MANTECH Publications
Recent Trends in Computer Science and Software Technology
Volume 7, Issue 3, September-December 2022

Identifying Spoofing Websites

K. M. Venkatachalam[1], M. Arumugam[2], M. A. K. Borelessa[3]
Students[1, 2, 3]
Department of Computer Science Engineering
Rajlakshmi Engineering College
Corresponding Author's E-mail:- *venkatachalamkm854@gmail.com[1]*

Abstract

Spoofing websites are forgeries made by hackers or intruders to get sensitive information from consumers. They appear to be legitimate websites. With the growing popularity of the internet, there is a growth in the number of individuals who rely on online services to shop, check their financial accounts, pay their bills, or even play games with anonymous pals. Spoofing attacks are a big security risk in today's society. It is a type of email-based assault. The goal is to fool the email recipient into thinking the message is something they want or need, such as a request from their bank or a note from someone in their company, and clicking a link or downloading an attachment will proceed with the collection of datasets, pre-processing of datasets, and creation of a URL. Machine learning algorithms are used to train the model. Finally, we determine whether the website is real or not. The study is presented in the form of a confusion matrix, along with the accuracy of all methods.

Keywords: *ML, URL, spoofing, SVM, Flask model*

INTRODUCTION

In all of our activities, we routinely make use of the internet. A large number of individuals participate in social networks such as Facebook and Twitter, and over the course of the past decade, the majority of governmental and financial institutions have made online services available to their customers. People are increasingly purchasing online, and when they do so, they provide their debit and credit card information, assuming that the URLs they are using are legitimate. Because of the fast advancement of technology and the

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

widespread adoption of digital systems, the data security of these systems has assumed a position of paramount significance [1, 2]. Intruders are now active on social media platforms, and they are using these platforms to get others to carry out a certain activity or provide specific information. Spoofing attacks are the most common type of social engineering that is carried out. The act of impersonating another individual in order to get access to their private data or bank accounts is known as "spoofing. The crime known as "spoofing" is a sort of identity theft. Websites that duplicate the web pages of legitimate websites are known as "spoofing" websites, and hackers are the ones who construct these bogus websites. In order to deceive their victims, hackers frequently generate web pages that are aesthetically extremely similar to the actual web pages that they are targeting. The proprietors of a spoofing online page may gain access to the victims' bank account information, passwords, credit card numbers, and other sensitive data if the victims fall victim to the spoofing web page. Another form of illegal conduct that may be seen taking place on the internet these days is known as "click fraud." [3] A recent study found that 57 million people who use the internet have reported receiving email that was linked to a spoofing site. Of those 57 million people, approximately 1.7 million are thought to have been the subject of too many attacks, which led to the disclosure of personal information from a variety of users. Based on their research, the members of the Anti-Spoofing Working Group (APWG) have come to the conclusion that spoofers have a chance of succeeding with up to five percent of all message receivers. As part of the spoofing process, hackers will steal the user's personal information by visiting and creating a fake webpage that looks like a genuine webpage containing one of a legitimate bank or company and asking the user to enter personal information such as their username, account number, password, credit card number, and so on. This will allow the hackers to steal the user's personal information. The end result of this is that the users are disclosing private information about themselves to unintended third parties, which might result in the loss of financial or other assets [4].

The primary objective of this study is to examine the potentially useful application of machine learning methods for the identification of the difficult problem of spoofing websites. The classification problem is strongly connected to this form of prediction in the field of machine

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

learning. For the purpose of categorization, the lengthy URLs, personalization, prefixes and suffixes, and other information obtained from the input websites provided by the users are employed. Client Application, Flask, and Trained Model are the three components that make up the entire system. This information is presented in Fig. 1, which can be found below.

The client application consists of a website with a web page that gives the user the ability to input the URL, which has to be examined to determine whether or not it is valid. Both HTML and CSS were used to create this page. Python is the programming language that is covered during the course.After the training process is complete, the model is saved in a pickle file. The use of a Pickle file is beneficial for both the storage of trained objects and the testing of trained objects [5]. The Flask model is utilised at this point to facilitate communication between the trained object and the data on the front end. The HTML pages that are generated by the Python programming language are rendered with the help of the Flask model. Therefore, the primary responsibility of the flask in this context is to render the HTML pages, transfer the data from the webpage to the trained object, and contribute to the presentation of the result about whether or not the URL is authentic. Due to the fact that the feature extraction model requires information on the URL that was input by the user, the system has to be connected to the internet.

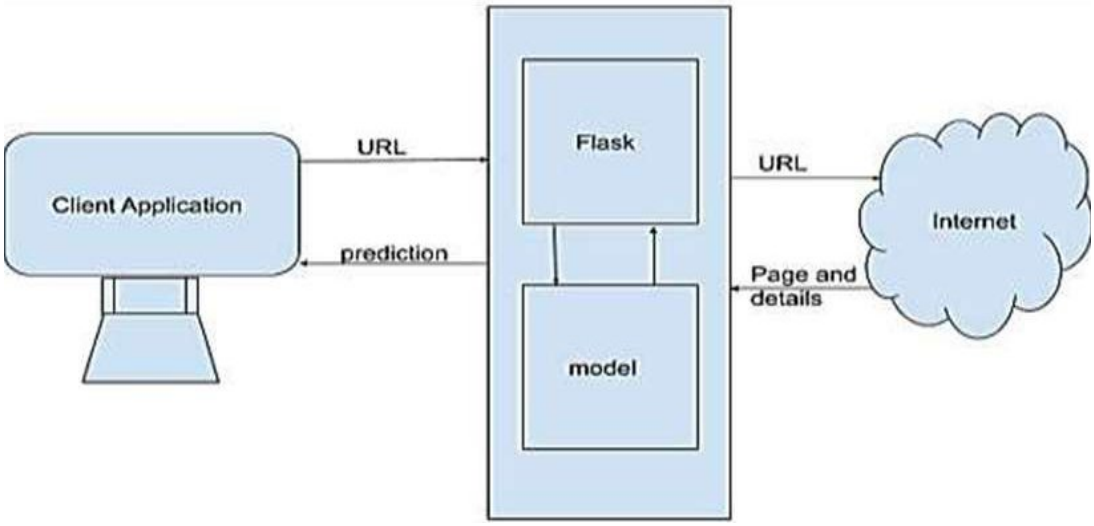


Figure 1: Block diagram

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Pre-Processing of Dataset and URL

The Uniform Resource Locator (URL) is what defines an address on the World Wide Web. The Uniform Resource Locator (URL) is used to identify networks in connection with web resources [6, 7]. The Uniform Resource Locator (URL) can be written down as protocol: hostname/other information.

Example: http://www.indiana.edu. The protocol that was used first in order to visit the page after then, the host server may be identified with the help of the domain name. It will include both the top-level domain and the registered domain name in its contents. Due to the requirement that it be registered with a domain name registrar, the domain name part is subject to certain restrictions. A domain name and a subdomain name are the two components that make up a hostname. Unauthorized users or hackers have complete authority over the subdomain parts and may assign whatever value they like to them. The invader is also able to manipulate the components of URLs that are referred to as the path and the file. The attacker has complete control over both the subdomain name and the path to the file. The portions of the URL that are discussed further on in this text are referred to by the name "free URL" by us. The perpetrator of the assault has the ability to register any domain name that has never been done so before. Spoofing allows for the modification of the free URL at any moment. Because it has a variety of domain names, it might be exceedingly challenging to identify domains that are used for spoofing. The URL of a website is the very first item that is examined in order to determine whether or not the website is spoofing. Some identifying characteristics may be found in the URLs of spoofing websites. When the URL is analysed, features that are connected to these points can be acquired. [Citation needed] Figure 2 illustrates the "Anatomy of a URL."



Figure 2: Anatomy of URL.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Spoofing URLs may be differentiated from illegal URLs thanks to the many elements contained within them. For instance, the adversary can provide an identical URL name while altering one character in the middle, which can lead to confusion among users [8]. Another method of attack allows the attacker to modify a single bit anywhere in the IP address. Attackers are also able to employ short domain names in their efforts.

1. The following categories have been created for the features that were gathered for the detection of spoofing domains:
2. URL-Based Features
3. Domain-Based Features
4. Page-Based Features
5. Content-Based Features

We have taken our dataset from the www.Kaggle.com website. This has around 11055 website's sample. Each website sample has around 32 attributes.
Such as;

1. IP address
2. Length of URL
3. Duplicate // symbol
4. Request URL
5. Length of domain registration
6. Different URL
7. Any random character as suffixing and prefixing etc.

In our project, these 22 different sorts of qualities are taken into consideration. This is a very time-consuming operation in order to extract the real-time characteristics from those attributes. The file will initially be saved in the.arff format, and then it will be converted to the.csv format. The converted file will be used in training. For the purpose of implementation, we made use of the Python programming language. In the Python programming language, there are a number of built-in functions that may be used to convert files from one format to another [9, 10].

Detection of Spoofing Websites

The model is constructed using machine learning principles. Machine learning provides many algorithms. Algorithms are used to forecast outcomes. Machine learning algorithms take input data, analyse it, and provide a statistical report. Here, both supervised and unsupervised techniques are applied. In a supervised algorithm, the user provides both input and expected output. The user will specify several features, after which the model will analyse and forecast the outcome [11, 12]. First, training must be finished. The user is not required to supply intended output in unsupervised methods. It employs an iterative methodology.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Data Flow Diagram

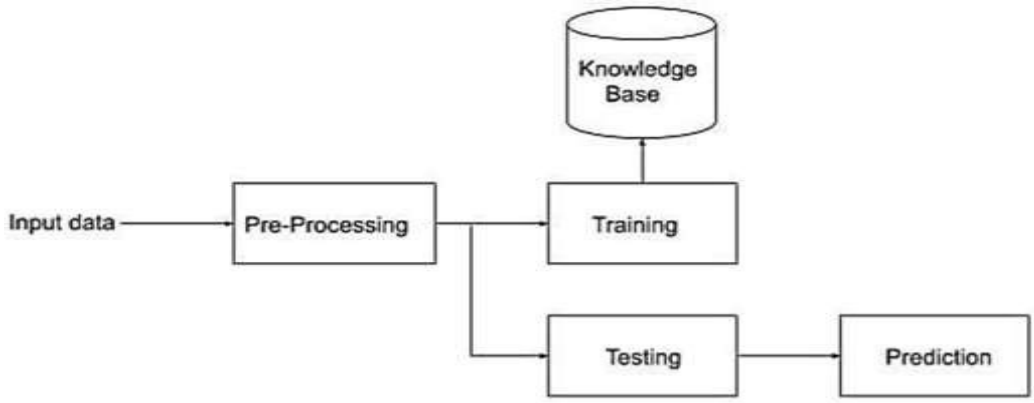


Figure 3: Data flow diagram

Unsupervised learning is used for more difficult problems than supervised learning. The raw data used to identify spoofing websites is gathered from many sources. To detect spoofing websites, we employ a categorization algorithm. We are classifying databases that contain samples of authentic and fraudulent websites. The datasets contain a variety of parameters. The datasets will be trained, which is the initial stage in the detection/identification process [13]. The samples obtained have been classified, and their class labels are known. The spoofing samples must be identified as spoofing or illegitimate. Similarly, the samples labelled as authentic must be identified as legitimateURLs. The system will not provide a good result if we do not categorise the gathered samples in the datasets. ML techniques are used to train the acquired dataset.

Following that, training and testing datasets are classified. The trained dataset is used to train the model. The testing dataset is used to validate the model. Test data can be used to determine the model's correctness. Both test and training data must be distinct. SVM, Random Forest, Decision Tree, and Neural Network methods are employed for classification in our research.

The initial input should be delivered to the system, as indicated in Fig. 3 of the suggested technique. The input should be in the form of a URL dataset. The URL provided by the user should be separated into features. Later, the values should be allocated based on the dataset's criteria, so that the prediction may be simply done using the dataset and the user-supplied URL. Once the dataset is complete, the model must be trained. There are several

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

machine-learning algorithms available. SVM, RF, DT, and NN are all being used. The confusion matrix is used for testing. Once the model has been trained using various algorithms, i.e., SVM, Random Forest, Decision Tree, and neural networks, an analysis is provided using accuracy measures such as the F1 score, recall, and precision, and it also forecasts whether the user input is valid or not.

Dataset Collection

The dataset must be gathered and pre-processed. The dataset is utilised in order to train the system.

Pre-Processing

The URL input from the user is analysed and transformed into values using rule-based approaches. These metrics can be used to predict whether or not a website is authentic.

Train the Module

We used SVM, Random Forest, Decision Tree, and Neural Network techniques in this case. The proposed model is depicted in the image below. Training is done first, followed by testing on the same datasets as the model. The confusion matrix displays the outcome of the experiment. The trained model is saved in the pickle file.The first stage, as indicated in Fig. 4 is to load the dataset into the system. The dataset should clearly distinguish between genuine and non-legitimate websites, and the attributes should be appropriately described. The training and testing datasets will then be separated. For such data sets, training and testing are both performed. Algorithms are trained in numerous ways.

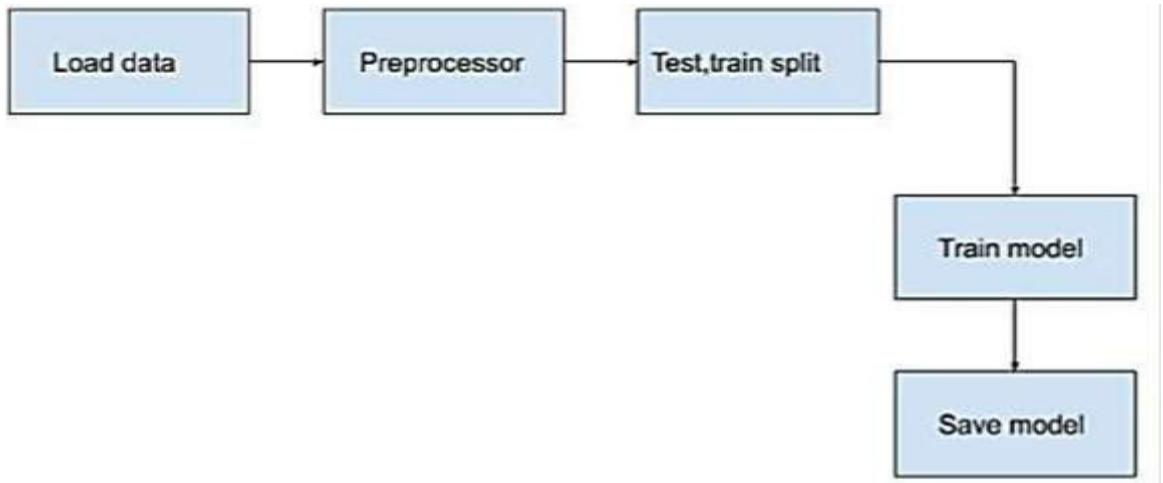


Figure 4: Training the model.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

ALGORITHM OF DETECTION WHETHER THE WEBSITE IS LEGITIMATE OR NOT

Input: Pre-processed data.

Output: object and display confusion matrix.

- Step 1: Load the pre-processed data.
- Step 2: Divide the data to training and testing data.
- Step 3: Use the training data to train the system using SVM, Random Forest, Neural network and decision tree.
- Step 4: Display the results using confusion matrix.
- Step 5: Save the trained object in the pickle file.
- Step 6: Predict whether the URL is legitimate or not.

RESULTS AND DISCUSSION

Results

The Fig. 5–8 shows confusion matrix of the obtained results.

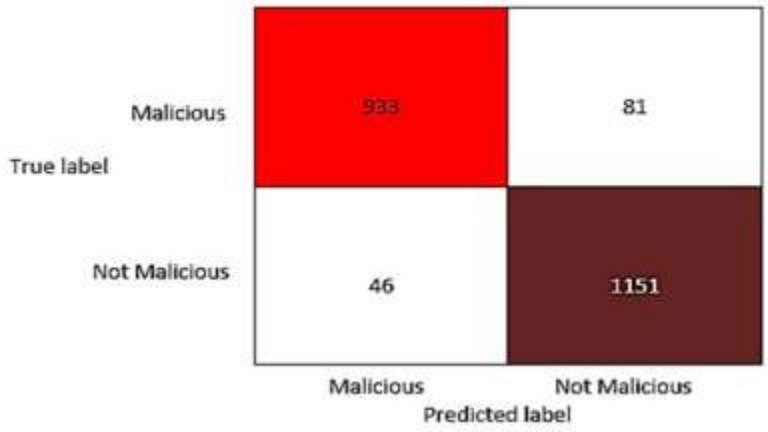


Figure 5: Random forest confusion matrix.

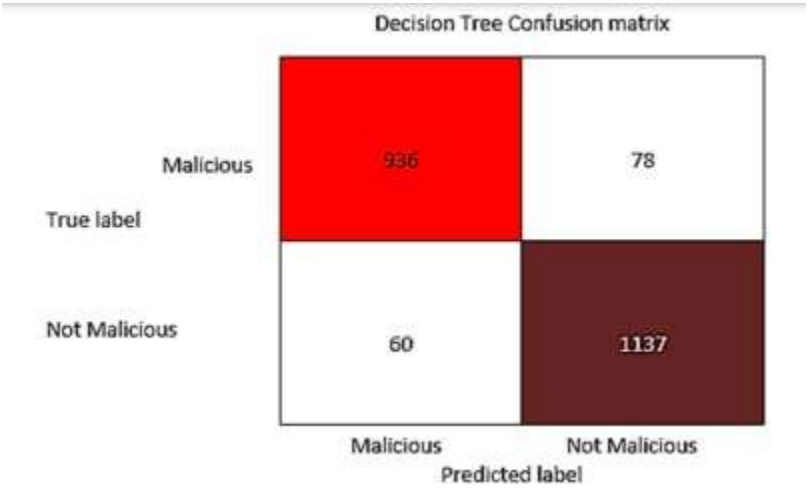


Figure 6: Decision tree confusion matrix.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

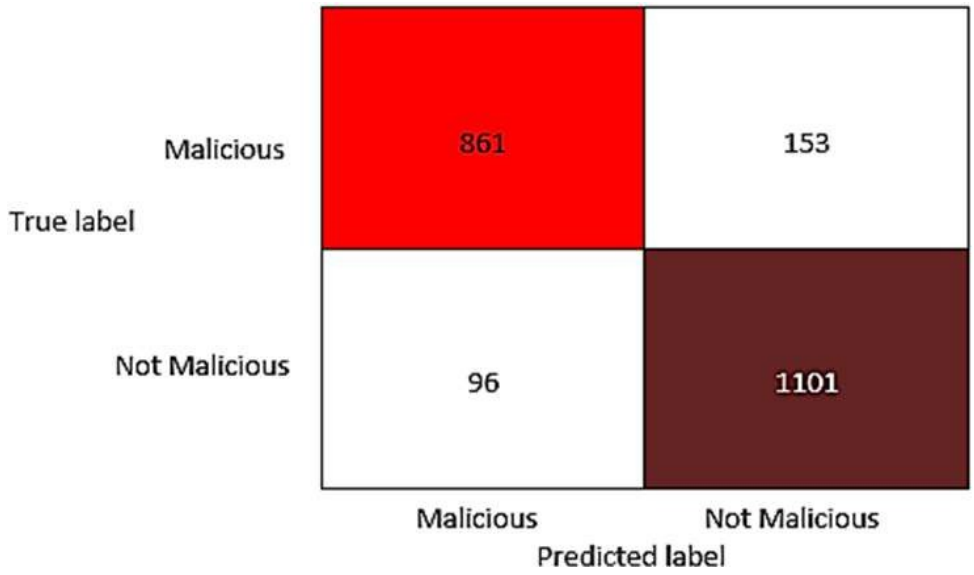


Figure 7: SVM confusion matrix.

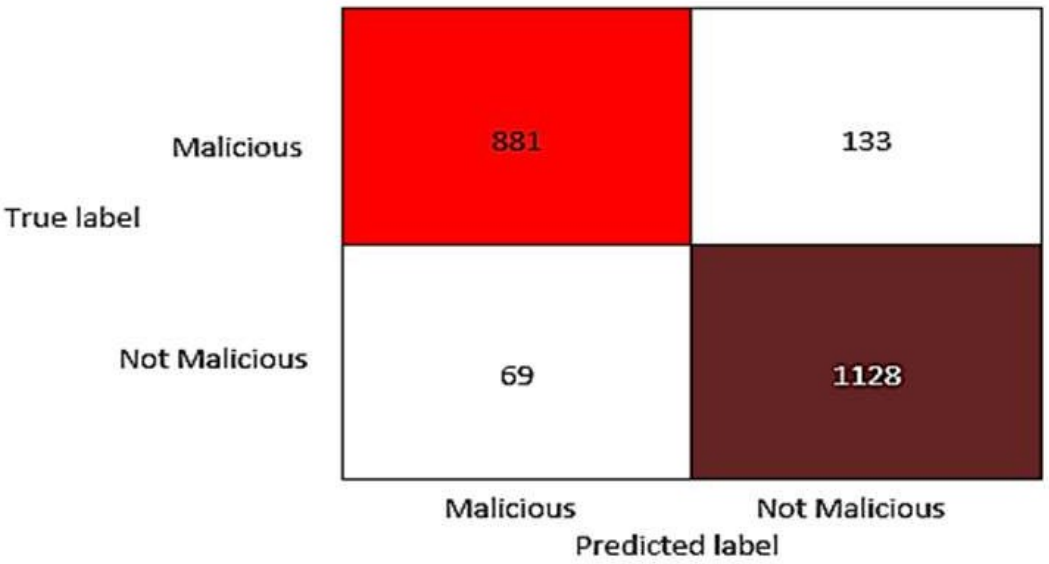


Figure 8: Neural network confusion matrix.

Analysis

The analysis is presented in two forms. One is the confusion and accuracy of all algorithms and the other is whether a user-provided input URL utilising a webpage is authentic or not. The confusion matrix is used to compute the accuracy parameters. The confusion matrix is represented by the many classes. True positive, True negative, False positive, False negative readings are used to compute accuracy parameters. The confusion matrix for several classes is shown in Tables 1–5.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Table 1: Confusion matrix for multiple classes with true positives.

Predictions →

Actual class ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

Table 2: Confusion matrix for multiple classes with true positives.

Predictions →

Actual class ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

Table 3: Confusion for multiple classes with true negative with respect. A

Predictions →

Actual class ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

Table 5: Confusion matrix for multiple classes with false negative with respect A.

Predictions →

Actual class ↓

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

PRECISION

Precision employs both positive and negative values. Precision provides an estimate of the total probability. It is a useful performance metric for detecting events.

The precision can be defined as follows: P =m/ (m + n)

Where, m stands for True Positive and n is the False Positive.

RECALL

Recall is the ratio of the number of articles that are rated properly (true positive) to the total number of articles in a certain category (false negative and true positive). This represents the likelihood of accurately recognising a positive test sample and is unaffected by class priors.

The recall is defined as follows:

R =m/ (m + t)

Where, m stands for True Positive and t stands for False Negative

F1 SCORE

F1Score will take precision and recall into account. It is calculated using the harmonic mean of precision and recall. If F1score is 1, it is the best value; if it is 0, it is the worst value. The F1Score number combines the recall and precision values into a single result.

F1=2∗ (P ∗R/ P + R)

where, P is the Precision and R stands for Recall.

Accuracy Result of Ml Algorithms

Table 6–8 gives the obtained results using Random forest, Decision Tree, SVM and Neural network. Accuracy of Random Forest is: 0.942559927635 Classification report for Random Forest:

Table 6: Accuracy of random forest.

	precision	recall	f1-score	support
-1	0.94	0.92	0.93	1014
1	0.94	0.95	0.94	1197
avg/total	0.94	0.94	0.94	2211

Confusion Matrix

[[936 78]
[60 1137]]

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Accuracy of SVM is: 0.937584803256 Classification report for SVM:

Table 7: Accuracy of decision tree

	precision	recall	f1-score	support
-1	0.94	0.92	0.93	1014
1	0.94	0.95	0.94	1197
avg/total	0.94	0.94	0.94	2211

Confusion Matrix
[[936 78]
[60 1137]]

Accuracy of SVM is: 0.937584803256 Classification report for SVM:

Table 8: Accuracy of support vector machine (SVM)

	precision	recall	f1-score	support
-1	0.90	0.85	0.87	1014
1	0.88	0.92	0.90	1197
avg/total	0.89	0.89	0.89	2211

Confusion Matrix
[[861 153]
[96 1101]]

The following Fig.9 shows classification efficiency of Random forest, Decision Tree, SVM and Neural network**.**

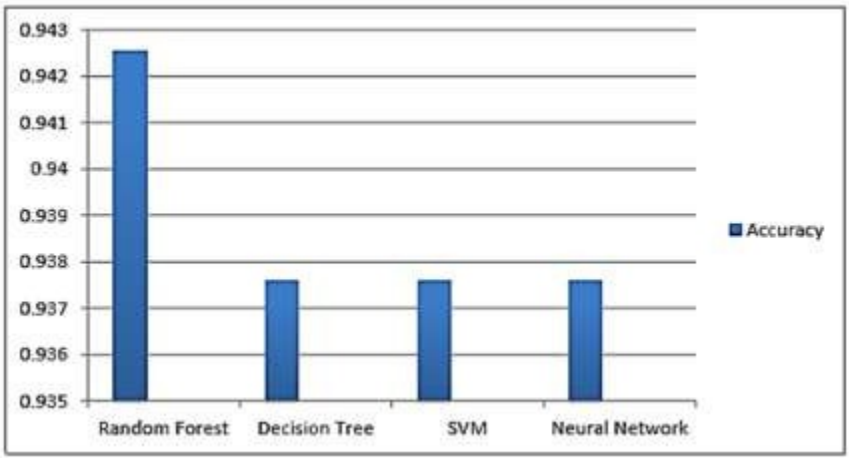


Figure 9: Bar chart of accuracy and algorithms.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

CONCLUSION

In this research, a mechanism for determining the validity of URLs is established. The suggested method trains and predicts if a URL is authentic or not using machine learning techniques. This can be utilised by a variety of users. It helps protect the user's credentials from the spoofing website. Among the SVM, decision tree, and neural network, we find that the random forest has the greatest accuracy. Our paper is also compatible with the Android and iOS platforms. The accuracy of the approaches may be improved by including more characteristics and using a more precise dataset, and the randomness of the process can be reduced by employing correct methodologies.

REFERENCES

1. D. Shen, Z. Chen, Q. Yang, H. Zeng,B. Zhang, Y. Lu, W. Ma (25–29 July, 2004), "Web-page classification through summarization", Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.0.26483/ijarcs.v8i7.4309.
2. U. Naresh, U. VidyaSagar, C.V. Madhusudan (2013), "Intelligent phishing website detection and prevention system by using link guard algorithm", IOSR J. of Com. Eng., Volume 14, Issue 3, pp. 28–36, DOI: 10.9790/0661-1432836.
3. Durgesh K. Srivactava, LekhaBhambhu (2010), "Data classification using support vector machine", J. of The. and App. Inf. Tech., Volume 12, Issue 1, pp. 1–7.
4. Moh'dIqbal AL Ajlouni, Wa'elHadi, JaberAlwedya (2013), "Detecting phishing websites using associative classification" Euro. J. of Busi. and Mgmt., Volume 5, Issue 15, pp. 36–40.
5. Hwanjo Yu, Sungchul Kim (2012), "SVM tutorial: classification, regression, and ranking," Handbook of Natural Computing, pp. 479–506.
6. Muhammet Baykara, Zahit Ziya Gürel (22–25 March, 2018), "Detection of phishing attacks", International Symposium on Digital Forensic and Security (ISDFS), Antalya, Turkey.
7. Sona Taheri, Musa Mammadov (2013), "Learning the naive bayes classifier with optimization models", Int. J. of Appl. Math. And Com. Sci., Volume 23, Issue 4, pp. 787–795, DOI: 10.2478/amcs-2013-0059.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved

8. Jesse Davis, Mark Goadrich (2006), "The relationship between precision recall and ROC curves", International Conference on Machine Learning, Pittsburgh, PA.
9. Jiawei Han (2006), "Data mining: concepts and techniques", Morgan Kaufinann, 2nd edition, pp. 559–569.
10. Thomas Landgrebe, PavelPaclík, Robert P. W. Duin (20–24 August, 2006), "Precision-recalloperating characteristic (P-ROC) curves in imprecise environments", International Conference on Pattern Recognition, Hong Kong, China.
11. Maher Aburrous, MohamadAlamgirHossain, Keshav Prasad Dahal, FadiThabtah (7–11 September, 2009), "Modeling intelligent phishing detection system for e banking using fuzzy data mining" International Conference on CyberWorlds, Bradford, West Yorkshire, UK.
12. R. Gowri, V Karamchand Gandhi, M. Suriakala (2017), "An efficient algorithm to identify phishing sitesusing url domain features", Int. J. of Adv. Res. in Com. Sci., Volume 8, Issue 7, pp. 508–510, DOI:
13. Leo Breiman (2001), "Random forests", Machine Learning, Volume 45, Issue 1, pp. 5–32.

Page 89-102 © MANTECH PUBLICATIONS 2022. All Rights Reserved