# IDENTIFYING PHISHING WEBSITES BY TECHNIQUES HYPER HEURISTIC AND MACHINE LEARNING

**Mahmoud Taalohi [1] , Nafise Langari [2], Hamid Tabatabaee [3,*]**

[1]Department of Computer Science and Technology , Islamic Azad University ,Mashhad ,Iran

[2]Department of Information Science and Technology , Birjand University , Birjand ,Iran

[3]*Young Researchers and Elite Club, Quchan Branch, Islamic Azad University, Quchan ,Iran

***ABSTRACT:*** *One of the most serious challenges confronted by IT development, especially in the field of E-Commerce is bank and identity information theft. This security threat is called phishing. Of the most common kinds of phishing is creation of fake websites of banks and financial institutions by the phishers in order to steal the information of consumers. Several ways for detection of phishing websites have been investigated and analyzed. In existing ways, short life-time of phishing websites, reduction of computational volume and possibility of analyzing and controlling a range of websites are not simultaneously considered. Thus, in this article a new approach is presented and implemented in order to simultaneously realize these three parameters and to create an efficient tool for supervisory authorities. In this approach, exploiting the characteristics of 8000 websites, the reduction of dynamics in evaluated features through detection of phishing websites are defined according to hyper heuristic gravitational search algorithms. Then, classification of the websites into two phishing website and legal websites is performed through support vector machine algorithm, a technique in machine learning. The presented approach has been implemented on standard data. The comparison of the results using the best existing algorithms has shown relative realization of desired objectives with accuracy rate of 87%, error rate of 7.8 and runtime of 8190ms. According to these figures, it is implied that this approach has a good grade among the other ones.*

**Key words:** *Phishing, gravitational search algorithm, support vector machine algorithm, classification*

## 1. INTRODUCTION

With the growing development of different websites in cyberspace, many matters including shopping, stock trading, electronic conferences, etc. are prospered and helped to ease doing things, but as always, the importance of security in these matters is high. One of the issues raised in the area of security is phishing, which generally means using various techniques to impersonate a valid website in order to access confidential information of users, such as credit cards, other institutions' financial accounts and etc. Therefore, trying to identify and deal with the issue is of a great importance. Phishing problem can be classified into three sub problems (identifying phishing websites, phishing e-mails, and social phishing).

In this paper, in order to identify and classify websites phishing, the gravitational search algorithm and Support vector machine algorithm is used. Gravitational search algorithm to select features that enhance the classification accuracy To accuracy, speed and acceptable error rate than other methods have done.

### 1.1. phishing websites

The advent of phishing attacks was via email, that users with clicking on the link provided in the email, and trusting it, disclosed their personal information [1]. So identifying phishing websites to protect users' personal information and reduce the losses caused by them, is an important issue which is noticed more than other issues today. From The constraints for identifying phishing websites, short lifetime, can be [2] noted which researches have demonstrated that the average lifespan of a phishing website varies from a few days to even a few hours for some of them, hence, the data set should be online and updated. In The following section, current methods are discussed.

To protect personal information against phishing attacks, objectives, such as accurate diagnosis of phishing websites in a short time [3,4],and reducing error rates[5] in detection must be fulfilled.

### 1.2. phishing emails

Increasing use of electronic mail in the world, because of its simplicity and low cost, has led many internet users to be interested in developing their work on the internet. In this environment, many companies using electronic mail are attracted to the idea of advertisements. This allows internet users' mailbox to be full of junk mail that are called Spam. Many internet users are faced emails, which caused wasting time to sort them, wasting bandwidth and inbox space. This was the starting point for the development of strategies for managing automated spam emails. Classifying emails based on proper classification was for the purpose of solving these problems.

In the following article, the second section, we will look at some previous works, in the third selection , we review the proposal for detecting phishing websites. In the fourth part we will come to evaluate this method and finally in the Fifth section we provide conclusion.

## 2. Related works

Website analysis technology has been widely utilized, using the extracted features of the site's content [6] such as:

1. Verbal expressions such as mean average of sentences, length and words' spelling
2. The number of slashes in the URL
3. saved in image formats
4. image pixel, such as determining the color of the pixel
5. links, such as the number of input and output links to website

Generally, a variety of indicators to detect phishing websites can be seen in the Table1.

**Table 1**

*phishing identification indicators [13]*

| Metric | Phishing indicators |
|---|---|
| URL & Domain Identity | • Using IP address<br>• Abnormal request URL<br>• Abnormal URL of anchor<br>• Abnormal DNS record<br>• Abnormal URL |
| Security & Encryption | • Using SSL Certificate<br>• Certificate authority<br>• Abnormal cookie<br>• Distinguished names certificate |
| Source Code & Java script | • Redirect pages<br>• Straddling attack<br>• Pharming attack<br>• On Mouse over to hide the Link<br>• Server Form Handler (SFH) |
| Page Style & Contents | • Spelling Errors<br>• Copying website<br>• Using from s with Submit button<br>• Using pop-ups windows<br>• Disabling right-click |
| Web Address Bar | • Long URL address<br>• Replacing similar char for URL<br>• Adding a prefix or suffix<br>• Using the @ Symbols to confuse<br>• Using hexadecimal char codes |
| Social Human Factor | • Emphasis on security<br>• Public generic salutation<br>• Buying time to access accounts |

Also, generally the identifying methods can be divided into four categories as follows [7].

**1.** black list
**2.** data mining
**3.** machine learning
**4.** algorithms based on innovative methods (heuristic)

### 2.1. Black List

Keeping a list of URLs related to phishing websites, helps identifying the phishing website whenever a website URL, agrees with the URL in the list. Due to the rapid growth of web sites, this list needs to be always up to date. Using anti-phishing tools in browsers, is a method for detecting phishing sites. These tools are based on characteristics such as the length of the URL [4], the popularity of the site [8], duration in which the site was registered and the site search in the blacklist[7].

These tools identify the phishing sites and in case of facing them, blocks the user activities and warn them. ID, Net Craft, EarthLink, Cloud mark, are among anti-phishing tools.

### 2.2. Data Mining Techniques

Fuzzy and associative classification techniques are data mining techniques. In Fuzzy data mining method, Maher Aburrous [4] uses a combination of data mining algorithms and fuzzy systems, in order to evaluate the risk of online banks that are exposed to phishing websites and through feature extraction detects the phishing site. In this method, a number of data mining classification techniques are used such as JRip, PART, Prism, and C4.5. The associative classification method [8] and MCAR ,CBA techniques have been used for identifying phishing websites in internet banking.

### 2.3. Machine Learning

Machine learning is an important branch of artificial intelligence that sets forward this question: how we can build a computer program of which can automatically improve itself through its experiences [9].. In many areas, creating a computer program for carrying out tasks of an activity, is very difficult (because they cannot be easily described), but they can be shown (input / output) through many examples. Machine learning techniques, by getting an abundance of data and samples of an especial activity, identifies their existing patterns and creates predictive models for them. The necessity of machine learning is most obvious when human lacks the required expertise or cannot describe it.

Machine learning methods evaluate the entry site based on phishing characteristics. When features of a site are similar to the ones of a phishing site, the entry site will be identified as a phishing one. Among machine learning methods, we can note the neuro-fuzzy based method [10], logistic regression classification method [11,12] the proposed method (page safe) [14], logical regression [15],support vector machine [15,16] , Bayesian additive regression tree [16], and random forest .[15,17]

### 2.3.1. Support vector machine algorithm

Support Vector Machine algorithm or SVM is used for classification of two groups. The machine follows the concept in such a way that nonlinear input vectors are mapped into a feature space with very large dimensions. In This features space there exists a linear decision-making boundary. The attempts are focused to choose a line with higher safety margin and in fact, the optimal dividing line. The equation is for finding the optimal line for data using a QP method which is useful in solving restricted problems.[18]

Unlike neural networks, support vector machine method is not stuck in a local maximum, and training them is also easier. They perform very well for high-dimensional data.[19]

### 2.4. Meta-Heuristic algorithms

Meta-Heuristic algorithms , are a kind of the optimization algorithms which have some properties to get out of local optimum. From these meta-heuristic methods that have been used in phishing detection, we can refer to [20,21] ant colony (ACO) particle swarm optimization (PSO) [20], Bacterial food-finding algorithm (BFOA) and improved bat algorithm (MBAT) [21,22].

The main advantage of the using methods based on the blacklist is their easy implementation [3].These tools are often dependent on a particular browser, and cannot be installed on multiple browsers [4].Methods based on black list, usually cannot identify all phishing web sites [7]. The list should be update, otherwise it cannot prevent the latest phishing threats. Unlike blacklist approaches, heuristic method can detect new phishing sites, and will not have the problems of blacklist-based approaches [19]. Methods based on machine learning and data mining suffer challenges of learning complexity and high computational volume and consequently, the failure of the operating time minimization [10]. In many provided methods, a certain number of features are intended to identify the phishing website and feature selection has been confirmed [6,23,20,21]. This fixation and lack of flexibility in the selection of features for identifying phishing websites, led us choose the features with GSA algorithm in this article to select the features with high detection rates.

The approaches based on hyper heuristic algorithms are less complicated and often, they reach the answer, quickly. Their simple structure and application in a wide range of problems are the cause for researchers to support these hyper heuristic algorithms. Therefore, these approaches do not have the defects of machine learning methods and give the answer more quickly. In this research, hence, gravitational search algorithm is considered and implemented as a hyper heuristic one for detection of phishing websites and selection of characteristics and optimizing the process of classifying phishing and legal website, due to its significant advantages such as quick convergence, not sticking in local maximum, reduction of computational volume and not needing for memory relative to other hyper heuristic algorithms.

## 3. The proposed method

This study involved detecting malicious URLs in social network environments attackers may use SNSs as vehicles, using compromised user accounts to post messages that contain malicious URLs. Social network users typically trust the information that their friends submit in posts and feeds; thus, they become the victims of social engineering attacks. Therefore, in addition to the traditional attributes of malicious URLs, social networking heuristics should be addressed to facilitate identifying malicious URLs in social networks.

Figure (1) shows the proposed system and Briefly explain each of the steps.

- **File features of web site** : 8000 contains features extracted from different websites(table 2)
- **Select the appropriate features GSA** : By gravitational properties of optimal algorithm for Classification the websites we use it.
- **Learning & Test Data** : To evaluate the Classification of the training data set is used.
- **Classification websites** : Classification of phishing websites and legal with support vector machine algorithm.

### 3.1. Gravitational Search Algorithm

Optimization in GSA algorithm is done with the aid of laws of gravity and motion, in an artificial system with discrete time. System environment is the same as problem definition range. According to law of gravity, any mass understands the location and status of other masses through the force of gravity. The system space will be determined in the first step. The space consists of a multi-dimensional coordinate system in the problem definition space. Every point of space, is one solution for the problem. Searching factors, are a set of masses. Every mass has four characteristics:

- a) The mass state b) active gravitational mass c) passive gravitational mass d) the inertial mass. [23]
- In this algorithm, first, N number masses in problem answers space that can have D dimension is randomly created, and then according to the location and fitness of masses, which is determined with the fitness function by equation (1), certain amount of mass would be given to each of them.

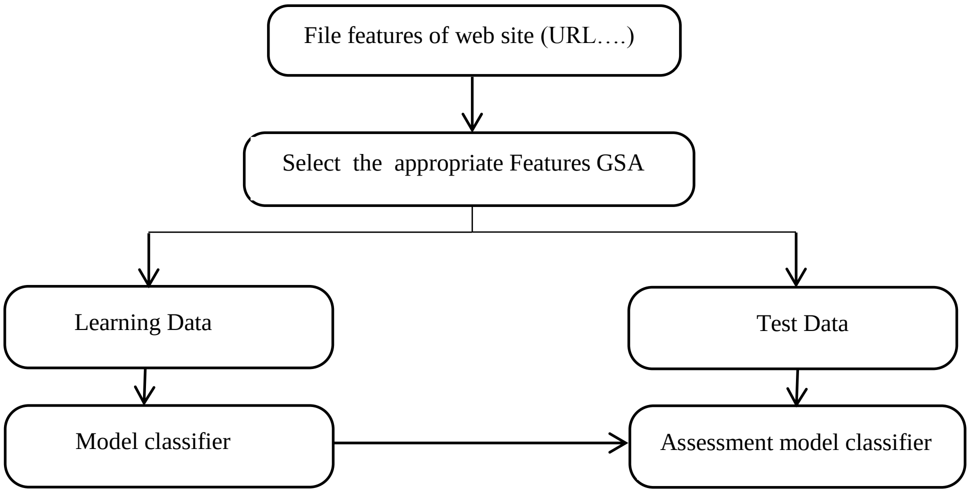
File features of web site (URL….)
Select the appropriate Features GSA
Learning Data
Test Data
Model classifier
Assessment model classifier

**Figure 1** The proposed system

(1) $$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)}$$

(2) $$m_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)}$$

$$worst(t) = \max_{j\in\{1..N\}} fit_j(t)$$

(3) $$best(t) = min_{j\in\{1..N\}} fit_j(t)$$

In equation (1), $\text{fit}_i(t)$ represents the mass fitness i at time t. In equation (2) mi is the mass of the inertial mass of the i. Minimum finding problems in equation (3) can be used to calculate the best and worst. Inspired by the law of gravity, masses, force each other, and by this, every mass take an amount of acceleration in a particular direction of which obtained acceleration vector is calculated by equation (4).

(4) $$a_i^d(t) = G(t)\sum_{j=1_{j\neq i}}^{N}\left[rand\frac{M_j(t)}{R_{ij}(t)+\varepsilon}\left(x_j^d(\, t) - x_i^d(t)\right)\right]$$

(5) $$G(t) = G_0 * e^{-\propto\frac{t}{T}}$$

In the above equation, R is the distance between the particle j, I at dimension d-th, and in iteration t-th of the algorithm and M, the mass of the j-th particle shows that particle i is under the force causing the acceleration applied to it. Sigma function is applied to collect the incoming momentum of all particles on particle i in each dimension and multiplied displacement vector specifies the direction of acceleration. Also, G (t) is considered as the gravitational constant in each iteration of the algorithm and is updated according to the given equation. Variable T represents the number of iterations of the algorithm and G0 and α is determined by the user.

Speed (vi) and position (xi) of each particle, due to the applied acceleration (ai) are updated by the following equations.

(6) $$v_i^d(t+1) = rand_j * v_i^d + a_i^d(t)$$
$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1)$$

By changing the position of each particle, again according to equation (1) the amount of fitness will be determined and previous procedure repeats, this process is executed for T iterations of the algorithm, finally, the best particle in Tth iteration will be extracted as the answer of algorithm.

### 3.2. Phishing Classifier With SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. For a linearly separable set of 2D-points which belong to one of two classes, find a separating straight line. [19

In the above picture you can see that there exist multiple lines that offer a solution to the problem. If any of them better than the others, we can intuitively define a criterion to estimate the worth of the lines: A line is bad if it passes too close to the points because it will be noise sensitive and it will not generalize correctly. Therefore, our goal should be to find the line passing as far as possible from all points. Then, the operation of the SVM algorithm is based on finding the hyper plane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of margin within SVM's theory. Therefore, the optimal separating hyper plane maximizes the margin of the training data

GSA algorithm has considerable ability to assist in making the correct data classification rate by the SVM. To simulate the results of this paper, MATLAB software is used and to identify phishing website, the data set of http://www.phishtank.com that contains a list of phishing websites URLs that are regularly updated.

In this article, we have selected variably 4, 8, 11 and 15 features among the feature sets of 8000 websites which have been used to prove the impact of gravitational search algorithm on selection of optimum features to classify them by support vector machine. We gave these features to the simulator with amounts of 11, 8, 4 and 15_the names

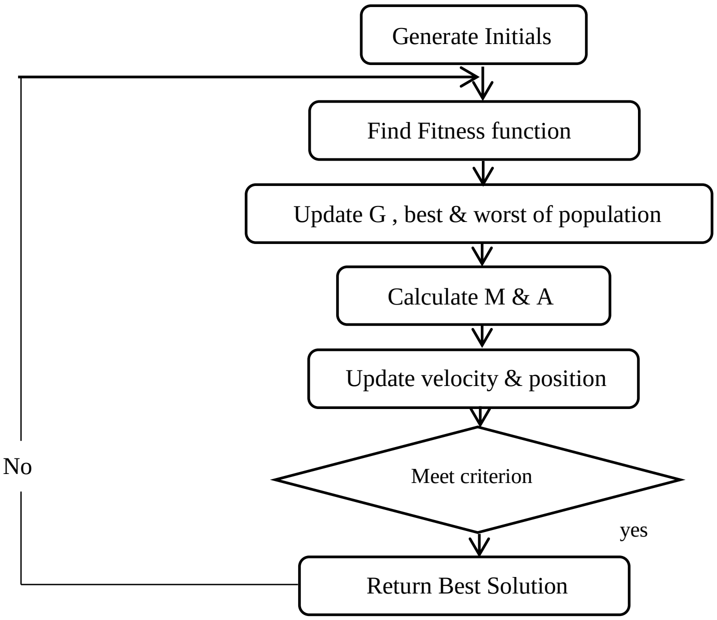

**Figure 2 Steps for gravitational search algorithm [24]**

**Table 2**
*selection of different features to learn*

| 15_Feature | 11_ Feature | 8_ Feature | 4_ Feature |
|---|---|---|---|
| • Right Click | • SFH | • SFH | • URL of Anchor |
| • URL of Anchor | • URL of Anchor | • URL of Anchor | • URL Length |
| • Support HTTPS | • SSL final State | • URL Length | • Pop Up Window |
| • SFH | • URL Length | • Pop Up Window | • Request URL |
| • URL Length | • Pop up Window | • Request URL | |
| • Having At Symbol | • Having IP Address | • Redirect | |
| • Pop Up Window | • Request URL | • Abnormal URL | |
| • Having IP Address | • Having Sub Domain | • SSL final State | |
| • Request URL | • Prefix Suffix | | |
| • Having Sub Domain | • Abnormal URL | | |
| • Prefix Suffix | • Redirect | | |
| • SSL final State | | | |
| • Redirect | | | |
| • On mouse over | | | |

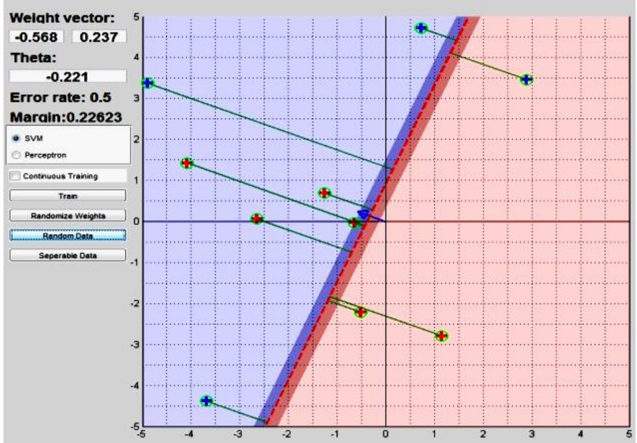
Weight vector:
-0.568 0.237
Theta:
-0.221
Error rate: 0.5
Margin:0.22623
● SVM
○ Perceptron
□ Continuous Training

**Figure3** (han,et al.,2006)

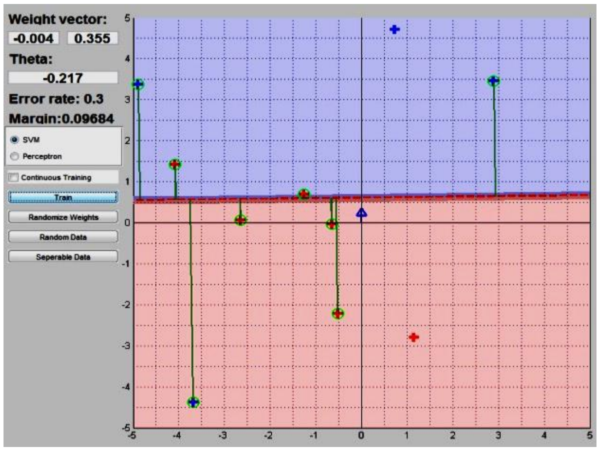
Weight vector:
-0.004 0.355
Theta:
-0.217
Error rate: 0.3
Margin:0.09684
● SVM
○ Perceptron
□ Continuous Training

**Figure 4 (han,et al.,2006**

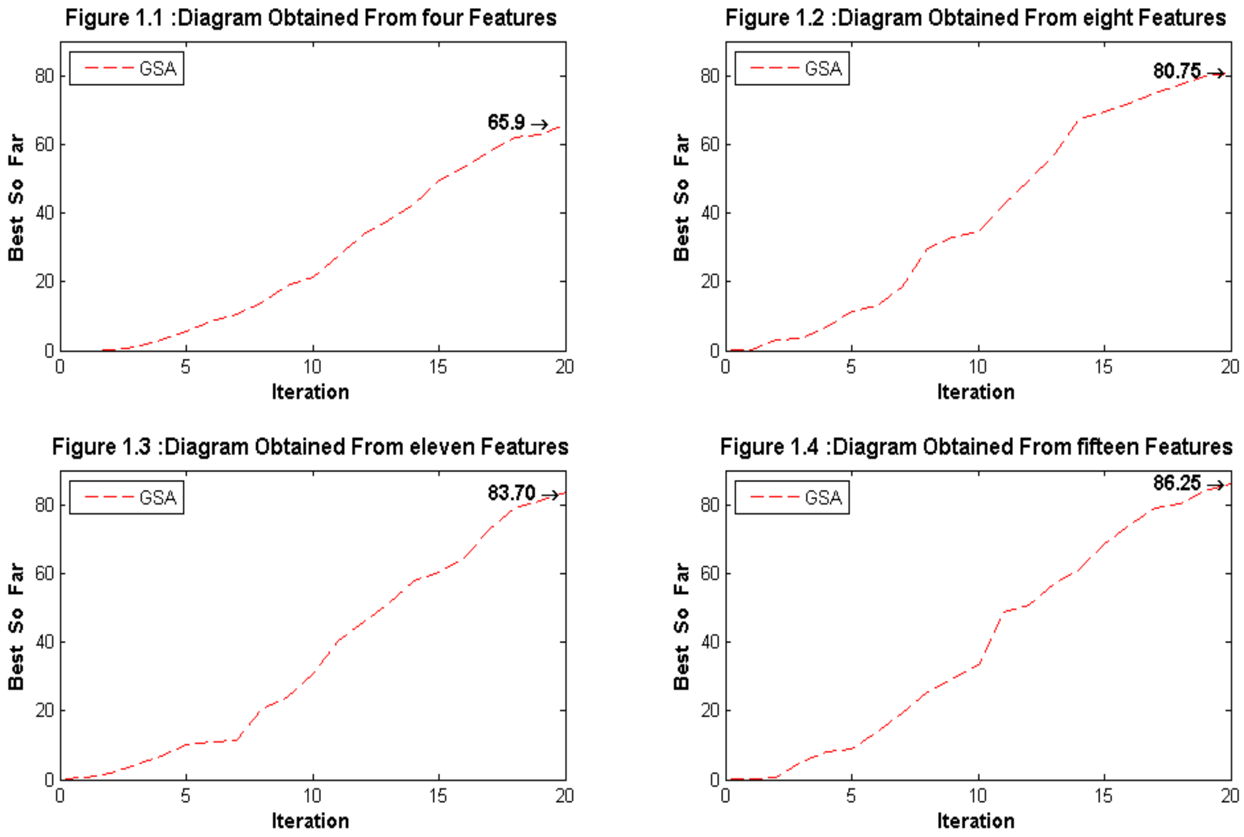

**Figure 5 Diagrams obtained from eleven features in program**

**Table 3**
*The results obtained from four different features in the program*

| feature | Classification without GSA | classification rate with GSA |
|---|---|---|
| 4_feature | 65 | 65.9 |
| 8_feature | 76.91 | 80.75 |
| 11_feature | 60.83 | 83.70 |
| 15_feature | 63.75 | 86.25 |

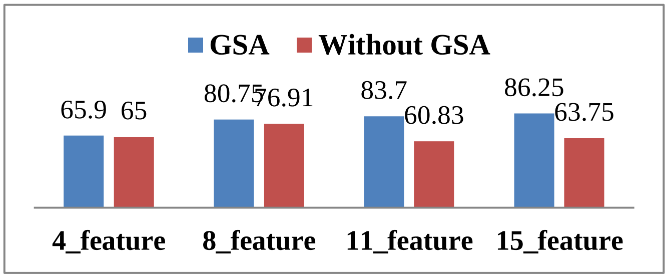

**Figure 6 Bar graph of the effect of GSA in categorization**

**Table4**
*Comparing the proposed method with other heuristic methods*

| | PSO | GSA | BFOA | ACO | MBAT |
|---|---|---|---|---|---|
| Error Rate in % | 9.13 | 7.80 | 8.67 | 10.23 | 8.56 |
| Time Taken in ms | 9578 | 8190 | 8211 | 11276 | 7902 |
| Accuracy | 92% | 87% | 97% | 89% | 98% |

of these features are shown in Table.2 in order for the rate of optimum selection to be determined by gravitational search algorithm.

## 4. Evaluation

Features extracted from the sites are shown in Table 2. Best features are selected by the GSA algorithm and are classified by SVM. To demonstrate the impact of features number in diagnosing the correct rate of classification, we will give the different primary features to the program and this is done by determining 70% as training data and 30% as learner data in the program.

As mentioned earlier, the columns of Table 2 show the features number. The best features are

selected by gravitational search algorithm and converted into some inputs to be classified by support vector machine. We accomplish the process by determining 70% of the selected features as training data and 30% of them as learner data. As we can see in Figure5, the rate of classification detection along with 4 of the selected features has shown different results. For example, rate of classification with 15 features is 63.75%, which increases up to 86.25% with gravitational search algorithm for optimally selecting the features.

The results of using gravitational search algorithm and not using that in selecting the results for 4 features are shown in Table2 , bar graph of Figure 6 and Table 3..

Also, in order to compare the approach presented in this research with the other methods in this field, Table.4 and bar graph of Figure.7, simulated from comparison of

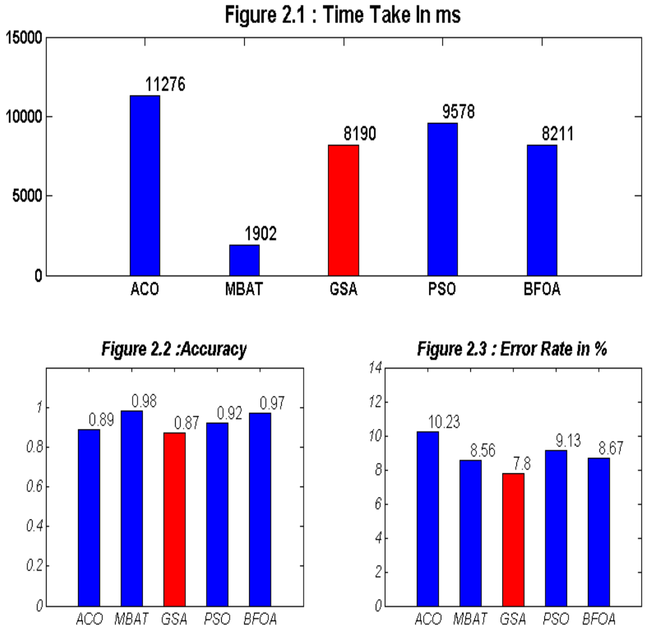

**Figure 7 Bar graphs of the Runtime ,Accuracy and Error rate**

approaches are presented below. In all existing approaches, the basis is comparison of 15 features. [13]

The results obtained show that the presented approach in this article has won the first place in the fields of accuracy rate (87%) and error rate (7.8) and won the second place in the field of runtime (8190ms)

As we imply from the results of simulation, the number of feature has a considerable impact on accurate rate of classification with gravitational search algorithm. According to Table.3, the rate of classification increase with increment of features number. A significant point in using gravitational search algorithm is that as the number of features increases, this algorithm shows its power more and more. The reason is that gravitational search algorithm will has more optimum selection, if it selects more features.

## 5. CONCLUSION

In this paper, to detect phishing websites and classifications by machine learning algorithms "support vector machine" and also to raise the rate of correct classification, heuristic algorithms "gravitational search" are used to select the appropriate feature from available features. The results for 15 features show that using gravitational search algorithm has significant impact in terms of runtime (8190ms) which is placed after the MABT algorithm (7902ms) and in terms of error rate with (7.8%) and accuracy with (0.87) in the first place, among the studied algorithms.

To classify the website in one of the phishing classes, a rule was used by SVM. As the most important challenges, using another classification can be noted. Another reason for the low accuracy rate, is feature selection issue which has a significant impact on the accuracy, and in future works, we could focus on the classification and selection of other features.

## REFERENCES

1. Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.
2. Garera, Sujata, et al. "A framework for detection and measurement of phishing attacks." Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007.
3. Aburrous, Maher, et al. "Associative classification techniques for predicting e-Banking phishing websites." Multimedia Computing and Information Technology (MCIT), 2010 International Conference on. IEEE, 2010.
4. Aburrous, Maher, et al. "Intelligent phishing detection system for e-banking using fuzzy data mining." Expert systems with applications 37.12 (2010): 7913-7921.
5. Blasi, Michael. "Techniques for detecting zero day phishing websites." (2009).
6. Rao, Hima Sampath, and SK Abdul Nabi. "A NOVEL APPROACH FOR PREDICTING PHISHING WEBSITES USING THE MAPREDUCE FRAMEWORK." (2014).
7. Zhang, Dongsong, et al. "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites." Information & Management 51.7 (2014): 845-853.
8. Ajlouni, Moh'D. Iqbal AL, Wa'el Hadi, and Jaber Alwedyan. "Detecting phishing websites using associative classification." European Journal of Business and Management 5.15 (2013): 36-40.
9. Mitchell, Tom. "The role of unlabeled data in supervised learning." Proceedings of the sixth international colloquium on cognitive science. 1999.
10. Barraclough, P. A., et al. "Intelligent phishing detection and protection scheme for online transactions." Expert Systems with Applications 40.11 (2013): 4697-4706.
11. DeMaris, Alfred, and Steven H. Selman. "Logistic regression." Converting Data into Evidence. Springer New York, 2013. 115-136.
12. Garera, Sujata, et al. "A framework for detection and measurement of phishing attacks." Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007.
13. Radha Damodaram, M. C. A., and M. L. Valarmathi. "Phishing Website Detection and Optimization Using Particle Swarm Optimization Technique."International Journal of Computer Science and Security (IJCSS) 5.5 (2011): 477
14. Sengar, P. K., and Vijay Kumar. "Client-side defense against phishing with pagesafe." International Journal of Computer Applications 4.4 (2010): 6-10.
15. Abu-Nimeh, Saeed, et al. "A comparison of machine learning techniques for phishing detection." Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit. ACM, 2007.
16. De Sa, JP Marques. Pattern recognition: concepts, methods, and applications. Springer Science & Business Media, 2001

17. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32
18. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.
19. Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.
20. Damodaram, Radha Damodaram, "Experimental Study on Meta Heuristic Optimization Algorithms for Fake Website Detection " International Association of Scientific Innovation and Research (IASIR) vol. 2 pp. 43-53 2012
21. Radha Damodaram, M. C. A., and M. L. Valarmathi. "Bacterial Foraging Optimization for Fake
25. .
22. Precup, Radu-Emil, et al. "Experiments in fuzzy controller tuning based on an adaptive gravitational search algorithm." PROCEEDINGS OF THE ROMANIAN ACADEMY SERIES A-MATHEMATICS PHYSICS TECHNICAL SCIENCES INFORMATION SCIENCE 14.4 (2013): 360-367.
23. Rashedi, Esmat, Hossein Nezamabadi-Pour, and Saeid Saryazdi. "GSA: a gravitational search algorithm." Information sciences 179.13 (2009): 2232-2248.
24. Ojugo, A. A., et al. "A Hybrid Artificial Neural Network Gravitational Search Algorithm for Rainfall Runoffs Modeling and Simulation in Hydrology." Progress in Intelligent Computing and Applications 2 (2013): 22-33