IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Phishing Website Detection on URLs and Content Based Features Using Machine Learning

Gaganaspoorthy R[1], Dr. Manjunatha S[2]

[1]PG Student, Dept. of Computer Science and Engineering, Cambridge Institute of Technology, Bangalore, India

[2]Associate Professor, Dept. of Computer Science and Engineering, Cambridge Institute of Technology, Bangalore, India

Abstract

Phishing is a prevalent and perilous form of cybercrime. The objective of this assault is to illicitly obtain user information by gaining unauthorized access to the credentials utilized by individuals and organizations. Phishing websites typically offer a variety of clues and information that may be found on the web. Phishing sites attempt to obtain the victim's sensitive information by tricking them into visiting a website that seems similar to a legitimate one. This is a common type of criminal attack on the internet. Phishing websites are a type of cyber threat that aim to get sensitive information, such as credit card details and social security numbers. Currently, there is no definitive solution available to identify phishing assaults that is both reliable and unpredictable. This is due to the presence of multiple factors and criteria that are always changing. This research aims to utilize Machine Learning techniques to classify features, specifically Phishing Websites Data, in the UC Irvine Machine Learning Repository database. Web pages can be classified into various categories based on their properties. Therefore, in order to thwart phishing attempts, it is imperative to utilize a distinct characteristic of web pages. We have presented a model that utilizes machine learning techniques, specifically Naïve Bayes, to identify and classify phishing web pages. Various machine learning algorithms, including Naïve Bayes (NB), ANN, Random Forest, Adaboost, and XGBoost, were compared for results assessment. It was found that the identified algorithm achieved a high accuracy score.

Keywords: *Phishing websites, Machine Learning, Content based features, URLs, Attacks*

1. Introduction

Throughout the years, several instances of phishing assaults have occurred, resulting in significant financial losses for many individuals who have fallen prey to these attacks. A phishing assault involves the sending of deceptive emails to users, falsely claiming to be from a reputable institution. These emails request users to provide personal information such as their name, telephone number, bank account details, and critical passwords. These emails prompt the consumer to visit a website where they input their personal information. These websites, sometimes referred to as phishing websites, are designed to illicitly obtain user information and facilitate illegal transactions, hence inflicting harm to the user. Phishing websites and their emails are regularly sent to a large number of people, making them a significant ongoing worry for cybersecurity.

Phishing is a type of cyber assault that lures people into visiting fraudulent websites and divulging sensitive personal information, such as their username and password. Phishing web pages are created by deceitful individuals with the intention of replicating an original web page. These fraudulent web pages closely resemble the authentic ones. Technical manipulation and psychological manipulation are closely intertwined in order to initiate a phishing attack. One crucial aspect of online security is safeguarding users against phishing attacks and fraudulent websites. Artificial intelligence techniques can be employed to create deceptive web sites. Consequently, both experienced and inexperienced internet users are susceptible to deception. Phishing attacks can be initiated by attackers who send emails that appear to be from reputable public or private organizations to deceive users.

Attackers deceive users into updating or verifying their information by enticing them to click on a hyperlink embedded within the email. Attackers can employ alternative techniques, such as file sharing, blogs, and forums, to carry out phishing attempts. Various strategies can be employed to combat phishing, such as implementing legislative measures, providing education, and utilizing technical solutions. In today's world, information and communication tools are utilized in a highly concentrated manner. Multiple solution approaches have been created for different types of problems to serve this objective.

Phishing is a deceptive tactic employed by cybercriminals with the aim of abusing the personal information of unsuspecting individuals. A phishing website is a fraudulent website that mimics the appearance of a legitimate website but redirects users to a different destination. The unsuspecting users submit their data under the assumption that these websites originate from reputable financial organizations. Multiple antiphishing strategies are always emerging, while phishers consistently develop new methods to bypass all existing antiphishing measures. Therefore, there is a requirement for an effective technique to predict phishing websites. Machine Learning (ML) techniques can be applied to develop applications for information security. An optimization, classification, prediction, and decision support system can offer significant advantages to the individual in charge of information security. Various attacks target the Information and Communication tools used to establish computer networks, each serving distinct aims. These assaults can be identified and appropriate measures should be implemented. The research of artificial intelligence appears to be accelerating as computer technology advances. The utilization of artificial intelligence techniques and research on information security are steadily growing. Intelligent systems offer significant advantages for information security experts when making decisions. Machine learning techniques can be employed for classification purposes across several domains. Classification is a method used to assess if a given piece of data belongs to one of the predefined classes in a dataset, based on specific rules.

The suggested approach aims to detect phishing emails by utilizing the identification and utilization of structural aspects of the emails. This paper employs artificial neural networks (ANN), Naïve Bayes, and classification techniques to accurately categorize phishing emails.

This technique relies on the structural qualities of emails and aims to expand the content properties to minimize errors in analysis. A suggested system has been developed that utilizes intelligent phishing website detection and prevention approaches, namely employing the link guard algorithm to safeguard hyperlinks. The categorization algorithm is insufficient in size and employs a singular methodology to identify suspicious emails, resulting in low levels of efficiency and scalability.

Firstly, the website features are gathered, extracted, and stored in a CSV file. Subsequently, we will utilize Machine Learning Algorithms to categorize the Phishing Websites. We will enhance the model by incorporating Random Forest, Adaboost, and XGBoost algorithms. We will then compare the performance of these algorithms with existing ones such as Naïve Bayes and Neural Network to demonstrate improved accuracy. A phishing website is a frequently used form of social engineering that imitates trustworthy URLs and webpages. The aim of this research is to utilize the dataset to train machine learning models and deep neural networks in order to

accurately identify the presence of phishing websites. A dataset is created by collecting both phishing and benign URLs of websites, from which URL and website content-based attributes are extracted. The performance level of each model is assessed and contrasted.

2. Related Work

Phishing sites, which expect to steal the victims' confidential data by redirecting them to a fake website page that looks exactly like the real thing, are another type of internet criminal act that is especially concerning in a variety of areas such as e-managing an account and retailing. Phishing site identification is a genuinely unexpected and elemental challenge with various unstable components and criteria. Because of the previous and additionally ambiguities in arranging sites as a result of the intelligent procedures programmers are utilising, some keen proactive strategies and powerful tools can be utilised, for example, fuzzy, neural system, and data mining methods can be a successful mechanism in distinguishing phishing sites [1].

The proposed approach is based on real-time automated phishing detection and machine learning. The majority of phishing URLs include links between the parts of the URL, indicating an inter-relationship, and by employing it, the properties of phishing URLs are retrieved. The collected characteristics are then used to detect phishing websites in real time using machine learning classification.

For a long time, Internet users have been subjected to phishing assaults, according to the authors of [2]. Attackers utilise maliciously built phishing websites to fool consumers and steal personal information such as bank account numbers, website usernames, and passwords. Many phishing detection systems have been developed in recent years, most of them rely on whitelists or blacklists, website content, or side channel-based strategies. However, due to the constant advancement of phishing technologies, current solutions are having problems obtaining efficient detection. As a result, in this research, we offer HinPhish, an effective phishing website detection technique. HinPhish collects diverse link associations from websites and builds a heterogeneous information network using domains and resource items. HinPhish uses a customised method to determine the phish-score of the target site on the webpage by leveraging the peculiarities of different link kinds. Furthermore, HinPhish not only enhances detection accuracy, but it may significantly increase the phishing cost for attackers. Extensive experimental data show that HinPhish has an accuracy of 0.9856 and an F1-score of 0.9858.

Authors [2] contributed to the study.

• An effective representative model—HinPhish models the domains and resource links on a webpage as an HIN to explain the relationship of the links, successfully preserving the deeper semantics between the connections.

• An effective phishing detection approach—HinPhish uses the modified algorithm to calculate the phish-score of the target domain in a webpage, then uses HIN and a machine learning algorithm to detect phishing attacks, significantly increasing the evasion cost.

• A sufficient experiment—in order to show the usefulness and superiority of the suggested strategy, we conducted a sufficient number of comparison tests using various cutting-edge methodologies. HinPhish performed admirably, with an accuracy of 0.9856 and an F1-score of 0.9858.

The Internet has become a vital part of our lives, but it has also offered chances for malevolent acts such as phishing to be carried out anonymously. Phishers attempt to trick their victims by using social engineering or constructing bogus websites in order to acquire personal and organisational information such as account IDs, usernames, and passwords. Although several methods for detecting phishing websites have been presented, phishers have updated their ways to avoid detection. Machine Learning is one of the most effective approaches for identifying these dangerous behaviours. This is because most Phishing assaults have some traits that machine learning systems can detect. We compared the outcomes of various machine learning approaches for predicting phishing websites in this article. [3].

According to the FBI's IC3 data for 2018, internet-based theft, fraud, and exploitation are still prevalent and were responsible for a stunning $2.7 billion in financial losses in 2018. The IC3 received 20,373 complaints in that year about business email breach (BEC) and email account compromise (EAC), with damages totaling more than $1.2 billion. According to the research, the number of these sophisticated attacks has increased in recent years. According to the Anti-Phishing Working Group (APWG), phishing attempts have increased in recent years. Phishing has caused significant harm to numerous organisations and the global economy. In the fourth quarter of 2019, APWG member OpSec Security discovered that SaaS and webmail sites were the most often targeted targets of phishing attacks. Phishers continue to collect credentials from these targets using BEC and get access to corporate SaaS accounts as a result. Many methods have been developed to detect phishing websites. Each of these measures, for example, network-level security, authentication, client-side tools, user education, server-side filters, and classifiers, is applicable at different phases of the attack flow. Although each sort of phishing assault has certain distinguishing characteristics, the majority of these attacks have significant commonalities and patterns. Since machine learning approaches have shown to be a strong tool for spotting patterns in data, these methods have made it possible to recognise some of the frequent phishing qualities, allowing phishing websites to be identified. We present a comparative and analytical evaluation of various machine learning approaches for detecting phishing websites in this research. We investigated Logistic Regression, Decision Tree, Random Forest, Ada-Boost, Support Vector Machine, KNN, Artificial Neural Networks, Gradient Boosting, and XGBoost as machine learning approaches.

Jain, A.K., and Gupta, B.B., [4] describe a unique technique for detecting phishing attacks by analysing hyperlinks included in the website's HTML source code. A phishing assault is carried out by exploiting the visual similarity between the fake and real web sites. The suggested method classified hyperlink-specific properties into 12 categories and used these categories to train machine learning algorithms. Using phishing and non-phishing website datasets, the author examined the efficacy of the proposed phishing detection strategy on several classification algorithms. [4].

A.K. Jain and B.B. Gupta, [5] Attackers steal sensitive information from internet users such as personal identification number (PIN), credit card data, login, password, and so on. The author suggested a machine learning-based anti-phishing solution based on Uniform Resource Locator (URL) properties in this study. To assess the efficacy of the suggested system, the author used 14 URL attributes to determine if a website was phishing or not. The suggested method is trained with SVM and Nave Bayes classifiers on over 33,000 phishing and authentic URLs. The results of the experiments reveal that the SVM classifier detects phishing websites with a high degree of accuracy. [5].

L. Machado and J. Gadge [6] Phishing sites are phoney websites constructed by phishers with the purpose of obtaining visitors' personal information and using it to commit fraud. The C4.5 decision tree methodology is used in this work to offer an efficient method for detecting phishing websites. For improved outcomes, the method suggested in this study employs different URL properties as well as the C4.5 decision tree methodology. [6].

P. Pujara and M.B. Chaudhari, [7] Phishing scams are the most common type of cybercrime nowadays. This research conducted a thorough literature review and developed a novel method for detecting phishing websites using features extraction and a machine learning algorithm. The author describes many approaches for phishing detection in this study, including the Blacklist method, Heuristic based method, Visual similarity, and Machine learning. The blacklist approach is utilised, in which a list of phishing URLs is recorded in a database, and if the URL is located in the database, it is recognised as a phishing URL and a warning is issued, otherwise it is known as a legal URL. The heuristic-based technique is an expansion of the blacklist that may identify new attacks by using characteristics taken from phishing sites to detect phishing attacks. The visual similarity approach deceives users by stealing images from trustworthy websites. The Machine Learning technique works well with huge datasets [7].

Rathore, S., Sharma, P.K., Loia, V., Jeong, Y.S., and Park, J.H. [8] give a thorough analysis of various security and privacy concerns that target all social networking site users. A Social Network Service (SNS) is a sort of online service that allows people with similar interests, backgrounds, and activities to interact virtually. SNSs have been a popular mode of communication in recent years. Every year, the number of people who utilise social media grows. This study focuses on numerous vulnerabilities that develop as a result of the sharing of multimedia information on a social networking site. This section describes three types of threats. Threats to multimedia content, traditional threats, and social threats [8].

Sahingoz, O.K., Buber, E., Demir, O., and Diri, B., [9] attempt to detect phishing sites using url in order to protect users' sensitive information. Computer users fall victim to phishing for five primary reasons:

• Users lack detailed knowledge of URLs;

• Users do not know which web pages can be trusted;

• Users do not see the entire address of the web page due to redirection or hidden URLs;

• Users do not have much time to consult the URL; and

• Users cannot distinguish phishing web pages from legitimate ones.

The suggested system classified phishing and non-phishing sites using NLP-based characteristics and Word features. Decision Tree, Adaboost, K-star, kNN(n=3), Random Forest, SMO (Squential Minimal Optimisation), and Naive Bayes were used for classification [9].

Phishing websites [11] are frequently used for online social engineering attacks and scams. Existing phishing defense mechanisms are insufficient in detecting new phishing attacks. A paper proposes an enhancement to phishing detection techniques using machine learning, specifically a learning-based aggregation analysis mechanism to determine page layout similarity. The experimental results demonstrate the accuracy and effectiveness of their approach in detecting phishing pages.
Limitations of the proposed approach include limited dataset size and quality, generalization to new and evolving phishing attack techniques, false positives, real-time detection, and potential vulnerability to adversarial attacks. The proposed approach, HinPhish, extracts link relationships from webpages to construct a heterogeneous information network using domains and resource objects. The algorithm calculates the phish-score of the target domain on the webpage, leveraging the characteristics of different link types.
Extensive experimental results show HinPhish achieves an accuracy of 0.9856 and F1-score of 0.9858. However, it may require significant computational resources to process and analyze the large amount of link relationships and construct the heterogeneous information network. The scalability of the approach to a large number of webpages and domains may need further investigation. HinPhish's adaptability to emerging sophisticated phishing techniques and dynamic changes in website structures and link behaviors may pose challenges.
Despite its high accuracy and F1-score, HinPhish's potential for false positives and false negatives in real-world scenarios should be thoroughly evaluated to assess its reliability. The ease of integration into existing cybersecurity systems and its usability for non-technical users may present usability and adoption challenges.
The paper [12] presents a machine learning-based phishing detection technique that can distinguish 95.66% of phishing websites from legitimate ones. When integrated with the Support Vector Machine classifier, the technique achieved high accuracy, despite using only 22.5% of its innovative functionality. The technique's performance was validated using standard phishing datasets from the University of California Irvine archive, yielding optimistic results. The technique is positioned as the preferred approach for phishing detection based on machine learning. However, the technique's capabilities may be limited to specific datasets used for validation. Further testing on diverse datasets or real-world scenarios could provide a clearer understanding of its practical applicability. The technique's scalability, adaptability, and false positive and false negative rates need further investigation. The technique's ease of integration and deployment into existing cybersecurity systems and its user-friendliness for security professionals should be considered for practical adoption. The paper

concludes that while the machine learning-based phishing detection technique offers significant promise, its limitations underscore the need for ongoing research and development to address real-world cybersecurity challenges comprehensively.

3. Proposed Methodology

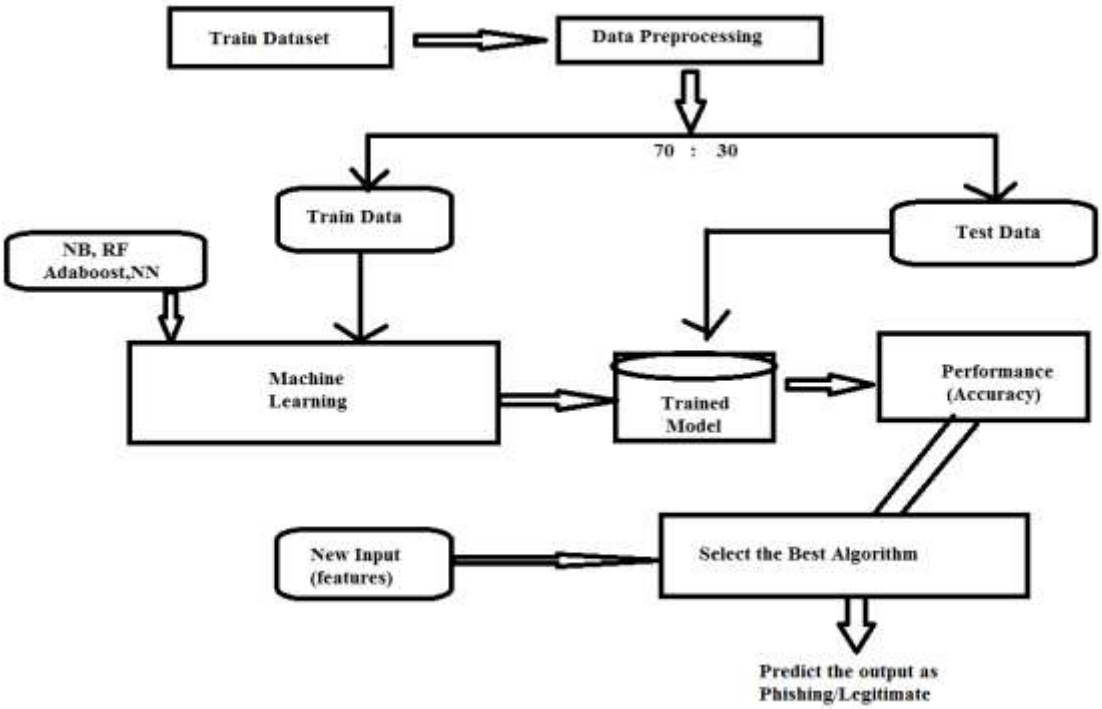

Figure 1: System Architecture

The Proposed model involves in developing a machine learning model to predict phishing websites by analyzing URLs and website content. Data Collected from the open Source repository- UCI, then extract the features of the dataset with URL based and Content Based, then Apply Different ML algorithms and compute the performance like accuracy and select the best model on the basis of accuracy for predicting the output for new unknown features with selected best model.

Following Describes the Modules of the System Architecture.

3.1 Feature Extraction on Dataset

3.1.1 URL-based Features:

1. **Length of URL**: Phishing URLs are often longer.
2. **Presence of IP address in URL**: Legitimate URLs usually do not contain IP addresses.
3. **Number of special characters**: Such as '@', '-', and '_'.
4. **Use of HTTPS**: Phishing sites might not use HTTPS.
5. **Domain age**: Older domains are usually more trustworthy.

3.1.2 Content-based Features:

1. **HTML Content Analysis**: Presence of certain keywords like "login", "verify", "update".
2. **Form handling URLs**: Checking where forms submit data.
3. **External links count**: Number of links leading to external sites.
4. **Script analysis**: Presence of obfuscated JavaScript code.

3.2 Data Preprocessing

1. **Cleaning**: Remove any irrelevant data.
2. **Labeling**: Mark the URLs as phishing or benign.
3. **Normalization**: Normalize the feature values for consistent scaling.

3.3 Machine Learning Models

3.3.1. Naive Bayes

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem. It is particularly well-suited for classification tasks. The "naive" aspect comes from the assumption that the features are independent of each other, which simplifies the computation of the probability.

Steps for Using Naive Bayes:

- **Training Phase:**

  1. Calculate the prior probabilities for each class.
  2. Calculate the likelihood of each feature given the class.

- **Prediction Phase:**

  1. For a new instance, calculate the posterior probability for each class using Bayes' Theorem.
  2. Assign the class with the highest posterior probability.

3.3.2. Random Forests

Random Forest is an ensemble learning method primarily used for classification and regression tasks. It builds multiple decision trees during training and merges their results to improve accuracy and control overfitting.

Steps for Using Random Forest:

- **Training Phase:**

  1. Create multiple bootstrap samples from the original dataset.
  2. For each bootstrap sample, grow a decision tree:

     - Randomly select a subset of features at each node.
     - Split the node based on the best feature from the subset.
     - Repeat the process until the maximum depth is reached or the node is pure.
     - Aggregate the predictions of all trees.

- **Prediction Phase:**

1. For classification, each tree in the forest casts a vote for the class label. The final prediction is the class with the majority vote.
2. For regression, the final prediction is the average of the predictions from all trees.

3.3.3. Neural Network

Neural networks are a class of machine learning algorithms inspired by the structure and function of the human brain. They are particularly well-suited for handling complex patterns and high-dimensional data, making them powerful tools for a variety of tasks, including image and speech recognition, natural language processing, and more.

Steps for Using Neural network:

Neuron:

1. The basic unit of a neural network. Each neuron receives input, processes it using a weighted sum and an activation function, and produces an output.
2. **Activation Function:** Determines if a neuron should be activated or not. Common functions include Sigmoid, Tanh, and ReLU (Rectified Linear Unit).

Layer:

1. Neural networks are composed of layers of neurons. There are three main types of layers:

    a) **Input Layer:** The first layer that receives the input data.
    b) **Hidden Layers:** Intermediate layers where computations are performed. A neural network can have multiple hidden layers, allowing it to learn complex representations.
    c) **Output Layer:** The final layer that produces the output of the network.

Weights and Biases:

1. **Weights:** Parameters that adjust the input's influence on the neuron's output. Each connection between neurons has an associated weight.
2. **Biases:** Additional parameters that shift the activation function, allowing the network to better fit the data.

Forward Propagation:

1. The process of passing input data through the network to obtain an output. Each layer's neurons compute their output based on the inputs, weights, biases, and activation functions.

Loss Function:

1. Measures the difference between the network's output and the actual target values. Common loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss for classification tasks.

Backpropagation:

- A training algorithm used to minimize the loss function by updating the network's weights and biases. It involves two steps:

    a) **Forward Pass:** Calculate the loss for a given input.
    b) **Backward Pass:** Compute the gradient of the loss function with respect to each weight and bias, then update them using a gradient descent optimization algorithm.

3.3. 4. Adaboost

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm designed to improve the performance of weak classifiers. It combines multiple weak learners to create a strong classifier with high accuracy. The key idea is to focus on the instances that are harder to classify by adjusting the weights of the training samples.

Steps for Using AdaBoost:

1. **Initialize Weights:**
    - Start with equal weights for all training samples.
2. **Train Weak Learners:**
    - For each iteration, train a weak learner on the weighted training data.
    - Evaluate the weak learner's performance and calculate the error rate.
3. **Update Weights:**
    - Increase the weights of misclassified samples and decrease the weights of correctly classified samples.
    - Calculate the weight of the weak learner based on its accuracy.
4. **Combine Weak Learners:**
    - The final strong classifier is a weighted combination of all weak learners. The weight of each weak learner is determined by its accuracy.
5. **Make Predictions:**
    - For a new instance, each weak learner makes a prediction. The final prediction is the weighted majority vote of all weak learners.

3.3.5. Xgboost

XGBoost (eXtreme Gradient Boost) is a powerful and efficient implementation of the gradient boosting framework, widely used for supervised learning tasks. It has gained popularity due to its performance, speed, and scalability, often leading to top results in machine learning competitions.

Steps for Using XGBoost:

1. **Initialization:**
    - Start with an initial prediction, usually the mean of the target variable for regression or the mode for classification.
2. **Compute Residuals:**
    - Calculate the residuals (errors) between the actual and predicted values.
3. **Train Base Learners:**
    - Fit a new decision tree to the residuals. This tree aims to predict the residuals of the previous model.
4. **Update Predictions:**
    - Update the predictions by adding the predictions from the new tree, scaled by a learning rate.

5. **Repeat:**
    - Repeat the process for a specified number of iterations or until the residuals are minimized.
6. **Combine Models:**
    - The final model is a weighted sum of the predictions from all the individual trees.

3.4 Performance

Comparison of Algorithms and shows the best algorithm based on performance metrics such as accuracy, precision, recall and f1-measure.

With the best algorithm we can predict the output for a new website feature as phishing or legitimate.

Accuracy,

$$Accuarcy\ A_c = \frac{T_P+T_n}{T_p+T_n+F_{P+}F_n} \quad (1)$$

For classification purposes, System efficiency was evaluated using commonly utilized metrics such as precision, recall, F1-measure, and accuracy (A), as outlined mathematically in Equations (2), (3), and (4).

Precision,

$$Precision\ P_r = \frac{T_P}{T_p+F_P} \quad (2)$$

Recall,

$$Recall\ \ R_C = \frac{T_P}{T_p+F_n} \quad (3)$$

F1-measure,

$$F1 - Measure(F1) = \frac{2* P_r * R_C}{P_r + R_C} \quad (4)$$

where T_P, T_n, F_P, and $F_n$are denoted as true positives, true negatives, false positives, and false negatives, respectively.

4. Results and Discussion

Based on the ML Models, the test data will classify if the input is phishing or not. Algorithm comparison that identifies the best algorithm based on performance criteria such as accuracy, precision, and recall.

We can anticipate if a new website feature is phishing or authentic using the best algorithm.

The accuracy comparison using graphs is shown below.:

- Accuracy of Random Forest Classifier: 96.86463671992765
- Accuracy of Adaboost Classifier: 91.34760325595418
- Accuracy of Neural Network Classifier: 96.86463671992765

The test data will classify whether the input is phishing or not based on the ML Models.

Domain_registeration_length	-0.225789
Shortining_Service	-0.067966
Abnormal_URL	-0.060488
HTTPS_token	-0.039854
double_slash_redirecting	-0.038608
Redirect	-0.020113
Iframe	-0.003394
Favicon	-0.000280
popUpWidnow	0.000086
index	0.000978
RightClick	0.012653
Submitting_to_email	0.018249
Links_pointing_to_page	0.032574
port	0.036419
on_mouseover	0.041838
having_At_Symbol	0.051330
URLURL_Length	0.057430
DNSRecord	0.075718
Statistical_report	0.079857
having_IPhaving_IP_Address	0.094160
Page_Rank	0.104645
age_of_domain	0.121496
Google_Index	0.128950
SFH	0.221419
Links_in_tags	0.248229
Request_URL	0.253372
having_Sub_Domain	0.298323
web_traffic	0.346103
Prefix_Suffix	0.348606
URL_of_Anchor	0.692935
SSLfinal_State	0.714741
Result	1.000000

Figure 2: Features Considered

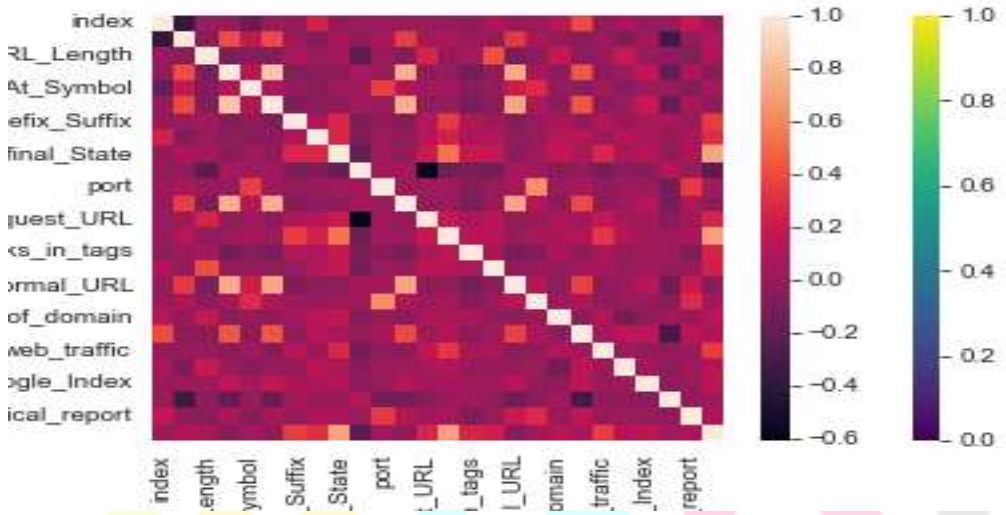

Figure 3: Confusion Matrix

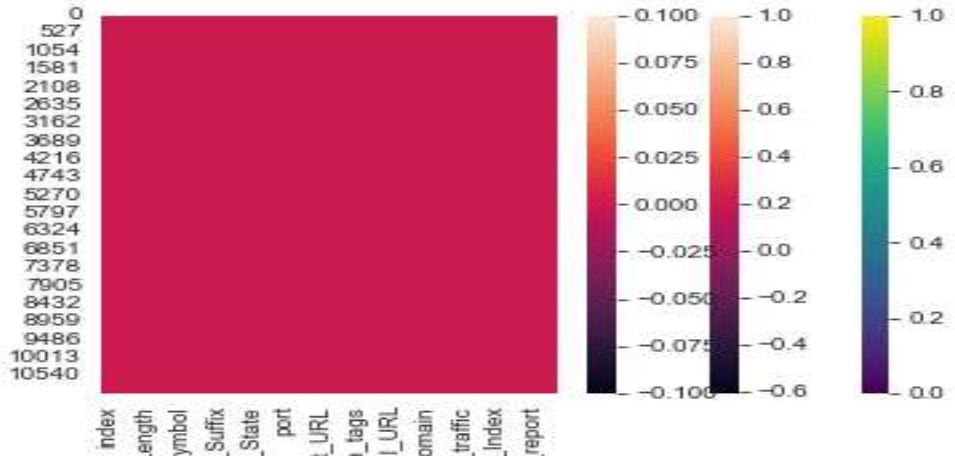

Figure 4: Heatmap After Clean

Figure 5: Features Considered from the dataset

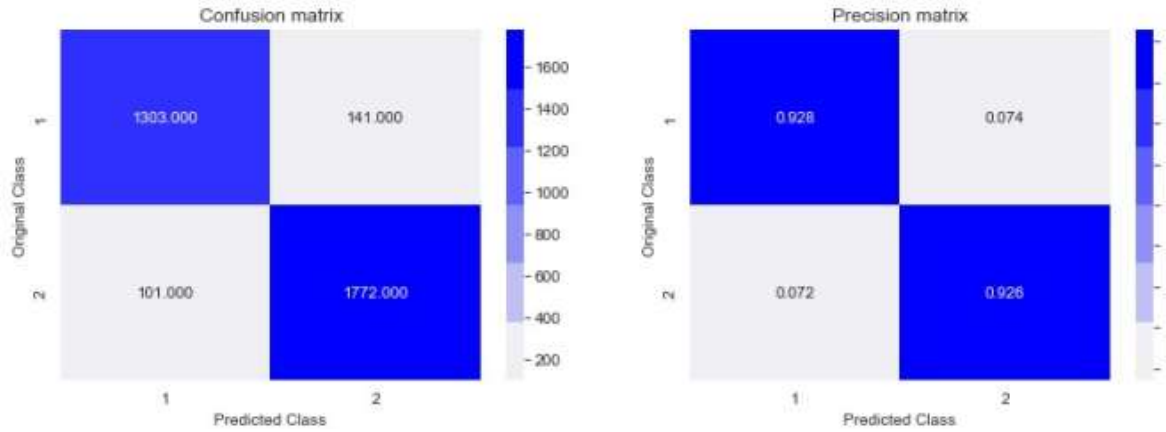

Figure 6: Confusion Matrix and Precision Matrix of the Algorithm

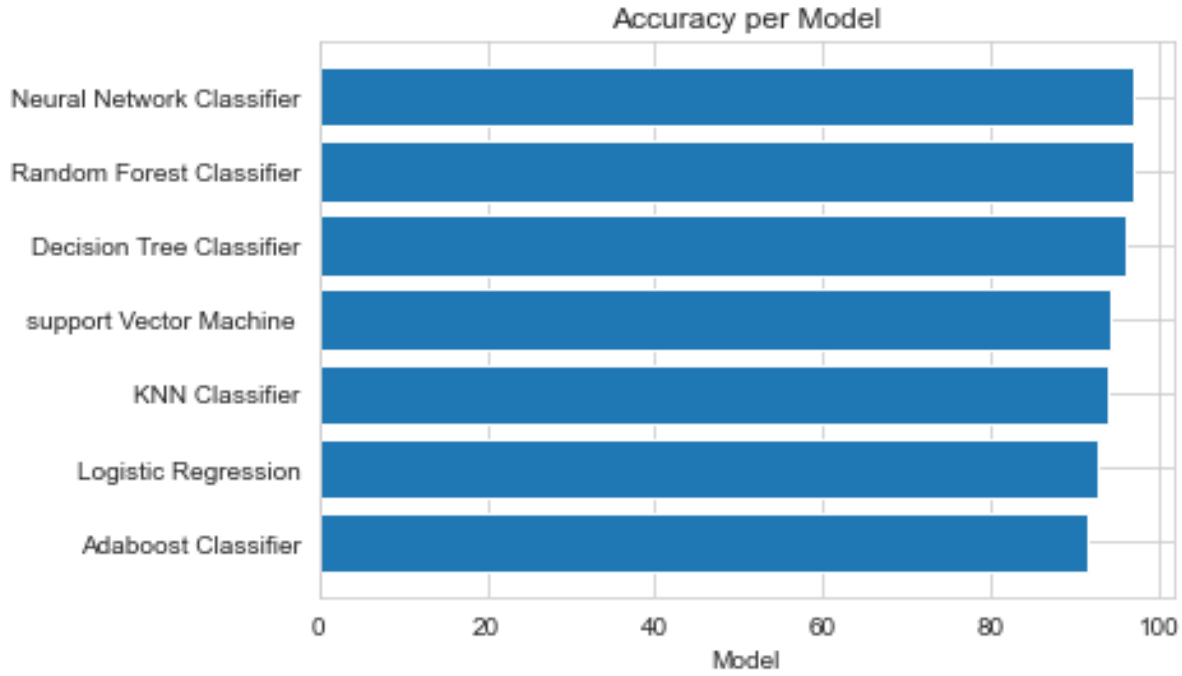

Figure 7: Comparison of Accuracy graph for the ML Algorithms

```
Logistic Regression Accuracy: 92.6439553813687
K-Nearest Neighbour Accuracy: 93.91015978293639
Decision Tree Classifier Accuracy: 95.93005728067531
Random Forest Classifier Accuracy: 96.86463671992765
support Vector Machine Accuracy: 94.36237564063913
Adaboost Classifier Accuracy: 91.34760325595418
Neural Network Classifier Accuracy: 96.86463671992765
```

Figure 8: Comparison of Accuracy Values.

5. Conclusion

Websites can be used to create a wide range of systems, including those for data entry and information processing. The provided data can be analyzed; the analyzed data can be acquired as output. In the present day, websites are utilized across various domains including scientific, technical, business, education, and economy. Due to its extensive utilization, it can also serve as a tool for hackers to carry out nefarious activities. One of the malevolent intentions manifests itself as a phishing attempt. Phishing assaults might replicate a website or webpage through the utilization of diverse techniques. These fraudulent websites or web pages can be used to acquire sensitive information such as users' credit card details and personal identities. The application's objective is to classify and identify one of the sorts of cyber dangers known as phishing assaults. Various machine learning

models, including Random Forest, Adaboost, Xgboost, Neural Network, and Naive Bayes, were employed for this task. For this investigation, we utilized a dataset obtained from the UCI website. The dataset consists of 30 input attributes and 1 output attribute. The input attributes can have three distinct values: 1, 0, and -1. The output attribute might have two distinct values: 1 and -1.

References

[1] Miss Sneha Mande1, Prof D.S.Thosar , "Detection of Phishing Web Sites Based On Extreme Machine Learning", 2017

[2] Bingyang Guo, Yunyi Zhang, Chengxi Xu ,Fan Shi , Yuwei Li and Min Zhang, "HinPhish: An Effective Phishing Detection Approach Based on Heterogeneous Information Networks",2021

[3] Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, "Phishing Detection Using Machine Learning Techniques" 2020

[4] Jain, A.K. and Gupta, B.B., 2019. A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, 10(5), pp.2015-2028.

[5] Jain, A.K. and Gupta, B.B., 2018. PHISH-SAFE: URL features- based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.

[6] Machado, L. and Gadge, J., 2017, August. Phishing Sites Detection Based on C4. 5 Decision Tree Algorithm. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

[7] Pujara, P. and Chaudhari, M.B., 2018. Phishing Website Detection using Machine Learning: A Review.

[8] Rathore, S., Sharma, P.K., Loia, V., Jeong, Y.S. and Park, J.H., 2017. Social network security: Issues, challenges, threats, and solutions. Information sciences, 421, pp.43-69.

[9] Sahingoz, O.K., Buber, E., Demir, O. and Diri, B., 2019. Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, pp.345-357

[10] J. Mao et al., "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," Procedia computer science, Jan. 01, 2018.

https://www.sciencedirect.com/science/article/pii/S187705091830276X

[11] B. Guo, Y. Zhang, C. Xu, F. Shi, Y. Li, and M. Zhang, "HinPhish: An Effective Phishing Detection Approach Based on Heterogeneous Information Networks," *Applied sciences*, Oct. 18, 2021. https://www.mdpi.com/2076-3417/11/20/9733

[12] J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," *2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, Riyadh, Saudi Arabia, 2020, pp. 43-46,

doi: 10.1109/SMART-TECH49988.2020.00026.

[13] R. Yang, K. Zheng, B. Wu, C. Wu, and X. Wang, "Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning," *Sensors*, Dec. 10, 2021. https://pubmed.ncbi.nlm.nih.gov/34960375/

[14] Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine Learning," *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.

https://ieeexplore.ieee.org/document/9824801

[15] Sadia Afroz; Rachel Greenstadt "PhishZoo: Detecting Phishing Websites by Looking at Them," *IEEE Conference Publication | IEEE Xplore*, Sep. 01, 2011. **https://ieeexplore.ieee.org/document/6061361**