PHISHING WEBSITE DETECTION by MACHINE LEARNING TECHNIQUES

CH. Sri Divya Naga Durga [1], T. Tejaswi[2],
[1]**Assistant professor(HOD) , MCA DEPT,** Dantuluri Narayana Raju College**, Bhimavaram, Andharapradesh**
Email:- nemovarma38@gmail.com
[2]**PG Student of MCA, Dantuluri** Narayana Raju College**, Bhimavaram, Andharapradesh**
Email:- talupuritejaswi@gmail.com

ABSTRACT

Phishing is one of the most popular and dangerous cybercrime techniques. The aim of these attacks is to steal information that people and businesses use to perform transactions. Phishing websites have a variety of clues in their content and web browser-based data. The aim of this study is to use random forest svm and logistic regression based classification to classify and predict phishing attacks for 30 features, including Data from Phishing Websites

K**eywords:** —Phishing websites, features, RandomForest, URLExtraction

1 INTRODUCTION

As a result of rapidly evolving technology, internet use has become an integral part of our everyday lives. Because of the rapid advancement of technology and the widespread use of digital systems, data protection has become increasingly important. The primary goal of information technology protection is to ensure that appropriate precautions are taken against threats and dangers that users can encounter when using these technologies [1]. Phishing is described as imitating trustworthy websites in order to obtain proprietary information such as usernames, passwords, and citizenship numbers that are entered into websites every day for various purposes. Phishing websites have a variety of clues in their material and web browser-based data [2-4]. Individuals committing the fraud give the fake website or e-mail information to the target address as though it came from a legitimate company, bank, or other trustworthy source. As the size and complexity of cyber security attacks continue to grow, social engineering techniques remain one of the easiest and most efficient ways to gain access to sensitive or confidential information. Phishing is described by the United States Computer Emergency Readiness Team (US-CERT) as a type of social engineering that involves posing as a trustworthy organization or entity and soliciting personal information from an individual or business through e-mails or malicious websites [1]. While organisations should train workers on how to spot phishing e-mails or links in order to protect themselves from the aforementioned types of attacks, Users can easily replicate entire websites for their own purposes using tools like HTTrack. As a consequence, even the most experienced users can be duped into disclosing private or confidential information by visiting a malicious website that appears to be legitimate. As a result of the aforementioned issue, computer-based solutions to protect against phishing attacks, as well as user education, are needed. A device with such a solution would be able to detect malicious websites and discourage users from communicating with them. The use of Uniform Resource Locators (URLs) is one way to identify phishing websites that aren't legitimate (URLs). A URL is a document's global address on the World Wide Web. One of the major challenges in designing machinelearning-based solutions for this issue is the scarcity of publicly accessible training data sets containing phishing URLs. As a result, studies evaluating the efficacy of machine-learning methods based on existing data sets are needed. This project aims to help

meet that need. The aim of this study is to compare the output of widely used machine learning algorithms on the same phishing data set. In this study, we use a data set from which features from data URLs have already been extracted, as well as class labels. We used Random Forest to evaluate basic machine learning algorithms for classifying URLs.

2.RELEATED WORK

Machine learning techniques for detecting phishing URLs typically evaluate a URL based on one or more features derived from it. In URLs, there are two types of features that can be extracted: host-based features and lexical features. Host-based features explain the website's characteristics, such as its location, who manages it, and when it was built. Lexical features, on the other hand, define the URL's textual properties. Since URLs are only text strings that can be broken down into subparts like protocol, hostname, and route, a system can evaluate a site's legitimacy based on any combination of those elements. Since URLs are only text strings that can be broken down into subparts like protocol, hostname, and route, a system can evaluate a site's legitimacy based on any combination of those elements. Ma et al. [3] treated URL classification as a binary classification problem and developed a URL classification system that processes a live stream of labelled URLs. It also gathers URL features from a wide Web mail provider in real time. Both lexical and host-based features were used. They were able to train an online classifier using a Confidence Weighted (CW) algorithm using the gathered features and labels. After reviewing 358 research papers in the field of phishing countermeasures and their efficacy, Parkait et al. [4] provide a detailed literature review. They divided antiphishing strategies into eight categories and highlighted advanced antiphishing techniques. Multi-label Classifier based on Associative Classification was developed by Abdelhamid et al. [5] for detecting phishing URLs (MCAC). They grouped URLs into three categories based on sixteen features: phishing, legitimate, and suspicious. The MCAC is a rule-based algorithm that extracts multiple label rules from phishing data. In their survey on malicious website detection techniques, Patil and Patil [6] presented a brief overview of different types of web-page attacks. Phishing is described as the act of imitating a legitimate company's website in order to steal personal information such as usernames, passwords, and social security numbers. Phishing websites use a number of clues in their content as well as browser-based security indicators that come with the website. To combat phishing, several solutions have been suggested. Nonetheless, there is no single silver bullet that can completely eliminate this hazard. Data mining, especially the induction of classification rules, is one of the promising techniques for predicting phishing attacks, as anti-phishing solutions aim to predict the website. The authors examine the main characteristics that differentiate phishing websites from legitimate websites, as well as the effectiveness of rule-based data mining classification techniques in predicting phishing websites and the classification technique has been shown to be more accurate. Attackers use various algorithms to gain access to confidential information stored in a website's database. V. Shreeram, M. Suban, P. Shanthi, and K. Manjula suggested an anti-phishing detection approach that uses a rule-based scheme derived from a genetic algorithm to detect phishing hyperlinks (GA). If a phishing connection matches the ruleset created by GA and stored in a database, it is detected [7]. Dridex malware is thought to have been used in this attack. It specialises in stealing bank credentials through the use of macros in Word or Excel documents. If Windows users open email attachments in Word or Excel that contain such a macro, the macro will begin downloading Dridex, which then infects computers and sets the stage for a banking robbery. In this case, a competent and alert employee or software that aids in the detection of such an attack would have been extremely beneficial [3]. To find hidden patterns in a

dataset, machine learning algorithms are commonly used. K-nearest neighbour, decision trees and help vector machine are the most popular algorithms [4]. Furthermore, a belief rule-based expert system [5] [6] will mine rules from the dataset. Hackers may insert secret links that lead to malicious pages into URLs where they are unlikely to be found. Susan Mengel and Mohammed Nazim Feroz suggest a tool to detect URL phishing using URL rating.They categorise and rate URLs using online URL reputation services [9]. They classify URLs based on lexical and host-based features. The method of DNS servers attempting to retrieve real IP addresses from outside the network can be bypassed by replacing hostnames in the host records. Due to corrupted IP associations in the server, this technique can poison the records, causing legitimate URLs that should lead to secure sites to instead lead to malicious pages. Suku Nair's Saeed Abu-Nimeh suggests a new DNS poisoning attack that can circumvent security toolbars and phishing filters. They generate fake results using spoofed DNS cache entries and successfully assault four well-known security toolbars as well as the phishing filters of three common browsers without being detected [10].

3.PROPOSED WORK AND ALOGRITHAM

Random forest is a popular Machine Learning algorithm that is used to solve a variety of problems. It's a supervised algorithm that can solve classification and regression problems. It is made up of a parallel decision tree that takes in data and generates a particular class. As a result, a large number of trees generate various groups. Finally, as the final output class, the sum of all classes is used.

WHAT IS THE RANDOM FOREST ALGORITHM AND HOW DOES IT WORK?

The random forest algorithm is carried out in the following general steps.

1. Select N columns at random from the dataset.
2. Build a decision tree using these N records.
3. Repeat steps 1 and 2 with the number of trees you want in your algorithm.
4. Finally, the sum of all the available classes is selected as the final class.

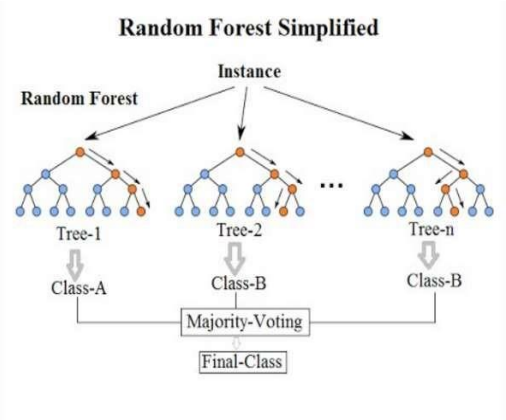

Fig1:- Random Forest Architure

Support Vector Machine:

The Support Vector Machine is one of the most widely used Machine Learning algorithms. The main goal of this algorithm is to find the best data split possible. It is used to solve problems involving classification and regression. It can solve both linear and nonlinear separable data, which is one of its main advantages. The separation line is known as the Hyper plane. Support vectors are the points on which the margins are built. The svm algorithm is depicted in the diagram below.

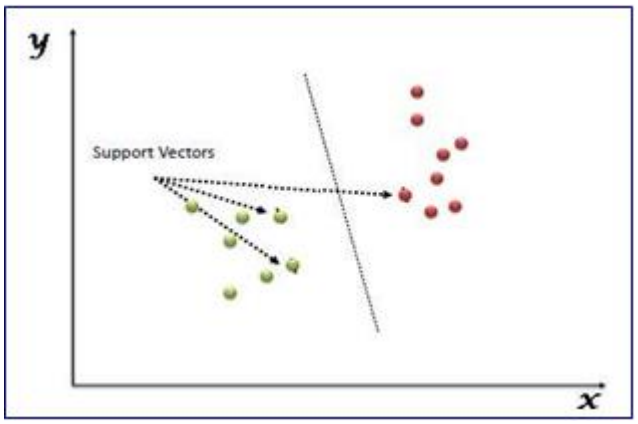

Fig2:- SVM Architure

4 METHODOLOGY

In this analysis, we used different classification algorithms such as andom forest

based classification for the following 30 features extracted from the UC Irvine Machine Learning Repository's website features. The following are the procedural steps for solving the classification problem:

4.1 Dataset

In the UC Irvine Machine Learning Repository database, there are approximately 11,000 data points containing 30 features extracted based on website features.

4.2 Modeling

After the data is ready to be processed, the learning algorithm modelling process begins. The model is essentially the construction of the production requirement based on the job qualifications.

Advantages:

• This research is thought to be a useful design for automated systems with high-performing classification against website phishing.

• Furthermore, this research is found to be high-performing in literature comparisons due to its high efficiency.

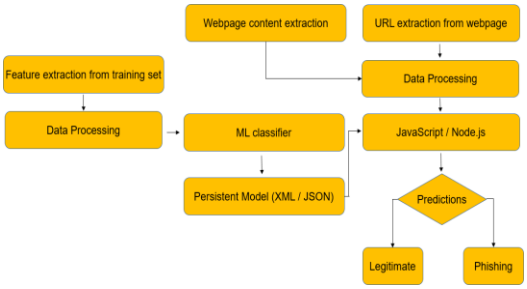

Fig. 3: proposed System Flow Diagram

5 RESULTS AND DISCUSSION

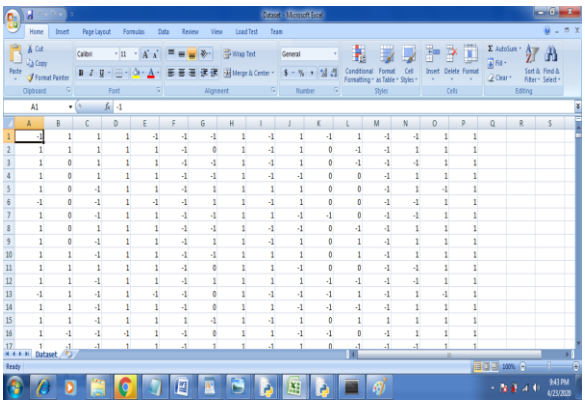

Fig4:-phishing binary dataset

```
accuracy = 87.34%
[[1205  250]
 [ 170 1692]]
(2L, 2L)
TP      FP      FN      TN      Sensitivity     Specificity
1205.0  170.0   250.0   1692.0  0.83            0.91
1692.0  250.0   170.0   1205.0  0.91            0.83
0.889589905362776
runtime = 8.33399987221 seconds
```

Fig5: SVM model accuracy and evolution metrics analysis report

```
accuracy = 89.63%
[[1293  162]
 [ 182 1680]]
(2L, 2L)
TP      FP      FN      TN      Sensitivity     Specificity
1293.0  182.0   162.0   1680.0  0.89            0.9
1680.0  162.0   182.0   1293.0  0.9             0.89
0.9071274298056156
runtime = 0.698999881744 seconds
```

Fig6: Random forest model accuracy and evolution metrics analysis report

TABLE : Performance matrix of classifiers

	Accuracy(%)	Specificity(%)	Sensitivity(%)
Artificial Neural Network	87.34	91	83
Random Forest	89.63	90	86
SVM	89.84	93	89

Fig7:-performance accuracy of all models

TABLE : SVM Confusion Matrix

	Predicted Phishing URLs	Predicted Legitimate URLs
Ground Truth Phishing URLs	1293	206
Ground Truth Legitimate URLs	131	1731

Fig8:- SVM confusion matrix

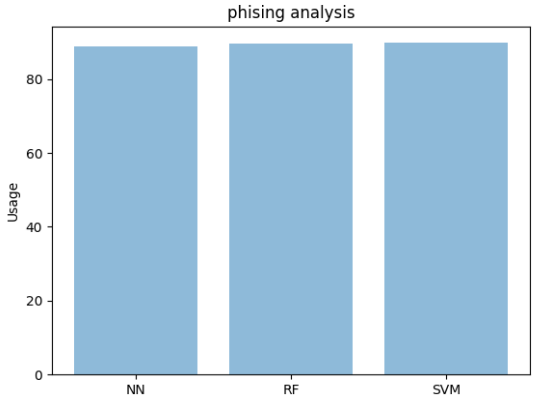

Fig9:- accuracy of all models bar graph

6. CONCLUSION AND FUTURE WORK

We addressed a method for determining whether a given website URL is phishing or not in this paper. We used two common machine learning algorithms, SVM and random forest, to accomplish this. Using advanced features of the URL, we will find a better way to find a phishing website in the future.

7. REFRENCES

1. N. Lord, "What is a Phishing Attack? Defining and Identifying Different Types of Phishing Attacks". https://digitalguardian.com/blog/whatphishing-attack-defining-and-identifying-different-types-phishingattacks, 2018.
2. N. Sadeh, A. Tomasic, and I Fette, "Learning to detect phishing emails", Proceedings ofthe16thinternational conference on world wide web, pp.649–656, 2007.
3. J. Ma, S. S. Savag, G. M. Voelker, "Learning to detect malicious URLs", ACM Transactions on Intelligent Systems and technology, vol. 2, no. 9, pp 30:1-30:24, 20
4. S. Purkait, "Phishing counter measures and their effectiveness–literature review", Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.
5. N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing Detection based Associative Classification", Data Mining. Expert Systems with Applications (ESWA), vol. 41, pp 5948-5959, 2014.
6. D. R. Patil and J. Patil, J., "Survey on malicious web pages detection techniques", International Journal of u-and e-Service, Science and Technology, vol. 8, no. 5, pp. 195–206, 2015.
7. W. Hadi, F. Aburrub, and S, Alhawari, "A new fast associative classification algorithm for detecting phishing websites", Applied Soft Computing vol. 48, pp 729-734, 2016
8. R. K. Nepali and Y. Wang, Y., "You look suspicious!! Leveraging visible attributes to classify malicious short urls on twitter", 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, pp. 2648–2655, 2016.
9. M. Kuyama, Y. Kakizaki, and R. Sasaki, "Method for detecting a malicious domain by using whois and dns features", The Third International Conference on Digital Security and Forensics (DigitalSec2016), p. 74, 2016.
10. D. Sahoo, C. Liu, and C. H. Hoi, "Malicious URL detection using machine learning: A Survey", https://arxiv.org/abs/1701.07179, 2017