IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86143

Detection of Website Phishing Attack Based on Deep Learning

Deepali B.Vaidya[1], Poonam R. Dholi[2]

Student, Department of Computer Engineering, Matoshri College of Engineering and Research Centre, Nashik, India[1]

Professor, Department of Computer Engineering, Matoshri College of Engineering and Research Centre, Nashik, India[2]

Abstract: Phishing sites which expect to take the users private data by attracting them to visit a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one amongst the particularly considerations toward varied areas as well as e-managing an account. Phishing website detection is really an unpredictable and component issue including numerous components and criteria that aren't stable. Proposed an intelligent model for detecting phishing web pages based on Deep Learning. Types of web content are completely different in terms of their features. Hence, we have to use a selected web page features set to avoid phishing attacks. We proposed a model which is based on Deep Learning techniques to detect phishing websites. We have done analysis of three models of algorithms and we have suggested some new rules to have efficient feature classification.

Keywords: Phishing websites, Machine Learning, SVM, NB, ELM.

I. INTRODUCTION

Technology is rapidly growing and with this rapid growing technology, internet has become a vital part of human's daily routines. Use of internet has grown because of the rapid growth of technology and intensive use of digital systems and hence data security has gets more importance. The main objective of maintaining security in information technologies is to confirm that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies. Phishing is the fraud commits to obtain personal data such as user id, password and details of cards by disguising as a trustworthy entity in an electronic transmission. Typically done by email spoofing or instant messaging, it usually directs users to enter personal details at a fake website, the look and feel of which is identical to the legitimate website. Information security threats are seen and developed through time along development within the internet and information systems. The impact is that the intrusion of information security through the compromise of personal data, and also the victims may lose money or different kinds of assets at last. Internet users will be affected from many types of cyber threats such as personal data loss, identity theft, and financial loss. Hence, victimization of the internet might suspect for home and official environments. Effective systems that may improve self-intervention should be formed using artificial intelligence-based information security management system at the time of associate attack. Phishing is an Internet-based attack that seduces end users to visit fake websites and give away personal information.

II. RELATED WORK

With the event of Information and Communication Technology, various varieties of information security threats may be seen. These threats are important within the prevention of damage to person or institution to guard data on computer systems. There are many phishing detection methods in literature review. In these studies, it is observed that ML is challenging techniques may be used. Santhana Lakshmi and Vijaya they used techniques of Machine learning to verify supervised learning algorithms and modelling the prediction task that Multi-Layer Perceptron. Decision tree and Naive Bayes classifications were used for observing technique for web Phishing Detection. It can detect. As compared to other learning algorithms the choice tree classifier is more accurate [2]. Zou Futai, Pei Bei and Panli projected uses Graph Mt some potential phishing which can't be detected by uniform resource locator analysis. It uses contact of user and website. To induce the dataset from the real traffic of an oversized ISP. After anonyms zing these data, they need cleansing dataset. Every record that includes eight fields: User node number (AD), Visiting URL (URL), User Agent (UA), User SRC IP (SRC-IP) access time (TS), Reference URL (REF), access server IP (DSTIP), User cookie (cookie)[3]. Kaytan and Hanbay proposed determining phishing websites supported neural network. Around 30 inputs attribute, and output attribute1 is used for that experiment. The values 1, 0, and -1 were used for input attributes and hence for output attribute Values 1, and -1 are used. To evaluate the system performance 5-fold cross validation method was used. The simplest classification accuracy has been measured as 92.45%. And hence the average accuracy has been measured as 90.61% [1]. Yasin Sonmez, Turker Tuncer perform Extreme learning Machine (ELM) for 30 features.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86143

That has phishing websites in database of machine learning repository. They compare ELM with SVM, Naives bayes. These are other methods of machine learning [6]. X. Chen, find the impact of phishing attacks as consider risk levels and potential market value that downs which is losses experienced by the target companies. It absolutely was analyzed around 1030 phishing alerts which are released on a public database, and financial data related to the targeted firms employing a hybrid method. This is the prediction that the attack was survive around 89% accuracy using supervised classification and text phrase extraction It's been identified some important textual and financial variables within the study. Impact the severity of the attacks and potential loss has been investigated [7]. Giovanni Armano and Samuel Marchal[4] proposed a system which is based upon minimum enclosing ball support vector machine (BVM) to find out phishing website. It has been aimed toward achieving high speed and high accuracy to detect phishing website. Studies were exhausted order to reinforce the integrity of the feature vectors. Firstly, an analysis of the topology structure of website was performed consistent with Document Object Model (DOM) tree. Then, the net crawler was accustomed extract 12 topological features of the web site. Finally the BVM classifier detects the feature vectors. When the proposed method is getting compared with DVM then observed that the proposed method has relatively high precision of detecting. Additionally it fully was discovered that the proposed methodology enhances the disadvantage of slow speed of convergence on large-scale data. It is been shown that the proposed method has better performance than SVM within the experimental results. Finally the proposed systems accuracy and validity has been evaluated. Gowtham and Krishnamurthi[5] studied the characteristics of legitimate and phishing web content thorough. Heuristics were proposed to extract 15 features from similar kinds of web pages supported the analysis. The heuristic results which were proposed are fed as an input to a trained machine learning algorithm to find out phishing websites. Before the applying the heuristics to the net pages, two preliminary screening modules were employed in the system. By the pre approved site identifier that is the primary module, sites were checked against a non-public white-list maintained by the user. By the login form finder that's the second module, web pages were tag as legitimate when there is no login forms present. Unnecessary computation within the system was reduced by helping the used modules. Additionally, the speed of false positives while not compromising on the false negatives was reduced by reduced to the used modules. To detect new output algorithms uses historical data as input. The extreme module websites having 0.4% false positive rate and 99.8% precision. It's been shown that the proposed method is efficient for safeguarding users from online identity attacks. The primary topic is concerning the computation of needed thresholds to describe the 3 email group conversation. And also the second topic is that the interpretations of the cost-sensitive characteristics of spam filtering. They calculate the decision-theoretic rough set model continue which are based on thresholds.

III. PROPOSED METHODOLOGY

The planned methodology that imports knowledge of phishing and legit URLs from the information and so the foreign data is pre-processed.

Detecting phishing website is performed supported four categories of URL features:

- Domain based
- Address based
- Abnormal based
- HTML, JavaScript features

These URL features are extracted with processed data and values for every URL attribute are generated. The analysis of URL is performed by using intelligent phishing detecting scheme algorithm which computes range value and the threshold value for URL attributes. Then it will distinguish between phishing and legitimate URL. The attribute values are worked using feature extraction of phishing websites and it is used to notify the value like range value and threshold value. Features of phishing website as attribute value range from {-1 , 0, 1} as values indicating low , medium and high. The classification of phishing and suspicious website is relies on the values of attributes extracted by getting four kinds of phishing classes and a deep learning approach.

The Figure shows system architecture wherein any given website is input. The features extracted are address based, Domain based HTML JavaScript based and Abnormal based. 2,000 phishing websites collected from Phishtank (http://www.phishtank.com/). Collected data sets carry label values, "legitimate" and "phishing". In this data set randomly selected 70% for training, 30% for the test. The training dataset is used to train the model and adjust the weight of the in the network, while the test dataset remains unchanged and used to evaluate the performance of the both models. After training, run the test data set on the optimized models to evaluate performance.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 6, June 2021
DOI: 10.17148/IARJSET.2021.86143

A. System Architecture

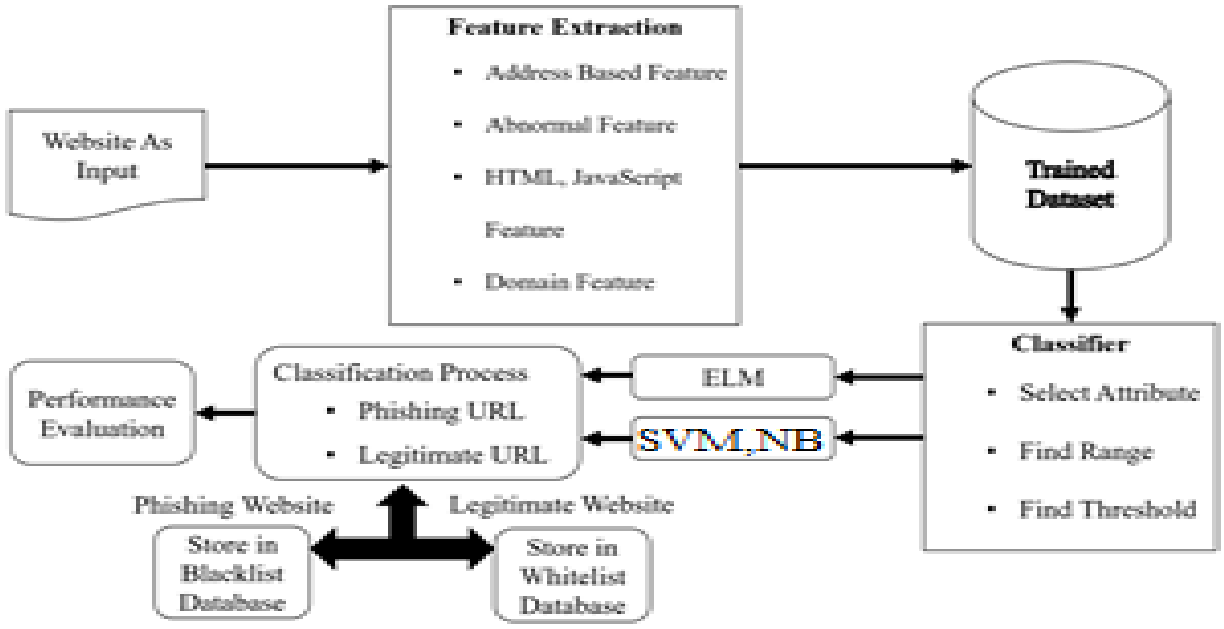

Fig 1: System Architecture

B. Algorithms

a. Extreme Deep Learning

Machine Learning is such that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Deep Learning algorithms use historical data as input to predict new output values. Extreme Learning Machine are feed forwards networks for classification, regression, clustering, sparse approximation, compression and feature learning with a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes (not just the weights connecting inputs to hidden nodes) need not be tuned. These hidden nodes can be randomly assigned and never updated (i.e. they are random projection but with nonlinear transforms), or can be inherited from their ancestors without being changed. In most cases, the output weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model. The name "extreme learning machine" (ELM) was given to such models by its main inventor Guang-Bin Huang.

C. Mathematical Model

Set Theory, Let the system be described as S, then dataset preprocessor, feature extraction, content based, classification and Deep Learning can be give as

S= {D, DP, FE, CB, C, ML}
Where S is the system
D: Set of Input Dataset.
DP: Dataset Preprocessing.
FE: Feature Extraction.
CB: Content Based.
WC: Classification.
ML: ELM
For the input dataset D
D= {d1, d2...dn}
F= {f1, f2...fn}
Y= {DP, FE, CB, C, ML}

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 6, June 2021
DOI: 10.17148/IARJSET.2021.86143

Where D: Set of Input Dataset.
F: Set of Function.
Y: Set of techniques use for Phishing Web Sites Features Classification.
Here,
Fn1: Source File.
Fn2: Data Preprocessing.
Fn3: Feature Extraction.
Fn4: Classification.
Fn5: ELM

D. Development Environment

1. Hardware Resources Required

• Hard Disk : 200 GB
• RAM: 8 GB
• Processor: Intel Pentium i5 and above

2.Software Resources Required

• Technology Used : Python
• IDE: Python IDE
• Operating System: Windows 7 or above

3. Dataset Required

• PhisTank

The set of phishing URLs are collected from open source service called PhishTank. This service provide a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly. From this dataset, 5000 random phishing URLs are collected to train the models.

IV.

V. RESULTS AND DISCUSSION

A. Analysis

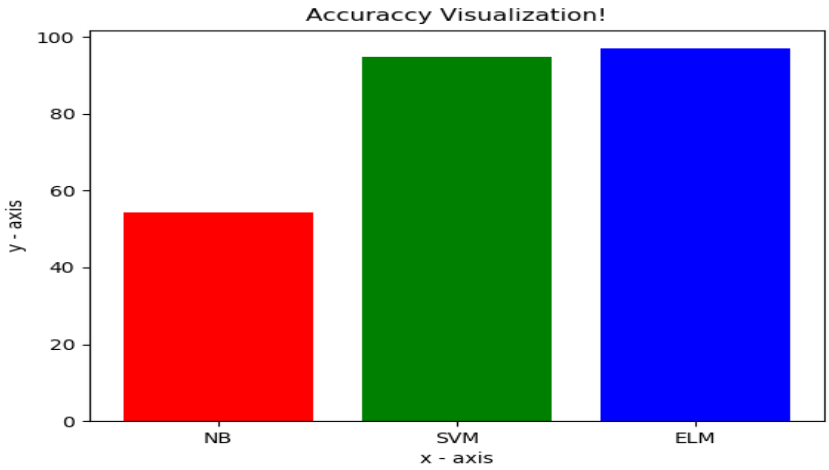

Fig 2: Accuracy Visualization

Three machine learning Algorithms are experimented to give accuracy of ELM to be the best and this used for further performance for detecting phishing websites.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86143

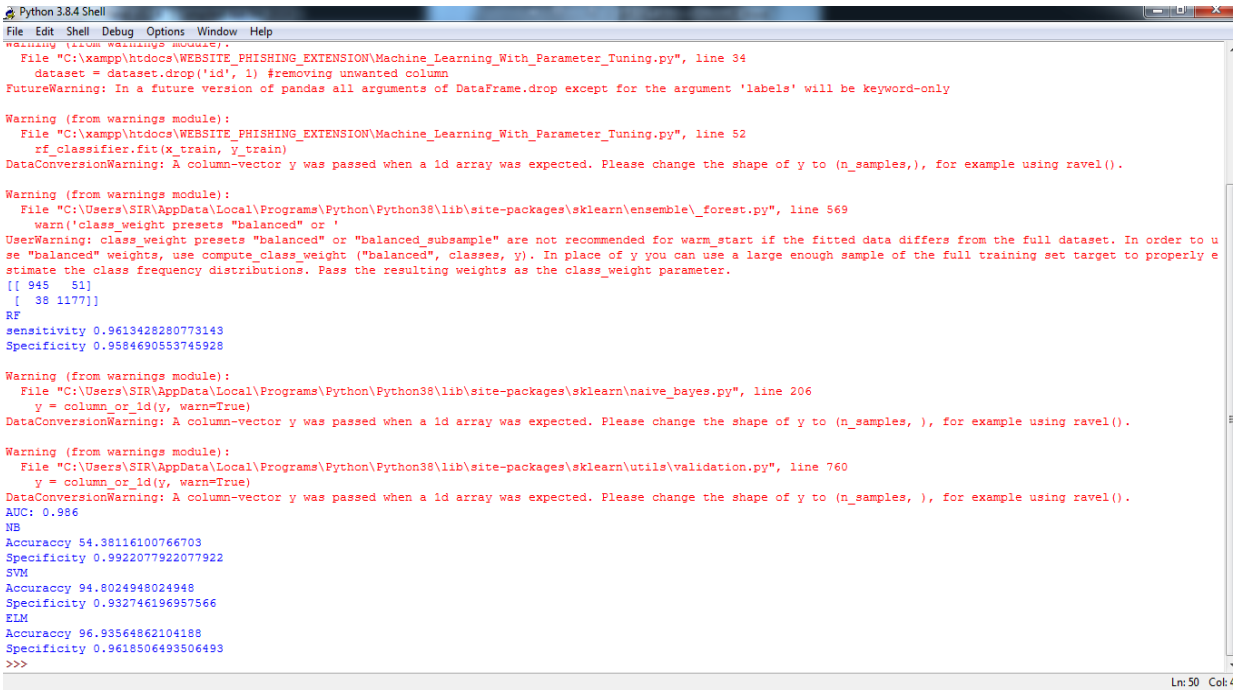

Fig 3: Accuracy Calculation

B. Results

The fig.4 first checks whether the website is legimate.Fig.5 shows screen shot with result that we can visit the site. The fig. 6 checks the site and fig 7 shows that it is phishing website and do not visit.

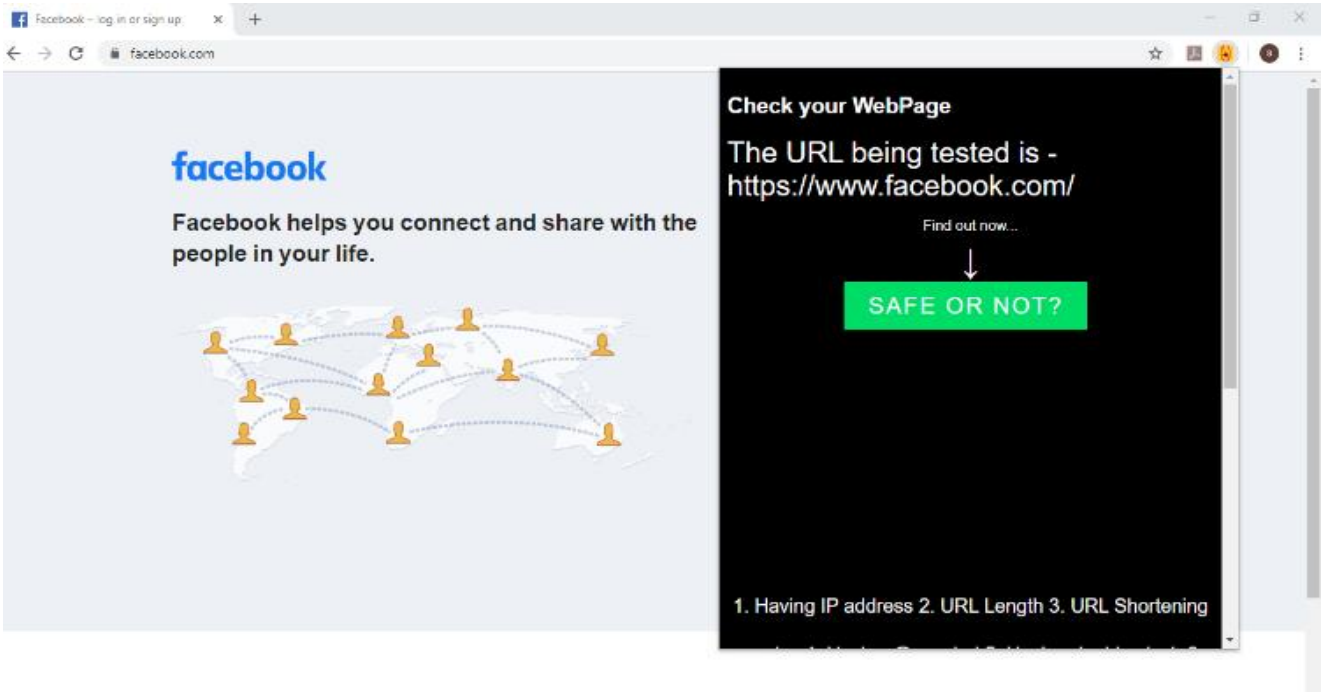

Fig 4: Face book Website

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86143

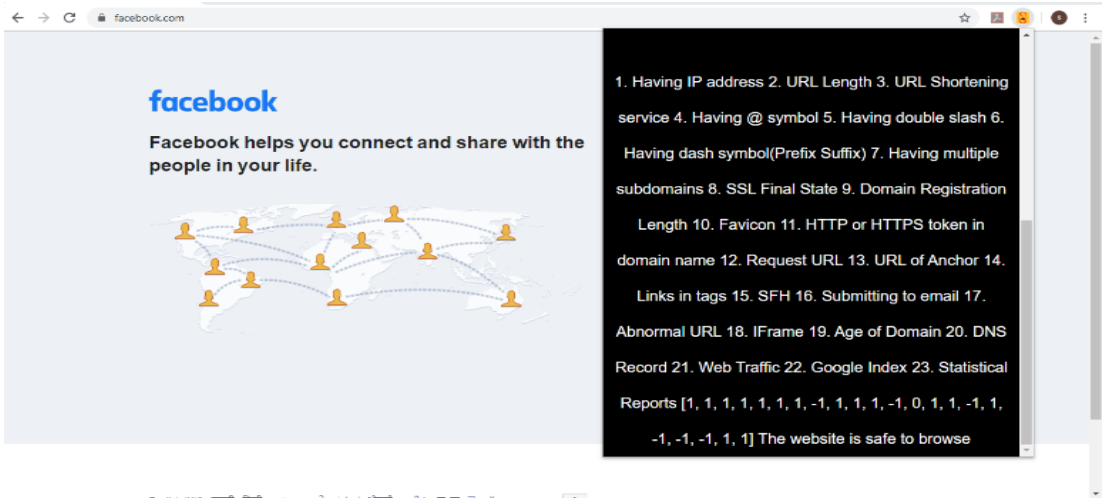

Fig 5: Facebook Website Legitimate

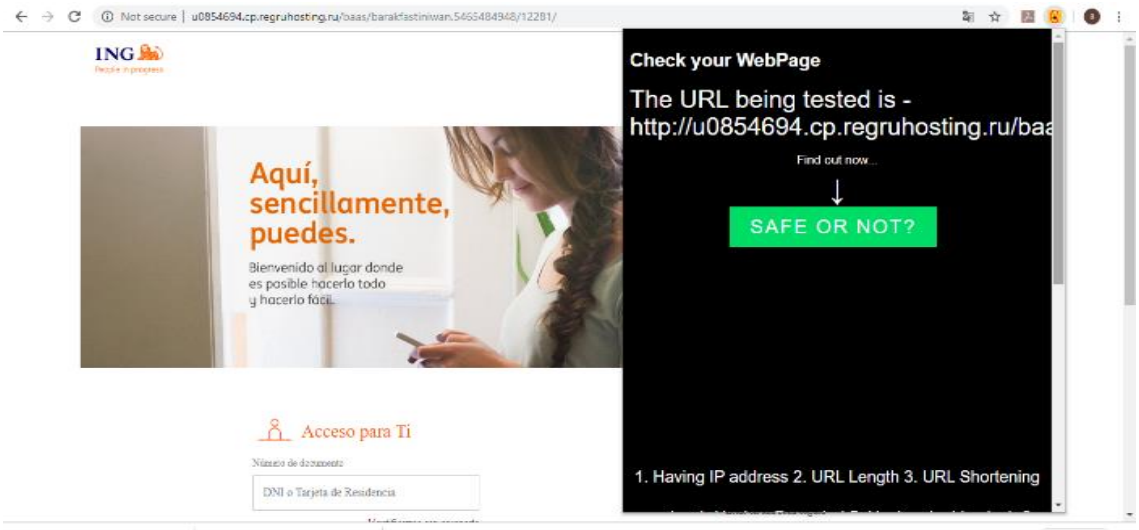

Fig 6: Other Website

Fig 7: Phishing Website

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.86143

VI. CONCLUSION

Systems varying from data entry to information processing applications can be made through websites. The entered information is processed then the processed information can be get as output. Nowadays, web sites are used in many fields such as scientific, technical, business, education, economy, etc. Because of this intensive use, it can be also used as a tool by hackers for malicious purposes. One of the malicious purposes emerges as a phishing attack. A website or a web page can be imitated by phishing attacks and using various methods. Some information such as user's credit card information, identity information can be obtained with these fake websites or the web pages. The purpose of the application is to make a classification for the determination of one of the types of attacks that cyber threats called phishing.

REFERENCES

[1] Mustafa KAYTAN and Davut HANBAY, "Effective Classification of Phishing Web Pages Based On New Rules By Using Extreme Learning Machines" , Anatolian Journal of Computer Sciences, Vol:2 No: 1, pp: 15-36, 2017

[2] V. Santhana Lakshmi and M. Vijaya, "Efficient prediction of phishing websites using supervised learning algorithms", Procedia Engineering, 30, pp.798-805, 2012.

[3] Zou Futai, Gang Yuxiang, Pei Bei , Pan Li, Li Linsen "Web Phishing Detection Based on Graph Mining " 2nd IEEE International Conference on Computer and Communications 978-1-4673-9026-2116 ©20 16 IEEE

[4] Giovanni Armano, Samuel Marchal, N. Asokan "Real-Time Client-Side Phishing Prevention Add-on" IEEE 36th International Conference on Distributed Computing Systems 1063-6927/16 © 2016 IEEE

[5] Ramesh Gowtham, Ilango Krishnamurthi "A comprehensive and efficacious architecture for detecting phishing webpages" researchgate.net/publication/259118063

[6] Yasin Sonmez, TurkerTuncer, based HuseyinGokal, EnginAvci, "Phishing Web Sites Features Classification Based On Extreme Learning Machine " IEEE 2018 6th International Symposium on Digital Forensic and Security (ISDFS), DOI: 10.1109/ISDFS.2018.8355342

[7] X. Chen, I. Bose, A. C. M. Leung and C. Guo, "Assessing the severity of phishing attacks: A hybrid data mining approach", Decision Support Systems, 50(4), pp.662-672, 2011

[8] Yi, P., Guan, Y., Zou, F., Yao, Y., Wang, W., & Zhu, T. (2018). Web phishing detection using a deep learning framework. *Wireless Communications and Mobile Computing, 2018*.

[9] Singh, C. (2020, March). Phishing Website Detection Based on Machine Learning: A Survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 398-404). IEEE.

[10] Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2020). Intelligent phishing detection scheme using deep learning algorithms. *Journal of Enterprise Information Management*.