Anatolian Journal of Computer Sciences

Volume: 2 No: 1 2017 © Anatolian Science

pp:15-36 ISSN:2548-1304

Effective Classification of Phishing Web Pages Based on New Rules by Using Extreme Learning Machines

Mustafa KAYTAN[a], Davut HANBAY[b]

[a]Computer Programming Program, Computer Technologies Department, Technical Sciences Vocational School, Harran University, 63250, Sanliurfa, TURKEY

[b]Computer Engineering Department, Engineering Faculty, Inonu University, 44280, Malatya, TURKEY

Abstract: Internet is an essential part of our life. Internet users can beaffectedfrom different types of cyber threats. Thus cyber threats may attack financial data, private information, online banking and e-commerce. Phishing is a type of cyber threats that is targeting to get private information such as credit cards information and social security numbers. There is not a specific solution that can detect whole phishing attacks. In this study, we proposed an intelligent model for detecting phishing web pages based on Extreme Learning Machine. Types of web pages are different in terms of their features. Hence, we must use a specific web page features set to prevent phishing attacks. We proposed a model based on machine learning techniques to detect phishing web pages.We have suggested some new rules to have efficient features. The model has 30 inputs and 1 output. In this application, the 10-fold cross-validation test has been performed. The average classification accuracy was measured as 95.05%.

Keywords — Machine Learning, Extreme Learning Machine, Phishing, Information Security.

1.Introduction

Information security threats have been seen and developed through time along development in the internet and information systems [1]. The impact is the intrusion of information security through the compromise of private data, and the victims may lose money or other kinds of assets at the end [2]. Internet users can be affectedfrom different types of cyber threats such as private information loss, identity theft, and financial damages [3]. Hence, using of the internet may suspect for home and official environments. Identify and defend against privacy leakage efficient analytical tools are required for users to reduce security threats [4]. Effective systemsthat can improve self-intervention must be formed using artificial intelligence-based information security management system at the time of an attack[5].

Phishing is an Internet-based attack that seduces end users to visit fake websites and give away personal information such as user id and password [6]. Phishing web pages are formed by fraudulent people to copy a web page from an original one. These phishing web pages are very similar to the original ones. Technical tricks and social engineering are extensively joined together for beginning a phishing attack [7]. An important view of online security is to protect users from phishing attacks and fake websites [8]. Intelligent methods can be used to develop fake web pages. For this reason internet users whether have enough experience in information security or not might be cheated. Phishing attacks can be launched via sending an e-mail that seems to be sent from a trusted public or private organization to users by attackers. Attackers get the users to update or verification their information by clicking a link within the e-mail. Other methods such as file sharing, blogs, and forums

can be used by attackers for phishing. There are many ways to fight phishing including legal solutions, education, and technical solution[9].A significant number of studies on the phishing have been done such as in [10], [11], [12].

Nowadays, information and communication tools are used in a manner that is very dense with information.For this purpose, various solution methods for various problem types have been developed.Machine Learning (ML) methods, can also be used in application development for information security.Optimization, classification, prediction and decision support system and great benefits can be provided to the person who is responsible for information security.Today, it has become an increasingly popular subject in developing intelligent applications.Non-intelligent application can cause losses in case the user is not required and can do a job that requires again.

There are attacks for different purposes to the Information and Communication tools that create computer networks.These attacks can be detected and the necessary precautions should be taken.For the study of artificial intelligence seems to gain speed as computer technology evolves.Artificial intelligence methods and studies on information security are increasing day by day.Intelligent systems provide great benefits in deciding to information security professionals[13].

ML methods can be used with classification purposes in various fields. Classification can be considered as a process to determine whether a data belong to one of the classes in the dataset organized according to certain rules. Classification which used in many fields and has an important place has a separate place for information security.

Neural nets models have been used in many areas such as data mining, medical applications, chemical industry, energy production, electrical and electronics industry, communications, nonlinear system modeling, pattern matching[14],[15].

In this study, an intelligent model for detecting phishing web pages based on Extreme Learning Machine is presented. We have suggested some new rules to have efficient features. The average classification accuracy as a result of the tests 95.05% evaluated. The paper is organized as follows, at first a brief of introduction for the study and related works about different phishing detection techniques are represented. Secondly the phishing threat, Extreme Learning Machines and details about the dataset that is used in intelligent model are summarized. Thirdly rules of used features and k-fold cross validation test briefly explained. Fourthly application of intelligent model is given in details. At last conclusions are given.

2.Related Works

With the development of Information and Communication Technology, various types of information security threats can be seen.These threats are important in the prevention of damage to person or institution to protect data on computer systems.Studies on various phishing detection methods have been seen when the literature is reviewed.In these studies, it is observed that ML is challenging techniques can be used.

Kaytan and Hanbay[13]proposeddetermining phishing websites based on neural network. UCI (University of California, Irvine) dataset was used for the study. 30 input attributes, and 1 output attribute were used for the experiment. The values 1, 0, and -1 were used for input attributes and the values 1, and -1 were used for output attribute. 5-fold cross validation method was used for evaluating the system performance. The best classification accuracy has beenmeasured as 92.45%. And the average accuracy has beenmeasured as 90.61%.

Santhana Lakshmi and Vijaya[16]used Machine-learning technique for modelling the prediction task and supervised learning algorithmsthat Multi-LayerPerceptron.Decision tree and Naïve bayes classifications were used for observing.It has been observed that the decision tree classifier predicts the phishing website more accurately then other learning algorithms.

Olivo et al.[17]proposed a methodthat yields the minimum set of relevant features providing reliability, good performance and flexibility for the phishing detection engine.It has been shown that the proposed method could be used to optimize the detection engine of the anti-phishing scheme.

Islam and Abawajy[18]proposed a new approach called multi-tier classification model for

phishing email filtering.It has a method for extracting the features of phishing email related to weighting of message content and message header and selects the features according to priority ranking.In addition, the impact of rescheduling the classifier algorithms in a multi-tier classification process to find out the optimum scheduling was examined.An empirical performance and analysis of the proposed algorithm have been presented.It has been shown that the proposed algorithm reduces the false positive problems substantially with lower complexity.

Chen et al.[19]evaluated intensity of phishing attacks in terms of risk levels and potential market value losses experienced by the target companies.It was analyzed 1030 phishing alerts released on a public database, and financial data related to thetargeted firms using a hybrid method.The severity of the attack was predicted with up to 89% accuracy usingtext phrase extraction and supervised classification.It has been identified some important textual andfinancial variables in the study.Impact the severity of the attacks and potential financial loss has been investigated.

Li et al.[20]proposed a novel approach based on minimum enclosing ball support vector machine (BVM) to detect phishing website.It has been aimed at achieving high speed and high accuracy to detect phishing website.Studies were done in order to enhance the integrity of the feature vectors.Firstly, an analysis of the topology structure of website was performed according to the Document Object Model (DOM) tree.Then, the web crawler was used to extract 12 topological features of the website.Later, the feature vectors weredetected by BVM classifier.The proposed method was compared to the DVM.It was observed that the proposed method has relatively high precision of detecting.In addition, it was observed that the proposed method complements the disadvantage of slow speed of convergence on large-scale data.It has been shown that the proposed method has better performance than SVM in the experimental results.The accuracy and validity of the proposed system has been evaluated.

Gowtham and Krishnamurthi[21]studied the characteristics of legitimate and phishing webpages in depth.Heuristics were proposed to extract 15 features from similar types of web pages based on the analysis.The proposed heuristic results were fed as an input to a trained machine learning algorithm to detect phishing websites.Before the applying the heuristics to the webpages, two preliminary screening modules were used in the system.By the preapproved site identifier that is the first module, webpages were checked against a private white-list maintained by the user.By the login form finder that is the second module, webpages were classified as legitimate when no login forms present.Unnecessary computation in the system was reduced by helping the used modules. Additionally, the rate of false positives without compromising on the false negatives was reduced by helping the used modules. By using the modules, webpages have been classified with 99.8% precision and a 0.4% of false positive rate.It has been shown that the proposed method is efficient for protecting users from online identity attacks.

Goh et al.[22]proposed amethodto improve the Web spam detection algorithms by including weight properties. Theymodified available Web spam detection algorithms with their methodon a Web spam dataset – WEBSPAM-UK2007 to measure the performances. The total performance observed up to 6.11% improvement at page level and 30.5% improvement at host level.The results showed the modified algorithms better than the benchmark algorithms.

Zhou et al.[23]studied on two topics. The first topic is about the computation of required thresholds to describe the three email groups. And the second topic is the interpretation of the cost-sensitive characteristics of spam filtering. Theyconsistently calculate the decision-theoretic rough set model based thresholds. The error rate of misclassification a legitimate email to spam is observed.And it has been seen that the new method reduces the error rate.The studyrepresents a better performance in order to the cost-sensitivity perspective.

3.Materials and Methods

3.1.Phishing

Phishing attack is one of the top ranked threats [24]so it has been addressed in this study.An overview of the threats that belong to 2014 and 2015, and the comparison of these threats are shown in the Table 1[24].

Table 1. An overview of the threats that belong to 2014 and 2015, and the comparison of these threats[24]

Ranking	Top Threats 2014	Trends 2014	Top Threats 2015	Trends 2015	Change in Ranking
1	Malicious code: Worms/Trojans	▲	Malware	▲	■
2	Web-based attacks	▲	Web-based attacks	▲	■
3	Web application/ Injection attacks	▲	Web application attacks	▲	■
4	Botnets	▼	Botnets	▼	■
5	Denial of service	▲	Denial of service	▲	■
6	Spam	▼	Physical damage/ theft/ loss	■	▲
7	Phishing	▲	Insider threat (malicious, accidental)	▲	▲
8	Exploit kits	▼	Phishing	■	▼
9	Data breaches	▲	Spam	▼	▼
10	Physical damage/ theft/ loss	▲	Exploit kits	▲	▼
11	Insider threat	■	Data breaches	■	▼
12	Information leakage	▲	Identity theft	■	▲
13	Identity theft/ fraud	▲	Information leakage	▲	▼
14	Cyber espionage	▲	Ransomware	▲	▲
15	Ransomware/ Rogueware/ Scareware	▼	Cyber espionage	▲	▼

Legend: ▲Increasing, ▼Decreasing, ■Same

In Fig.1.[25], industrial sectors that most exposed to phishing attacks are shown for the 3rd quarter of 2015.As can be seen from the Fig. 1, while the beginning of the sectors targeted by phishing attacks through the Internet Service Provider industry 24.35%, the financial sector is located in the 2nd with 20.43%.

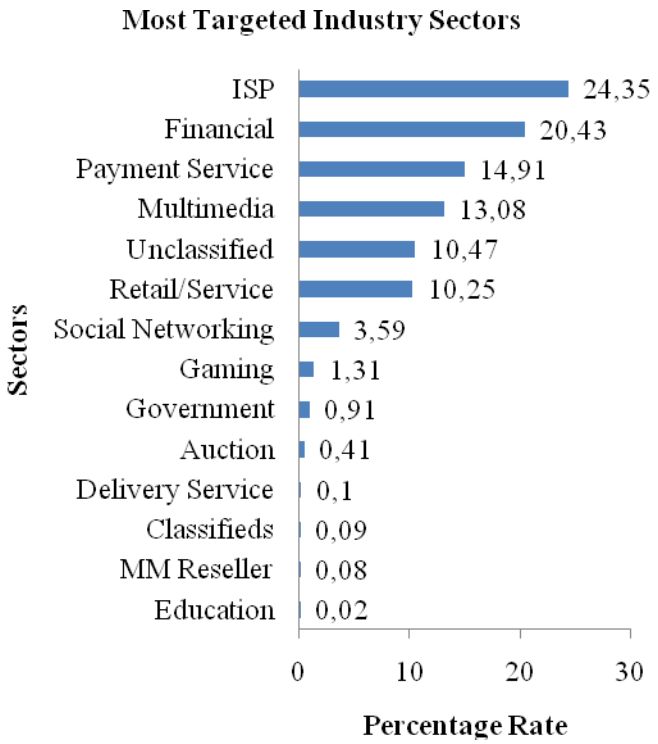

Fig.1. The industry sectors most exposed to phishing attacks through the 3rd quarter of 2015[25]

Countries with the Highest Infection Rates in phishing attacks for 3rd quarter of 2015 are shown in the Fig.0[25].As can be seen from the Fig. 2, while China is the country with the highest infection rate with 45.35% of phishing attacks, Turkey is located in the 2nd with 42.89%.

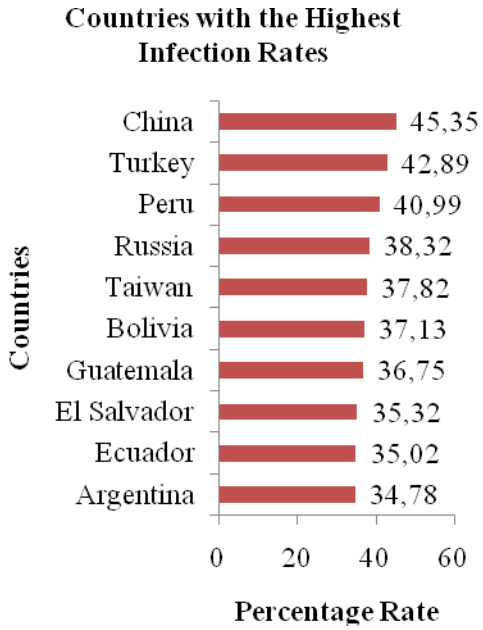

Fig.0. Countries with the Highest Infection Rates in phishing attacks for 3rd quarter of 2015[25]

3.2.Extreme Learning Machine

Huang et al.[26]have proposed a learning algorithm called as Extreme Learning Machine (ELM).The algorithm has been presented for Single-hidden Layer Feedforward Neural Network (SLFN).The netschoose the hidden nodes randomly and determinethe output weights analytically.Theoretically the algorithm tends to provide a good generalization performance.The algorithm also makes this performance with an extremely fast learning.The experimental results have shown that this algorithm provides good generalization performance.It has been expressed that the new algorithm could learn thousands of times faster than known algorithms for feedforward networks.

ELM is used for SLFN model[27]. ELM has been proposed as a single hidden layer feedforward neural network in primarily studies [26]. In later studies,the algorithm has been developed as a generalized feedforward neural network whichis not a single hidden layer in the neural network[28]. Weights of neurons in the input layer and threshold of neurons in the hidden layer of the neural network are generatedrandomly; weights of neurons in the output layer are measured analytically by ELM[29]. Hardlimit, Gaussian, and sigmoid etc. activation functions are used in the hidden layer, linear function is used in the output layer [30].

ELM has been implemented successfully in a wide range of applications such as document classification[31], bioinformatics [32], semantic concept detection [33], security assessment [34], face recognition [35], image super-resolution [36].

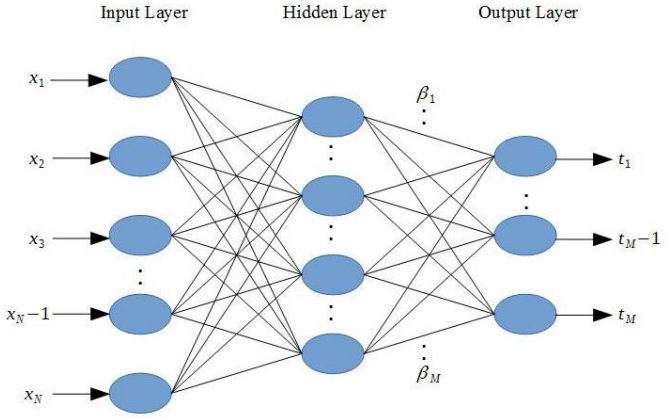

Fig. 3. Network model for ELM

For a standard SLFN (Single Layer Feed-forward Network) model in

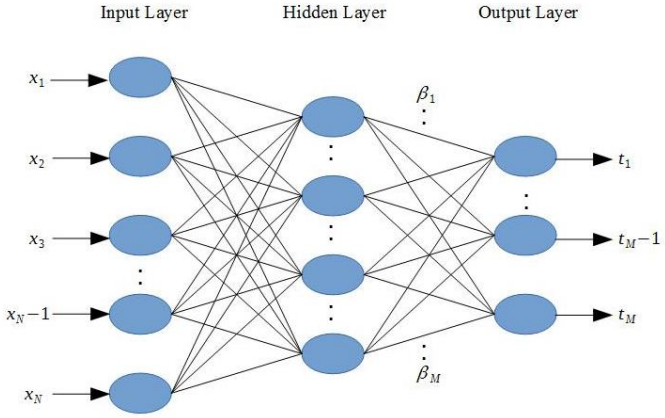

Fig**. 3.**, while, N: training samples, m: number of classes, $i = 1,2,\dots,N$, x_i: input, t_i: desired output,L: number of hidden layer nodes, $g(x)$: activation function, w_i: randomly chosen input weight vector between i'th hidden neuron and input neurons, β_i: weight vector between i'th hidden neuron and

output neurons, b_i: randomly chosen bias of i'th hidden node, o_j: actual output, $w_i . x_j$: inner product of w_i and x_j.With these parameters ELM can be mathematically modelled in Eq. 1[26]

$$\sum_{i=1}^{L} \beta_i g_i(x_j) = \sum_{i=1}^{L} \beta_i g(w_i . x_j + b_i) = o_j, \ j = 1,2, \dots, N \quad (1)$$

Output nodes are chosen linear.

It is aimed to minimize the relative error by ELM.

For this; it is expressed mathematically as;

$$\sum_{i=1}^{L} \beta_i g(w_i . x_j + b_i) = t_j, \ j = 1,2, \dots, N \quad (2)$$

Nequations in Eq. (2) can be mathematically modelled as

$$H\beta = T \quad (3)$$

where

H: hidden layer output matrix that named by Huang et al.[37]

$$H = \begin{bmatrix} g(w_1 . x_1 + b_1) & \cdots & g(w_L . x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(w_1 . x_N + b_1) & \cdots & g(w_L . x_N + b_L) \end{bmatrix}_{NxL} \quad (4)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{Lxm}$$

$$T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{Nxm}$$

Optimal output weight matrix can be obtained by solving Eq. (3) as follows:

$$\hat{\beta} = H^{\dagger}T \quad (5)$$

$H^{\dagger}$: Moore-Penrose generalized inverse[38] of hidden layer output matrix H.

ELM algorithm:

Step 1: randomly generate hidden node parameters and randomly assign hidden nodes.

(w_i, b_i), $i = 1,2, \dots, L$

Step 2: calculate hidden layer output matrix.

H

Step 3: calculate output weight matrix.

$\hat{\beta} = H^{\dagger}T$

3.3.Dataset

According to the source number [39], phishing websites dataset has been collected mainly from PhishTank archive, MillerSmiles archive and Google search operators.One of the challenges that are encountered in the study has been the lack of reliable training data sets.This challenge has been a difficulty faced by any researchers who want to work in this area.Recently many studies have been done on the prediction of phishing websites.In spite of this, a reliable training data set has not been published so far.For this reason, a consensus on the defining features as possible that identifies phishing webpages has not been achieved.Therefore, it has not seen easy to create a data set comprising all possible features.It has been focused on key features in the dataset.This dataset has been proven to be effective in predicting phishing websites.In addition to this dataset, some new features were proposed.It was contributed with the studies [40], [9] and [41]on the creation of this dataset and rules.

Table2. Top 10 samples in the dataset[43]

IP_address	URL_length	tiny_URL	@_symbol	//_symbol	prefix_suffix	sub_domain	SSL_state	domain_length	favicon	port	HTTPS_token	request_URL	anchor_URL	tags_links	SFH	email_submit	abnormal_URL	redirect	onMouseOver	right_click	popup_window	iframe	domain_age	DNS_record	web_traffic	Page_Rank	Google_Index	links_pointing	statistical_report	result

-1	1	1	1	-1	-1	-1	-1	-1	1	1	-1	1	-1	1	-1	-1	-1	0	1	1	1	1	-1	-1	-1	-1	1	1	-1	-1
1	1	1	1	1	-1	0	1	-1	1	1	-1	1	0	-1	-1	1	1	0	1	1	1	1	-1	-1	0	-1	1	1	1	-1
1	0	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	-1	-1	-1	-1	0	1	1	1	1	1	-1	1	-1	1	0	-1	-1
1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	0	-1	1	1	0	1	1	1	1	-1	-1	1	-1	1	-1	1	-1
1	0	-1	1	1	-1	1	1	-1	1	1	1	1	0	0	-1	1	1	0	-1	1	-1	1	-1	-1	0	-1	1	1	1	1
-1	0	-1	1	-1	-1	1	1	-1	1	1	-1	1	0	0	-1	-1	-1	0	1	1	1	1	1	1	1	-1	1	-1	-1	1
1	0	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	0	-1	-1	-1	0	1	1	1	1	1	-1	-1	-1	1	0	-1	-1
1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	-1	-1	1	1	0	1	1	1	1	-1	-1	0	-1	1	0	1	-1
1	0	-1	1	1	-1	1	1	-1	1	1	-1	1	0	1	-1	1	1	0	1	1	1	1	1	-1	1	1	1	0	1	1
1	1	-1	1	1	-1	-1	1	-1	1	1	1	1	0	1	-1	1	1	0	1	1	1	1	1	-1	0	-1	1	0	1	-1

"Phishing website features" and "training data set" files were used from the source [42]for this study.In addition to previous studies [40], [9], and [41]some features added and some corrections were made in "Phishing website features" file.

There are attributes and values for input dataset, class and values for output dataset and samples.Input dataset is consisting of 30 attributes.Input dataset attributes, according to the established rules can take the value 1, 0 or -1.In this way, the attributes of the generated input dataset can take 2 or 3 different values.Class in the output dataset can take the value 1 or -1.Result of the output dataset obtained in this way may take twodifferent values[43].

There are 11055 samples in the dataset.The first 10 samples are shown in the Table2.[43].Each row is shown for a sample in the dataset.The first 30 values in a row are represented for input data that are attributes.The last value in a row is represented for output data that is the result of the class.

A website is considered as legitimate, suspicious or phishing in the generated rules for input attributes of the dataset.A classification has been done for the output in the form of phishing or legitimate in the dataset.The values of 1 for legitimate, 0 for suspicious and -1 for phishing were used in this study. Table 3 shows the method that is applied to the values in the dataset.

Table 3. The method applied to the values in the data set

1.	2. Legitimate	3. Suspicious	4. Phishing
5. Input dataset values	6. 1	7. 0	8. -1
9. Output dataset values	10. 1	11. N/A	12. -1

For each of the attributes; the legitimate, suspicious and phishing number of samples have been identified in this study.Some of the attributes have two, some have three values are observed.The reason for this based on from the rules generated for the attributes.A total of 11055 samples are located in the dataset.The dataset is comprised of 30 attributes.The number of legitimate, suspicious, and phishingsamples are shown in Fig.4 for input dataset attributes.

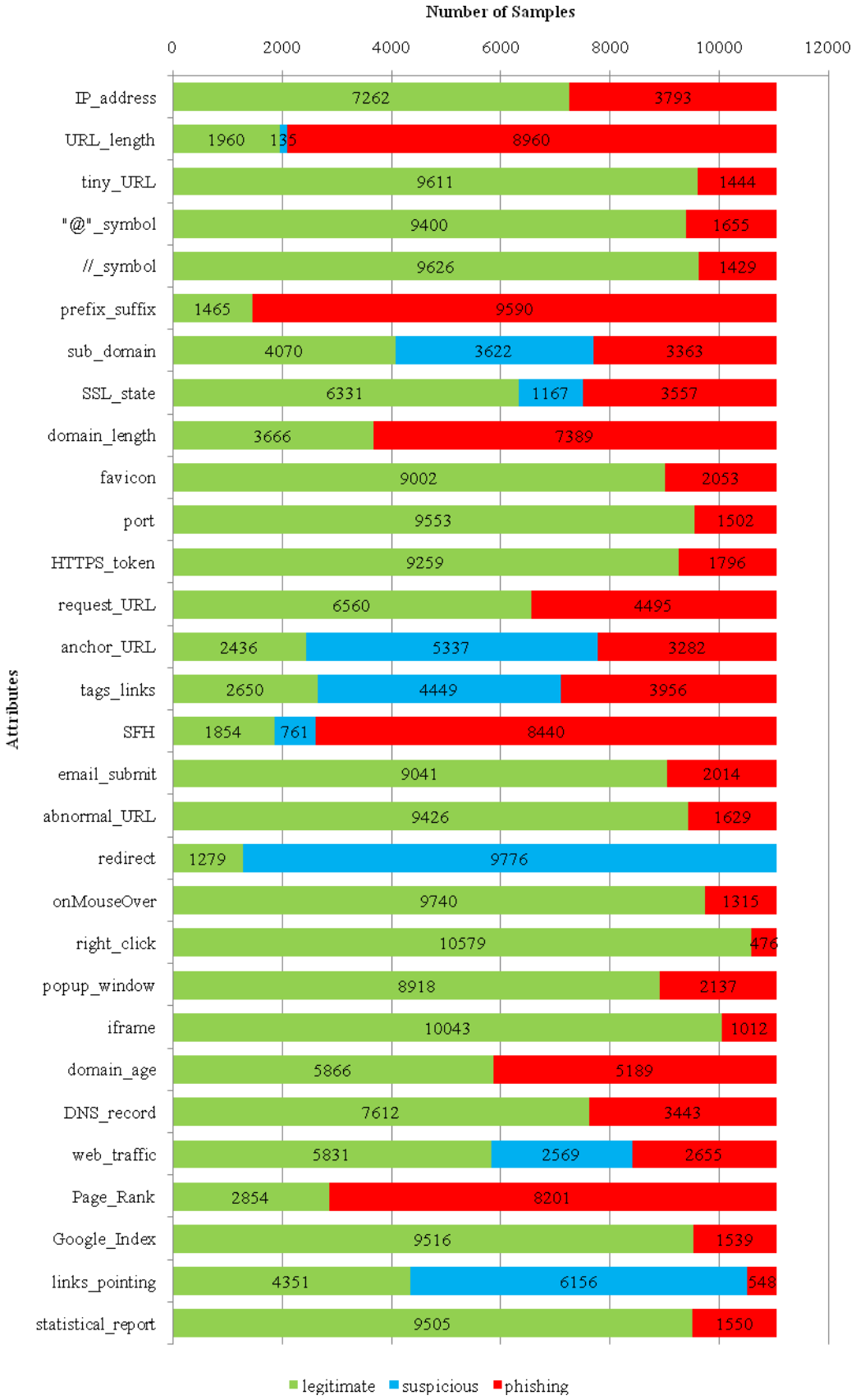

Fig.4. Number of legitimate, suspicious, and phishing for input dataset attributes.

For the result of classification; the legitimate, and phishing number of samples have been identified in this study.It is observed that the classification result of the two values.Classification results are made only for legitimate or phishing.Suspicious classification has not been included in the result of classification.The number of legitimate and phishing samples are shown in the Fig.5for output dataset classification result.

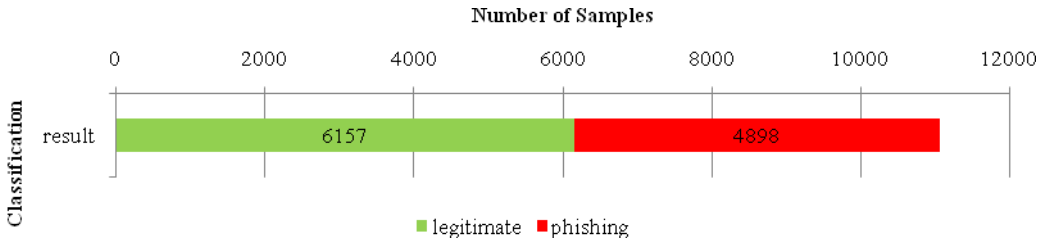

Fig.5. Number of legitimate and phishing for output dataset classification result.

There are totally 11055 samples in the dataset.However, because of the 10-fold cross validation test is applied to the data set; the first 11050 samples of the dataset were used in this study.The last 5 samples were not used in this study. TheTable provides some information on the dataset used in this study.90% of the dataset was used for the training dataset.Training dataset consists of 9945 samples in this way.In addition, 10% of the dataset was used for the test dataset.Test dataset consists of 1105 samples in this way.

Table 4. The dataset used for model

13. Dataset	**14. Rate (%)**	**15. Number of Samples**
16. Total	17. 100	18. 11050
19. Training	20. 90	21. 9945
22. Test	23. 10	24. 1105

3.4.Attributes and Rules

According to the ref. number [44] as one of the consequences of the problems, there wereno reliable training datasets.Those who work in this field have encountered difficulties arising from this issue.Despite these difficulties, many articles have been published related to the prediction of phishing websites that used recent data mining techniques.A reliable training dataset has not been published so far.As a result of possible causes, it has not reached a consensus in the literature on the specific properties that describe phishing websites.Because of this, it has not been easily seen to create a data set that covers all possible features.

It is intended to explain and prove the important feature, and prediction of the websites in the source number [44].In addition, some new features were proposed.Experimentally, it was appointed the new rules for some well-known properties.Updates were made to some other features.30 rules created for the attributes of the prepared data set examined.

*Using the IP address: Feature 1:****As an alternative, an IP address in the URL domain name can be used.Sometimes an IP address can be converted into radix 16 code[44].***

Rule:IP address exist in domain →phishing, otherwise → legitimate

*URLlength: Feature 2:****The average URL length has been calculated.If the number of URL characters is equal to 54 or greater than 54 then URL has been classified as phishing[44].***

Rule:URL length < 54 → legitimate, URL length ≥ 54 and ≤ 75 → suspicious, otherwise → phishing

*Using TinyURL: Feature 3:****URL length can be shortened and even a web page can be opened in this way.Short URL domain name, which depends on behalf of the Long URL domain,can be performed with HTTP Redirection [44].***

Rule: TinyURL is used → phishing, otherwise → legitimate

*Using "@" symbol: Feature 4:****It has been said that the previous part of "@" symbol in URL is ignored by the browser.It has been said that the next part of "@" symbol in URL is often the real address [44].***

Rule:URL has @ symbol → phishing, otherwise → legitimate

*Using "//" symbol: Feature 5:****The user may be directed to another web site using "//"in URL.If URL starts with "HTTP" then "//" symbol must be in the sixth position. If URL starts with "HTTPS" then "//" symbol must be in the seventh position [44].***

Rule:the last seen position of "//" symbol in URL > 7 → phishing, otherwise→ legitimate

*Using "-" symbol: Feature 6:****The dash symbol is rarely used in the legitimate URL.In this way users think that they are using a legitimate web page [44].***

Rule:"-" symbol exists in domain name → phishing, otherwise → legitimate

*Sub domain and multi sub domain: Feature 7:****"www." and country code in the URL are ignored.The remaining points are counted in the URL.If the number of dots is equal to 1 then web site has been classified as "legitimate".If the number of dots is equal to 2, then web site has been classified as "suspicious".If the number of dots is greater than2 then web site has been classified as "phishing"[44].***

Rule:number of dots in domain = 1 → legitimate, number of dots in domain = 2 → suspicious, otherwise → phishing

*Using HTTPS: Feature 8:****The authors [40], [41]have been suggested checking the certificate including HTTPS used, trusted certificate issuer, and the certificate age.It has been found that the minimum age of a certificate was 2 years [44].***

Rule: Using HTTPS, trusted security certificate providers, age of certificate ≥ 1 year → legitimate
Using HTTPS, untrusted security certificate providers → suspicious, otherwise → phishing

*Domain registration length: Feature 9:****It has been found that the longest fake domains have been used for one year only in the dataset [44].***

Rule:domains expires on ≤ 1 year → phishing,otherwise→ legitimate

*Favicon: Feature 10:****If a web page that contains the favicon is loaded from a domain different from the domain shown in the address bar, then the web page has been classified as "phishing" [44].***

Rule:favicon loaded from external domain → phishing, otherwise → legitimate

*Standard port status: Feature 11:****It has been investigated open or closed status of the service on a server with this feature[44].The port number, service name, description, and preferred status are shown in the Table5regarding some of the ports that are used in general.***

Table 5. General used ports

Port Number	Service Name	Description	Preferred Status
21	FTP	File Transfer Protocol	Close
22	SSH	Secure Shell	Close
80	HTTP	Hyper Text Transfer Protocol	Open
443	HTTPS	HTTP Secure	Open
445	SMB	Server Message Block	Close
3389	RDP	Remote Desktop Protocol	Close

Rule:port number is out of the preferred status → phishing, otherwise→ legitimate

*Using HTTPS token: Feature 12:**HTTPS token can be added to a part of domain of URL by attackers[44].***

Rule: Using HTTPS token in part of domain of URL → phishing, otherwise → legitimate

*Request URL: Feature 13:**Web page address and most of the objects which are embedded in web pages may share the same domain in a legitimate web page [44].***

Rule:% of request URL<22% → legitimate, % of request URL ≥22% and<61% → suspicious, otherwise → phishing

*URL of anchor: Feature 14:**Anchor has been identified as a member indicated by <a> tag.<a> tags and the web site may have different domain names.The anchor element may not be a connection to any web page[44].***

Rule:% of URL of anchor<31% → legitimate, % of URL of anchor ≥ 31% and ≤ 67% → suspicious, otherwise → phishing

*Links in <meta>, <script> ve <link>: Feature 15:**These tags are expected to be connected to the same domain on a web page.<meta> tag is used to retrieve metadata about the HTML (Hyper Text Markup Language) document recommendation.<script> tag is used to create client-side script.<link> tag is used to get other web resources [44].***

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

*Server Form Handler: Feature 16:**SFH (Server Form Handler) that contain an empty string or about: blank classified as "phishing".If the domain name in SFH is different from the domain name of the webpage then classified as "suspicious" [44].***

Rule: SFH is "about: blank" or empty → phishing, SFH refers to a different domain → suspicious, otherwise → legitimate

*Submitting information to e-mail: Feature 17**: A web form is used to send a user's personal information to a server."mail()" function can be used by using a server-side language and "mailto" can be used by using a client-side language [44].***

Rule:using "mail()" or "mailto:" → phishing, otherwise → legitimate

*Abnormal URL: Feature 18:****This feature could be extracted from the WHOIS database.Identity is typically part of its URL for a legitimate website [44].***

Rule: Host name is not in URL → phishing, otherwise → legitimate

*Website forwarding: Feature 19:****It has been found that legitimate websites are redirecting mostly once, andphishing websites are redirecting at least 4 timesin the dataset [44].***

Rule:number of redirect page ≤ 1 → legitimate, number of redirect page ≥ 2 and < 4 → suspicious, otherwise → phishing

*Status bar customization: Feature 20:****A fake URL can be displayed to the users in the status bar by the attackers.JavaScript can be used for this purpose.Especially "onMouseOver" event was focused on[44].***

Rule:onMouseOver changes status bar → phishing, otherwise → legitimate

*Disabling right click: Feature 21:****JavaScript can be used for this purpose.The source code of a web page could not be displayed and recorded by the user in this way."event.button==2" event has been investigated in a source code of webpage[44].***

Rule:right click disabled → phishing, otherwise → legitimate

*Using pop-up window: Feature 22:****Request to send the users' personal information in a pop-up window on a legitimate website is not regarded as a normal situation.This feature can be used in some legitimate websites for specific purposes[44].***

Rule:popup window contains text field → phishing, otherwise → legitimate

*Iframe redirection: Feature 23:****It has been said that to show an extra webpage the iframe tag is used[44].***

Rule:using iframe → phishing, otherwise → legitimate

*Age of domain: Feature 24:****This feature could be extracted from the WHOIS****[45]****database.It is observed thatan age oflegitimate domain is at least 6 months[44].***

Rule:age of domain ≥ 6 months → legitimate, otherwise → phishing

*DNS record: Feature 25:****An identity of phishing website is not recognized or no records are found for the host name in the WHOIS database[45],[46]. If the DNS (Domain Name System) record does not exist or has not been found, then website is classified as "phishing".Otherwise it is classified as "legitimate"[44].***

Rule:no DNS record for domain → phishing, otherwise → legitimate

*Website traffic: Feature 26:****This feature is measured interest in a website.Because of phishing websites live for a short period of time they may not be recognized by the Alexa database[47]. It was found that the legitimate websites are among the top in the ranking of 100.000.If the domain has no traffic or it is not recognized by the Alexa database,then it has been classified as "phishing".Otherwise it has been classified as "suspicious"[44]. The values of Alexa Traffic Ranks are shown for http://www.ucla.edu/ website in the Fig.6. The Traffic Ranking values were measured for Global and The United States in 2026 and 662 respectively.***

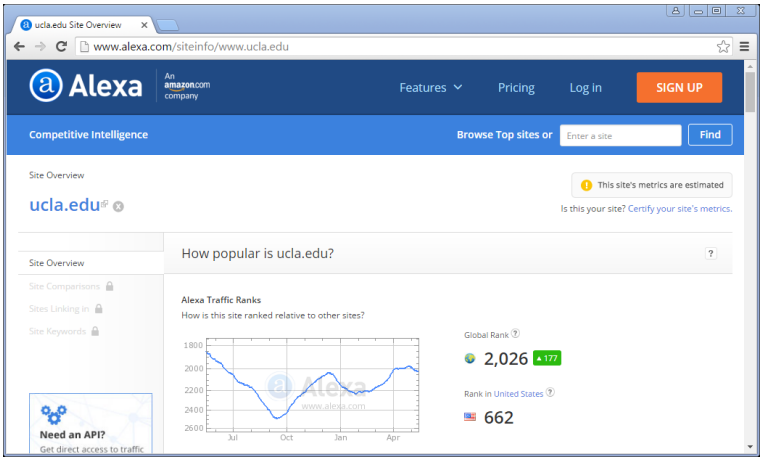

Fig.6. Alexa Traffic Ranks for ucla.edu

Rule:website rank < 100.000 → legitimate, website rank > 100.000 → suspicious, otherwise → phishing

*PageRank: Feature 27:**It has been said that PageRank is a value from 0 to 1.It has been found that 5% of phishing webpages may reach a PageRank value up to "0.2" [44]. The values between 0 and 1 in the PageRank algorithm, the values between 1 and 10 in the Google Toolbar PageRank tool are used [48].***

Example:
As aresult of searching the PageRank value of the web site was measured as 8 in Fig.7..

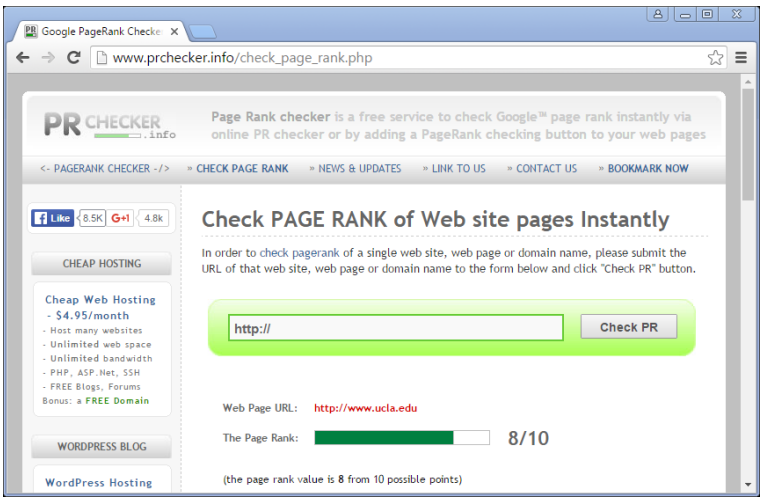

Fig.7. PageRank value for http://www.ucla.edu

Rule: PageRank < 0,2 → phishing, otherwise → legitimate

*Google Index: Feature 28:**A site is displayed on search results [49]when it is indexed by Google.Because of phishing webpages that can be accessed for a short period generally, many phishing webpages may not be found in the Google Index[44].***

Example:
In the

Fig.**8.**,search result page for info:www.ucla.edu is seen on Google.Five different information linksare shown for the URL on the page.As an alternative to thus links, cache:, related:, link:, site: operators can also be used respectively.Expression “www.ucla.edu” (including “”) is searched with the last link.

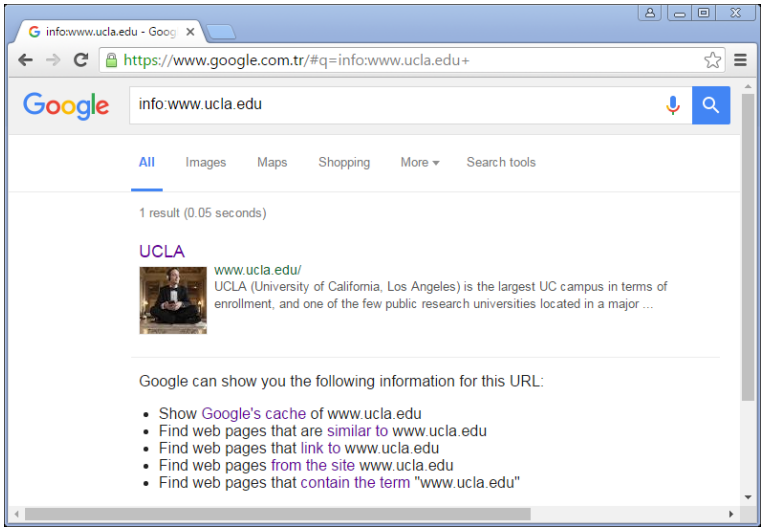

Fig.8. The information about an URL on Google

Rule:webpage indexed by Google → legitimate, otherwise → phishing

*Number of links pointing to page: Feature 29:**This feature has been defined about legitimate level even if some links are on the same domain[50].It has been observed that legitimate websites have at least 2 external links pointing to them in the dataset[44].***

Rule:number of links pointing to webpage = 0 → legitimate, number of links pointing to webpage> 0 and ≤ 2 → suspicious, otherwise → legitimate

*Statistical reports: Feature 30:**Many statistical reports on phishing websites have been defined for period of times by PhishTank[51] and StopBadware[52].Two types of top ten of PhishTank have been used in the study.These types are “top 10 domains” and “top 10 IPs”.“Top 50 IPs” of StopBadware have been used[44].***

Rule:host in top 10 phishing IPs or domains → phishing, otherwise → legitimate

3.5.k-foldCross Validation Test

As a result of the operation of a system, the accuracy of the results can be measured, and the success of the system can be evaluated.One of the methods used for this purpose is k-fold cross validation test.A dataset is divided into the sub datasetswhere each subset contains same number of data.If the dataset is represented as V, and the number of sub datasets is represented as k, then the sub datasets can be represented as V_1, V_2, V_3,…, Vkrespectively.For the implementation of the method, one of the sub datasets is selectedand this sub dataset is used as test datawhile the rest of the sub datasets are used as training data.This implementation is repeated for all other sub datasets.In this way, the number of k as the measure of success is obtained.To calculate the overall success of the system, the average is taken[13].

4.System Architecture

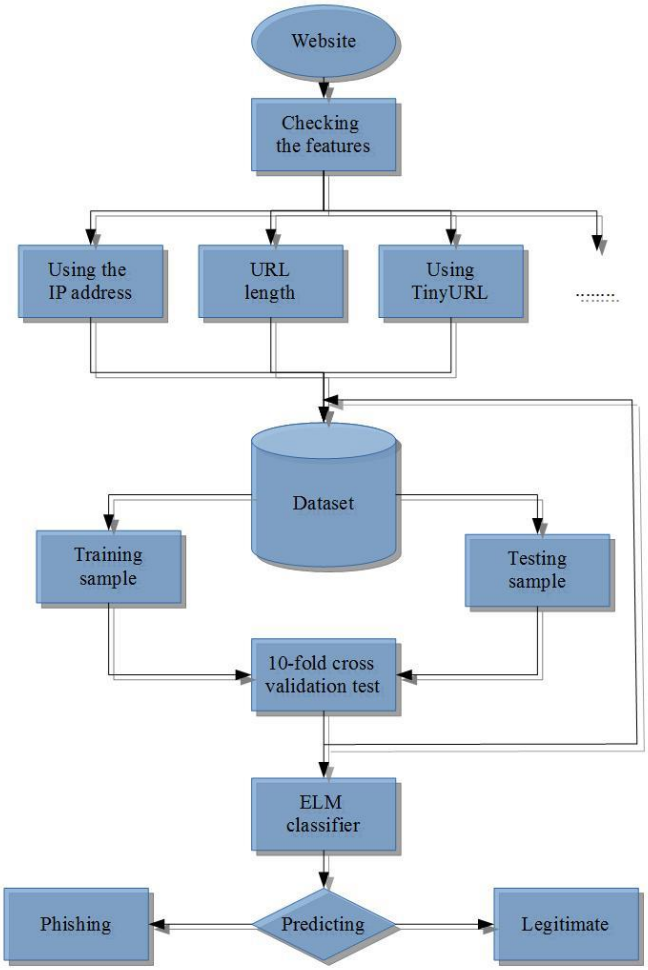

Fig. 9. System architecture for predicting a phishing website

General system architecture for prediction of a website is shown in Fig. 9. An ELM based phishing website classification procedure can be demonstrated as follows:

Step1.Visiting a website or a webpage.
Step 2.Checking the 30 input attributes according to the features and their rules.
Step 3.Collecting samples to the dataset.
Step 4.Randomly chosen 90% training samples and 10% testing samples of the dataset.
Step 5.Classification by using ELM.
Step 6.Prediction for phishing or legitimate.

5.Application

The purpose of this application is to create an intelligent system for detecting fake websites which one of the cyber-threats and is known as phishing.For this purpose, a model was created.

A model was developed for the detection of phishing websites using ELM.In this study, the average classification accuracy was observed as 95.05%.

5.1.Proposed Rules

A website is classified as legitimate, suspicious and phishing in the rule which is created for Feature 13[44]. However, in this study, it has been seen that the attribute had the values of 1, and -1 as a result of the investigation in the dataset.Feature 13 “Request URL” consists of a total of 11055 samples which are 6560 legitimate and 4495 phishing.Therefore, a new rule has been proposedfor the Feature 13.This proposed rule was used in this study.

A website is classified as legitimate, suspicious and phishing in the rule which is created for Feature 19[44]. However, in this study, it was seen that the attribute had the values of 1, and 0 as a result of the investigation in the dataset.Feature 19 that is “Website Forwarding” consists of a total of 11055 samples which are 1279 legitimate and 9776 suspicious.Therefore, a new rule has been proposed for the Feature 19.This proposed rule was used in this study.

Request URL: Feature 13:
Rule:% of request URL < 22% → legitimate, % of request URL ≥ 22% and <61% → suspicious, otherwise → phishing
Proposed Rule:% of request URL < 22% → legitimate, otherwise → phishing

Websiteforwarding: Feature 19:
Rule:number of redirect page ≤ 1 → legitimate, number of redirect page ≥ 2 and < 4 → suspicious, otherwise → phishing
Proposed Rule:number of redirect page ≤ 1 → legitimate, otherwise → suspicious

5.2.Model Parameters

ELM model parameters and descriptions within the scope of this study have been given in the Table 6.

Table 6.ELM model parametersanddescriptions

25. Parameter	**26. Description**
27. Network type	28. Feed forward
29. Number of layers	30. 3
31. Number of neurons for input layer	32. 30
33. Number of neurons for hidden layer	34. 850
35. Number of neurons for output layer	36. 2
37. Activation function	38. Hyperbolic tangent
39. Number of testing data	40. 1105
41. Number of training data	42. 9945
43. Number of total data	44. 11050

10-fold cross validation test was implemented to verify the system.90% of the dataset wasused for training, and 10% of the dataset was used for testing.This ratio was used for training and testing without changing other portions of the dataset.

5.3.Performance Results

In this study, a model was generated by the ELM.Average classification accuracy was observed as 95.05%.The highest classification accuracy is obtained with 95.93%.Classification accuracy rates are shown in the Table 7.

Table 7. Rates of accuracy and error of classification of neural networks.

Neural Net	**Accuracy (%)**	**Error (%)**
1	95,1131	4,16290
2	95,1131	4,34389
3	95,9276	4,34389
4	94,2986	5,79186
5	95,6561	3,61991
6	95,2036	5,24887
7	94,1176	5,79186
8	95,0226	5,06787
9	94,8416	5,06787
10	95,2036	4,79638
Average	**95,0498**	**4,82353**

Conclusions

Systems varying from dataentry to information processing applications can be made through websites.The entered information can be processed; the processed information can be obtained 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 webpage can be imitated by phishing attacks and using various methods.Some information such as users' credit card information, identity information can be obtained with these fake websites or webpages.

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.Extreme Learning Machine is used for this purpose.In this study, we used a data set taken from UCI website.In this dataset, input attributes are determined in 30, and the output attribute is determined in 1.Input attributes can take 3 different values which are 1, 0, and -1.Output attribute can take 2 different values which are 1, and -1.10-fold cross validation test has been implemented for measuring the performance of generated system in this study.As a result of the study, the average classification accuracy was measured as 95.05%, and the highest classification accuracy was measured as 95.93%.

When the dataset is examined, it has been observed that the rule created for feature 13 where are classified in the form of legitimate, suspicious, and phishing[44].When the dataset was examined by us, it was observed that 13th attribute values were consisted of 1 and -1.It was detected by us that the 13th attribute has 6560 legitimate and 4495 phishing samples which are 11055 samples totally.

For this reason, a new rule has been proposed by us for the Feature 13.This proposed rule is used for applications in this study.

When the dataset is examined, it has been observed that the rule created for feature 19 where are classified in the form of legitimate, suspicious, and phishing[44].When the dataset was examined by us, it was observed that 19th attribute values were consisted of 1 and 0.It was detected by us that the 19th attribute has 1279 legitimate, and 9776 suspicious samples which are 11055 samples totally. For this reason, a new rule has been proposed by us for the Feature 19.This proposed rule is used for applications in this study.

Funding:
This study has not been funded by any institution.

Conflict of Interest:
No conflict exists.
Author Davut HANBAY declares that he has no conflict of interest. Author Mustafa KAYTAN that he has no conflict of interest.

Ethical approval:
This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] G. Spanos and L. Angelis, "The impact of information security events to the stock market: A systematic literature review", Computers & Security, 58, pp.216-229, 2016.

[2] M. Aburrous, M. Hossain, K. Dahal and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining", Expert Systems with Applications, 37(12), pp.7913-7921, 2010.

[3] N. Abdelhamid, A. Ayesh and F. Thabtah, "Phishing detection based Associative Classification data mining", Expert Systems with Applications, 41(13), pp.5948-5959, 2014.

[4] S. Wu, P. Wang, X. Li and Y. Zhang, "Effective detection of android malware based on the usage of data flow APIs and machine learning", Information and Software Technology, 75, pp.17-25, 2016.

[5] M. Kaytan and D. Hanbay, "Kurumsal Bilgi Güvenliğine Yönelik Tehditler ve Alınması Önerilen Tedbirler", 1st International Symposium on Digital Forensics and Security, ISDFS'13, pp.267-270, 2013, Fırat University, Elazığ.

[6] H. Shahriar and M. Zulkernine, "Trustworthiness testing of phishing websites: A behavior model-based approach", Future Generation Computer Systems, 28(8), pp.1258-1271, 2012.

[7] R. M. Mohammad, F. Thabtah and L. McCluskey, "Tutorial and critical analysis of phishing websites methods", Computer Science Review, 17, pp.1-24, 2015.

[8] M. Alsharnouby, F. Alaca and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks", International Journal of Human-Computer Studies, 82, pp.69-82,

2015.

[9] R. M. Mohammad, F. Thabtah and L. McCluskey, "Predicting phishing websites based on self-structuring neural network", Neural Computing and Applications, 25(2), pp.443-458, 2014.

[10] V. Ramanathan and H. Wechsler, "Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation", Computers & Security, 34, pp.123-139, 2013.

[11] I. R. A. Hamid and J. H. Abawajy, "An approach for profiling phishing activities", Computers & Security, 45, pp.27-41, 2014.

[12] C. Konradt, A. Schilling and B. Werners, "Phishing: An economic analysis of cybercrime perpetrators", Computers & Security, 58, pp.39-46, 2016.

[13] M. Kaytan and D. Hanbay, "The Determining with Artificial Neural Network Based Intelligent System Against The Attacks to The Internet Sites by Phishing Method", International Conference on Natural Science and Engineering, ICNASE'16, pp.3221-3226, 2016, Kilis 7 Aralık University, Kilis.

[14] D. Hanbay, I. Turkoglu and Y. Demir, "An expert system based on wavelet decomposition and neural network for modeling Chua's circuit", Expert Systems with Applications, 34(4), pp.2278-2283, 2008.

[15] D. Hanbay, I. Turkoglu and Y. Demir, "Modeling switched circuits based on wavelet decomposition and neural networks", Journal of the Franklin Institute, 347(3), pp.607-617, 2010.

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

[17] C. K. Olivo, A. O. Santin and L. S. Oliveira, "Obtaining the threat model for e-mail phishing", Applied Soft Computing, 13(12), pp.4841-4848, 2013.

[18] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach", Journal of Network and Computer Applications, 36(1), pp.324-335, 2013.

[19] 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.

[20] Y. Li, L. Yang and J. Ding, "A minimum enclosing ball-based support vector machine approach for detection of phishing websites", Optik, 127(1), pp.345-351, 2016.

[21] R. Gowtham and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages", Computers & Security, 40, pp.23-37, 2014.

[22] K. L. Goh, R. K. Patchmuthu and A. K. Singh, "Link-based web spam detection using weight properties", Journal of Intelligent Information Systems, 43(1), pp.129–145, 2014.

[23] B. Zhou, Y. Yao and J. Luo, "Cost-sensitive three-way email spam filtering", Journal of Intelligent Information Systems, 42(1), pp.19–45, 2014.

[24] ENISA Threat Landscape 2015, European Union Agency for Network and Information Security (ENISA), Ocak 2016.

[25] Phishing Activity Trends Report, Anti Phishing Working Group (APWG), 1st-3rd Quarters 2015.

[26] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: Theory and applications", Neurocomputing, 70(1-3), pp.489-501, 2006.

[27] M. Luo and K. Zhang, "A hybrid approach combining extreme learning machine and sparse representation for image classification", Engineering Applications of Artificial Intelligence, 27, pp.228-235, 2014.

[28] G.-B. Huang and L. Chen, "Convex incremental extreme learning machine", Neurocomputing, 70(16-18), pp.3056-3062, 2007.

[29] J. Tang, C. Deng, G.-B. Huang and B. Zhao, "Compressed-Domain Ship Detection on Spaceborne Optical Image Using Deep Neural Network and Extreme Learning Machine", IEEE Transactions on Geoscience and Remote Sensing, 53(3), pp.1174-1185, 2015.

[30] G.-B. Huang, "An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels", Cognitive Computation, 6(3), pp.376-390, 2014.

[31] X.-g. Zhao, G. Wang, X. Bi, P. Gong and Y. Zhao, "XML document classification based on ELM", Neurocomputing, 74(16), pp.2444-2451, 2011.

[32] G. Wang, Y. Zhao and D. Wang, "A protein secondary structure prediction framework based on the Extreme Learning Machine", Neurocomputing, 72(1-3), pp.262-268, 2008.

[33] B. Lu, G. Wang, Y. Yuan and D. Han, "Semantic concept detection for video based on extreme learning machine", Neurocomputing, 102, pp.176-183, 2013.

[34] Y. Xu, Z. Y. Dong, J. H. Zhao, P. Zhang and K. P. Wong, "A Reliable Intelligent System for Real-Time Dynamic Security Assessment of Power Systems", IEEE Transactions on Power Systems, 27(3), pp.1253-1263, 2012.

[35] K. Choi, K.-A. Toh and H. Byun, "Incremental face recognition for large-scale social network services", Pattern Recognition, 45(8), pp.2868-2883, 2012.

[36] L. An and B. Bhanu, "Image Super-Resolution by Extreme Learning Machine", 19th IEEE (Institute of Electrical and Electronics Engineers) International Conference on Image Processing (ICIP), pp.2209-2212, 2012, Orlando, ABD.

[37] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks", Proceedings of IEEE (Institute of Electrical and Electronics

Engineers) International Joint Conference on Neural Networks, 2, pp.985-990, 2004.

[38] D. Serre, Matrices:Theory and Applications, Springer, New York, 2002.

[39] İnternet:http://archive.ics.uci.edu/ml/datasets/Phishing+Websites#, Accessed 24 03 2016.

[40] R. M. Mohammad, F. Thabtah and L. McCluskey, "An Assessment of Features Related to Phishing Websites using an Automated Technique", The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012), pp.492-497, 2012, London.

[41] R. M. Mohammad, F. Thabtah and L. McCluskey, "Intelligent rule-based phishing websites classification", IET Information Security, 8(3), pp.153-160, 2014.

[42] İnternet:http://archive.ics.uci.edu/ml/machine-learning-databases/00327/, Accessed 24 03 2016.

[43] İnternet:http://archive.ics.uci.edu/ml/machine-learning-databases/00327/Training%20Dataset.arff, Accessed 24 03 2016.

[44] İnternet:http://archive.ics.uci.edu/ml/machine-learning-databases/00327/Phishing%20Websites%20Features.docx, Accessed 24 03 2016.

[45] İnternet:http://who.is/, Accessed 19 04 2016.

[46] Y. Pan and X. Ding, "Anomaly Based Web Phishing Page Detection", 22nd Annual Computer Security Applications Conference (ACSAC'06), IEEE (Institute of Electrical and Electronics Engineers) Conference Publications, pp.381-392, 2006, Miami Beach, Florida, USA.

[47] İnternet:http://www.alexa.com/, Accessed 14 04 2016.

[48] İnternet:https://en.wikipedia.org/wiki/PageRank, Accessed 22 04 2016.

[49] İnternet:https://support.google.com/webmasters/answer/40052?hl=en, Accessed 14 04 2016.

[50] İnternet:http://backlinko.com/google-ranking-factors, Accessed 14 04 2016.

[51] İnternet:http://www.phishtank.com/stats.php, Accessed 19 04 2016.

[52] İnternet:https://www.stopbadware.org/top-50, Accessed 19 04 2016.