Modern Applied Science; Vol. 13, No. 5; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Using Sequential Minimal Optimization for Phishing Attack Detection

Ali Mohammad H. Al-Ibrahim[1]

[1] Computer Science Department, Faculty of Information Technology, The World Islamic Sciences and Education University (WISE), P.O. Box 1101, Amman 11947, Jordan

Correspondence: Ali Mohammad H. Al-Ibrahim, Computer Science Department, Faculty of Information Technology, The World Islamic Sciences and Education University (WISE), P.O. Box 1101, Amman 11947, Jordan. Tel: 962-560-0230 (Extra 2363)/962-799-112-999. E-mail: ali.alibrahim@wise.edu.jo

Received: October 14, 2018 Accepted: November 28, 2018 Online Published: April 30, 2019

doi:10.5539/mas.v13n5p114 URL: https://doi.org/10.5539/mas.v13n5p114

Abstract

With the development of Internet technology and electronic transactions, the problem of software security has become a reality that must be confronted and is no longer an option that can be abandoned. For this reason, software must be protected in all available ways. Where attackers use many methods to enable them to penetrate systems, especially those that rely on the Internet and hackers try to identify the vulnerabilities in the programs and exploit them to enter the database and steal sensitive information.

Electronic phishing is a form of illegal access to information, such as user names, passwords, credit card details, etc. Where attackers use different types of tricks to reveal confidential user information. Where attacks appear as links and phishing is done by clicking on the links contained in them. This leads to obtaining confidential information by using those false emails, redirecting the user without his knowledge to a site similar to the site he wants to access and capturing information. The main purpose of this paper is to protect users from malicious pages that are intended to steal personal information. Therefore, an electronic phishing detection algorithm called the SMO algorithm, which deals only with the properties of links, has been used.

Weka was used in the classification process. The samples were the characteristics of the links and they contain a number of sites which were 8266 and the number of phishing sites 4116 and legitimate sites 4150 sites and results were found to be new for the previous algorithms where the real classification rate 99.0202% in the time of 1.68 seconds.

Keywords: Phishing, SMO (Sequential Minimal Optimization), Attackers, Classification, URL, Password, Weka

1. Introduction

The need for information security has increased significantly in this time with the development of technology and the emergence of new technologies to protect data and exchange it in different ways or transmit it over the network, which puts sensitive information at risk, and presents it Risk and loss.

As well as applications that operate on the network have been attacked and infiltrated by attackers, disrupting the commercial and economic movement of many institutions, and it became clear that there is a need to protect data, applications and programs that operate on the network of theft to ensure the trust of customers and continuity of services, Can lead to data loss Phishing: It is a form of obtaining sensitive information in illegal ways, such as for usernames, passwords, credit card details, etc. [1].

Migrants use different kinds of tricks to detect user information, where phishing attacks appear mostly as phony e-mails that appear as links and are phishing by clicking on the links contained in e-mail messages. This type of e-mail is called e-mail because its recipients use e-mail as bait to hunt for secret numbers and other sensitive personal data.

Phishing is a modern problem; however, it has a significant impact on the financial and commercial sectors. The trolls send fake e-mails asking web users to visit one of the websites so that the user is asked to update their data. These sites are designed only to steal user information, and the password to enter his e-mail without knowing that the data was accessed [1]

2. Electronic Phishing

Is a form of obtaining illegally sensitive information, such as usernames, passwords, and credit card details. The

attackers use different methods to deceive the user in revealing confidential information? Phishing attacks appear as false e-mail messages and are deceived by clicking on links in e-mail. This type of e-mail is called electronic phishing message because its senders use e-mail message as bait to hunt password. And other sensitive personal data. [4] It is also known as one of the deceptive methods used to write the content of a message, namely that attempts to access the personal account of the future have been exhausted and that the form must be filled Found in the link in the message and this link to the site spoofed designed to match the design of the original site in terms of look and feel and the name of the counterfeit location and the purpose is to try to deceive the future and to steal the data entry to the electronic account that Usually the user name and password, and the e-mail message itself may be provided with a form and the recipient is asked to fill it [1].

2.1 Types of Phishing

Phishing was based on the use of false e-mail messages containing links to Web sites aimed at obtaining Internet user information.

Phishing techniques have evolved to include new techniques including:

1. Deceptive messages
2. Malware- based phishing
3. Session hijacking
4. Hosts file poisoning
5. Data theft
6. DNS- based phishing (pharming)
7. Search engine phishing

2.2 Prevention of Phishing

1. Protect your computer using anti viruses and must be constantly updated.

2. Make sure to update the user's Internet browser

3. Make sure to use the secure website incase enter private information by making sure that the site starting by "https: //"

4. be wary of links in emails that lead to electronic (web) pages

5. Avoid filling out relevant forms that ask you for personal information.

6. Avoid giving private information such as PINs or passwords when talking on the phone with banks or financial institutions because they do not require this information over the phone but require personal presence.

3. Methodology

3.1 Data Collection

The training data were obtained from the Internet from a well-known international site known as UCI, which provides a set of test samples (datasets) used by researchers in their experiments and laboratory tests. This site is referred to in the reference list for clarification.

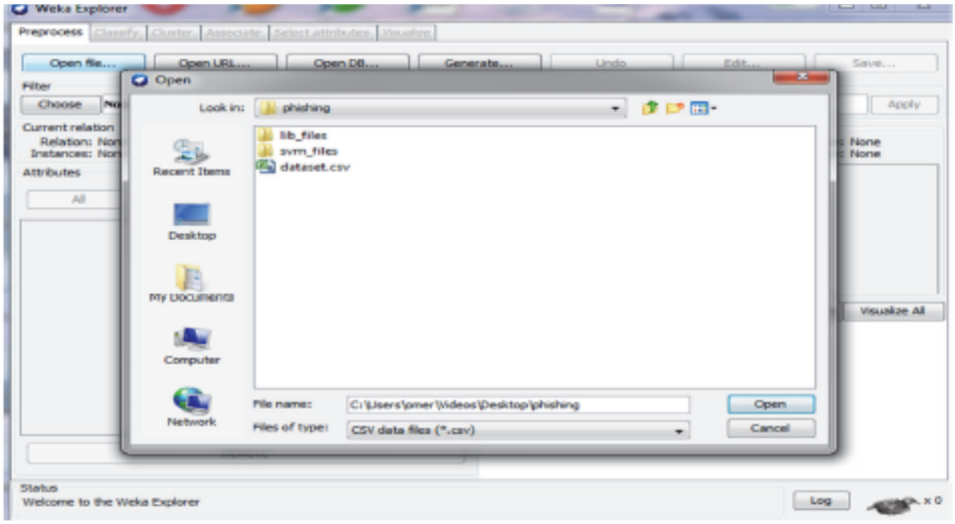

Figure 1. The Interface of the Introduction of Dataset. Csv in the Weka Program

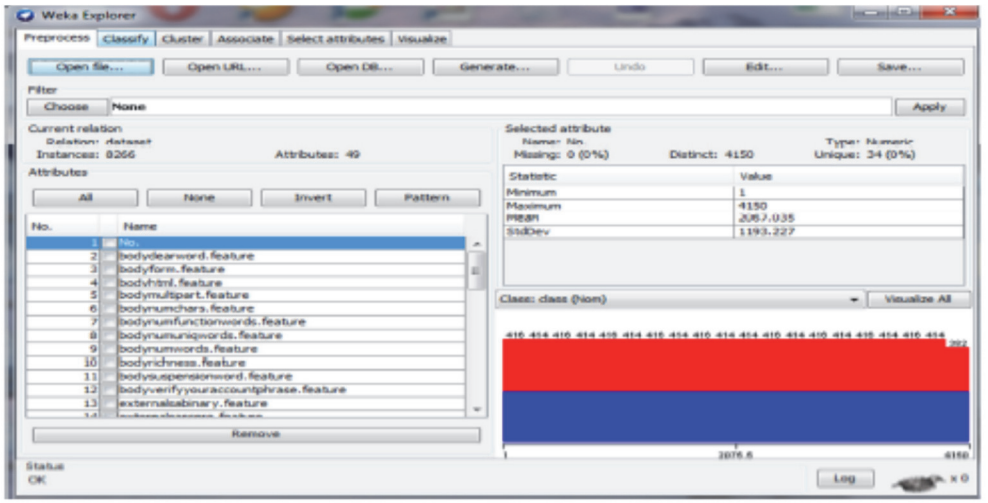

Figure 2. pening the sample file (dataset) and the appearance of features within the Weka program

3.2 Measurements

To discover sites and check them, whether legitimate sites or phishing sites. Modern site verification techniques must be used. Thus, the Weka program was used as a machine learning software that contain set of machine learning algorithms, the observer and the non-observer (3). The method of observing learning has been used in this paper in addition to the use of the algorithm of sequential minimal optimization (SMO), which is one of the machine learning algorithms.

This algorithm is based on a number of situations, so the best case, the worst case, and the middle case will be dealt with [2]. Therefore, we will calculate these cases only for the application of the algorithm in the WIKA program according to the percentage spilt, which is the percentage of learning algorithm where the ratios used in the training were 80%, 70% and 60%. The classification was based on what is known as the correct classification instance and incorrect classification instance in order to distinguish between sites whether legitimate sites or phishing sites, and based on comparisons between the results of this algorithm and the previous algorithms in the same subject, whether or not this paper has contributed.

4. Results

This paper uses the SMO algorithm, which is one of the algorithms used to detect and categorize URL links to determine whether this link is to a phishing site or a legitimate site. The total number of samples (dataset) is 8266

site, where 4116 phishing sites and 4150 legal sites. The correct TP was 99.0201% and the wrong rating was 0.97%. The correct TP rate was 0.99%, the FP rate was 0.01%, precision 0.99%, which is the average rate of TP & FP.

All this indicates that the results of the algorithm are good compared to the previous studies and we will provide a presentation of the results after the implementation of the algorithm detailed in the table below.

4.1 Percentage split 80%

Table 1. Classification Results when use Education Percentage 80%

	Terms	Results average
1	Correctly classified instances	98.67%
2	Incorrectly classified	1.33%
3	Accuracy	1.64%
4	Misclassification	-0.64
5	Total number of instance	1653

Table1 above shows that the correct classification of sites is 98.36% and be a good result. The error rate, ie, the incorrect classification rate was 1.3309%. Accuracy rate is also calculated using the previous equation. The percentage of impurities in the selected samples (Misclassification) was also calculated and was - 0.644 (negative). The samples were reduced from their total number through the algorithm because of the properties identified to the 1653 site as shown in Table1.

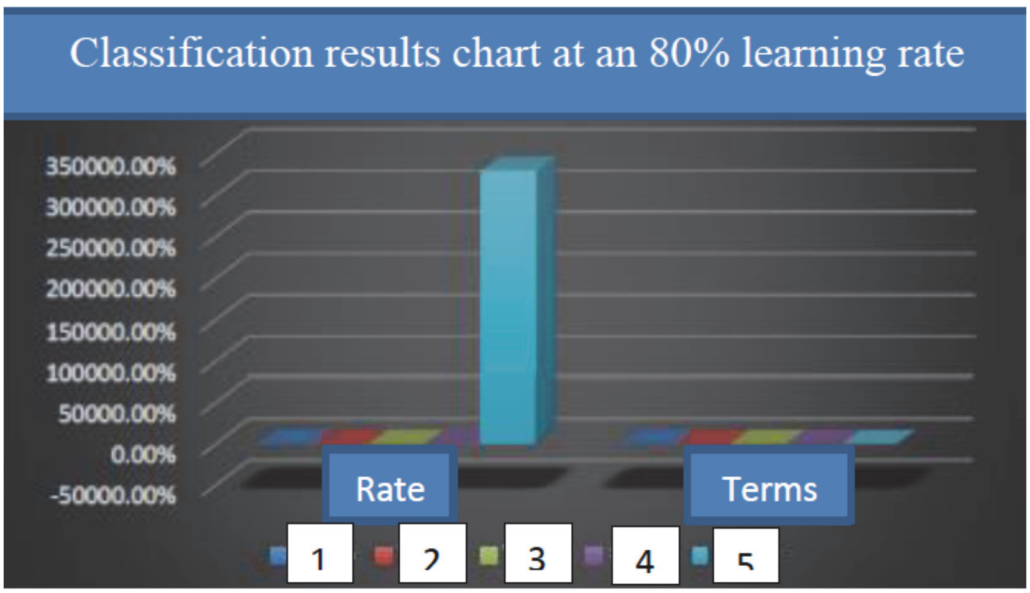

Figure 3. Classification Results Chart at an 80% Learning Rate.

4.2 Percentage Split 70%.

Table 1. Classification Results when use Education Percentage 70%

	Terms	Results average
1	Correctly classified instances	98.67%
2	Incorrectly classified	1.33%
3	Accuracy	2.00%
4	Misclassification	-1.001
5	Total number of instance	2480

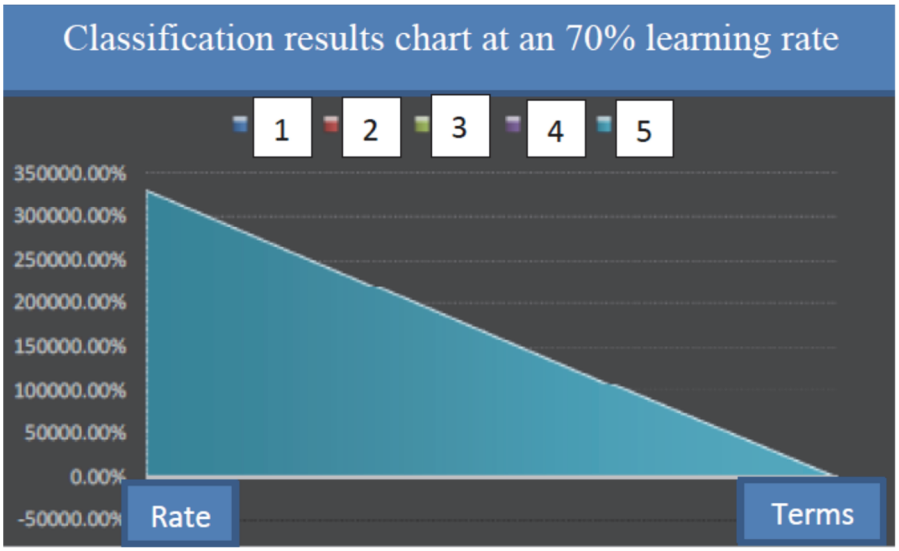

Figure 4. Classification Results Chart at 70% Learning Rate

When the percentage split was determined by 70%, the algorithm achieved very excellent classification rates. The correct classification rate was 98.6694%. The error rate, i.e., the wrong classification rate was 1.3306% and Accuracy rate was also 2.002%. The miscalculation was also calculated in the selected samples and the rate of -1.002 the samples were reduced from the total number through the algorithm because of the characteristics identified to 2480 sites as shown in Table 2 above. This indicates that the algorithm in the case of learning ratio of 70% gives better results than learning by 80% as shown in the Figure 4.

4.3 Percentage Split 60%

Table 1. Classification Results when use Education Percentage 60%

	Terms	Results average
1	Correctly classified instances	98.58%
2	Incorrectly classified	1.42%
3	Accuracy	2.08%
4	Misclassification	-1.08
5	Total number of instance	3306

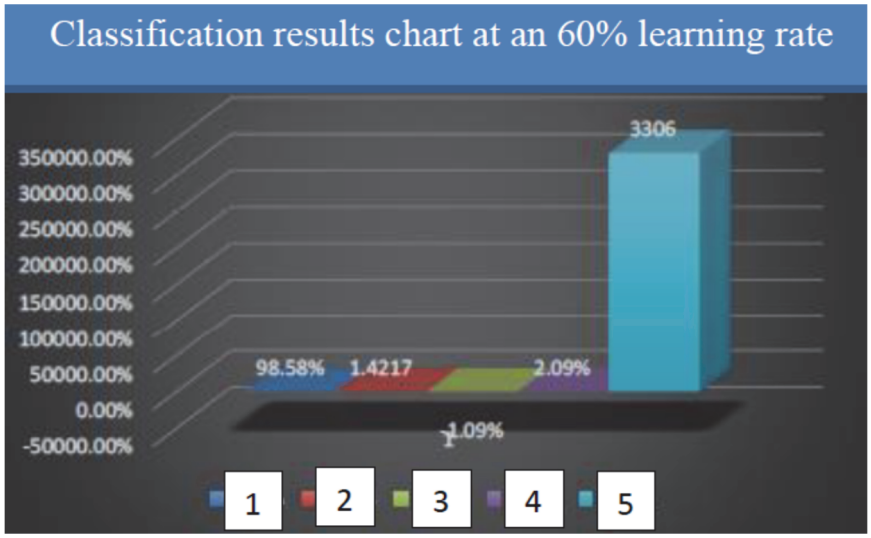

Figure 5. Classification Results Chart at an 60% Learning Rate

When the percentage split was determined by 60%, the algorithm achieved very excellent classification rates. The correct classification rate was 98.578%. The error rate, i.e., the wrong classification rate was 1.42% and Accuracy was also 2.088%. The miscalculation was also calculated in the selected samples with rate of -1.0788 the samples were reduced from the total number through the algorithm because of the characteristics identified to 3306 sites as shown in Table 3.

4.4 Comparison of Sample Classification Results

The results were compared when the percentage split of different percentages was determined. The ratio was divided between the selection and training. The percentage of selection was 80%, the second time was 70% and the third time was 60%. Where the best results were at percentage 70% which achieve the highest rate of classification with correct classification rate 98.67% and the following table shows previous results.

Table 4. Previous Results

Percentages	Classified Accuracy	Incorrectly percentage
80%	98.669%	1.3309%
70%	98.6694%	1.3306%
60%	98.578%	1.4217%

From Table4, we find that the best results were reached when we selected the 70% selection ratio and the 30% training rate. These results can be illustrated by the following chart showing the results in the form of charts:

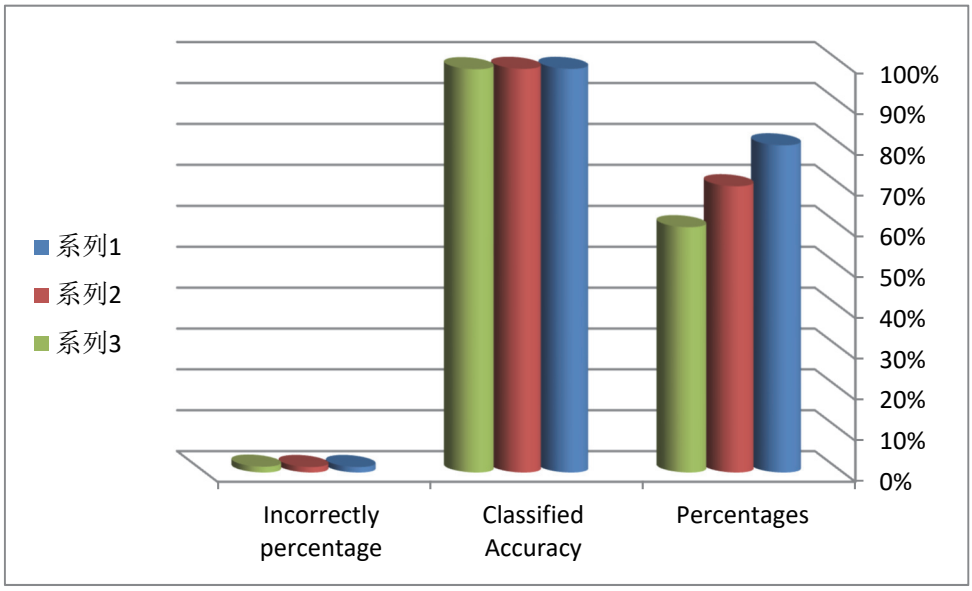

Figure 6. Comparison Results Classification of the sample (Dataset)

5.4 Comparison Results of Previous Studies with The Result of Classification of The Algorithm SMO

Table 5. Comparison of the Results of Previous Studies

Algorithm	TP	TN	FP	FN
SVM[4]	97.00%	94.90%	5.10%	3.00%
Decision Tree[7]	96.90%	96.90%	3.10%	3.10%
Naïve Bayes[5]	90.90%	95.10%	4.90%	9.10%
SMO Algorithm	99.02%		1.33%	

4.6 Comparison Results of Previous Studies with The Result of The Classification of The Algorithm SMO

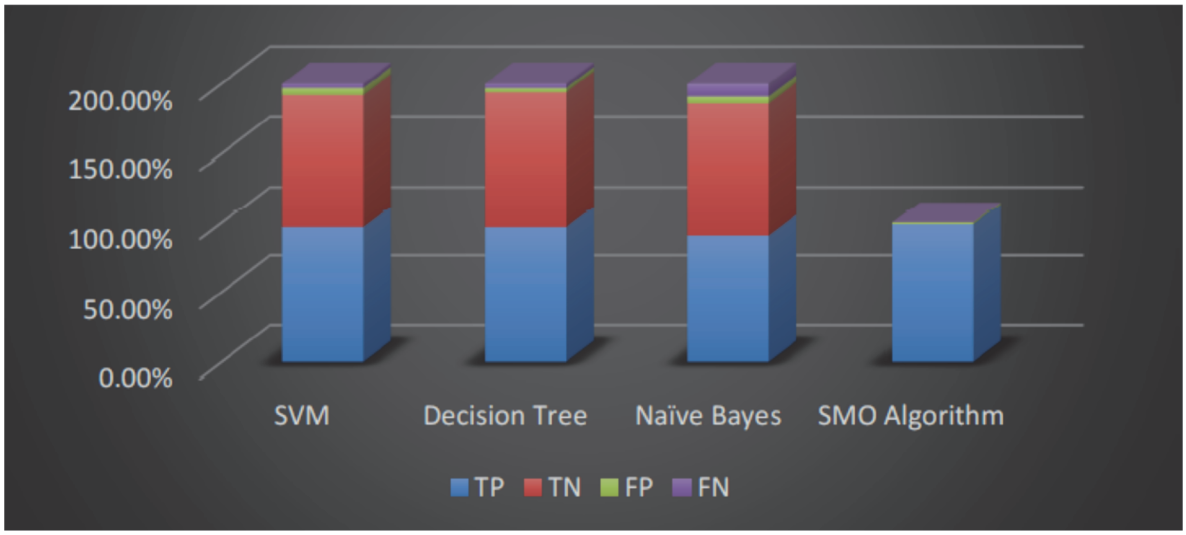

Figure 7. Comparison of the Results of Previous Studies with the Algorithm SMO

Of the above results in Table 2 and Figure 7, we note that Algorithm SMO achieves better results than the results of other algorithms in terms of correct and false prediction rate FP, TP as shown in Figure 6.

5. Conclusions

In this paper, an algorithm was used from detect algorithm or classify links to determine the type of site, whether it was a phishing site or a legitimate site. The SMO algorithm was chosen and applied through Weka program and 8266 training data(Dataset) were used. Where 4116 of them is Phishing site and 4150 legitimate sites. The correct classification rate was 99.02%. The error rate was 0.97. The average rate of TP was 0.99 and the average rate of FP was 0.01%. However, the results of the algorithm are good.

After the classification of all dataset, the best feature of the dataset was determined by an algorithm of determined features algorithms, this algorithm is called infGainAttribute, which extracts the best features. Then, the percentage split was determined and three ratios were calculated. The results were mixed between the three ratios. In the first instance, 80% was given. The number of dataset was 1653 feature. The classification was done with the SMO algorithm. The correct classification rate was 98.667%, the incorrect rate was 1.33%, and the TP rate was 0.987% and FP 0.013 and accuracy rate was also calculated with a score of 2.001% and a misclassification was calculated with a result of -1.001.

In addition, at 70% and 60% of the percentages of learning were obtained with relatively different ratios on different ratio. Those who found the best ratios were at 70%. When comparing the results of this algorithm with the previous studies in the field of classification, as shown in the results. We found that this algorithm achieve the best result.

Acknowledgment

This work is encouraged by the World Islamic Science and Education University (WISE), Amman Jordan.

References

Bhutan, D. D. (2014). A Hybrid Model to Detect Phishing-Sites using Clustering and Bayesian Approach. International Conference for Convergence for Technology. https://doi.org/10.1109/I2CT.2014.7092141

Christopher, J. C. B. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Jin, L. L., Dong, H. K., & Chang, H. (2015). Heuristic-based Approach for Phishing Site Detection Using URL features. Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET. https://doi.org/10.15224/ 978-1-63248-056-9-84

Khalid, A. (2013). *Analysis and Design of Algorithms.* Retrieved from https://www.tutorialspoint.com/design_and_analysis_of_algorithms/index.htm

Khalid, S. al-G., & Sulayman, A. H. (2012). Electronic Methods and Countermeasures. IEEE AUTOTESTCON Proceedings. https://doi.org/10.1109/AUTEST.2012.6334540

Rohit, G., & Kaushal, S. D. (2017). *Machine-learning-databases.* Retrieved from https://towardsdatascience.com/machine-learning-in-your-database-the-case-for-and-against-bigquery-ml-4f2309282fda?gi=e38c0e844ad0

Zdravko, M. (2006). An Introduction to the WEKA Data Mining System. ITiCSE. https://doi.org/10.1145/1140124.1140127