JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Framework for Detecting Phishing Website by using ExtremeLearning Machine Algorithm

[1]Dr.N.RAHUL PAL, [2]SISTI.VAMSI

[1]ASSISTANT PROFESSOR, [2]STUDENT
[1,2] DEPARTMENT OF COMPUTER SCIENCE ENGINEERING, UNIVERSITY COLLEGE OF ENGINEERING
[1,2]ADIKAVI NANNAYA UNIVERSITY, RAJAHMUNDRY, ANDHRA PRADESH, INDIA

Abstract- The Internet is a crucial part of our life. Internet customers may be affected by different sorts of cyber threats. Thus, cyber threats may additionally attack monetary statistics, non-public statistics, online banking, and e-trade. Phishing is a form of cyber threat that is focused on getting non-public information such as credit card statistics and social safety numbers. This undertaking proposes a unique technique for detecting phishing websites with the usage of an Extreme Learning Machine (ELM). The ELM is a gadget learning set of rules that is acknowledged for its rapid-getting-to-know velocity and precise generalization performance. We also compared the overall performance metrics of ELM with SVM, Gradient Boosting Classifier, and Gaussian Naïve Bayes. The proposed approach uses a function extraction approach to extract crucial capabilities from the URLs of websites. The extracted features are then fed into the ELM classifier to differentiate between legitimate and phishing websites. The technique evaluates the use of a dataset of real-international phishing websites and valid websites, and the outcomes display that the proposed approach outperforms numerous other latest phishing internet site detection strategies. Overall, this assignment gives a promising technique for detecting phishing websites with the usage of the ELM set of rules. The proposed approach may be beneficial in improving the security of online customers by using preventing them from gaining access to phishing websites and protecting their touchy data.

Keywords: *Internet, Phishing, ELM, SVM, Gradient Boosting Classifier, Naïve Bayes, Legitimate.*

I. Introduction

Internet use has emerged as a crucial a part of our daily activities due to hastily growing technology. Due to this fast boom of the era and extensive use of virtual structures, facts security of those systems has won excellent importance. The number one objective of keeping security in information technologies is to make sure that necessary precautions are taken towards threats and dangers probably to be faced by customers at some point during the usage of those technologies. Phishing is described as imitating reliable websites to be able to obtain the proprietary statistics entered into websites each day for numerous purposes, such as usernames, passwords, and citizenship numbers. Phishing websites contain numerous tips amongst their contents and web browser-primarily based records. The individual committing the fraud sends the faux internet site or e-mail records to the target address as though it comes from a business enterprise, financial institution or every other dependable supply that performs reliable transactions.

Contents of the website or the email consist of requests aiming to trap the individuals to enter or replace their non-public statistics or to alternate their passwords as well as links to websites that seem like precise copies of the websites of the agencies worried.

II. Existence System

Phishing has been a key apparatus utilized by cybercriminals to gain information. There aren't numerous procedures to resist this phishing. A few of them are users who are aware of phishing websites, which isn't truly conceivable since a few clients are not well-informed about phishing. Also, a few third-party websites offer an interface that employments SVM models and an angle-boosting classifier to decide whether a URL is phished or not. They don't give more productive results.

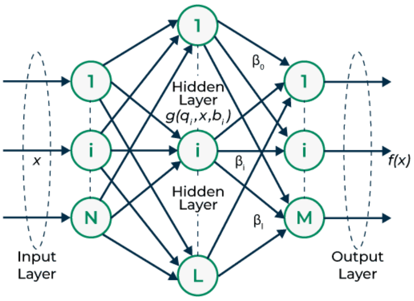
III. Proposed System

Phishing website detection using ELM is to develop a system that can accurately distinguish between legitimate and phishing websites. This requires analyzing a large number of features that can be used to distinguish between the two types of websites. These features might include factors such as the URL structure, the presence ofcertain keywords, the use of SSL certificates, or the content of the website itself. This requires a large amount of training data, as well as sophisticated machine learningalgorithms that can analyze and classify web pages quickly and accurately.

IV. Methodology

A.Extreme Learning Machine (ELM)

An extreme learning machine is known as a feedforward neural network. In neural networks, neurons are part of Artificial neural networks in machine learning. It is a single-layer neural network (SLFN) with randomly generated neurons. The randomly chosen hidden nodes produce the output weight. This algorithm provides a good performance and fast learning process. This is process produces the new algorithms the number of times fast learning. The overall performance of the model is better. This process is called as feed-forward network. In the SLFN model is not a single hidden layer in a neural network. That is the weight of the neuron is the input layer and passes the values through the channel to the hidden layer, then calculating the output layer neuron weight uses ELM and otherwise



uses linear function. ELM has been implemented excellent in many applications,these are semantic concept,google assistant,document classification,image super solution bioinformatics,face recognition etc. Above are the applications of neural networks that are used in real time applications.

Figure-1 Extreme Learning Machine Architecture

V. Phishing Website Features

1. **Using IP address**: IP address in the URL domain nameis used. Few times IP address is to be converted into radix16 code . Proposed rule - If IP address exist in domain→ phishing, otherwise→ legitimate

2. **Using URL length**: The URL length can be calculated.In next the length of given URL characters is equal to 54or greater than 54 then given URL is determined as phishing. Proposed rule – If the URL_len < 54 → legitimate, URL_len ≥ 54 and ≤ 75 → suspicious and otherwise → phishing

3.**Tiny URL**: URL length can be shortened and if web page can be opened in this way. Short URL domain name,which Depends on behalf of the Long URL domain, perform With HTTP Redirection. Proposed rule – Tiny URL is used→phishing, otherwise → legitimate

4.**Using "@" symbol**: If pervious part of the "@" symbolin URL then Browser ignores it it says that the next part of"@"Symbol in URL often the real address. Proposed rule -URL having the @ symbol →phishing, otherwise→ Legitimate

5.**Using "//" symbol**: If the any user directs the website by using"//" symbol in URL. And URL starts with "HTTP" Then "//" symbol must be in the sixth position. IfURL starts with "HTTPS" then "//" symbol in the 7th position . Proposed rule – The "//" symbol seen at last position in URL. Then URL > 7 → phishing, otherwise →legitimate

6.**Sub domain and multi sub domain**: If the given URL is "www." And specific country code in URL is ignored. The remaining points are counted in the URL. If the number of dots is equal To 1 then web site are classified as"legitimate". If the number of dots is equal to 2, then web site are to be classified as "suspicious". If the number of dots is greater than 2 then web site has been classified as "phishing" . Proposed Rule - The given URL number of dots in domain = 1 → legitimate and number of dots in domain = 2 → suspicious, otherwise → phishing

7.**Using HTTPS**: Checking the certificate with using HTTPS trusted certificate issuer and certificate age. Then
It is found that the minimum age of a certificate was 2 years. Proposed Rule - In URL there is HTTPS, trusted provides security certificate, age of given certificate ≥ 1 year → legitimate when uses HTTPS, un trusted securitycertificate providers → suspicious, otherwise → phishing

8. **Domain registration length**: It checks the fake domainvalidity that is one year or more in given dataset. ProposedRule - domain expires on ≤ 1 year → phishing, otherwise→ legitimate

9. **Favicon**: If a web page contains the favicon is loaded from a domain which can be different from the domain shown in the address bar, then web page is to be classified as "phishing". Proposed Rule - If the given URL favicon isto be loaded from external domain then → phishing, otherwise → legitimate

10. Using HTTPS Token: HTTPS token is added to the part of domain of URL by attackers . Proposed Rule - By using HTTPS token in part of domain of URL is seen then
→ phishing, otherwise → legitimate

11. **Request URL**: Web page address and mostly the objects are embedded in web pages it may share the same domain in a legitimate web page . Proposed Rule - If the %of request URL< 22% then → legitimate, % of request URL ≥ 22% and < 61% then → suspicious, otherwise → phishing

12. **URL of Anchor**: it identifies as a member indicated by tag. tags and the web site may have another domain names. The anchor element may not be a connection to any web page . Proposed Rule - If the % of URL of anchor < 31% then → legitimate, % of URL of anchor ≥31% and ≤ 67% → suspicious, otherwise →
Phishing.

13. **Links in <Meta>, <script> <link>:** These tags are expects to be connected to the same domain on a web page. <Meta> tag is

used for the retrieving metadata about the HTML (Hyper Text Markup Language) document recommendation. <Script> tag is used for creating the client-side script. <Link> tag is used to get another web resources [14]. Proposed Rule - % of links in <Meta>, <script> and <link> tags <17% → legitimate, % of links in <Meta>, <script> and <link> tags ≥ 17% and ≤ 81% → suspicious, otherwise → phishing

14.**Server Form Handler (SFH):** SFH (Server Form Handler) contains an empty string or about: blank that classified as “phishing”. If the domain name in SFH is various from the domain name of the webpage then it classifies the “suspicious” [14]. Proposed Rule - SFH is "about: blank" or empty then → phishing, SFH refers to the other domain then→ suspicious, otherwise →legitimate

15.**Submitting the information to e-mail**: this web form is used to send a user's personal information to the server. “Mail ()” function is used by using a server-side language and “mailto” can be used by using a client-side language [14]. Proposed Rule - using "mail ()" or "mailto:" → phishing, otherwise → legitimate

16.**Abnormal URL:** It can be extracted from the WHOIS database. Identity is the part of its URL for a legitimate website [14]. Proposed Rule - If the Host name is not in UR then → phishing, otherwise → legitimate

17. **Website forwarding:** It seen that legitimate websites are redirecting mostly once, and phishing websites are redirecting at least 4 times in the dataset [14]. Proposed Rule - The number of redirect page ≤ 1 then → legitimate, number of redirect page ≥ 2 and < 4 then → suspicious, otherwise → phishing

18. **The status bar customization:** The fake URL can be displayed to the users in the status bar by the attackers. For this purpose JavaScript can be used. Especially “on Mouse over” event was focused on [14]. Proposed Rule - on Mouse over changes status bar then → phishing, otherwise → legitimate

19. **Using pop-up window:** It sends the Request to user’spersonal information in pop-up window on the legitimatewebsite is not regarded as in the normal situation. This feature is to be used in some legitimate websites for precise purposes [14]. Proposed Rule – popup window contains the text field then phishing, otherwise → legitimate

20. **Iframe redirection:** It has been said that to show an extra webpage the Iframe tag is used [9]. Proposed Rule: using Iframe → phishing, otherwise → legitimate

21.Age of domain: It can be extracted from the WHOIS database. It is seens that an age of legitimate domain is at least 6 months Proposed Rule - age of domain ≥ 6 months → legitimate, otherwise → phishing

VI. Modeling and Analysis

System Design-

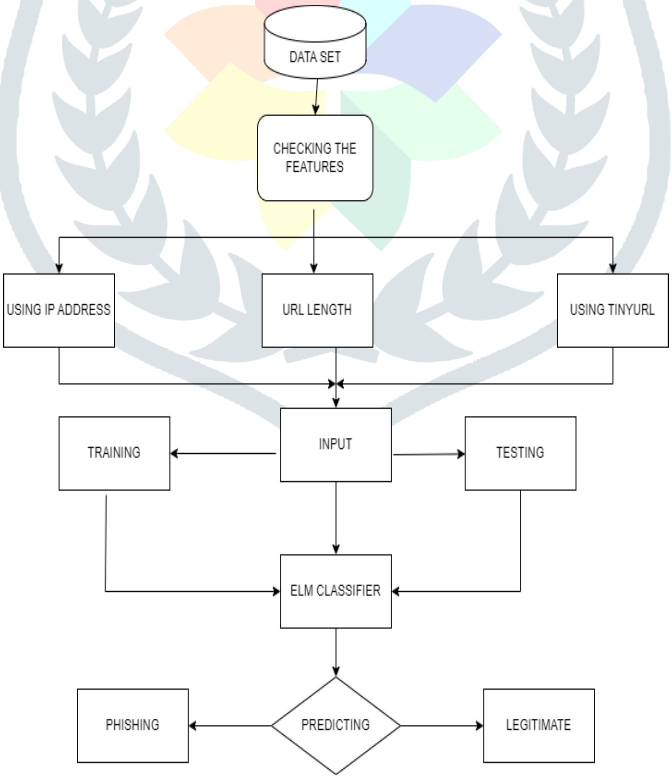


Figure-2 System Architecture

Step 1. Visiting the webpages
Step-2.Checking all given 30 attributes with their featuresStep 3. Collecting the dataset samples
Step 4. Randomly chose 80% training samples and 20%testing samples.
Step 5. ELM classification.
Step 6. Prediction of webpages are phishing or legitimate.

This methodology imports the dataset for checking the URLs are phishing or legitimate from database then imports the data and data can be pre-processed. Phishing web pages can be identified with the help of some categories of URLs features: domain name, address. We can compute the attribute values by extracting the features of phishing web pages and we get the threshold value and range value {-1,0,1} is the range for each phishing attribute as {low, medium, high].we can identify the websites are phishing or legitimate on the basis ofextracted attribute values using machine learning.

VI. Data Set Analysis

1. Data Set-

Approximately 11,000 data containing the 30 features extracted based on the features of websites in UC Irvine Machine learning Repository database.

2. Data Preprocessing

The dataset is in the csv file format. In this step cleanthe data, to remove null values. Dataset contains string values which are not fit to the system for making predictions, by using label encoder we convert string values into numerical values.

3. Data Splitting

It is standard in Machine Learning to split data into training and test sets. Separating data into training and testing sets is important part of evaluating data mining models. Typically, when separate a data set into a training set and testing set, most of data is used for training, and a smaller portion of the data is used for testing. In this project the training and testing datasets are divided in 80:20 ratio. This means 80% of the dataset for training and 20%dataset for testing.

4. Model Training

The Processed dataset is used to train the model by using machine learning algorithms that can predict the website. In this project the model is trained by the Extreme Learning Machine, Support Vector Machine, Gradient Boosting Classifier, Gaussian Naïve Bayes.

5. Model Evaluation

The performance of the model can be evaluated on a separate testing set, to ensure that it is able to make accurate predictions on new data. The evaluation metrics include for this classification task are precision, recall, F1-score, accuracy.

VII. Observations

Methods	Train Accuracy	Test Accuracy
ELM	100%	96.56%
SVM	100%	93.53%
GBC	100%	93.98%
GNB	100%	89.10%

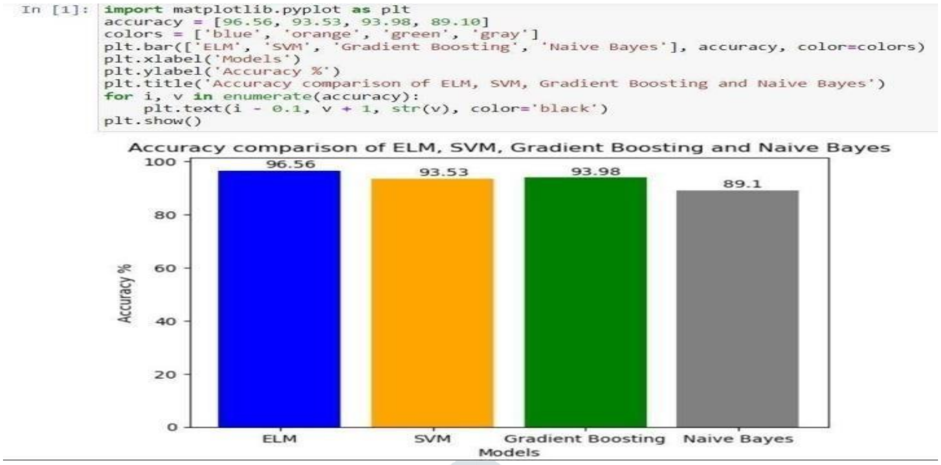
Table-1 Accuracy Comparison

Methods	Train Accuracy	Test MSE
ELM	100%	13.75%
SVM	100%	25.87%
GBC	100%	24.06%
GNB	100%	43.60%

Table-2 Mean Score Error Comparison

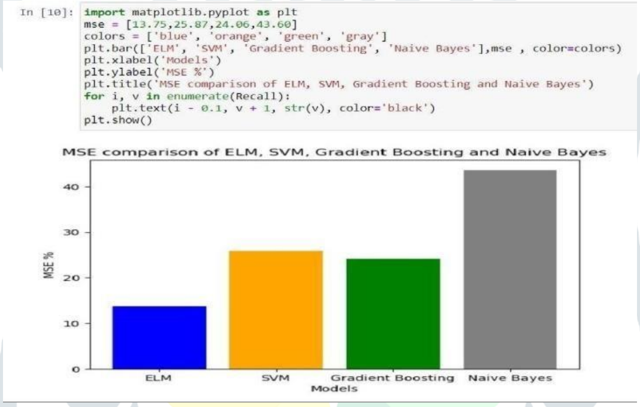
VIII. Experimental Results

In this study, features in the database created for phishingwebsites are classified by determining the input and output parameters for the ELM classifier. Results obtained by ELM show that ELM has higher achievement compared to other classifier (SVM ,GBC, GNB) methods. This study is considered to be an applicable design in automated systems with highperforming classification against the phishingactivity of websites. Furthermore in this project ELM has highest accuracy and low mean score error compared with remaining models (SVM,GBC,GNB).



```
In [1]: import matplotlib.pyplot as plt
        accuracy = [96.56, 93.53, 93.98, 89.10]
        colors = ['blue', 'orange', 'green', 'gray']
        plt.bar(['ELM', 'SVM', 'Gradient Boosting', 'Naive Bayes'], accuracy, color=colors)
        plt.xlabel('Models')
        plt.ylabel('Accuracy %')
        plt.title('Accuracy comparison of ELM, SVM, Gradient Boosting and Naive Bayes')
        for i, v in enumerate(accuracy):
            plt.text(i - 0.1, v + 1, str(v), color='black')
        plt.show()
```

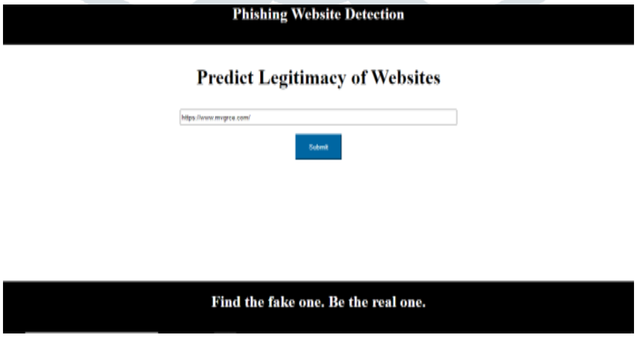
Figure-3 Bar Graph of Accuracy Comparison



```
In [10]: import matplotlib.pyplot as plt
         mse = [13.75,25.87,24.06,43.60]
         colors = ['blue', 'orange', 'green', 'gray']
         plt.bar(['ELM', 'SVM', 'Gradient Boosting', 'Naive Bayes'],mse , color=colors)
         plt.xlabel('Models')
         plt.ylabel('MSE %')
         plt.title('MSE comparison of ELM, SVM, Gradient Boosting and Naive Bayes')
         for i, v in enumerate(Recall):
             plt.text(i - 0.1, v + 1, str(v), color='black')
         plt.show()
```

Figure-4 Bar Graph of MSE Comparison

IX. Results



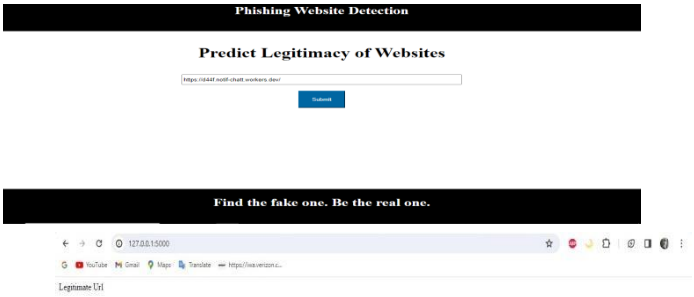
Phishing Website Detection

Predict Legitimacy of Websites

https://www.mvgrce.com/

Find the fake one. Be the real one.

Figure:5 Legimate URL



Phishing Website Detection

Predict Legitimacy of Websites

https://d44f.notif-chatt.workers.dev/

Find the fake one. Be the real one.

127.0.0.1:5000

Legitimate Url

Figure6:Phishing URL

X. **Conclusion**

Using Extreme Learning Machines (ELM) for phishing website detection is a promising approach that has the potential to achieve high accuracy and fast detection times. ELM has several advantages over other machine learning algorithms, such as its fast training time, high accuracy, and good generalization ability. These properties make it a suitable choice for detecting phishingwebsites, where fast and accurate classification is essential. However, the success of any machine learning approach heavily relies on the quality of the data used for training and testing. Therefore, it is crucial to have a diverse and representative dataset to ensure that the ELM model can accurately generalize to unseen phishing websites. Moreover, it is important to incorporate other techniques such as feature selection, data preprocessing, and ensemble methods to improve the performance of the ELM model. Overall, ELM can be a valuable tool for phishing website detection, but it should be used in conjunction with other approaches and techniques to increase the effectiveness and reliability of the detection system.