IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 7, July 2025

DOI: 10.17148/IARJSET.2025.12702

DETECTING PHISHING WEBSITES WITH AN ENSEMBLE MACHINE LEARNING METHOD

Mrs. G. Suvetha[1], Dr. T. Jaya[2], Dr. Z. Mary Livinsa[3], R.Ranjith Kumar[4], Mohammed Ajis[5]

Department of Electronics and Communication Engineering,

Vels Institute of Science, Technology & Advanced Studies, (VISTAS), Chennai, India[1]

Department of Electronics and Communication Engineering,

Vels Institute of Science, Technology & Advanced Studies, (VISTAS), Chennai, India[2]

Department of Electronics and Communication Engineering,

Vels Institute of Science, Technology & Advanced Studies, (VISTAS), Chennai, India[3]

Department of Electronics and Communication Engineering,

Vels Institute of Science, Technology & Advanced Studies, (VISTAS), Chennai, India[4]

Department of Electronics and Communication Engineering,

Vels Institute of Science, Technology & Advanced Studies, (VISTAS), Chennai, India[5]

Abstract: phishing websites now pose a critical threat to digital infrastructure across industries. They frequently serve as the initial vector for various cyber intrusions that steal, change, or gain access to both customer and company data. This study presents a new way to find phishing websites by combining attribute selection and data point derivation methods with an ensemble-based machine learning algorithm. It does this after looking at all the research that is already out there. The proposed method uses a carefully chosen dataset to build and test ensemble models that can accurately predict phishing activity.

I. INTRODUCTION

Phishing websites are deceptive platforms crafted to mimic legitimate websites, in order to trick users into giving secrets, like credit card values, usernames, passwords, and other personal information. Phishing websites are typically created by cybercriminals who use a variety of techniques to make the website appear legitimate, such as copying the design and layout of a real website, or using a domain name that is similar to the legitimate website. Cybercriminals often redirect users to phishing platforms via multiple channels, including email, social media, instant messaging, or even search engine results. The phishing message may claim that the user's account is about to expire or that there is a problem with their account, and urge the user to click on a link or open an attachment to fix the problem. to defend effectively against phishing-based intrusions, users should be cautious when clicking on links or opening attachments in emails, especially from unknown or suspicious sources In addition, users need to verify a website's reliability via a look at the the URL and looking for signs of a secure connection, such as the padlock icon or "https" in the address bar. Additionally, users should enable two-factor authentication and regularly update their passwords to reduce the risk of their accounts being compromised.

II. LITERATURE SURVEY

Title: "A Novel Hybrid System for Phishing Website Detection Based on Unsupervised Machine Learning Techniques", 2021

Authors: Taher Abualsaud, Tarek Gaber, Hamada Ghaleb, Ashraf Darwish

In this paper, the authors propose a hybrid system for phishing website detection based on unsupervised machine learning techniques. The proposed system aims to detect phishing websites without relying on labeled data, which is often difficult and expensive to obtain. The clustering module uses several unsupervised learning algorithms, including k-means, DBSCAN, and Gaussian mixture models, to cluster the websites based on their features. The authors point out that more study is required to refine the attribute selection and data point derivation and clustering modules and assess the system on bigger and more varied datasets.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 7, July 2025

DOI: 10.17148/IARJSET.2025.12702

Title: "A Deep Learning Approach for Phishing Website Detection Based on Domain Name and Visual Features", 2021
Authors: Lingfeng Zhang, Yibin Hou, Yuanyuan Li
In this paper, the authors propose a deep learning approach for phishing website detection based on both domain name and visual features. The proposed system aims to leverage the strengths of deep learning in attribute selection and data point derivation and classification to improve the accuracy of phishing website detection. The proposed system consists of two main components: a domain-based attribute selection and data point derivation module and a visual-based attribute selection and data point derivation module. The authors note that DNNs can be useful in this regard, as they can learn complex representations of the data and generalize well to new examples. The results show that the proposed system achieves higher accuracy in detecting phishing websites than the existing methods.

Title: "Phishing Website Detection Based on a Rule-Based Method and its Application on Mobile Devices", 2021
Authors: Kai Zhu, Jiancheng Li, Yiqun Liu, et al.
In this paper, the authors propose a rule-based approach for phishing website detection and its application on mobile devices. The proposed system aims to detect phishing websites based on a set of rules that represent known phishing patterns and characteristics. The proposed system consists of two main components: a rule-based module and a mobile application module. Overall, the proposed system shows promise in detecting phishing websites using a rule-based approach, which can be useful in situations where labeled data is not available or where the features are difficult to extract using traditional methods. The author's note that further research is needed to optimize the rule-based module and to evaluate the system on larger and more diverse datasets.

Title: "A Hybrid Strategy for Machine Learning-Based Phishing Website Diagnosis"
", 2021
Authors: Sumathi Raman, Sreeramana Aithal, Anusha Prabhu
In this paper, the authors propose a hybrid approach for phishing website detection using machine learning and rule-based techniques. The proposed system aims to leverage the strengths of both approaches to improve the accuracy of phishing website detection. The proposed system consists of two main components: a machine learning-based module and a rule-based module. Overall, the proposed system shows promise in detecting phishing websites using a hybrid approach, which combines the strengths of machine learning and rule-based techniques. The author's note that further research is needed to optimize the attribute selection and data point derivation and classification modules and to evaluate the system on larger and more diverse datasets.

Title: "A Graph-Based Phishing Detection System Using Multi-Objective Optimization", 2021
Authors: Guohao LAN, Haiyun Xu, Shuang Xu, et al.
In this paper, the authors propose a graph-based phishing detection system using multi-objective optimization by defining the website as a graph and using graph-based features to spot phishing movements, this suggested method seeks to detect phishing websites.
The proposed system consists of two main components: a graph construction module and a multi-objective optimization module. The graph construction module uses a set of features extracted from the website, such as the URL structure, the presence of redirects, and the similarity to known phishing websites, to construct a graph representation of the website. Overall, the proposed system shows promise in detecting phishing websites using a graph-based approach, which can capture the complex relationships between the website components and identify phishing patterns that are difficult to detect using traditional methods. The author's note that further research is needed to optimize the graph construction and multi-objective optimization modules and to evaluate the system on larger and more diverse datasets.

Title: "A Graph-Based Approach for Phishing Website Detection", 2020
Authors: Rui Li, Yongsheng Ou, and Gang Li
In this paper, the authors propose a system for detecting phishing websites using a graph-based approach. The proposed system aims to leverage the structural properties of web pages to distinguish between phishing and legitimate websites. The graph construction module constructs a graph representation of the web page, where each node represents a structural element of the web page, such as HTML tags or hyperlinks, and each edge represents the relationship between two nodes. The graph-based classification module uses a set of graph-based features, such as the degree distribution and clustering coefficient, to classify the web page as phishing or legitimate.

Title: "Phishing Website Detection using Behavioral Analysis",2020
Authors: Abdulaziz Alnajim, Nasser Alsaedi, and Hatim Alharbi
In this paper, the authors propose a system for detecting phishing websites using a behavioral-based approach. The proposed system aims to leverage the behavioral characteristics of the web user to distinguish between phishing and legitimate websites. The proposed system consists of two main components: a user behavior collection module and a

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 7, July 2025

DOI: 10.17148/IARJSET.2025.12702

machine learning-based classification module. The performance of the proposed system was evaluated on a dataset of phishing and legitimate websites, and the results showed that the proposed system achieved high accuracy, precision, and recall in detecting phishing websites. The proposed system also outperformed several baseline methods that use traditional machine learning algorithms.

Title: "A Hybrid Approach for Phishing Detection based on Machine Learning and User Behavior Analysis", 2019
Authors: Xi Li, Xiaobin Tan, Li Chen, and Hongwei Li
In this paper, the authors propose a system for detecting phishing websites using a hybrid approach that combines machine learning and user behavior analysis. The proposed system aims to leverage both the structural properties of web pages and the behavioral characteristics of web users to distinguish between phishing and reliable websites. the proposed system consists of a two main components: a web page analysis module and a user behavior analysis module. The web page analysis module extracts a set of structural features from the web page, such as the number of hyperlinks and the length of the URL, while the user behavior analysis module collects a set of user behavior data, such as the mouse movement and keystroke dynamics, while the user interacts with the web page.

III. PROPOSED SYSTEM

Collect a publicly available dataset that includes both legitimate and malicious websites. write automation code to derive key features from the URL list. Use EDA techniques to analyse and pre-process the dataset. Divide the dataset into training and testing sets. Apply specific machine learning techniques to the dataset, such as Random Forest and Decision Tree. Write code that takes accuracy measures into account when displaying the evaluation result. Determine which of the trained models' results is superior by comparing them. Forecasting with the Flask framework. Gathering and pre-processing data.

IV. METHODOLOGY

DATA COLLECTION:

Suitable URLs were collected from the dataset published by https://www.unb.ca/cic/datasets/url-2016.html Ten thousand URLs are chosen at random from the collection. Phishing URLs are collected via the open- source service phish Tank .This site offers an hourly-updated collection of phishing URLs in several formats, including csv, json, and others.

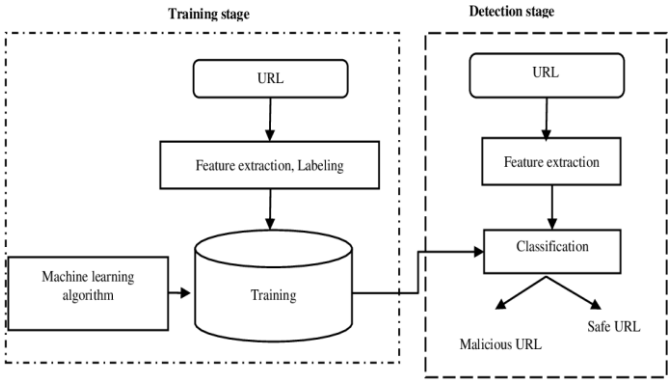

Fig 1. Architecture Diagram of Phishing Website Deduction.

DATA CLEANING / PREPROCESSING

Checks uses the is zero () and sum () functions from the Panda's library to check the data for missing or null values. Uses the pandas remove () method to eliminate the 'Domain' columns in the information. The characteristics and targeted columns from the data can be separated into x and y variables using the drop () function and a shape parameter of the panda's library.

FEATURE EXTRACTIONS

There are many various techniques and kinds of data used in phishing URL monitoring research and commercial programs.

A Phishing URL and its accompanying website differ from harmful URLs in a number of ways.
For example, an attacker might create a long, complicated domain name to hide the real domain name. A range of features used in machine learning algorithms are used in academic phishing detection techniques

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 7, July 2025

DOI: 10.17148/IARJSET.2025.12702

MODELS AND TRAINING

Preparing for training the ML method, Two thousand testing samples and eight thousand training samples make up the data. The dataset makes it evident that supervised machine learning play a part in this problem come into two primary categories: classification and regression. There is a sorting problem with this data set because the input URL is categorized as either phishing (0) or legal (0). For this project's dataset training, the following supervised machine learning models were investigated: random forest and decision tree.

DECISION TREE CLASSIFIER

decision tree algorithms are commonly applied in solving classification and regression tasks. They are essentially taught a series of if/else questions that lead to a decision. In order to learn a decision tree, The sequence of if/else question which offer the correct response in the shortest amount of time must be committed to memory. To create a tree, each test is iterated through to see which offers the most insight into the target variable

RANDOM FOREST CLASSIFIER

Classification and regression is a common method in random forests of machine learning methods. A grouping of somewhat different decision trees is all that a random forest. random forest algorithms assume uniform importance of individual decision trees will almost certainly over fit on some data, even if it projects rather well. They are quite powerful, don't require data scalability, and frequently run effectively with minimum parameter change.

ENSEMBLE ALGORITHM

ensemble learning involves aggregating several models to boost predictive reliability to outperform any one model alone. To increase the accuracy of the outcome in this situation, the predictions of two models can be combined to construct an ensemble of decision tree and random forest classifiers.

A series of if/else questions that culminate in a choice is learned by the decision tree classifier. It picks up on the tests that lead to the right response. Conversely, a group of decision trees called a random forest classifier is employed to lessen the model's overfitting.

The model's accuracy can be increased and overfitting can be decreased by combining decision tree and random forest classifiers. The random forest classifier reduces variance by mixing several decision trees, while the decision tree classifier finds the most informative variables.

Compared to employing a single model, the use of a combination of decision tree and random forest classifiers has the advantage of potentially producing more accurate predictions. By merging several models with distinct decision limits, it might help lessen the model's overfitting. This method's drawback is that it may be computationally costly and necessitate additional resources for deployment and training.
Stacking is an ensemble learning strategy that enhances predicted accuracy by combining several regression or classification models. In a stacking classifier, a meta-classifier uses the predictions of base classifiers as input characteristics.
Here's how to use Random Forest, Decision Trees, and Logistic Regression to build a stacking classifier.

Base Classifiers:
Train a Random Forest classifier.
Train a Decision Tree classifier.
Train a Logistic Regression classifier.

Meta-Classifier:
Combine the predictions from the Random Forest, Decision Tree, and Logistic Regression classifiers.
Use these predictions as input features to train a meta-classifier, which can be another Decision Tree, Random Forest, Logistic Regression, or any other classifier.

Stacking:
Combine the predictions of the base classifiers along with the original features.
Train the meta-classifier on this combined dataset.

Prediction:
To make predictions, pass the new data through the base classifiers to generate predictions.
Use these predictions along with the original features as input to the trained meta-classifier to get the final prediction.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 7, July 2025

DOI: 10.17148/IARJSET.2025.12702

Stacking helps to capture the diverse patterns learned by each base classifier and can potentially improve the overall predictive performance compared to using individual classifiers alone.

V. RESULT AND DISCUSSION

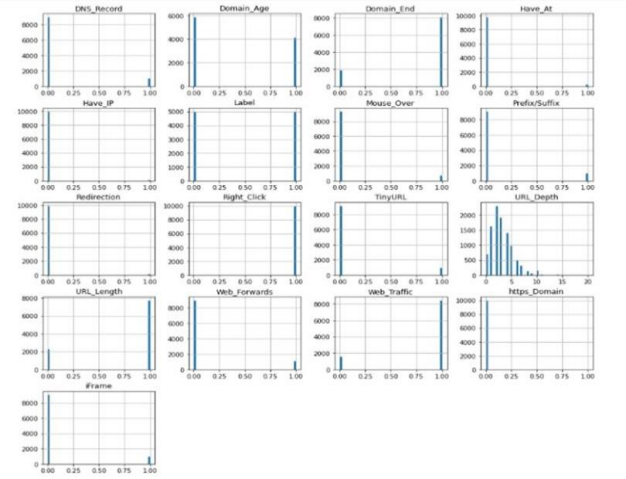

Fig 2. Feature Distribution

Fig 2 Represents the Feature Distribution of the Phishing website deduction using ensemble machine learning approach.

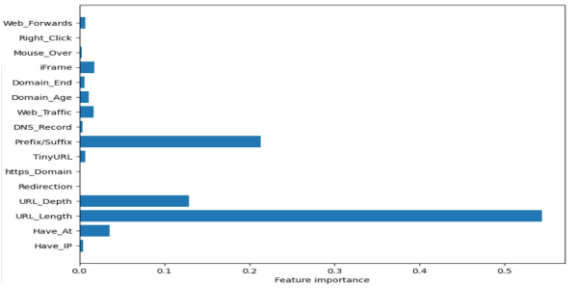

Fig 3. Feature Importance

Fig 3 Represents the Feature Importance of the Phishing website deduction using ensemble machine learning approach.

Fig 4. Screenshot of Phishing website first page.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 7, July 2025

DOI: 10.17148/IARJSET.2025.12702

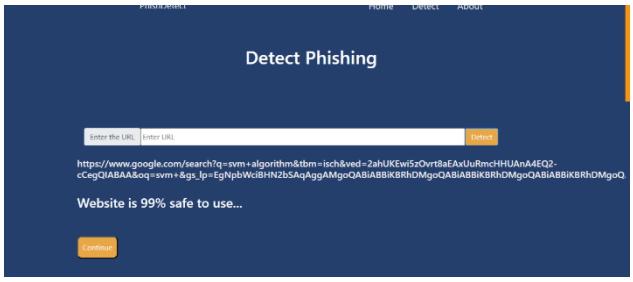

Fig 5. Screenshot shows the beginning of Phishing deduction website

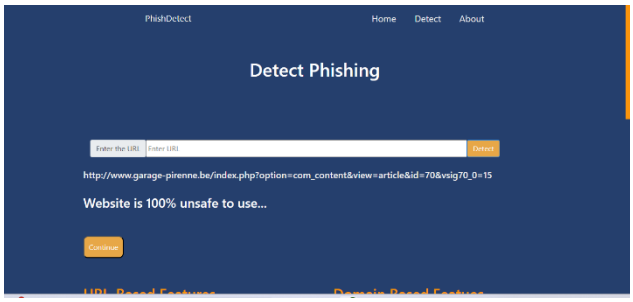

Fig 6. Screenshot shows the output of Phishing deduction website

Our study involves using machine learning methods-specifically, ensemble models-to identify phishing websites. By conducting a thorough literature review and proposing a novel method that involves attribute selection and data point derivation, The objective we have is to improve phishing detection's precision and effectiveness. using machine learning presents a viable way to get over the drawbacks of earlier techniques and successfully predict phishing assaults additional tests and performance verification with the compiled data sample will be crucial to assess the proposed method's performance and its potential impact on improving web security.

REFERENCES

[1]. APWG | Linking the National Cybercrime Response online at: https://apwg.org/ (n.d.)
[2]. IT administrators should be aware of the following 14 types of phishing attacks [online] in 2021. The website in question can be accessed at https://www.blog.syscloud.com.types of phishing
[3]. The paper "Detecting phishing sites using a novel machine learning fusion approach" by Lakshmanarao, A (2021) was presented at the 2021 international conference on machine learning and smart systems (ICAIS),1164-1169
[4]. H. Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier", 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 383-388, 2019, July
[5]. In D., Suwetha (2021),Vaishnavi presented "An Envalution of Machine Learning(ML) Strategies on malicious URL Prediction," at the 5th world conference on intelligent calculating and control system (ICICCS), 1398-140
[6]. Microsoft's consumer safety training .The Microsoft safety index illustrates the impact of inadequate internet safety measures in Singapore https://news.microsoft.com/ensg/2014/02/11/sm.001xdu50tlxsej410r11kqvksu4nz
[7]. Internal Revenue Service, IRS E-mail Schemes. Available at https://www.irs.gov/uac/newsroom/consumers-warnedof-new-surge-in-irs-email-schemes-during-2016-tax-season-tax-industry-also-targeted.
[8]. Nappa, D.(2007), as it A comparison of phishing detection methods using machine learning.
[9]. ECrime 07: proceedings of the second annual ECrime researchers summit of Anti-phishing working groups, doi:10.1145/1299015.1299021.
[10]. Phishing URL Detection: A Machine Learning and web mining-based approach, E., B., K., T.(2015) Journal of computer Applications international, 123(13),46-50doi:10.5120/ijca2015905665.
A Static Malicious JavaScript detection using SVM, Wang Wei-Hong proceedings of the 2nd international conference on computer science and electrical engineering (ICCSEE 2013)
[11]. Ningxia, For Zhang, phishing prevention using neural networks, proceedings of the international workshop on neural information processing, PP. 714-719. Springs, Heidelberg university,2004.
[12]. Rama Basnet, et al., "prevention of phishing incidents: A Machine Learning-Based Approach," proceedings of the international world wide web meeting (the web),2003. SVM library, sci-kit learn.SVM.html http://scikit-learn.org/stable/modules/svm.html.