Detection of Phishing Website using Machine Learning

Mrs. Keerthana Shankar[1], U. Rithika[2], Sathya M[3], Vaishnavi Chitapur[4], Tejaswini J[5]
[1]*Assistant. Professor, Department of Computer Science and Engineering,*
[2,3,4,5] *UG Students, Department of Computer Science and Engineering*
[1,2,3,4,5]*Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka India*

Abstract – Offenders looking for touchy data develop illicit clones of real websites and mail accounts. The email will be made up of genuine firm logos and mottos. When a client clicks on an interface given by these programmers, the programmers pick up get to all user's private data, counting bank account data, individual login passwords, and pictures. Irregular Woodland and Choice Tree calculations are intensely utilized in display frameworks, and their precision should be enhanced. The existing models have more inactivity. Existing frameworks don't have a specific user interface. Within the current framework, distinctive calculations are not compared. Buyers are driven to a fake site that shows up to be from the true company when the e-mails or the joins given are opened. The models are utilized to identify phishing Websites based on URL centrality highlights and to discover and actualize the ideal machine learning show. The Random Forest method will compare the accuracy and the result.

Index Terms-Features, Machine Learning dataset, URL, Phishing.

I. INTRODUCTION

In today's digital connectivity, phishing attacks have become a most powerful threat to users of the internet, posing a high risk to their tactful information. Machine learning offers a solution to all the challenges. It provides a more sophisticated and adaptable approach for the identification of websites that are phishing by analyzing the features that are extracted from the website, machine learning algorithms can learn to distinguish between legitimate and phishing websites with high accuracy the machine learning algorithm that is random forest has emerged as a most powerful tool for phishing website detection.

An ensemble learning technique is the random forest algorithm using several decision trees together to provide predictions. Every decision tree produces a varied collection of classifiers by being trained on a random subset for the characteristics and data. The decision trees vote to predict the final prediction by a majority method. By using an ensemble technique, the model's overall accuracy is increased and the possibility of overfitting is decreased. As random forests can manage enormous datasets, handle missing data, and discern intricate correlations between features, they are especially well-suited for phishing detection.

Adding on to that, they are not vulnerable to misconceptions regarding the underlying data distribution due to their non-parametric structure.

II. PROBLEM DESCRIPTION

An ensemble learning technique called the random forest algorithm uses several decision trees together to provide predictions. Every decision tree produces a varied collection of classifiers by being trained on a random subset of the characteristics and data. The decision trees determine the final prediction by a majority method. By using an ensemble technique, the model's overall accuracy is increased and the likelihood of overfitting is decreased.

Because random forests can manage enormous datasets, handle missing data, and discern intricate correlations between features, they are especially well-suited for phishing detection.

In addition, they are not so vulnerable to misconceptions regarding the underlying data distribution due to their non-parametric structure.

Creating a scalable and effective method for identifying and stopping phishing attacks is essential to protecting internet users and their private data. Current methods, which rely on manually created rules and blacklists, frequently fall short since phishing websites are constantly changing to avoid being discovered. A viable substitute is machine learning, which offers a flexible and data-driven method for detecting and categorizing phishing websites.

III. LITERATURE SURVEY

Phishing attacks are still a danger to cybersecurity because they take advantage of gullible people by using trickery to obtain private data. To address this dynamic dilemma, scholars have investigated diverse approaches for identifying and averting phishing assaults. The purpose of this literature review is to present a thorough analysis of current methods, with an emphasis on using ML techniques in the recognize of phishing websites.

Qasem Abu Al-Haija, and Ahmad Al Badawi, [1] use traditional supervised learning techniques, such as Support Vector Machines (SVMs) and Naiev Bayse classifirs, which is the main emphasis of early machine learning techniques. These approaches were constrained by their reliance on manually designed characteristics and their incapacity to adjust to new phishing strategies, even though they achieved respectable accuracy.

Jordan Stobbs, Biju Issac, and Seibu Mary Jacob, [2] stressed the significance of feature engineering and feature selection as phishing websites become more complex. While feature selection seeks to identify the most discriminative elements that aid in accurate classification, feature engineering concentrates on extracting pertinent and useful aspects from websites. Methods for ensemble learning, such as gradient-boosting machines and Random Forests, have become effective instruments for phishing detection. These techniques combine several classifiers, each trained on subsets of the data, to increase robustness and total accuracy.

Noor Faisal Abedin, Rosemary Bawm, Tawsif Sarwar, Mohammed Saifuddin, Mohammad Azizur Rahman, Sohrab Hossain, [3] When compared to typical machnie learinng techniques, deep learinng techniques particularly neural networks have shown higher effectiveness in phishing detection. Given the ability of identifying advanced patterns and correlations in data, neural networks are ideally equipped to handle the complex nature of phishing websites.

Shihabuz Zaman, Shekh Minhaz Uddin Deep, [4] To address the ever-evolving nature of phishing techniques, researchers have looked into ways to improve the adaptability of ML-based phishing detection systems. This entails implementing methods for ongoing education, adjusting to novel functionalities, and managing conceptual drift.

Dr.G.K.Kamalam, Dr.P.Suresh, R.Nivash, A. Ramya, G.Raviprasath, [5] Particularly regarding phishing detection, the interpretability of model has grown in significance. Stakeholders can comprehend the variables influencing the system's classifications through Explainble AI (XAI) approaches, which offer more insight into the ML models' decision-making process.

Preethi P. Pokal Ramadevi; K Akshaya, Sangamitra SD; Pritikha A P [6] Systems for detecting phishing in real time are essential for safeguarding individuals when they browse the internet. Researchers have looked into strategies for effective data processing and model optimisation as well as real-time phishing detection using machine learning.

Anu Vazhayil, Vinaya Kumar R, and Soman KP etal. [7] focus on combining CNN with the Convolutoinal Neural Newtork, Long Short Term Memry to predict the accuracy of categorizing phishing URLs. CNN assists in identifying unique information between the characters while LSTM retrieves sequential information. CNN used to discover the unique character relationships.

IV. URL-BASED DETECTION

Phishers are the unseen wolves that lurk among trustworthy websites in the digital wilderness. We can examine their internet addresses and URLs to uncover who they are. We look for discrepancies, such as suspicious keywords, misspelled domains, and excessive dashes, much like detectives analyzing clues. Blacklists provide hints of previous deceptions, and Machine learning assists us in identifying potential threats. Despite its vigilance, URL-based detection is not the only defense.

A. Phishing

The cyber security attack is where hackers create believable identities to trick people into revealing their personal information, usernames, passwords, or bank account information. Phishing websites are fraudulent online platforms that eventually mimic trustworthy websites in an attempt to trick the users and tend to reveal their personal information. These fraudulent websites will use social engineering techniques, that take advantage of the casualty’s confidence and sense of urgency. In the digital era,

protecting private and organizational security requires an awareness and response regarding the phishing risks.

B. Machine Learning Approach

Machine learning transforms the detection of Phishing websites by identifying patterns that suggest malicious intent. To distinguish between genuine and phishing websites in this situation, machine learning algorithms examine various factors, including URL architecture, attributes, and behavioral patterns. These models the capacity to generalize through training on chosen datasets. ML-driven approaches strengthen cybersecurity efforts and, increase user and organization resilience against phishing attacks.

C. Dataset

The quality and representativeness of the training data have a key role in how well machine learning models detect phishiing websites. Our machine-learning models for detection of phishing websites will be trained and evaluated using this dataset as the foundation. We follow ethical standards while using data, and privacy laws are followed in the acquisition and management of the dataset.

Categorical Data: The dataset mostly employs features that are classified as "-1," "1," and "0," which denote the presence or lack of particular attributes. Because of this, the data is appropriate for machine learning models, which can conclude links and patterns among different categories.

Binary Classification: The website's classification as either authentic (1) or phishing (-1) is indicated by the binary variable "Class," which is the last attribute. As a result, we can train machine learning models to anticipate new websites' classes based on their traits.

Discovering Lexical Clues: Functions such as "LongURL," "Symbol@," and "Prefix Suffix-" probe the website's URL in an attempt to find patterns or characters that raise red flags for phishing attempts Examining

Infrastructure:"HTTPS,""DomainRegLen," and "Favicon" look at the technical features of the website, checking for discrepancies in its security protocols, age of domain registration, and visual components.

D. Feature Selection

To enhance model performance, feature selection entails selecting the most significant and illuminating features from the dataset. The majority of the features are. Domain and categorical, with values like -1, 1, and 0. The analysis focuses on whether or not the URL contains particular qualities.

ATTRIBUTES	VALUES
UsingIP	{ -1,1 }
LongURL	{ 1,0,-1 }
ShortURL	{ 1,-1 }
Symbol@	{ 1,-1 }
Redirecting	{ -1,1 }
PrefixSuffix-	{ -1,1 }
SubDomains	{ -1,0,1 }
HTTPS	{ -1,1,0 }
DomainRegLen	{ -1,1 }
Favicon	{ 1,-1 }
NonStdPort	{ 1,-1 }
HTTPSDomainURL	{ -1,1 }
RequestURL	{ 1,-1 }
AnchorURL	{ -1,0,1 }
LinksInScriptTags	{ -1,0,1 }
ServerFormHandler	{ -1,0,1 }
InfoEmail	{ -1,1 }
AbnormalURL	{ -1,1 }
WebsiteForwarding	{ 0,1 }
StatusBarCust	{ -1,1 }
DisableRightClick	{ -1,1 }

IV. IMPLEMENTATION

Utilization is the application or carrying out of a positive idea, action, plan, model, estimation, system, or specific. It comprises using programming and programming action to promote and enforce a computer or programming technique. A given specification or standard may also have different executions or implementations.

Modules

- Data collection
- Data preprocessing
- Feature extraction
- Module selection
- Analysis

Description of the modules

Data collection

Data collection is a fundamental phase of the development of a machine-learning phishing detection system. The quality of significant data affect the model's ability to accurately identify and classify phishing sites. Different sources can be used in collecting the data related to phishing sites

Publicly accessible databases of phishing websites: These databases contain lists of known phishing incident websites, often classified by purpose Sectors, or types of fraud. Examples included OpenDNS PhishTank, Phishing.org, and URLhaus.

Crowdsourcing platforms: Platforms like Kaggle and GitHub provide phishing records and legal websites created by researchers and enthusiasts. These records may contain other functions and notes.

Web Scraping Techniques: Web Scraping methods can be used to extract data from websites, including phishing sites, to build custom records. This approach allows for individual customization data collection based on specific requirements.

Data preprocessing

The raw website data must be handled by the Preprocessing Module to get it ready for faeture extraction and model training. This module completes several essential tasks, like data cleaning. We can reduce noise and bias in the model by eliminating inconsistent data points by using method of data cleaning. Handles missing values in the data by substituting suitable values, like the mean, median, or mode, for them, finds and eliminates aberrant data items that drastically differ from the remaining data and havethe potential to skew the model's learning process. Ensures that features with bigger sizes do not dominate the model's learning process by scaling the data to a consistent range.

Feature extraction

By preprocessing the website data, this module is responsible for extracting essential features. The model inputs and shows the traits that set phishing websites apart from legitimate ones. The traits consist of URL features and examine the subdomain name, domain name, query string, and path parameters, of the websitee to find patterns linked to phishng websites.

Module selection

This module is in charge of model selection and training and is also in charge of selecting a particular machine learning method, the model is being trained using the features that are being retrieved, later this is optimized with its hyperparameters. This module completes several critical tasks, such as:

Algorithm Selection: By Reviewing various machine learning algorithms, like Random Forest, etc, They Support naïve Bayes classifiers and Vector Machines, depending on how well they work, and compare datasets that are suitable for the goal of phishing detection.

The training of the model involves dividing data into testing and training sets, using this training data the chosen algorithm needs to be trained. Here features and website classifications are related to one another.

Hyperparameter Optimization the model eventually adjusts the hyperparameters to enhance their performance on the test data.

Analysis

How well the machine learning model is performing after training is checked by the Analysis Module using various measures, including precision, accuracy, and F1-score and recall. This module helps us to improve the performance by providing insights into the model's disadvantages and advantages to the areas being pinpointed.

Accuracy: The total percentage of precise classifications the model is indicated in this module.

Precision: Calculates the percentage of phishing-classified websites that are genuinely phishing.

Recall: The percentage of real phishing websites is calculated appropriately and labeled as.

The efficiency is being determined using the Model Assessment Module in mL model additional optimization or algorithm selection by assessing these indicators.

V. CONCLUSION

Phishnig websites pose a massive hazard to cyber-security, tricking users into divulging sensitive information and exposing them to monetary losses, identification theft, and reputational damage. Machine learning has emerged as a powerful tool for combating phishing attacks, offering an adaptable and data-driven approach to identifying and classifying phishing websites. This project explored the effectiveness of machine learning, particularly the Randmo Foerst algorithm, in detecting phishing websites. By reading a complete dataset of phishing and legitimate websites, the Randmo Foerst model achieved an accuracy of

96.82%, demonstrating its ability to accurately Distinguish between the 2 classes. The project also investigated various feature Extraction strategies to discover relevant traits that distinguish phishing websites from valid ones. URL features, content features, and visual features are extrcated from the url of the websites are the input to the machine learning model. The result is the differentiation between legitimate websites from the phishing ones. Also the high accuracy of the Randmo Foerst model, the project also highlighted its adaptability and robustness. The model was able to generalize well to new data and maintained its effectiveness against evolving phishing techniques. This adaptability is crucial in the ever-changing landscape of cyber-security threats.

REFERENCE

[1] Qasem Abu Al-Haija, Ahmad Al-Badawi, "URL-based Phishing Websites Detection via Machine Learning", 2021.

[2] Jordan Stobbs, Biju Issac, Seibu Mary Jacob, "Phishing Web Page Detection Using Optimised Machine Learning", 2020

[3] Noor Faisal Abedin, Rosemary Bawm, Tawsif Sarwar, Mohammed Saifuddin, Mohammad Azizur Rahman, Sohrab Hossain, "Phishing Attack Detection using Machine Learning Classification Techniques", 2020.

[4] Shihabuz Zaman;Shekh Minhaz Uddin Deep;Zul Kawsar;Md. Ashaduzzaman; Ahmed Iqbal Pritom," phishing website detection using effective
classifiers", 2021

[5] Dr.G.K.Kamalam, Dr.P.Suresh, R.Nivash, A.Ramya, G.Raviprasath, "Detection of Phishing Websites Using Machine Learning", 2022 International Conference on Computer Communication and Informatics (ICCCI - 2022), IEEE Conference, 2022.

[6] Preethi P; Pokal Ramadevi; K Akshaya; Sangamitra SD; Pritikha A P "phishing website detection using feature selection technique", 2023

[7] Anu Vazhayil, Vinaya Kumar R. and Soman KP, "Comparative Study Of The Detection Of Malicious URLs Using Shallow and Deep Networks "2018

[8] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Many Understanding and Assisting Users' Online Choices with Nudges for Privacy and Security 50(3), Article No. 44, ACM Computing Surveys, 2017.

[9] Helena Matute, Mara M. Moreno-Fernández, Fernando Blanco, Pablo Garaizar I'm looking for phishers. To combat electronic fraud, Internet users' sensitivity to visual deception indicators should be improved. pp.421- 436 in Computers in Human Behavior, Vol.69, 2017.

[10] F.J. Overink, M. Junger, L. Montoya. Preventing social engineering assaults with priming and warnings does not work. pp.75-87 in Computers in Human Behavior, Vol.66, 2017. 2017.

[11] Retting, "Caution: a comment on Alena Erke's red light for red-light cameras? Ameta- analysis of the effects of 2017

[12] Anu Vazhayil, Vinaya Kumar R and Soman KP, "Comparative Study Of The Detection Of Malicious URLs Using Shallow and Deep Networks "2018

[13] Amani Alswailem, Bashayr Alabdullah, and Norah Alrumayh, "Detecting Phishing Websites Using Machine Learning"2019

[14] Varsharani Ramdas Hawanna, V. Y. Kulakarni and R.A. Rane, "A Novel Algorithm to Detect Phishing URLs", 2016

[15] Mohamed Alqahtani." Phishing Websites Classification Using Association Classification (ATWCAC)", 2019

[16] Martyn Weedon, Dimitris Tsaptsinos and James Denholm-Price, "Random Forest Explorations for URL Classification", 2017