DETECTION AND CLASSIFICATION OF PHISHING WEBSITES USING MACHINE LEARNING

Dr. Pampapathi B M[1*], Shruthi S M[2]
[*1]Associate Professor, Department of Computer Science & Engineering, RYM Engineering College,Ballari,VTU Belagavi ,Karnataka, India,
Department of Computer Science & Engineering, RYM Engineering College,Ballari,VTU Belagavi ,Karnataka, India.

***Abstract*: Phishing attacks have recently become one of the most prominent attacks against internet users, governments, and service provider organizations. In a phishing attack, attackers use phishing emails or fake websites to collect sensitive customer information (such as user account login details, credit/debit card numbers, etc.). Phishing websites is a common online manipulation attack, including many website scams. In such attacks, attackers create web pages that mimic the behavior of legitimate websites and send URLs to targeted victims via spam, text messages, or social networks. For a comprehensive understanding of phishing attacks, this article provides a literature review of artificial intelligence techniques: machine learning, deep learning, hybrid learning and scenario-based techniques to detect phishing attacks. This article also compares different studies on phishing attack detection for each AI technique and explores the features and shortcomings of these methods. In addition, this article provides a comprehensive collection of current challenges in phishing attacks and future research directions in this area.**

1. Introduction

Phishing is an unethical practice that uses both social engineering and technical methods to obtain sensitive information and credentials such as banking credentials from users. Some social engineering techniques employ spam emails that masquerade as real companies or organizations and are specifically meant to direct users to bogus websites that manipulate receivers into falling into traps that steal financial credentials such as user-ids and passwords. Technical intrigue methods implant malicious software onto systems to directly collect data, frequently using systems to intercept users' online account usernames and passwords.

In general, two ways are commonly utilized in detecting phishing websites. The first method is often based on a blacklist, in which the given URL is compared to the URLs in the blacklist. Another aspect of this strategy is that the blacklist cannot always recognize all phishing sites, thus a new bogus website is built. The alternative or second strategy is known as heuristic based methods, in which only a few features are collected from the sites to determine whether it is phishing or authentic.

A heuristic-based technique, as opposed to a blacklist approach, can detect newly constructed phishing sites. The heuristic-based approach's accuracy is dependent on identifying a set of chosen features that may aid in classifying the type of the website. Data mining techniques are some of the research domains that can use knowledge aspects that promise the nature, reliability, and completeness, as well as shorten the duration of knowledge achievement. In data mining, there are two types of rules-induction approaches: associative techniques and classification-rule techniques. This work is concerned with the application of classification rules. The classification task's goal is to assign each test data point to one of the test dataset's specified classes. Various studies on phishing website identification based on website attributes have been undertaken, but these studies were unable to detect the exact or precise phishing website.

Compared to most previous approaches, researchers focus on identifying harmful substances URLs from a large pool of URLs. The objectives of the study are as follows:

- Develop a new approach to detect malicious URLs and alert users.
- Apply ML techniques in the proposed app.

2. Problem Statement.

An The Internet has ruled the world by bringing half of the world's population into the cyber realm at an exponential rate. Because of the anonymity provided by the internet and the booming of internet transactions, hackers attempt to trap end-users through various forms such as phishing, SQL injection, malware, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Among all these attacks, phishing appears to be the most deceptive. The primary goal of this work is to classify a phishing website using multiple machine learning approaches to achieve maximum accuracy and a succinct model.

A. Proposed System

Our proposed system begins with data extraction, where we will collect the data. Based on the data collected, we will train the model using machine learning algorithms, and then we will design the front end. After designing the front end, we will go through flask integration and then deploy the model on a cloud platform to make it available to end users.

Advantages of proposed system:
- High detection rate
- Fast computation.

B. Objective

In recent years, phishing has been a big source of concern for security professionals because it is quite easy to create a phishing website that appears to be identical to a legitimate website. Although professionals can recognize bogus websites, not all users can, and as a result, some people fall prey to phishing assaults. The attacker's primary purpose is to obtain bank account information. Every year, businesses in the United States lose $2 billion because of their customers falling prey to phishing. According to the third Microsoft Computing Safer Index Report, released in February 2014, the annual global impact of phishing might be as high as $5 billion. Phishing attacks are growing increasingly successful because of user ignorance.

3. Literature Survey

Rishikesh Mahajan, Irfan Siddavatam , International Journal Computer Applications , October 2018. "Phishing Website Detection using Machine Learning Algorithms". This paper explains how phishing is the easiest way to get one. Sensitive information from unsuspecting consumers. The purpose of data capture is to obtain sensitive information such as usernames, passwords, and bank accounts. People working in cyber security are increasingly looking for reliable and consistent detection strategies for phishing sites. The purpose of this work is using machine learning to identify phishing URLs by extracting and evaluating different URLs aspects of genuine and phishing URLs. Data traps are detected using decision tree, random forest, and Support Vector machine algorithms. The purpose of the survey is to identify phishing URLs and narrow down the best machine learning method by analyzing each algorithm. accuracy rate, false positive rate, and false negative rate. This paper attempts to improve detection using machine learning techniques. Mechanism of phishing sites using a random forest approach, we did it to obtain a detection accuracy of 97.14 percent with the lowest false positive rate. Furthermore, the results show that the classifiers perform better when there is more data used as training data.

Mangala Kini, Deekshitha, International Journal of Engineering Research & Technology (IJERT) Vol. 1.2, No. 6, 2021. "A Review Paper on Detection of Phishing Websites using Machine Learning". Phishing is an attempt to steal personal information such as usernames, passwords and credit card numbers of individuals or organizations as a reliable source in electronic contact. Phishing attacks cause a serious threat to user privacy and security. This provides an overview of various phishing attempts and information security approaching it also covers the basics of Extreme Learning Machine (ELM). Classification of 30 features in the UC Irvine machine learning repository database, including information about phishing sites. This report included three main parts of the study: the theory of phishing crimes, a review and study of anti- phishing techniques proposed by various studies research gaps While you can never eliminate phishing, it does happen before finding a solution, it is important to understand the crime. we are gone through the many features of phishing attacks and in many ways to detect phishing sites. The next step is to research how to build phishing detection in a system that focuses on phishing sites because it is the most popular method attack, we can use artificial neural networks or random forest classifiers Instead of Naive Bayes method for more accurate results. This detection tool help users avoid phishing attacks in the future.

Ma et al. [3,4] A manuscript on Website phishing Identification.
Compared different group-based learning algorithms used to classify phishing sites and found that a combination of host-based and lexical features provides the highest classification accuracy. In addition, they also compare the performance of cluster-based algorithms with network-based algorithms using complete functions and find that network-based algorithms, especially confidence-weighted (CW), outperform set-based algorithms. Attributes include presence of red flag keywords on the website, attributes based on Google PageRank and Google Website Quality Guidelines. Can't directly compare without access to the same sites and properties.

Kunju et al. (Kunju et al., 2019) used a survey method to detect phishing attacks. The study offers several solutions and methods to detect phishing attacks. According to the study, many of the proposed solutions proved to be insufficient to find solutions to phishing attacks. The literature of this work contains only 14 studies from 2007-2019. The study only deals with machine learning techniques to detect phishing sites. conducted a systematic review to analyze different approaches by other researchers to detect phishing attacks using deep learning algorithms. In conclusion, there is still a significant gap in the field of deep learning algorithms to detect phishing attacks. The literature for this work includes only 19 studies published between 2014 and 2019. This article only covers research papers that deal with the main topics of phishing and deep learning.

Atulya and Praveen (Atulya and Praveen, 2020) discussed different phishing attacks, latest phishing tactics, phishing and anti-phishing strategies. In addition, the article aims to raise awareness of phishing attacks and the strategies used to detect phishing. According to this study, the best way to prevent phishing attacks is to educate users about different types of phishing attacks. Users can choose the best security software tools or applications like anti-phishing browser extensions to detect phishing attacks. The literature of this work is based on nine research topics. The study does not include deep learning techniques to identify phishing sites. reports on research into artificial intelligence-based phishing detection techniques. The authors used statistical phishing reports to examine flaws and trends in phishing attempts. The article classifies anti-phishing assessments into four categories: machine learning, hybrid learning, scenario-based and deep learning. Research shows that machine learning methods provide the best results compared to other approaches. The work is based on the literature published in the last ten years and analyzes only 21 research objects.

Arshad et al. (Arshad et al., 2021) presented various phishing and anti-phishing techniques in their study. According to SLR, the most common phishing techniques used are phone phishing, email spoofing, phishing, and email manipulation. According to this study, the highest accuracy was achieved using machine learning methods. The study is limited by the fact that it is based on only 20 studies. proposed a review paper to select features that can be used in URL-based phishing detection systems. The purpose of this study is to create a general research resource for researchers working on website classification or network security. A limitation of this study is that the article only considered five literature studies. presented a framework to detect and prevent various phishing attacks. According to this study, algorithms based on machine learning are effective in identifying true positives. The limitations of this study are as follows: only 11 studies were considered in the literature of this work, and the study does not include deep learning techniques to reduce phishing sites.

Catal et al. (Catal et al., 2022) worked on a systematic literature review that answered nine research questions. The main goal of the research is to identify, evaluate and synthesize the results of deep learning approaches for phishing detection. According to this study, supervised ML algorithms were used in 42 out of 43 studies. The most used algorithm was DNN, and the best performance was given by DNN and Hybrid DL algorithms. The article only deals with research related to Deep Learning for Phishing Detection. Recently worked on conventional and automated phishing detection techniques. Traditional anti-phishing methods include raising awareness, educating users, organizing periodic trainings or workshops, and using a legal perspective. Computer-based or automated anti-phishing approaches refer to list-based and machine learning-based techniques. More importantly, the paper compares the similarities, positives and negatives of these approaches from a user and performance perspective. According to this study, machine learning and rule extraction are suitable for combating phishing attacks. The limitations of this work are as follows: the review is based on 67 research topics and the study does not include deep learning techniques to detect phishing sites.

4. Methodology.

This paper uses various Machine learning algorithms and techniques to detect whether the website is phishing website or its legitimate. Initially the data is collected, and data is trained using various machine learning algorithms and train model and then comes the designing of the front end in data Collection the user will select whether to trust the website, a trusted source is required to train the model. Data is gathered from a website called Phish Tank. Cisco Talos Intelligence Group (Talos) is trusted and maintained. As a result, the data has been retrieved from https://www.phishtank.com/ as well as cleaning and formatting the data accordingly. The project's features have been extracted.

A. Data Flow

The technique consists of website data collected through host-specific, page-specific and feature extraction. The first step is to collect phishing and benign websites. In this approach, attributes are extracted based on admiration to form a database of values. This database consists of data mined using various machine learning techniques. When evaluating algorithms, a selection classifier is selected, which is implemented in Python.

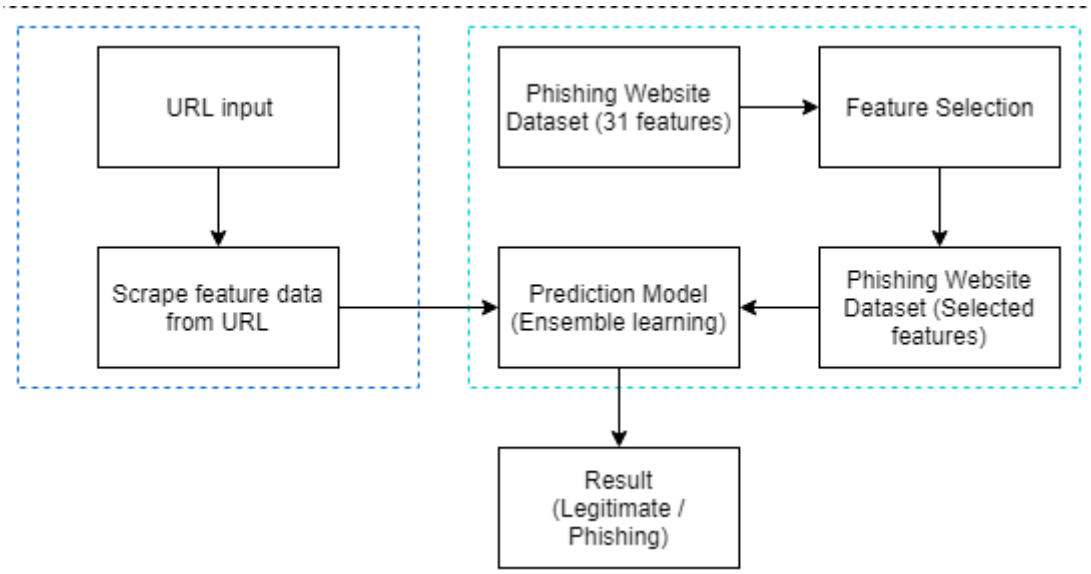


Fig 1. Data Flow Diagram

B. Feature Extraction

Feature separation is based on sixteen different features. Information is segregated exactly URLs for phishing sites are marked with 1 and benign sites URLs are marked with the number 0. [1] Presence of IP address: If a URL contains IP addresses, it will be marked as 1 or phishing website According to the collected information, most of the phishing sites contain IP addresses instead of the domain name.

Presence of @: If a URL contains the @ symbol, it will be marked as 1, a phishing site. Because if the URL contains the @ symbol, the browser will take the address after that symbol, and it is redirected to the attacker's desired website.

More dots („."): If the URL contains more dots (."), the URL is flagged for example 1. A quality website contains up to 3 points in terms of information.

URL length: If the URL is longer than 54 characters, The URL is labelled 1. Because attackers use large URLs to hide the domain name the actual website.

URL depth: URL depth can be thought of as subpages that the website contains It can be identified by the number "/" in the URL.

URL Redirection: The URL must contain a single "//" after "HTTPS:" if it contains more if one, it means it redirects to another site that may not be related to the site where we go. So, according to the dataset, most URLs contain "//" more than one is engaged in phishing. So a URL containing more than "//" is marked with 1

Prefix or Suffix: Domain names usually do not contain a "-" prefix or suffix. If there is a "-" in the domain name, then it is marked as 1, and if not, as 0.

Web Traffic: The Alexa ranking is an analyzed ranking of web traffic. Sites that are ranked Alexa cannot

have phishing sites. So, if the Alexa rank of a URL is less than 100,000, it is marked as 0 unless 1 is marked. Time to register a domain: Legitimate websites often mind their own business domain registration and stay informed. So, if the domain registration is less than 1 in, it is indicated by 1 and if not by 0.

Domain Age: Domain age less than 6 months is classified as 1 and if more than 6 months, it is marked as 0.

DNS Record: If a website is legitimate, it should have registered who it is database. So, if the URL domain name is registered which is, it will be marked as 0 and if not specified 1.

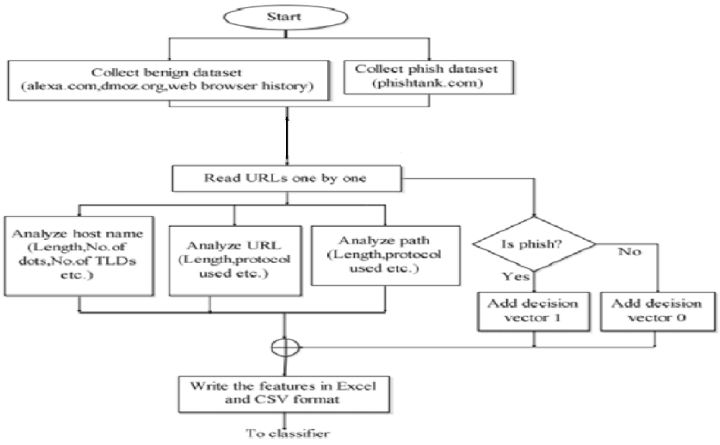
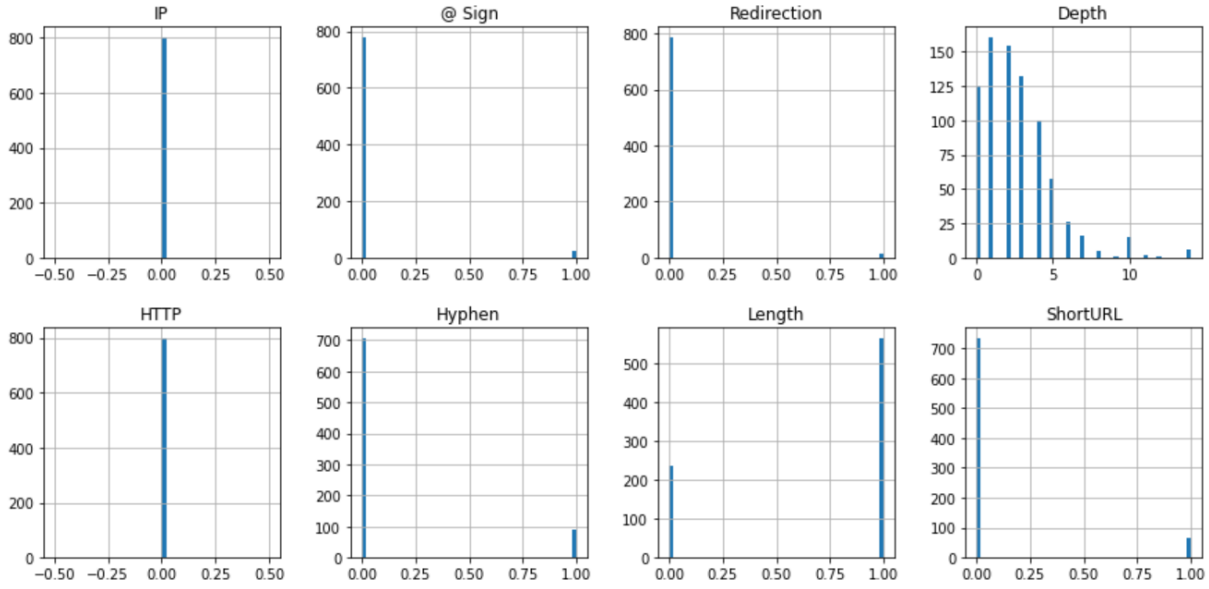


Fig 2. Feature Extraction Process.

C. Visualization of Data

Data visualization is a discipline that seeks to understand data by placing it in a visual context to reveal patterns, trends, and correlations that might not otherwise be observed. Python provides some great graphics libraries packed with many different features. Whether you want to create interactive or very custom storyboards, Python has a great library for you Data visualization is a discipline that seeks to understand data by placing it in a visual context to reveal patterns, trends, and correlations that might not otherwise be observed. Python provides some great graphics libraries packed with many different features. Whether you want to create interactive or very custom storyboards



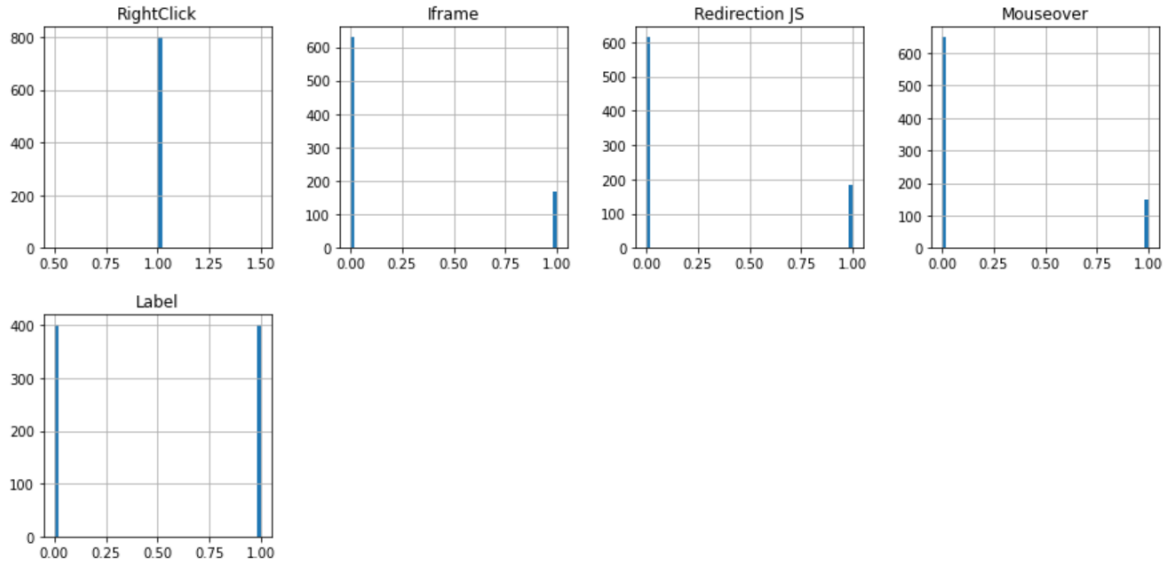


Fig 3. Data visualization

D. MACHINE LEARNING ALGORITHMS/MODELS

The model is created with five main classification algorithms/models K-means Classifier, Naive bayes classifier, Logistic Regression, Decision Tree Model, Random Forest.

K-means Classifier: It is a method for finding subgroups of observations, is widely used in applications such as market segmentation, where we try to find structure in data. Although clusters are an unsupervised machine learning technique, they can be used as features in a supervised machine learning model. Clustering is a type of unsupervised machine learning that aims to find homogeneous subgroups so that objects in the same group (clusters) are more like each other. K Means is a clustering algorithm that divides observations into k clusters. Since we can dictate the number of clusters, this can easily be used in classification, where we divide the data into clusters that can be as large or larger than the number of classes. The K-Means function applies a K-Means cluster to the train data as the number of clusters formed by the number of classes and creates labels for both the train and test data. The parameter output controls how we want to use these new labels, "add" adds the labels as features to the dataset, and "replace" uses the labels instead of the training and test data to train our classification model.

Naïve-Bayes Classifier: The Naive Bayes algorithm is a supervised learning algorithm based on Bayes' theorem used to solve classification problems. It is mainly used in text classification, which contains high-dimensional training data. The Naive Bayes classifier is one of the simplest and most powerful classification algorithms that helps build fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts the target based on probability. Some popular examples of Naive Bayes algorithm are spam filtering, sentiment analysis and article classification. Using the data set, it decides and follows the following steps.

- Convert the given data set into frequency tables.
- Create a probability table by finding the probabilities for the given properties. Then using Bayes theorem to calculate the posterior probability.

Logistic Regression: Logistic regression is a supervised machine learning algorithm used primarily for classification tasks that aim to predict the probability that a case belongs to a certain class. It is used for classification algorithms; it is called logistic regression. it is called regression because it takes as input the output of a linear regression function and uses a sigmoid function to estimate the probability of a given class. The difference between linear regression and logistic regression is that linear regression produces a continuous value that can be anything, while logistic regression predicts the probability that a case belongs to a certain class or not. It is used to predict a categorical dependent variable using a given set of independent variables.

Decision Tree: A decision tree is a supervised learning technique that can be used for both classification and regression tasks but is mostly recommended for classification tasks. It is a tree-structured classifier, where internal nodes represent features of the dataset, branches represent decision rules, and each leaf node represents an outcome. A decision tree has two nodes which are the decision node and the leaf node. Decision nodes are used to make any decisions and have multiple branches, while leaf nodes are the outputs of those decisions and contain no other branches. Decisions or tests are made based on the properties of a particular material. It is a graphical representation that provides all possible solutions to a problem/decision under given conditions.

Random Forest: It is used to solve both regression and classification problems. It is a set of decision trees. Depending on the data, parameters how to calculate decision trees, depth of trees, etc., parameters are given to the function as arguments so that the accuracy of the model can increase. Each decision tree delivers value The set of values is averaged, and that average is randomly returned forest algorithm.

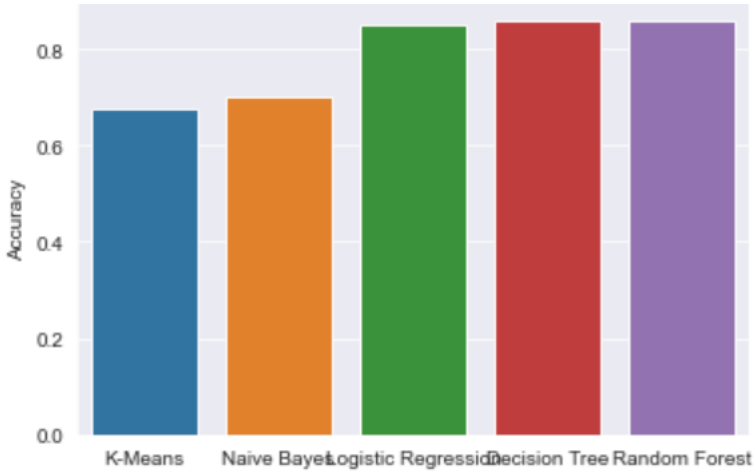


Fig 4. Machine Learning Model Scores.

5.Results

Phishing sites and their domains exhibit characteristics that differ from other sites and domains. (For example, Google; www.google.com and some random phishing site is example: www.googlee.com). Phishing Uniform Resource Locator websites and "domain names" tend to have different lengths compared to other websites and domain names.

The Table below shows the training and testing accuracies of all models. The difference between training and test accuracy values indicates that the models do not fit the large data set. We consider the accuracy of different classifiers and found Random Forest as the best classifiers provides ultimate accuracy.

Table: Classifier performance.

ML Model	Train Accuracy	Test Accuracy
K-Means Classifier	0.634375	0.675000
Naive Bayes Classifier	0.740625	0.7
Logistic Regression	0.86875	0.85
Decision Tree Model	0.887	0.863
Random Forest	0.881	0.863

Table 1. Shows the classifier performance of each Classifier.

If the URL entered by the user is found a small pop-up will appear on the phishing site to warn the user about this malicious website. It is when the user will decide needs to access to the information on this website or not.

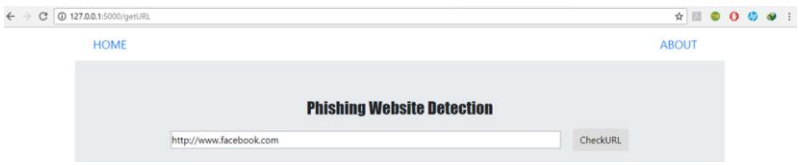


Fig 5. A simple interface to enter a website URL.

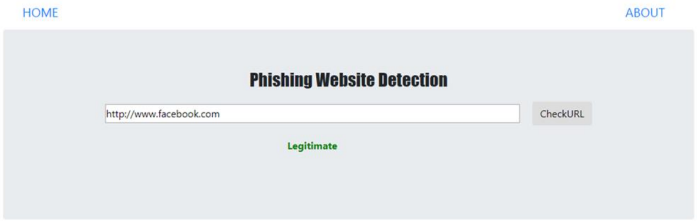


Fig 6. Output of Legitimate URL

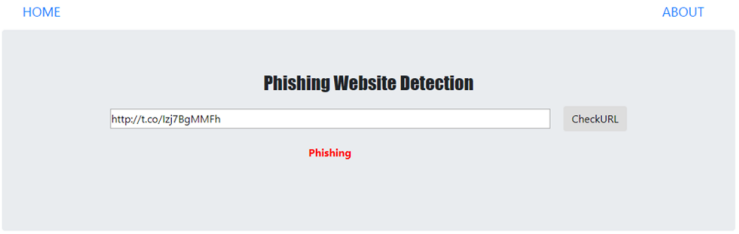


Fig 7. Output of Phishing URL

6. Conclusion

In conclusion, we have seen how phishing is a a serious threat to network security and how phishing detection is a major problem area. We have looked at some traditional phishing methods spotting, i.e. blacklist and heuristic evaluation methods, and their disadvantages. We tested five machine learning phishing website database algorithms and changed their results. We then chose the best algorithm based on this performance and created an additional extension for detection phishing sites. The extension enables easy deployment our phishing detection model for end users. We have identified phishing sites using the Random Forest algorithm with an accuracy of 88.01%. For future improvements we plan to build a scalable phishing detection system online service that includes online learning new phishing attack models can be easily learned and improve the accuracy of our models with better features mining.

7.References

Rishikesh Mahajan, Irfan Siddavatam , International Journal Computer Applications , October 2018. "Phishing Website Detection using Machine Learning Algorithms". This paper explains how phishing is the easiest way to get one. Sensitive information from unsuspecting consumers

Mangala Kini, Deekshitha, International Journal of Engineering Research & Technology (IJERT) Vol. 1.2, No. 6, 2021. "A Review Paper on Detection of Phishing Websites using Machine Learning". Phishing is an attempt to steal personal information such as usernames, passwords and credit card numbers of individuals or organizations as a reliable source in electronic contact.

Maet al. A manuscript on Website phishing Identification. Compared different group-based learning algorithms used to classify phishing sites and found that a combination of host-based and lexical features provides the highest classification accuracy.

Kunju et al. (Kunju et al., 2019) used a survey method to detect phishing attacks. The study offers several solutions and methods to detect phishing attacks.

Atulya and Praveen (Atulya and Praveen, 2020) discussed different phishing attacks, latest phishing tactics, phishing, and anti-phishing strategies. In addition, the article aims to raise awareness of phishing attacks and the strategies used to detect phishing.

Arshad et al. (Arshad et al., 2021) presented various phishing and anti-phishing techniques in their study. According to SLR, the most common phishing techniques used are phone phishing, email spoofing, phishing, and email manipulation.

Catal et al. (Catal et al., 2022) worked on a systematic literature review that answered nine research questions. The main goal of the research is to identify, evaluate and synthesize the results of deep learning approaches for phishing detection.

Pampapathi B M , Chandana Murthy , Supritha Kumar , Pooja M , Supriya K "Survey on IOT Based Medical Box for Elderly People" in International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE) ISSN 2278-3091, Vol.10 No.3 (May – June 2021 issue), https://doi.org/10.30534/ijatcse/2021/531032021.

Pampapathi B M, A Madhuri, Chennareddy Nikhil, Amar Gouda Patil "Water Monitoring And Purification Of Waste Water For Agriculture Using Iot" in Journal For Basic Sciences Volume 23, Issue 4, 2023 , https://doi.org/10.37896/JBSV23.4/2050.

Pampapathi B M, Mohammad Moshin P , Mohammed Kareemuddin Saqlain, Prajwal Marthur, K Md Ibrahim hussain "Wireless Fire Detection Systems Using Iot" in NOVYI MIR Research Journal , Volume 8 Issue 4 2023 https://doi.org/16.10098.NMRJ.2022.V8I4.256342.37538.

X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017. [9] L.

MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018