International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Phishing Website Detection With Semantic Features Based on Machine Learning Classifiers:
A Comparative Study

Ammar Almomani, Research and Innovation Department, Skyline University College, UAE & IT Department, Al-Huson University College, Al-Balqa Applied University, Jordan*

https://orcid.org/0000-0002-8808-6114

Mohammad Alauthman, University of Petra, Jordan

https://orcid.org/0000-0003-0319-1968

Mohd Taib Shatnawi, Al-Balqa Applied University, Jordan

Mohammed Alweshah, Al-Balqa Applied University, Jordan

Ayat Alrosan, School of Information Technology, Skyline University College, UAE

Waleed Alomoush, School of Information Technology, Skyline University College, UAE

Brij B. Gupta, Department of Computer Engineering, National Institute of Technology Kurukshetra, India & Asia University, Taiwan

ABSTRACT

The phishing attack is one of the main cybersecurity threats in web phishing and spear phishing. Phishing websites continue to be a problem. One of the main contributions to the study was working and extracting the URL and domain identity feature, abnormal features, HTML and JavaScript features, and domain features as semantic features to detect phishing websites, which makes the process of classification using those semantic features more controllable and more effective. The current study used the machine learning model algorithms to detect phishing websites, and comparisons were made. The authors have used 16 machine learning models adopted with 10 semantic features that represent the most effective features for the detection of phishing webpages extracted from two datasets. The GradientBoostingClassifier and RandomForestClassifier had the best accuracy based on the comparison results (i.e., about 97%). In contrast, GaussianNB and the stochastic gradient descent (SGD) classifier represent the lowest accuracy results, 84% and 81% respectively, in comparison with other classifiers.

KEYWORDS

Machine Learning Models, Phishing Website, Semantic Classification, Semantic Features

1. INTRODUCTION

Phishing is an illegal tool used to identify information about customers' identity and financial institution passwords. Social engineering techniques employ spoofed e-mails from lawful companies and agencies. Those emails are designed to enable users to reveal financial data, including usernames

DOI: 10.4018/IJSWIS.297032 *Corresponding Author

and passwords on fake websites. Computer subterfuge programs place offenders on servers to deliberately access data by using devices that retrieve usernames or passwords from online accounts. Corrupt local browsers misdirect customers to fake websites (or legitimate Internet sites). They use pipe-controlled proxies to track and capture keystrokes by consumers(Al-Momani et al., 2011; Ammar Almomani et al., 2013; Ammar Almomani, Obeidat, Alsaedi, Obaida, & Al-Betar, 2015; Ammar Almomani, Wan, Altaher, et al., 2012; Ammar ALmomani, Wan, Manasrah, et al., 2012; A Almomani et al., 2013; B. B. Gupta, Arachchilage, & Psannis, 2018; B. B. Gupta, Tewari, Jain, & Agrawal, 2017)

Recently,phishing detection based on Semantic Link Network (SLN) and semantic features, semantically organizing web resources, identify a phishing web page and its phishing target, become most popular techniques in recent years(R. M. Mohammad & AbuMansour, 2017; Verma & Hossain, 2013; Wenyin, Fang, Quan, Qiu, & Liu, 2010).A significant number of our everyday activities (e.g. activities on social networks, online banking activities and electronic business activities) have been receiving much attention. That is attributed to the growth of world networking and communication technologies. The free, transparent and unrestricted internet infrastructure creates an attractive environment for cyber-attacks and critical network vulnerabilities, including seasoned software users. Although the user's knowledge and expertise are significant, users cannot completely stop the phishing scam (Al-Nawasrah, Almomani, Atawneh, & Alauthman, 2020; Alauthman, Almomani, Alweshah, Omoush, & Alieyan, 2019; A Almomani, Alauthman, Omar, & Firas, 2017)

Attackers often take into account the personality characteristics of the end-user to increase the effectiveness of phishing attacks. They consider these characteristics to trick the users who are relatively experienced(Alauthman et al., 2019). It should be noted that end-user-specific cyber-attacks cause massive losses in sensitive information and cash for individuals. Such loss is represented in billions of dollars each year (Alauthman, Aslam, Al-Kasassbeh, Khan, Al-Qerem, Choo, et al., 2020).

The metaphor used in the term (phishing attacks) is derived from 'fishing, fishing' for targets. Investigators have received a lot of attention in recent years.Carrying out phishing attacks is enticing and tempting for hackers, who open some fake websites that are built just like the common and legal websites on the internet. Although these sites have identical visual user interfaces, there is a need for URLs that are different from the URLs of the original page. A patient and a knowledgeable client can easily detect most of these malicious sites through browsing the URLs.

End-users most often forget to examine their entire website address, usually conveyed via other web pages, social-network apps or just through e-mails, as is defined in figure 1. A phisher aims to obtain confidential and personal information (e.g. financial data) through using such malicious URLs. When entering such a fraudulent website, users can simply enter their information without concern. That is because users assume that the website is legitimate

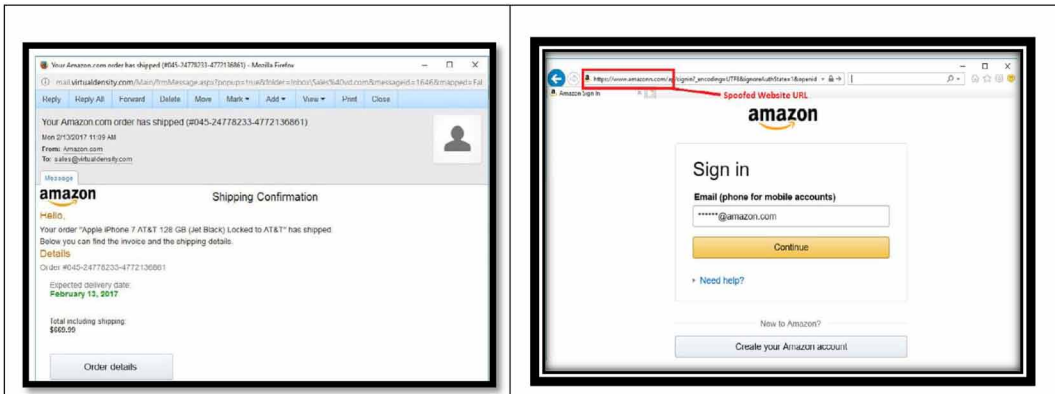
The client does not understand which web pages to trust, due to several reasons, which are: a)- users do not see a full web-page link, because of re-directions and hidden URLs. b)- users do not have time to consult the URL; and accidentally, users can not access those web pages. c)- Users can not discern phishing web pages from legitimate ones. d)- Users are unable to differentiate phishing pages from the phishing client due to the main reasons (Sahingoz, Buber, Demir, & Diri, 2019).

suggests that e-mail negatively affects connectivity and teamwork in business and daily life. E-mail-based abuse has been increasing and developing. The driving force rapidly behind rapidly developing socially engineered and increasingly dangerous email attacks are represented in the new generation of digital offender's organizations.

The world is expected to have cybercrime damage of 6 trillion dollars per year by 2021 (Alkhalil, Hewage, Nawaf, & Khan, 2021). According to the cybersecurity entry report, it is up from 3 trillion dollars in 2015 (Morgan, 2019). Phishing attacks are the most common kind of cybersecurity violation, as the official cybersecurity infringement statistics of the UK 2020 survey say that the attacks affect groups as well as people (Johns, 2020).

A report was published in the first quarter of 2020 by the Anti-Phishing Working Group about phishing attacks (APWG, 2020). The most significant class of phishing is a software as a service

Figure 1. Example of a web-page and E-mail examples Phishing Web Page(Sahingoz et al., 2019)



(SaaS). Webmail services are the most commonly used for phishing.The number of phishing locations identified by APWG increased significantly over the third and fourth quarters of 2018. Phone phishing has become more common in Brazil in recent years, with the phishers primarily targeting SaaS providers.Phishers, on the other hand, have infiltrated phishing firms. 36% of all phishing attacks occur through using payment services. Cybercriminals also use ransomware to hack the websites of several banks simultaneously.

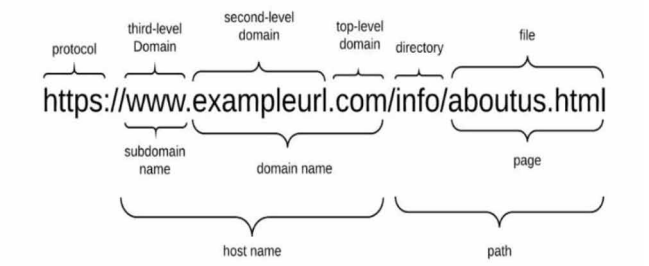
URLs and Attacker Tactics: Attackers employ a variety of tactics to prevent security systems or system administrators from detecting them. Some of these techniques are listed in this section. Firstly, components of URLs should be examined to explore the approach adopted by attackers. The fundamental structure of the URL is presented in figure 2.

In the default type, a URL begins with the name of the website protocol. The second level is a domain name (SLD), which typically refers to the server hosting the organization's name. It is located in the third level domain, and lastly, the top-level domain name (TLD), which refers to the domains of the Internet root zone of DNS. The last section is represented by the webpage's domain name. The internal address is shown by the server path and the HTML page name.

Since the name of the SLD typically refers to the operation form or business name, it is easy to locate the phishing attacker. The SLD name can only be described once at the start. Nevertheless, an attacker can create an infinite number of URLs with the path and file name extension of the SLD. That is because the inner address is determined by the attackers.

A combination of the TLD and the SLD (i.e. the domain name) is the most important component of a URL. Cybersecurity firms go to great lengths to identify the fake domains used to launch phishing attacks by name.The IP address should be blocked if a domain name is classified as phishing. In this

Figure 2. URL structure(Sahingoz et al., 2019)



case, one can't reach the web pages found there. The attacker employs essential methods to increase the attack efficiency and steal sensitive information. He/she employs essential methods to increase the victim's vulnerability. Such methods include random characters, mixed word use, cybersquatting and type-squatting. The mechanisms used for detecting such attacks take these methods into account (Sahingoz et al., 2019).

Annotations extracted from the quality of the tools are referred to as semantic features. We are inspired to demonstrate the utility of using semantic features to identify phishing websites. Semantic features enable us to determine conceptual similarity. This similarity enables us to organize, extract and group knowledge based on principles rather than subsequent or regularly expressed phrases. Semantic features are content-based metrics derived from the text (email, posts, and websites) using data-driven latent topic models. These latent topics are groups of words that appear frequently in the text. In a phishing email, we can expect the terms "click" and "account" to appear together, while in normal financial emails, the words "market," "prices," and "plan" can co-occur. Latent topic models create such features by leveraging the co-occurrence of words in a training collection of emails. Typical latent subject models do not account for several types of documents, such as phishing or non-phishing.

Most researchers proposed depending on semantic features, which use email content to detect phishing websites only. However, one of the main contributions to our study was working and extracting URL features as semantic features to detect phishing websites, which makes the semantic features more controlled and more effective for the classification process. Nevertheless, several strategies are used for detecting phishing attacks. For instance, machine learning is used in several fields to come up with automated solutions. Several articles have been written to shed more light on machine learning. Several approaches have been proposed by researchers to identify phishing attacks by using machine learning techniques. Through this article, we aimed to investigate training methods to come up with effective methods for identifying phishing sites. We aimed to shed light on aggregation analysis and its role in creating rules automatically to evaluate website format similarities and identify pages for phishing. However, this paper aims to investigate the possibility of using the semantics of URL features based on AI classifier models.

The structure of the following parts is presented in the following: Sec.2: It presents some related works. These studies shed light on the methods for detecting phishing web pages and their role. Sec.3: It presents a comparison study based on 16 AI algorithms to detect phishing websites. Sec.4: It presents data about the experiments. It presents the results. Sec. 5: It presents the conclusion and the implications.

2. RELATED WORK

This section presents several approaches adopted by researchers for detecting phishing sites. Instead of focusing on external web functions such as URLs, the researchers focused on Phishing detection that is based on list based detection systems semantic website features and Semantic machine learning algorithm.

2.1. List-based Detection Systems

List-based phishing identification mechanisms use white and blacklists. Blacklists are generated by URLs that are recognized as phishing sites. These sources include spam detection systems, user alerts and third-party companies. Using blacklists prevents attackers from attacking again by using the same IP address or URL. The protection framework updates the black-list through using malicious IPs detection methods, or users can obtain the malicious IPs from the server immediately and support the systems to defeat the attacks. However, the blacklist-based approach can't detect a real attack as a zero-day attack. Based on the false-positive rate, the performance of these attack detection systems is lower than the machine-based learning methods in terms of the false-positive rate(Jain & Gupta, 2021).

Approximately, 20% of the black-list phishing attack detection programs are effective(Khonji, Iraqi, & Jones, 2013). Thus, the blacklist-based approach isn't effective for detecting attacks. Many organizations have black-list based phishing attack detection systems (e.g. Google Safe Browsing API, and PhishNet)(Prakash, Kumar, Kompella, & Gupta, 2010). Service: These systems use a similar approximate algorithm for checking whether the suspicious URL is listed in the black-list or not. Black-list strategies need to be revised regularly. Moreover, the fast growth in the black-list items requires having many machine resources(B. B. Gupta et al., 2021).

For the detection of legal web pages and phishing, whitelist-based phishing detection systems provide information about secure and legitimate websites. Any website not included in the whitelist is considered suspicious. In Ref. (Cao, Han, & Le, 2008), the researcher developed a white-list program that shares each site's IP address with the Login user on the other side accessed by the user. When the user visits a website, the method warns the user if the registration data of the website is incompatible. However, in case the legitimate site is visited by the customer for the first time, this approach is considered ineffective. In ref. (Jain & Gupta, 2016), the researcher developed a system that alerts web users to an automatically modified white-list of legitimate websites. The latter system consists of two phases (i.e. the modules for matching the domain IP address and eliminating connection features in the source code). In this study, the real positive value that is achieved is 86.02%. In this study, the false-negative rate is 1, 48%. Those rates were reached through the experiment.

2.2. The Phishing Detection Methods That are Based on Pagefeatures:

CANTINA (Y. Zhang, Hong, & Cranor, 2007)detects phishing pages based on "term frequency-inverse document frequency (TF-IDF)." (Dunlop, Groat, & Shelly, 2010)are used to obtain site data. It employs optical feature recognition. Then, it employs search engines to evaluate if the quality of the website refers to the page content or detects phishing sites. Presents a general semantic text problem selection method, based on the statistical t-test and WordNet, and show its efficiency in the detection of phishing e-mail by designing classification systems that combine semantics and statistics in email text analysis(Verma & Hossain, 2013).

In ref.(W. Zhang, Lu, Xu, & Yang, 2013), the researcher used the spatial website templates and used as a guide to evaluating the resemblance of the site. In ref.(Moghimi & Varjani, 2016), the researchers discovered a rules-based system by utilizing two different feature sets for internet banking phishing. One of the feature sets is used to determine the identification of site services. The other feature set is used to define a protocol for entry. The phishingalarm employs powerful and reliable CSS design tools for the identification of websites for phishing. The article focuses on the way of learning automatically from CSS functionality to distinguish different pages to identify new features as the basis for phishing sensing. Through carrying out the analysis, it was found that the NON-linear HS-based regression generates results that are better than the results of SVM. A framework is proposed for identifying phishing websites that employ the non-linear meta-heuristic, non-linear regression algorithm and implement a function choice strategy. To identify the websites with URL and HTML capabilities, the stacking template is used. As for usability, the device develops lightweight URLs, HTML and HTML without the use of third-party providers. Ref(Adebowale, Lwin, Sanchez, & Hossain, 2019)presents an Adaptive Neuro-Fuzzy Inference System (ANFIS) that is based on a robust scheme. It employs script, photographs, and frames optimized technologies for identifying web-phishing.

2.3. Phishing Detection Methods Are Based On Machine Learning Algorithms.

Machine learning has been used to detect phishing e-mail and web page. In ref.(Xiang, Hong, Rose, Cranor, & Security, 2011), the researchers proposed CANTINA+ Take 15 elements, including HTTP, DOM, Third Party Companies, HTML User Object Model and search engines. Such apps were trained to detect phishing attacks by using a Support Vector Machine (SVM). In ref.(Abdelhamid, Ayesh, & Thabtah, 2014; Anupam & Kar, 2021), the researchers proposed an associative phishing identification

system that is based on multi-label ranking for the website. In ref.(Jain & Gupta, 2019), the researchers presented a method to detect phishing attacks by investigating the hyperlinks included within the HTML source code of the website. The suggested solution includes several different hyperlinks to prevent phishing attacks.

In ref. (Islam & Abawajy, 2013)the researchers took into account the title of the message and priority rating of the incoming message. To filter the post, they established a multi-layered classification method. Based on the experimental tests, the false-positive rate was calculated. The researchers identified the extracted functions for transport layer protection and the URL features like length, number and location of the slashes in URLs and subdomain names. Jeeva and Rajsingh extracted features. Apriori algorithm was employed to set rules for the detection process by using the extracted features. Research has shown that 93% of Phishing URLs were detected.

(Le, Markopoulou, & Faloutsos, 2011)developed a method for detecting phishing websites through using URLs features (e.g. unique character number, and domain and file names). The researchers used a support vector machine for offline classification. Weighted confidence and online perceptron are employed for online detection. Based on the results, the application of adaptive regularity increases the classification efficiency level.(Selvakumari, Sowjanya, Das, & Padmavathi, 2021).

(Sahingoz et al., 2019)developed a phishing detection method by employing 209-phrase vectors and 17NLP features. The influence of the NLP features was observed in the detailed analysis. However, vectors of the word and NLP features number must be increased. Therefore, the researchers of the present study emphasized this problem. The accuracy rate is 7%. In addition, the researchers extended the research by comparing three separate machine learning algorithms in terms of accuracy.

(Babagoli, Aghababa, & Solouk, 2019) have recently used phishing for non-linear regression. They prefer harmony quest and the use of metaheuristic algorithms for the training of the network port vector machine. According to those researchers, the rates of accuracy in train and check processes are 94.13 and 92.80%, respectively, when about 11.0 0 0 web pages are used.

While most researchers are concerned with URL phishing detection, some researchers aimed to detect phishing e-mails by analyzing the data contained in e-mail packages. (Smadi, Aslam, & Zhang, 2018)merged the methodology of the neural network with classification reinforcement learning to detect phishing attacks. The proposed framework includes 50 features. Those features are divided into four groups (e.g. headers of e-mail, web URLs and main text). The proposed system is based on e-mails and analytical processes. The accurate rate is 98.6%. The incorrect positive rate is 1.8%.

In this paper, the confidentiality policies and mechanisms of semantic web applications are investigated, and a smart model based on NN is introduced to classify relevant web services (R. M. Mohammad & AbuMansour, 2017).(Feng et al., 2018) proposed a novel model of a neural network with high accuracy and strong generalization capabilities. This model aims to detect phishing websites. Unlike the conventional neural network, this method includes the principle of risk minimization design and the Monte Carlo algorithm. They reached a high accuracy rate with different features derived from the URLs.

(Jain & Gupta, 2018) developed an anti-phishing strategy by using machine learning. This strategy aims to differentiate between phishing websites and legitimate ones across 19 features on the customer's side. They used 2141 phishing pages. The positive rates of the proposed strategy are 99%, 39% with the implementation of computer training.

In the literature, natural language processing (NLP) isn't used much. In a recent study by (Peng, Harris, & Sawa, 2018), phishing e-mails are detected through NLP. Through analyzing the content of e-mails (as plain text) semantically, the researchers identified malicious intentions. They aimed to use NLP to capture request and order phrases. For the identification of phishing attacks, a common blacklist of word pairs is used. For training and checking the program, the researchers used 5009 phishing and 5000 legitimate e-mails. The precision rate is 95%.

The authors in (X. Zhang, Zeng, Jin, Yan, & Geng, 2017) proposed a phishing detection model that effectively detects phishing output by the use of semantic characteristics for word insertion,

semanticized characteristics and multiple statistical characteristics on Chinese websites. Eleven features were extracted to obtain statistical features of web pages and grouped into five groups. For the implementation of model learning and testing, AdaBoost, Bagging, Random Forest and SMO are used. The Anti-Phishing Alliance of China obtained legitimate URLs from direct-industry web guides and phishing info. The study shows that the phishing sites with high efficiency and the fusion model have been well established only with semic characteristics and achieved the best output detection. The model is specific to Chinese websites and depends on some languages.

In(Machado & Gadge, 2017) provide an efficient method usingthe C4.5 decision tree approach for deteriorating phishing URL pages. This method calculates heuristic values by extracting features from the pages. The c4.5 decision tree algorithm was used to decide if the site was phishing or not using these values. The data was gathered from PhishTank and Google. This procedure is divided into two stages: pre-processing and detection. During the pre-processing step, features were extracted based on rules, and the features and their valued values were fed into the c4.5 algorithm, which yielded an accuracy of 89.40%.

In the study by (Rao, Pais, & Applications, 2019), the researchers adopted a hybrid approach through machine learning and image surveillance. A significant limitation to picture / visual phishers detection concerns the need for an initial database for pictures or a previous awareness (web history) of the website. The three types of accessibility were utilized: third-party apps, hyper-link features and URL obscenity. Since the use of resources by third parties increases the time of detection, the accuracy of the system increases to reach 99.55%.However, the hybrid technique was in many research work such as (Juvanna, Aravindan, Kumar, & Vignesh, 2021).

In the previous results, it was discovered that the anti-phishing technology causes the web browser to respond slowly.This means that the user inputs their information into the suspicious website with confidence and subsequently becomes aware of the site's type. It is a hard job to inform web users of the website category. This means that the anti-phishing tool should operate quickly enough, which can only be done if the programming codes are precise and easy to implement.

In this research, we conducted several experiments based on several datasets to detect phishing websites using semantic features. However, the main concern of the research is to evaluate and select the best classification algorithm to be used for phishing detection. We still need to do more experiments on phishing website detection based on semantic features and deep learning algorithms, to show which algorithms represent the best selection for phishing website detection.

3. A COMPARATIVE STUDY

The main aim of the comparison study is to isolate and sequentially compile and sort phishing websites. In the proposed comparison study, 16 machine-learning classifier algorithms were modified to recognize correlations in the data set between more than characteristics in dataset one (R. Mohammad, Thabtah, & McCluskey, 2015). and 48 Semantic URL features selected as phishing web from dataset two.

A block diagram of our plan, as shown in Figure 3. There are three phases in the proposed approach. The first stage is the pre-processing stage. Through this stage, characteristics and sub-functions are derived from phishing and related websites. The second stage contains the classification of machine learning. Such classification represents the basis of laws. In the third stage, the system classifies the webpages into phishing or normal webpages.Semantic features refer to annotations that are derived from the URLs and content of the resources. We are motivated to show the value of using semantic features as a means of detecting phishing webpages.

3.1. Pre-Processing and Features Extraction

The pre-processing stage involves features selection and extraction of thin comparison vector creation. The lack of reliable learning databases is a challenge that faced the researchers of the present study.

Every scholar specialized in the area faces this challenge. Nonetheless, several papers were conducted about the detection of phishing sites through using data mining techniques. However, no credible research database was published. That may be because there is not much agreement among scholars about the characteristics of phishing websites. Through the present study, the researchers aimed to shed a light on the key features of phishing websites. They aimed to evaluate the effectiveness of machine learning in detecting phishing websites. They aimed to suggest a technique through checking address bar-based features.

However, the semantic features used in previous studies depend on content features only, while our study focused on building semantic features based on URL & Domain Identity, Abnormal Based, HTML and JavaScript-based Features, Domain-based Features, which makes the number of features low compared with other studies and the speed of classification fast if we use any adaptive machine learning.

Selection of the Feature is a way of searching for a subset of important features from the original set, reducing the number of irrelevant data set iterations to improve the efficiency of classification and memory storage. Selection of features helps to understand data to minimize the impact of the dimensional curse, to reduce measuring requirements, to improve accuracy and to distinguish features that can apply in a particular problem(Alauthaman, Aslam, Zhang, Alasem, & Hossain, 2018; Alauthman, Aslam, Al-kasassbeh, Khan, Al-Qerem, & Raymond Choo, 2020). There are several methods of feature selection and in this paper,we have used Information Gain Attribute Evaluation to select the best features from the dataset.By measuring the information gap calculated for the target class, this method measures the significance of the attribute.The formula can be used to calculate it(Ammar Almomani, 2013):

Table 1 presents the features of phishing websites that are targeted through the present study. They have considered the ones use the most in artificial intelligent classifiers. Table 1 presents the name of the features only. The full description is presented in the work conducted by (R. Mohammad

$$\text{InfoGain}\left(\text{Class.Attribute}\right) = \text{H}\left(\text{Class}\right) - \text{H}(\text{Class} \mid \text{Attribute}) \tag{3.1}$$

et al., 2015).

Figure 4 presents the way in which the system will extract the features vectorbased on the features matrix which depends on the dataset used in the study's experiments. After this stage, the researchers have adopted 16 machine learning models for detecting phishing website as it's illustrated below.

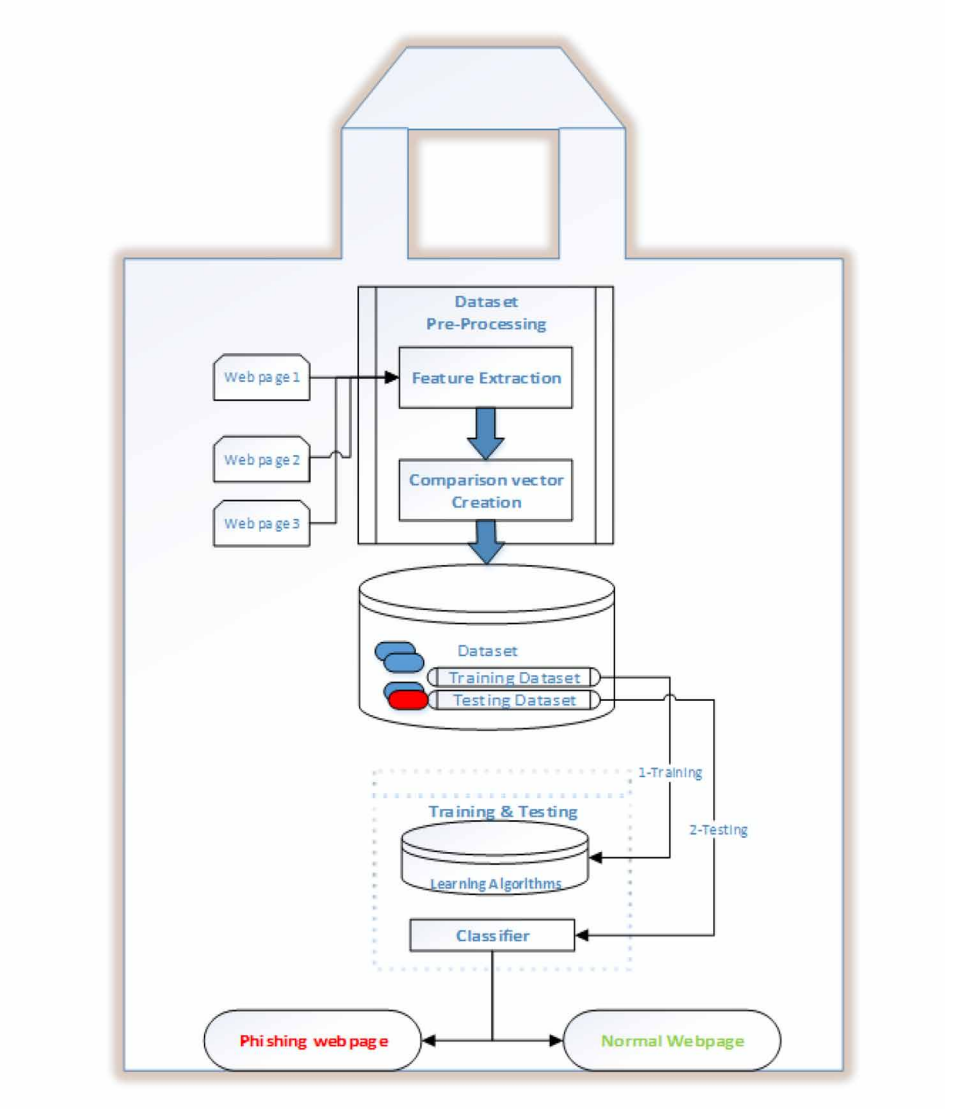
3.2. Machine Learning Models

In this phase, the researchers used 16 AI algorithms for detecting phishing websites. Table 1 presents full details about the references related to each algorithm. Through the present study, the researchers employed a machine learning algorithm which is available in the sickie-learn library Pedregosa et al. (2011). Moreover, all the experiments conducted in this study were conducted on a desktop computer with Intel(R) Core(TM) i7-2600, CPU 3.40 GHz, and 8GB RAM.

This study compares the predictive accuracy of several machine learning methods for predicting phishing websites, including Random Forests (RF), Classification and Regression Trees (CART), Logistic Regression (LR), Support Vector Machines (SVM), and Neural Networks (NNet), Bayesian Additive Regression Trees (BART), and more AI algorithm models.

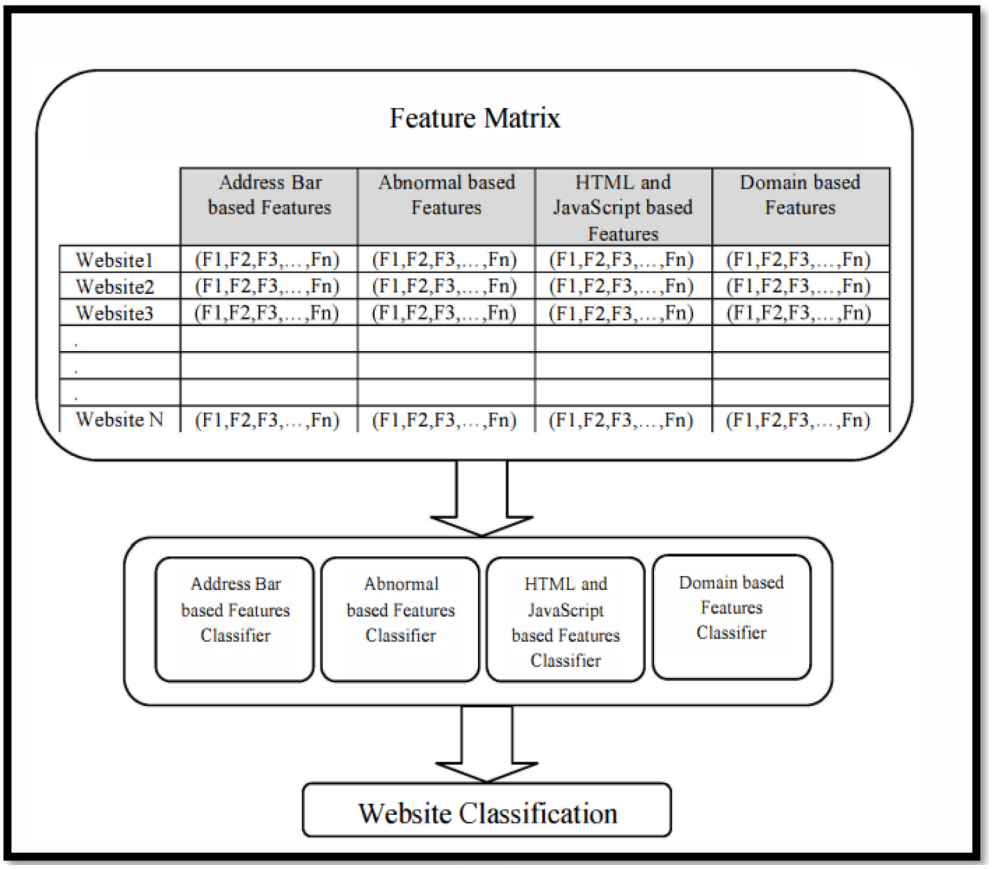
The following are the reasons for employing these AI classifiers:

Figure 3. Phishing website detection using the machine learning algorithms



1- These 16 AI classifiers were the most can be adaptive in phishing website and email-based on last studies
2- Many of the AI classifiers in our study were the first ones used in phishing website detection, such as the Ridge Classifier and CV BernoulliNB models.
3- Many features that were used in our study were adopted for the first time in phishing website detection. as a semantic feature.

Figure 4. Semantic phishing website detection using the machine learning algorithms



4. RESULTS AND DISCUSSION

4.1. Experiment 1:

This experiment was conducted using CPU: I7, 8GB Memory (RAM), Python 3.6.0 Programming Language in implementations. The two classes training and testing based on the 10-fold cross-validation. The Collected primarily from archive PhishTank, archive MillerSmiles, query operators Google. (R. M. Mohammad, Thabtah, McCluskey, & Engineering, 2015), started in 2012 and last modified in 2018, validation by random method. The researchers of the present study used the Huddersfield university dataset(R. M. Mohammad, Thabtah, & McCluskey, 2012). The result was built based on 70% training and testing.

Table 1. The first dataset characteristics:(R. M. Mohammad et al., 2012).

Attribute Characteristics:	Integer
The number of samples:	2456
A number of features:	30
A number of Web Hits:	119690

Table 2. Semantic Phishing Websites Features selected in our study (R. M. Mohammad et al., 2012; R. M. Mohammad et al., 2015)

Features Group	Phishing Websites Features	
URL & Domain Identity	1	Using the IP Address
	2	Long URL to Hide the Suspicious Part
	3	Using URL Shortening Services ”TinyURL”
	4	RL’s having ”@” Symbol
	5	Redirecting using ”//”
	6	Adding Prefix or Suffix Separated by (-) to the Domain
	7	Sub Domain and Multi-Sub Domains
	8	HTTPS (HyperText Transfer Protocol with Secure Sockets Layer)
	9	Domain Registration Length
	10	Favicon
	11	Using Non-Standard Port
	12	The Existence of ”HTTPS” Token in the Domain Part of the URL
Abnormal Based Features	13	Request URL
	14	URL of Anchor
	15	Links in <Meta>, <Script>and <Link>tags
	16	Server Form Handler (SFH)
	17	Submitting Information to E-mail
	18	Abnormal URL
HTML and JavaScript-based Features	19	Website Forwarding
	20	Status Bar Customization
	21	Disabling Right Click
	22	Using Pop-up Window
	23	IFrame Redirection
Domain-based Features	24	Age of Domain
	25	DNS Record
	26	Website Traffic
	27	PageRank
	28	Google Index
	29	Number of Links Pointing to Page
	30	Statistical-Reports Based Feature

Table 1: shows the first dataset used based on 30 semantic features,while Table 2 shows the details of features selected in our study.

Table 3 shows 16 Machine learning model algorithms used to detect phishing webpages, whileTable 4 shows the results of a comparison between 16 Semantic classifiers. It shows that the best algorithm results are based on a random forest classifier to detect phishing website based on semantic features with an accuracy of about 99% in the training phase and 96% in the testing phase. While the

Table 3. 16 Machine learning models algorithms used to detect Semantic phishing webpages

Number	Machine learning algorithms (MLA)	References
1	Random Forest Classifier	(Breiman, 2001)
2	Bagging Classifier	(Breiman, 1996)
3	Decision Tree Classifier	(Swain & Hauska, 1977)
4	Extra Tree Classifier	(Geurts, Ernst, & Wehenkel, 2006)
5	Gradient Boosting Classifier	(Bansal & Kaur, 2018)
6	SVC	(Müller & Guido, 2016)
7	K Neighbors Classifier	(Sarkar & Leong, 2000)
8	AdaBoost Classifier	(Hastie, Rosset, Zhu, & Zou, 2009)
9	Linear SVC	(Platt, 1999)
10	Logistic Regression CV	(de Melo & Banzhaf, 2016)
11	Ridge Classifier CV	(Kowsher, Tahabilder, & Murad, 2020)
12	Perceptron	(Stephen, 1990)
13	BernoulliNB	(Müller & Guido, 2016)
14	Passive Aggressive Classifier	(Lu, Zhao, & Hoi, 2016)
15	SGD Classifier	(Dongari, 2014)
16	GaussianNB	(Dongari, 2014)

lost accuracy is based on guessing the NB algorithm with 60% training and 61% testing phase. Table 4, figure 5 and figure 6 show respectively the results of accuracy in the training and testing phase.

Those results are reached by using 16 Machine learning algorithms.

as a classifier which can be classified as phishing webpage or normal webpage, and features represent the most effective features in the phishing web page. Based on the result, the highest classifying quality (99% accuracy) is seen through employing the Random Forest algorithm. That can be seen through this table. While figure 7 shows the results of the ROC curve comparison between the same classifiers in the training and testing phase.

4.2. Experiment 2:

We used different datasets in this study to investigate the ML algorithms' performance as well as the attribute importance within these datasets. Dataset 2(Tan, 2018) has 48 different attributes gathered from 5000 different phishing and legal websites. The webpages were downloaded between January and May 2015 to June 2017 (Tan, 2018). The binary labels in this dataset are 0 for legitimateand 1 for phishing.

This dataset contains 48 Semantic URL features that are extracted from five thousand phishing web pages and five thousand genuine webpages. Those webpages were accessed between January andMay and June 2015 and May to June 2017. The app optimization (e.g., Selenium WebDriver) is more reliable and stable than the parsing approach that is based on regular expressions. It employs an improved function extraction methodology. Its database is eligible for WEKA. Start for Phishing: PhishTank, OpenPhish. Legitimate web pages: Alexa, Popular Crawl, anti-phishing investigators, and experts can consider this database useful for the evaluation of phishing features, fast proof of concept tests, and phishing classification models (Tan, 2018), as shown in table 5.

Table 4. Experimental results 1: Comparison results between 16 Machine learning algorithms as a classifier

No.	Machine learning algorithms (MLA)	MLA Train Accuracy	MLA Test Accuracy	MLA Precision	MLA Recall	MLA AUC
1	RandomForestClassifier	**0.9906**	**0.9650**	0.962183	0.974822	0.963921
2	BaggingClassifier	0.9900	0.9638	0.962101	0.972633	0.962826
3	DecisionTreeClassifier	0.9912	0.9572	0.960131	0.962233	0.956620
4	ExtraTreeClassifier	0.9912	0.9512	0.947340	0.964970	0.949599
5	GradientBoostingClassifier	0.9539	0.9454	0.935911	0.967159	0.942976
6	SVC	0.9510	0.9403	0.9269	0.963875	0.937642
7	KNeighborsClassifier	0.9646	0.9355	0.939989	0.9476	0.934625
8	AdaBoostClassifier	0.9389	0.9346	0.925026	0.958949	0.931824
9	LinearSVC	0.9293	0.9267	0.920382	0.949097	0.924213
10	LogisticRegressionCV	0.92	0.9261	0.919851	0.948550	0.923604
11	RidgeClassifierCV	0.9219	0.9162	0.910005	0.940887	0.913396
12	Perceptron	0.9059	0.9053	0.887353	0.948550	0.900449
13	BernoulliNB	0.9095	0.9032	0.906587	0.918993	0.901443
14	PassiveAggressiveClassifier	0.8864	0.8842	0.883165	0.910235	0.881292
15	SGDClassifier	0.8563	0.8453	0.786150	0.987958	0.829214
16	GaussianNB	**0.6024**	**0.6262**	0.994941	0.322934	0.660460

Figure 5. Experiment 1: Comparison results between 16 Machine learning algorithms as a classifiers-Training phase

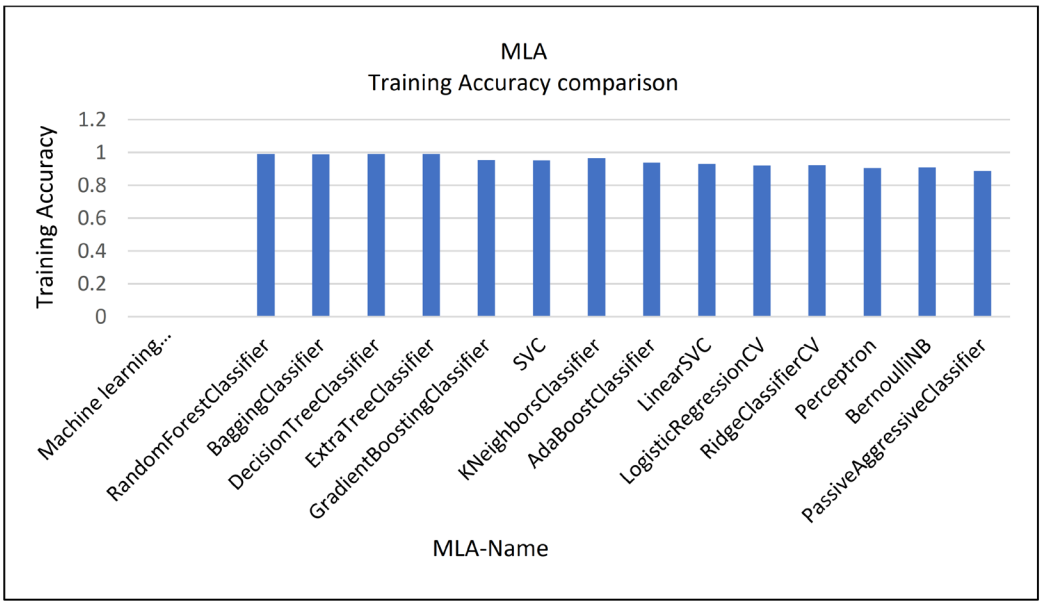


Figure 6. Experiment 1: Comparison results between 16 Machine learning algorithms as a classifiers-Testing phase

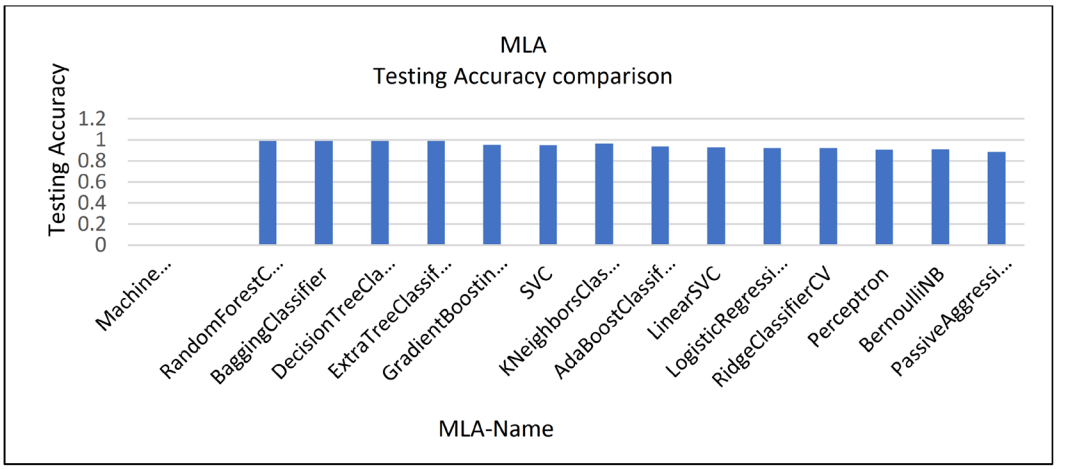


Figure 7. Experiment 1: ROC curve comparison between 16 Machine learning algorithms as a classifier

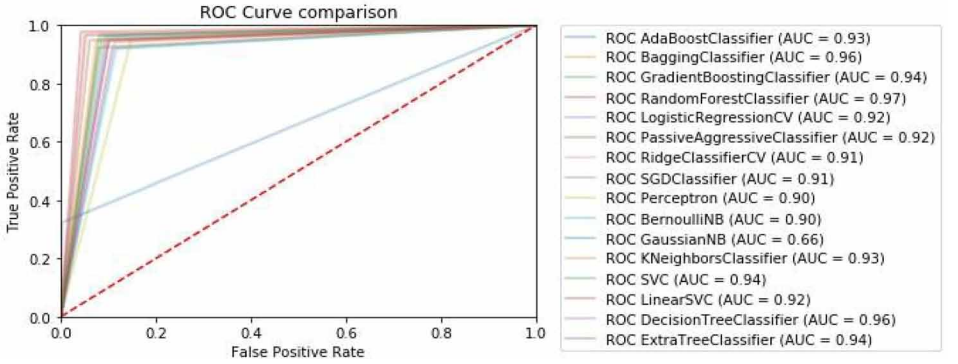


Table 6 present the results of a comparison between 16 classifiers. It presents the best algorithm results were based on a random forest classifier to detect phishing website with an accuracy of about 99.83% on the training phase and 97.68% on the testing phase. While the lost accuracy based on guessing NB algorithm with 81.73% on training and 82.62% on testing phase

Table 6, figure 8 and figure 9 respectively show the results of a comparison between 16 Machine learning algorithms as a classifiers-in testing phase. Table 5 and figure 8 show the accuracy reached through using 16 Machine learning algorithms as a classifier. 48 Semantic features represent the most useful features in the phishing URL web page. Based on the table, the highest classification output (99% precision) is reached by using the Random Forest algorithm. The effect of the characteristics. Figure 10 shows the ROC curve comparison between the same classifiers results in the training and testing phase.

Table 5. Second dataset characteristics all features are numeric (Tan, 2018)

Features	Features
Num of Dots	NumSensitiveWords
Sub domain Level	EmbeddedBrandName
Path Level	PctExtHyperlinks
Url Length	PctExtResourceUrls
Num of Dash	ExtFavicon
Num of Dash In Hostname	InsecureForms
At Symbol	RelativeFormAction
Tilde Symbol	ExtFormAction
Num of Underscore	AbnormalFormAction
Num of Percent	PctNullSelfRedirectHyperlinks
Num of Query Components	FrequentDomainNameMismatch
Num of Ampersand	FakeLinkInStatusBar
Numof Hash	RightClickDisabled
Num of NumericChars	PopUpWindow
No Https	SubmitInfoToEmail
RandomString	IframeOrFrame
IpAddress	MissingTitle
DomainInSubdomains	ImagesOnlyInForm
DomainInPaths	SubdomainLevelRT
HttpsInHostname	UrlLengthRT
HostnameLength	PctExtResourceUrlsRT
PathLength	AbnormalExtFormActionR
QueryLength	ExtMetaScriptLinkRT
DoubleSlashInPath	PctExtNullSelfRedirectHyperlinksRT

Table 6. Experimental results 2: Comparison results between 16 Machine learning algorithms as a classifier.

Num	MLA Name	MLA Train Accuracy	MLA Test Accuracy	Mean squared error	MLA Precision	MLA Recall	MLA AUC
1	GradientBoostingClassifier	0.9845	0.9770	0.020	0.9773	0.976334	0.976995
2	RandomForestClassifier	0.9983	0.9768	0.02325	0.9846	0.968278	0.976691
3	BaggingClassifier	0.9975	0.9758	0.02425	0.9806	0.970292	0.975712
4	AdaBoostClassifier	0.9727	0.9722	0.02775	0.9756	0.968278	0.972222
5	DecisionTreeClassifier	1.0000	0.9668	0.03325	0.9653	0.967774	0.966757
6	ExtraTreeClassifier	1.0000	0.9412	0.05875	0.9441	0.937059	0.941221
7	LogisticRegressionCV	0.9450	0.9405	0.05950	0.9432	0.936556	0.940473
8	RidgeClassifierCV	0.9400	0.9375	0.06250	0.9469	0.925982	0.937420
9	LinearSVC	0.9282	0.9275	0.07250	0.9076	0.950655	0.927661
10	BernoulliNB	0.9207	0.9205	0.07950	0.9246	0.914401	0.920458
11	SVC	0.9613	0.9010	0.09900	0.8878	0.916415	0.901107
12	Perceptron	0.8838	0.8782	0.12175	0.8704	0.886707	0.8789
13	PassiveAggressiveClassifier	0.8590	0.8632	0.13675	0.9460	0.768379	0.862591
14	KNeighborsClassifier	0.9038	0.86	0.13700	0.8836	0.833837	0.862797
15	GaussianNB	0.8437	0.8415	0.15850	0.7828	0.942095	0.842199
16	SGDClassifier	0.8173	0.8262	0.17375	0.9473	0.688318	0.825291

Figure 8. Experiment 2: Comparison results between 16 Machine learning algorithms as a classifiers-in Training phase

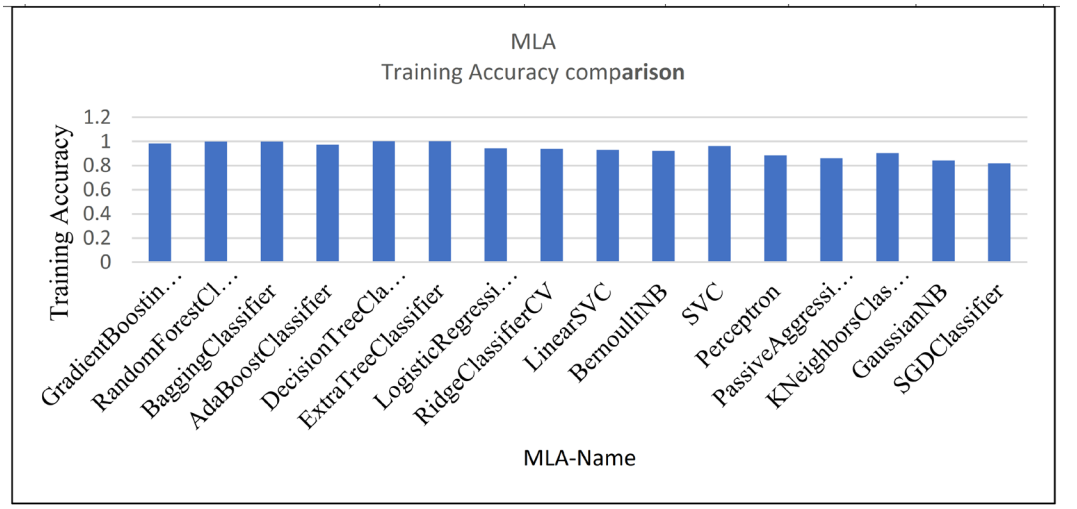


Figure 9. Experiment 2: Comparison results between 16 Machine learning algorithms as a classifiers-in testing phase

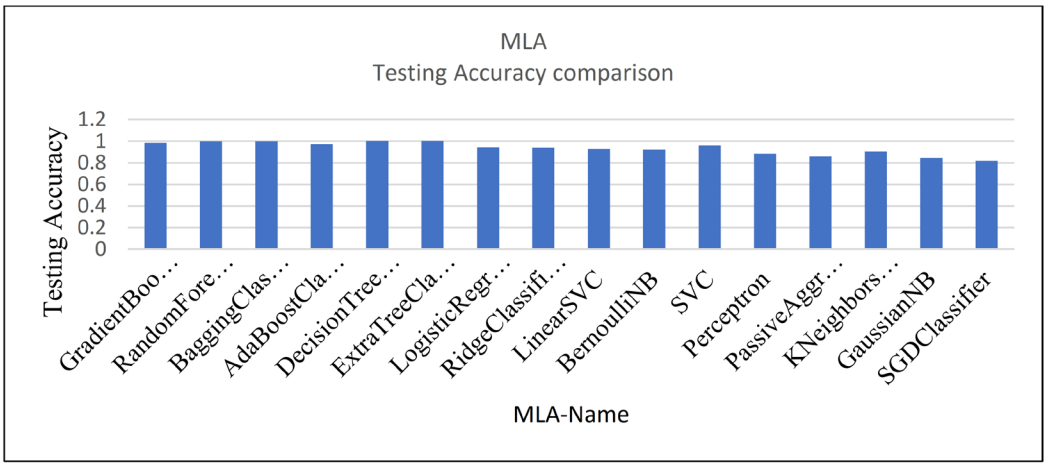


Figure 10. Experiment 2: ROC Curve comparison between 16 Machine learning algorithms as a classifiers

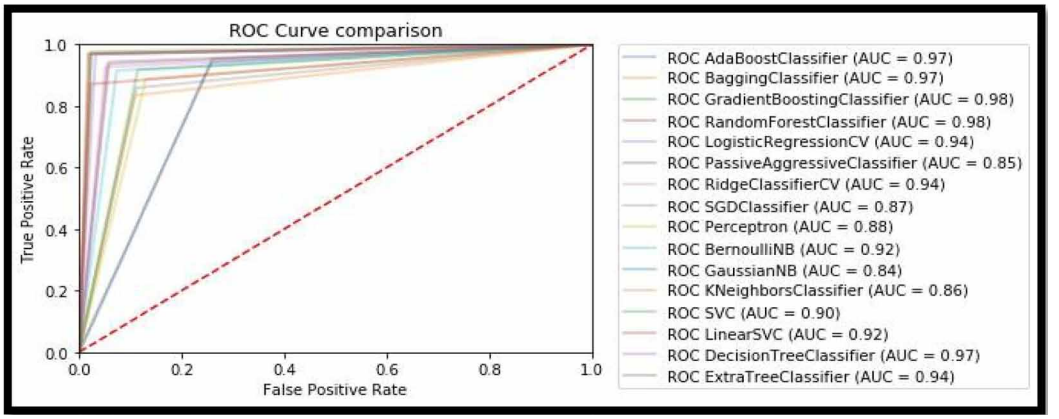


Table 7. *Summary of the existing works compares with our comparisonstudy*

Authors	Contribution Summary	Weakness	Mechanism	Algorithms
(Abu-Nimeh, Nappa, Wang, & Nair, 2009)	Prove that there is no standard classifiers for phishing email predection	More features consume more time and memory	Compared six classifiers relating to machine learning	LR,CART,SVM,NNET,BART,RF
(Miyamoto, Hazeyama, & Kadobayashi, 2009)	comparison of machine learning algorithms to detect phishing	The observed F-measure is still low	Detect phishing website based on 3000 website data	adaBoost
(R. Gupta, 2016)	a number ofanti-phishing toolbars have been discussed and proposed a system model to tackle the phishing attack.	BIG number of features, consumingtime and cost	The proposed anti-phishing system is based on the development of the Plug-in tool for the web browser	Random Forest, Nearest NeighborClassification (NNC), Bayesian Classifier (BC).
Our Comparison Study	working and extracting URL & Domain Identity feature Abnormal Based Features, HTML andJavaScriptbased Features and Domain based Features as asemantic features to detect phishing website,	The complexity of selection AI classifiers	used 16 machine learning models that have more than 48 semantic features represent the most effective features for the detection of phishing webpage extracted from two datsets	16 AI CLASSIFIERS, Discussed in this article

5.CONCLUSION

In this study, we examined the predictive accuracy of 16 classification systems and other measures based on semantic URL features. URL & Domain Identity feature Abnormal Based Features, HTML and JavaScript-based Features and Domain-based Features are semantic features to detect phishing websites, which makes the semantic features more controlled and more effective for the classification process. Major cybersecurity threats include web phishing and spear phishing. Applications of these comparisons can identify phishing sites effectively. No user intervention. These roles are automatically retrieved and used by computer-developed devices. Ten characteristics that distinguish genuine websites from phishing websites using semantic features have been compiled and analyzed.To detect phishing web pages, our study employed 16 AI classifiers with two datasets and over 48 semantic features. Based on the results of the comparison, GradientBoostingClassifier and RandomForestClassifier have the highest accuracy (i.e. about 97%). In contrast, GaussianNB and the stochastic gradient descent (SGD) classifier represent the lowest accuracy results (84) (81) respectively in comparison with other classifiers. To give more conclusive results concerning the predictive accuracy of classifications, we proposed accuracy, recall and AUC measures. The results

motivate future work to consider the inclusion of further variables in the data set, which could improve classification predictive accuracy. Analysis of URL features, for example, has demonstrated that they improve prediction ability and reduce classification error rates.

The limitations of our study include the large number of machine learning algorithms that can be used in our study, as well as the large number of features that can be used as semantic features. For future work, we propose that cost-sensitive measures are taken to give more conclusive results on the provision of classification accuracy and we suggest working with ensemble learning techniques because RandomForestClassifier is one of these techniques. The most precise and reliable ML methods are ensemble and hybrid ML. Ensemble methods are developed using various methods of grouping, such as boosting or bagging, to use several ML classification systems.

ACKNOWLEDGMENT

This research is supported by Al-Balqa Applied University, Jordan, Grant Number: DSR-2018-#4.

REFERENCES

Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). *Phishing detection based associative classification data mining*. Academic Press.

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2009). *Distributed Phishing Detection by Applying Variable Selection Using Bayesian Additive Regression Trees.* Paper presented at the 2009 IEEE International Conference on Communications. doi:10.1109/ICC.2009.5198931

Adebowale, M. A., Lwin, K. T., Sanchez, E., & Hossain, M. A. (2019). *Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text*. Academic Press.

Al-Momani, A., Wan, T.-C., Al-Saedi, K., Altaher, A., Ramadass, S., Manasrah, A., . . . Anbar, M. (2011). An online model on evolving phishing e-mail detection and classification method. *Journal of Applied Science, 11*(18), 3301-3307.

Al-Nawasrah, A., Almomani, A. A., Atawneh, S., & Alauthman, M. (2020). A Survey of Fast Flux Botnet Detection With Fast Flux Cloud Computing. *International Journal of Cloud Applications and Computing*, *10*(3), 17–53. doi:10.4018/IJCAC.2020070102

Alauthaman, M., Aslam, N., Zhang, L., Alasem, R., & Hossain, M. A. (2018). A P2P Botnet detection scheme based on decision tree and adaptive multilayer neural networks. *Neural Computing & Applications*, *29*(11), 991–1004. doi:10.1007/s00521-016-2564-5 PMID:29769759

Alauthman, M., Almomani, A., Alweshah, M., Omoush, W., & Alieyan, K. (2019). Machine learning for phishing detection and mitigation. In *Machine Learning for Computer and Cyber Security* (pp. 48–74). CRC Press. doi:10.1201/9780429504044-2

Alauthman, M., Aslam, N., Al-Kasassbeh, M., Khan, S., Al-Qerem, A., & Choo, K.-K. R. (2020). *An efficient reinforcement learning-based Botnet detection approach*. Academic Press.

Alauthman, M., Aslam, N., Al-kasassbeh, M., Khan, S., Al-Qerem, A., & Raymond Choo, K.-K. (2020). An efficient reinforcement learning-based Botnet detection approach. *Journal of Network and Computer Applications*, *150*, 102479. doi:10.1016/j.jnca.2019.102479

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers of Computer Science*, *3*(6), 563060. Advance online publication. doi:10.3389/fcomp.2021.563060

Almomani, A., Wan, T.-C., Manasrah, A., Altaher, A., Almomani, E., Al-Saedi, K., . . . Ramadass, S. (2012). A survey of learning based techniques of phishing email filtering. *International Journal of Digital Content Technology and its Applications, 6*(18), 119.

Almomani, A., Alauthman, M., Omar, A., & Firas, A. (2017). *A Proposed Framework for Botnet Spam-email Filtering Using Neucube.* Paper presented at the The International Arab Conference on Information Technology, Yassmine Hammamet, Tunisia.

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). *A survey of phishing email filtering techniques*. Academic Press.

Almomani, A., Obeidat, A., Alsaedi, K., Obaida, M. A.-H., & Al-Betar, M. (2015). Spam e-mail filtering using ECOS algorithms. *Indian Journal of Science and Technology*, *8*(S9), 260–272. doi:10.17485/ijst/2015/v8iS9/55320

Almomani, A., Wan, T., Manasrah, A., Altaher, A., Baklizi, M., & Ramadass, S. (2013). An enhanced online phishing e-mail detection framework based on evolving connectionist system. *International Journal of Innovative Computing, Information, & Control*, *9*(3), 169–175.

Almomani, A., Wan, T.-C., Altaher, A., Manasrah, A., Almomani, E., Anbar, M., . . . Ramadass, S. (2012). *Evolving fuzzy neural network for phishing emails detection*. Academic Press.

Ammar Almomani, T.-C. W. (2013). An enhanced online phishing e-mail detection framework based on "Evolving connectionist system. *International Journal of Innovative Computing, Information, & Control*, *9*(3), 1065–1086.

Anupam, S., & Kar, A. K. (2021). Phishing website detection using support vector machines and nature-inspired optimization algorithms. *Telecommunication Systems*, *76*(1), 17–32. doi:10.1007/s11235-020-00739-w

APWG. (2020). *Phishing Activity Trends Report 1st Quarter 2020*. Retrieved from https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf

Babagoli, M., Aghababa, M. P., & Solouk, V. (2019). *Heuristic nonlinear regression strategy for detecting phishing websites*. Academic Press.

Bansal, A., & Kaur, S. (2018). *Extreme Gradient Boosting Based Tuning for Classification in Intrusion Detection Systems.* Paper presented at the Advances in Computing and Data Sciences, Singapore.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi:10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Cao, Y., Han, W., & Le, Y. (2008). Anti-phishing based on automated individual white-list. *Proceedings of the 4th ACM Workshop on Digital Identity Management*. doi:10.1145/1456424.1456434

de Melo, V. V., & Banzhaf, W. (2016). Improving logistic regression classification of credit approval with features constructed by Kaizen programming. *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. doi:10.1145/2908961.2908963

Dongari, A. R., & Saidi, M. (2014). Suspicious URL detection system using SGD algorithm for Twitter stream. *Int. J. Comput. Sci. Inf. Eng. Technol.*, *2*(4), 1–6.

Dunlop, M., Groat, S., & Shelly, D. (2010). *Goldphish: Using images for content-based phishing analysis.* Paper presented at the 2010 Fifth International Conference on Internet Monitoring and Protection. doi:10.1109/ICIMP.2010.24

Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. (2018). *The application of a novel neural network in the detection of phishing websites*. Academic Press.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. doi:10.1007/s10994-006-6226-1

Gupta, B. B., Arachchilage, N. A., & Psannis, K. E. (2018). Defending against phishing attacks: Taxonomy of methods, current issues and future directions. *Telecommunication Systems*, *67*(2), 247–267. doi:10.1007/s11235-017-0334-z

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2017). Fighting against phishing attacks: State of the art and future challenges. *Neural Computing & Applications*, *28*(12), 3629–3654. doi:10.1007/s00521-016-2275-y

Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*, *175*, 47–57. doi:10.1016/j.comcom.2021.04.023

Gupta, R. (2016). Comparison of classification algorithms to detect phishing web pages using feature selection and extraction. *International Journal of Research-Granthaalayah*, *4*(8), 118–135. doi:10.29121/granthaalayah.v4.i8.2016.2570

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and Its Interface*, *2*(3), 349–360. doi:10.4310/SII.2009.v2.n3.a8

Islam, R., & Abawajy, J. (2013). A multi-tier phishing detection and filtering approach. *Journal of Network and Computer Applications*, *36*(1), 324–335. doi:10.1016/j.jnca.2012.05.009

Jain, A. K., & Gupta, B. (2021). A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, 1–39. doi:10.1080/17517575.2021.1896786

Jain, A. K., & Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. *J EURASIP Journal on Information Security*, *2016*(1), 1–11. doi:10.1186/s13635-016-0034-3

Jain, A. K., & Gupta, B. B. (2018). Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, *68*(4), 687–700. doi:10.1007/s11235-017-0414-0

Jain, A. K., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, *10*(5), 2015–2028. doi:10.1007/s12652-018-0798-z

Johns, E., Williams, H., Clark, L., Leggett, O., & Shah, J. N. (2020). *Cyber security breaches survey 2020: Statistical release*. Academic Press.

Juvanna, I., Aravindan, K., Kumar, C. D., & Vignesh, S. (2021). Phishing Website Detection Using Hybrid Multi-Feature Classification. *Design Engineering (London)*, 1436–1451.

Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys and Tutorials*, *15*(4), 2091–2121. doi:10.1109/SURV.2013.032213.00009

Kowsher, M., Tahabilder, A., & Murad, S. A. (2020). Impact-learning: a robust machine learning algorithm. *Proceedings of the 8th International Conference on Computer and Communications Management*.

Le, A., Markopoulou, A., & Faloutsos, M. (2011). Phishdef: Url names say it all. *Proceedings IEEE INFOCOM*. doi:10.1109/INFCOM.2011.5934995

Lu, J., Zhao, P., & Hoi, S. C. H. (2016). Online Passive-Aggressive Active learning. *Machine Learning*, *103*(2), 141–183. doi:10.1007/s10994-016-5555-y

Machado, L., & Gadge, J. (2017). *Phishing Sites Detection Based on C4.5 Decision Tree Algorithm.* Paper presented at the 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). doi:10.1109/ICCUBEA.2017.8463818

Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2009). *An Evaluation of Machine Learning-Based Methods for Detection of Phishing Sites.* Paper presented at the Advances in Neuro-Information Processing, Berlin, Germany.

Moghimi, M., & Varjani, A. Y. (2016). New rule-based phishing detection method. *Expert Systems with Applications*, *53*, 231–242. doi:10.1016/j.eswa.2016.01.028

Mohammad, R., Thabtah, F. A., & McCluskey, T. (2015). *Phishing websites dataset*. Academic Press.

Mohammad, R. M., & AbuMansour, H. Y. (2017). *An intelligent model for trustworthiness evaluation in semantic web applications.* Paper presented at the 2017 8th International Conference on Information and Communication Systems (ICICS). doi:10.1109/IACS.2017.7921999

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012). *An assessment of features related to phishing websites using an automated technique.* Paper presented at the 2012 International Conference for Internet Technology and Secured Transactions.

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). *Phishing websites features*. Academic Press.

Morgan, S. (2019). *Official annual cybercrime report. USA, UK, Canada.* Retrieved from https://www.herjavecgroup.com/wp-content/uploads/2018/12/CV-HG-2019-Official-Annual-Cybercrime-Report.pdf

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc.

Peng, T., Harris, I., & Sawa, Y. (2018). *Detecting Phishing Attacks Using Natural Language Processing and Machine Learning.* Paper presented at the 2018 IEEE 12th International Conference on Semantic Computing (ICSC).

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, 10*(3), 61-74.

Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010). Phishnet: Predictive blacklisting to detect phishing attacks. *Proceedings IEEE INFOCOM*. doi:10.1109/INFCOM.2010.5462216

Rao, R. S., & Pais, A. (2019). *Detection of phishing websites using an efficient feature-based machine learning framework*. Academic Press.

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. doi:10.1016/j.eswa.2018.09.029

Sarkar, M., & Leong, T.-Y. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. *Proceedings of the AMIA Symposium*.

Selvakumari, M., Sowjanya, M., Das, S., & Padmavathi, S. (2021). *Phishing website detection using machine learning and deep learning techniques*. Paper presented at the Journal of Physics: Conference Series. doi:10.1088/1742-6596/1916/1/012169

Smadi, S., Aslam, N., & Zhang, L. (2018). Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. *Decision Support Systems*, *107*, 88–102. doi:10.1016/j.dss.2018.01.001

Stephen, I. (1990). *Perceptron-based learning algorithms*. Academic Press.

Swain, P. H., & Hauska, H. (1977). *The decision tree classifier: Design and potential*. Academic Press.

Verma, R., & Hossain, N. (2013). *Semantic feature selection for text with application to phishing email detection.* Paper presented at the International Conference on Information Security and Cryptology.

Wenyin, L., Fang, N., Quan, X., Qiu, B., & Liu, G. (2010). *Discovering phishing target based on semantic link network*. Academic Press.

Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). *Cantina+ a feature-rich machine learning framework for detecting phishing web sites*. Academic Press.

Zhang, W., Lu, H., Xu, B., & Yang, H. (2013). *Web phishing detection based on page spatial layout similarity*. Academic Press.

Zhang, X., Zeng, Y., Jin, X., Yan, Z., & Geng, G. (2017). *Boosting the phishing detection performance by semantic analysis.* Paper presented at the 2017 IEEE International Conference on Big Data (Big Data).

Zhang, Y., Hong, J. I., & Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing web sites. *Proceedings of the 16th international conference on World Wide Web*. doi:10.1145/1242572.1242659

Ammar Almomani received his Ph.D. from Universiti Sains Malaysia (USM) in 2013. He is the author of over 75 research papers in renowned International Journals and Conferences including IEEE, Elsevier, ACM, Springer, Inderscience, with countless of international awards, He has toured various nations in presenting his valued research work, and has revised 10s Journals in IEEE, Springer, Wiley, and Taylor & Francis. He has been lecturing for 17 years covering over 40 different subjects in computer science, networks, and cybersecurity, and programming language. He is a holder of numerous international certificates, and has contributed in various projects and specialized scientific courses. As an ardent researcher, he has been exploring cybersecurity, advanced Internet security, and monitoring. Dr. Ammar Almomani is currently serving the post of senior lecturer at Al- Balqa Applied University. At present, he leads the research and innovation department in SKYLINE university college-SHARJAH-UAE.

Mohammad Alauthman received his PhD degree from Northumbria University in Newcastle, UK, in 2016. He received a B.Sc. degree in Computer Science from Hashemite University, Jordan, in 2002, and received an M.Sc. degree in Computer Science from Amman Arab University, Jordan, in 2004. He is currently an Assistant Professor in the Department of Information Security at Petra University in Jordan.His research interests include cyber-security, Cyber Forensics, advanced machine learning and data science applications.

Mohammed Alweshah received his BSc in Computer Sciences in 1993 from Al-Mustansiriah University, Iraq and his Master's in Computer Sciences in 2005 from Al-Balqa Applied University, Jordan. He received his PhD from the Computer Science Department, School of Information Technology, National University of Malaysia (UKM), Malaysia in 2013 where he was under the supervision of Prof. Salwani Abdullah. Currently, he is an Assistant Professor with the Prince Abdullah Bin Ghazi Faculty of Information Technology, Al-Balqa Applied University, Al-Salt, Jordan. His research interests include meta-heuristic algorithms in optimisation areas that involve different real-world applications, such as data mining problems.

Ayat Alrosan received a PhD degree from Universiti Sains Islam Malaysia (USIM) in 2017. She has published many research papers in International Journals and Conferences of high repute. Currently, she is an assistant professor at School of Information Technology, Skyline University College, Sharjah P.O. Box 1797, United Arab Emirates. His research interest includes image processing, data clustering, and optimization.

Waleed Alomoush received a Ph.D. degree from University Kebangsaan Malaysia (UKM) in 2015. He has published many research papers in International Journals and Conferences of high repute. Currently, he is an assistant professor at the School of Information Technology, Skyline University College. His research interest includes, but is not limited to, Data clustering and optimization.

B. B. Gupta received PhD degree from Indian Institute of Technology Roorkee, India in the area of information security. He has published more than 250 research papers in international journals and conferences of high repute. He has visited several countries to present his research work. His biography has published in the Marquis Who's Who in the World, 2012. At present, he is working as an Assistant Professor in the Department of Computer Engineering, National Institute of Technology Kurukshetra, India. His research interest includes information security, cyber security, cloud computing, web security, intrusion detection, computer networks and phishing.