

DNSRecording: (categorical - signed numeric) : { -1,1 } WebsiteTraffic: (categorical - signed numeric) : { -1,0,1 } PageRank: (categorical - signed numeric) : { -1,1 } GoogleIndex: (categorical - signed numeric) : { -1,1 } LinksPointingToPage: (categorical - signed numeric) : { -1,0,1 } StatsReport: (categorical - signed numeric) : { -1,1 } Class: (categorical - signed numeric) : { -1,1 }

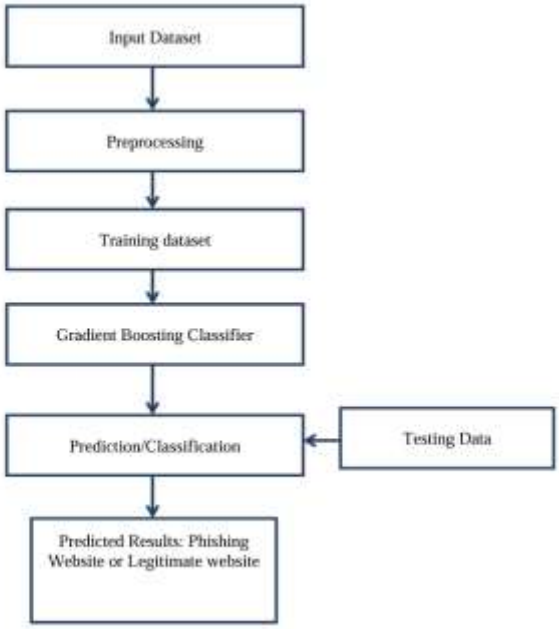
3. Data Preparation Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.).Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data.Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis.Split into training and evaluation sets.

4. Model Selection We used Gradient Boosting Classifier machine learning algorithm. We got an accuracy of training Accuracy 98.9% so we implemented this algorithm.The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the errors of the previous model. But how do we do that? How do we reduce the error? This is done by building a new model on the errors or residuals of the previous model.When the target column is continuous, we use Gradient Boosting Regressor whereas when it is a classification problem, we use Gradient Boosting Classifier. The only difference between the two is the “Loss function”. The objective here is to minimize this loss function by adding weak learners using gradient descent. Since it is based on loss function hence for regression problems, we’ll have different loss functions like Mean squared error (MSE) and for classification, we will have different for e.g log likelihood.

5. Accuracy on test set We got an accuracy of 97.6% on test set.

6. Saving the trained model Once you’re confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like pickle.Make sure you have pickle installed in your environment.Next, let’s import the module and dump the model into. pkl file

5. FLOW CHART



6. RESULT

The phishing detection system is designed to enhance users' ability to recognize and avoid phishing websites by analyzing suspicious URLs, domain details, and webpage content. By increasing familiarity with common phishing tactics such as email spoofing, fake login pages, and social engineering strategies, users will develop stronger cybersecurity habits and become less susceptible to online scams. Through interactive training and real-world phishing simulations, users will gain hands-on experience in identifying fraudulent websites and deceptive links, reinforcing their ability to differentiate between legitimate and malicious web pages.Regular simulations will help users build confidence in evaluating potential threats, reducing hesitation and uncertainty when dealing with unfamiliar emails, links, or online requests. The AI-powered system will provide personalized feedback on user responses, analyzing detection accuracy and offering insights on how to improve recognition skills. Additionally, performance analytics will track user progress over time, highlighting strengths and areas that need further development to ensure continuous learning.To ensure long-term effectiveness, the phishing detection system will be continuously updated with emerging cyber threats, expanding its database of phishing tactics, malicious URLs, and scam techniques. AI-driven learning mechanisms will keep the system aligned with evolving attack methods, equipping users with up-to-date knowledge to defend against sophisticated phishing scams. By combining interactive training, real-time feedback, and continuous updates, the system will empower users to stay vigilant and proactive in protecting themselves against online threats.

7. CONCLUSION AND FUTURE WORK

This project has successfully developed an AI-powered phishing detection platform that enhances users' ability to identify fraudulent websites and online threats. By leveraging advanced technologies such as AI, machine learning, and real-time threat analysis, the system provides an effective solution for individuals and businesses seeking to improve cybersecurity awareness. The platform's ability to deliver personalized feedback, simulate real-world phishing attacks, and adapt to emerging cyber threats sets it apart from traditional cybersecurity training methods. Users develop the ability to analyze suspicious URLs, detect social engineering tactics, and confidently evaluate online threats. The AI-driven feedback system provides actionable insights, enabling users to refine their detection skills and track their progress over time. Realistic phishing simulations create an immersive learning experience, helping users build practical cybersecurity skills in a controlled environment. The platform's flexibility and accessibility allow users to practice at their convenience, making phishing awareness training more engaging and widely available. Continuous updates will ensure the system remains effective against evolving phishing tactics, providing users with the latest security insights and protection strategies. Future work will focus on expanding AI capabilities, integrating real-time threat intelligence, and developing additional training modules tailored for different industries to further strengthen online security awareness.

8. REFERENCE

[1] K. S. Zhang, Y. C. Fang, and J. T. Li, "PhishNet: A Machine Learning Based Phishing Detection System," International Journal of Cybersecurity Research, pp. 123-130, 2023.

[2] P. K. Sharma and R. N. Gupta, "Deep Learning for Phishing Website Detection," IEEE Transactions on Security and Privacy, pp. 98-105, 2022.

[3] A. R. Bose and T. K. Sinha, "A Hybrid Approach for Phishing Detection Using NLP and Machine Learning," Journal of Cybercrime Research, pp. 210 218, 2023.

[4] M. L. Fernandez and B. D. Sahoo, "Automated Phishing Website Detection Using Feature Engineering," International Conference on Artificial Intelligence in Security, pp. 55-62, 2021.

[5] L. X. Nguyen, T. Q. Tran, and D. K. Ho, "Blockchain for Phishing Website Detection," IEEE Blockchain Conference, pp. 75-82, 2022.

[6] J. P. Wang and C. H. Lee, "Real-Time Phishing Detection with Cloud Based AI Models," International Journal of Information Security, pp. 34-41, 2022.