A Comparative Analysis on Artificial Intelligence Techniques for Web Phishing Classification

IOP Conf. Series: Materials Science and Engineering **769** (2020) 012073 doi:10.1088/1757-899X/769/1/012073

A Comparative Analysis on Artificial Intelligence Techniques for Web Phishing Classification

Tengku Balqis binti Tengku Abd Rashid, Jamaludin bin Sallim and Yusnita binti Muhamad Noor

Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, 26300 Kuantan Pahang Malaysia

E-mail: jamal@ump.edu.my

Abstract. Over the last few years, the web has been expanded to serve millions of users for various purposes all over the world. The web content filtering is essential to filter offensive, unwanted web content from web pages, reduced inappropriate content to prevent access to content which could compromise the network and spread malware. It also to tightened network security where web content filtering adds a much-needed layer of security to the network by blocking access to sites that raise an alarm. However, there are lack of comparison between classification techniques in previous studies in order to find the best classifier for the web page classification and the analysis related to it. Thus, the purpose of this study was to apply web page classification techniques and their performances is compared as it is the initial step in data mining before going to web filtering. In this project, three classifiers called Artificial Neural Network, J48 Decision Tree and Support Vector Machine were used to web phishing dataset in order to find the best possible classifier with small computational efforts that will give the best result in classifying the web page.

Keywords. Web page classification, Artificial Neural Network, SVM (Support Vector Machine), J48 Decision Tree.

1. Introduction

In a matter of time, the web has changed the way people do business and communicate by becoming a very powerful platform that influence everyday life. According to [1], growth of internet user from 1995 until today increase by 55% and for internet host, it growth more than almost double the amount a decade prior where the statistic shows around 1.03 billion internet hosts were available on the DNS by January 2018 [2].

From the statistic we can see the rapid growth of the Internet usage as source of information which make the classification of web pages becoming a significant process and challenging task to do as the number of web pages also increasing [3]. Web page classification is important especially in education institute for web content filtering to control the access to unwanted data, inappropriate web pages and to manage the traffic [4]. Classification is the most crucial stage in web filtering as it is the initial stage of the process.

According to [5], classification can be explained as organized learning problem where the labelled dataset is used to train classifier to before applying it to new dataset for categories determination. Web

page classification based on [6] are tasks to determine either a certain web page is in a category or categories which also known as web page categorization. It is in the area of machine learning that can provide ways to classify web pages to certain categories. There are many ways and approaches to classify web pages and this project will be focusing on classification algorithms and techniques used for classifying web pages.

Classification related problems arise as the rapid growth of Internet host and web pages happen. First problem discovered is institutions are lacking in terms of web page classification and identification without the content of web page. Most of institution still can't classify and identify the web page can be access accurately as the classifier used is selected randomly. This happen because lack of comparison between classification techniques used in order to find the best classifier for the web page classification and the analysis related to it [7].

Besides that, another problem arises is about manual classification of web pages where the classification process of web page is done manually. According to [8], manual classification is costly and need a very intensive labor. Other than it is cost more than automated process [9], have a heavy labor-intensive process, manual classification is an extremely time-consuming process that does not guarantee with the latest results [10]. Manual classification process also has limitation where it is not always reliable and adequate as it need for automated keywords prediction and tags recommendation through machine learning [11].

Lack of reliability in existing studies of web page classification also one of the problems discovered as lot of research discussing more about way to improve the existing classification algorithms and lot of the research is focusing on web page content rather than discovering new features for classifying web pages without the content of web pages [12].

Based on the problem as explained above, in the next section, which is section 2 will summarize the related technique used in previous studies. Section 3 will explain about selected techniques for this study. Section 4 is about the methodology, Section 5 will discuss about the result and will summarize the conclusion in Section 6.

2. Related works

[13] using URL based approach for web page classification where they don't need to download the whole content of the web page to do the classification. They proposed an automated way of learning universal dictionary as through this way the challenge to manage a big scale data can be avoided with the training set is made independent. They also applying this technique on another dataset gain search results gained from Google and the final result of the research by implementing the universal dictionary rather than to use dataset-specific term dictionary show that the difference is not statistically significant, but the classification process is faster, and the bandwidth can be saved as they use URL features instead of taking the web page content.

[14] stated that automatic web classifier is required urgently in this fast growth of information in world wide web as manual process is time-consuming. In this research they proposed a way to reduced dimensionality of thousands of inputs by introducing two novel feature selection approaches for web page classification. As classifier need to manage massive number of web pages with thousands features and categories, it faces many problems in the implementation. So, it is important to reduce the dimensionality of the data for classification process work smoothly without facing problems. To decrease the number of input dimensionality they proposed a fuzzy ranking analysis paradigm together with a novel relevance measure and discriminating power measure (DPM). For the result of this research, analysis on fuzzy ranking show that this technique has successfully validate uncertain behavior of each relevance measure and DPM method capable in setting up a better classifier as it can decrease amount of redundancy and noise features.

Next, [15] recommended automatic recognition method using classification rule by combining content, structure and uniform resource locator (URL) attributes for news web page classification. They are focusing on classifying new web page as it is different from other web pages to retrieve only useful news web pages. In this research they used Naïve Bayes algorithm to classify news articles from other

non-news articles. Based on their research show that Naïve Bayes algorithm give competent accuracy of classification with different dataset.

Also, [16] proposed that to overcome web page classification problem it is important to use the most efficient way in order to select best feature and reducing the feature space. In this research, for web page classification best features selection they used firefly algorithm (FA) for features subset selection and J48 classifier for selected features fitness evaluation. The dataset is accurately classified in a short time by applying the FA algorithm. The result in this research show that the process of web page classification is done faster with high quality classification after reducing the features by removing unnecessary features.

In their study, [17] explained about phishing website classification based on URL and their feature extraction process. First, they extract the features of website URL then analyze the techniques used for feature selection and classification to detect the phishing website. As phishing technology has becoming more complicated and advanced, the anti-phishing program can be pass easily by them. So, in this research they proposed to increase the number of URL features extracted and to analyze the techniques for feature selection that have not been used yet in phishing websites detection based on URL. They choose Naïve Bayes and SMO algorithms for the experiment and the result show that SMO algorithms have the best accuracy in feature selection and classification of phishing website.

Next [18] pointed out about the need of web page classification to get efficient indexing, search and retrieval as web information repository has increase drastically. In this research they use SVM classifier for classification to build web classifier with highest accuracy using structure information including META tags, title, descriptions of links and alternative texts of images. Based on their research show that there is improvement in term of accuracy by combining the full text with structure information compared to traditional methods and the compatibility of SVM classifier with web classification is very high.

[19] in their research propose a novel approach for web page classification that uses the HTML information present in a web page for its classification. There are many ways of achieving classification of web pages into various domains. This paper proposes an entirely new dimension towards web page classification using Artificial Neural Networks (ANN). The presence of additional information provided by the HTML tags and the hyperlinks gives the researchers idea of exploring new techniques for representing Web sites for automatic classification. As conclusion they proposed a solution for web page classification using HTML elements of a web page. The proposed model will provide the necessary web page classification technique for fast and efficient working of the search engines. Further, it is also expected to obtain results with high classification accuracy.

Next [20] have proposed a technique for web page categorization using artificial neural network (ANN) through automatic feature extraction is proposed. The main objective behind this task is to provide an efficient way for classification of web pages. This will facilitate the different search engines to classify the web pages more efficiently and also to provide a rich web directory. The end user will also be facilitated to find the web page of their desired classes. From [21], they firstly research the SVM classifiers for Web page classification with different features based on same data set. On the Web page presentation, the HTML structure is considered and separately tested on different combination. Then they study the classification performance of SVM classifier based on the polynomial kernel function and the radius basis function (RBF) kernel function. Then, NB classifier is used to investigate the different performance of the SVM classifier and NB classifier changing with the number of the feature dimension. Finally, the results are proved that the SVM classifier has better performances.

Based on [22], three different data mining algorithms have been discussed for the analysis of anti-phishing website data sets. Theses algorithms are Random Forest (RF), Nearest Neighbor Classification (NNC), Bayesian Classifier (BC). The Random Forest shows around 68 percentage of successful result when the training data is split to 75 percentages. The Nearest Neighbor Classification technique gives better and accurate result when the checking conditions are less. The result of Bayesian Classification shows the accuracy rate is around 88 percentages for finding the phishing websites. With the comparison of all these algorithms, the Bayesian classification is more accurate and shows fast response to the system.

[23] stated that medical professionals need a reliable prediction methodology to diagnose hematological data comments. There are large quantities of information about patients and their medical conditions. In this paper they are studying the various classification algorithms. Their main aims are to show the comparison of different classification algorithms using Waikato Environment for Knowledge Analysis or in short, WEKA and find out which algorithm is most suitable for user working on hematological data. The best algorithm based on the hematological data is J48 classifier with an accuracy of 97.16% and the total time taken to build the model is at 0.03 seconds. Naïve Bayes classifier has the lowest average error at 29.71% compared to others.

[24] investigated the features selection aiming to determine the effective set of features in terms of classification performance. They compare features selection and classification methods in order to determine the least set of features of phishing detection using data mining. Experimental tests on large number of features data set have been done using Information Gain and Correlation Features set methods. Further, five data mining algorithms Naïve Bayes, KNN, Random Forest, SVM and j48 have been used to classify the web phishing data set, analyze the results and identify the efficient technique to classify the web page phishing data set. In this research work, web page's phishing data sets are used to analyze the various classification techniques and find out the efficient classifier. The research show that Random forest model shows better performance than KNN, SVM, J48 and Naïve Bayes classification models.

3. Selected techniques

Based on the related work as explained and analyzed in section 2, there are three (3) techniques were selected based on their advantages which are Artificial Neural Network, J48 Decision Tree and Support Vector Machine. Each of this technique is explained in the next section.

3.1. Artificial Neural Network

Neural networks try to replicate biological systems which is human brain where in the human brain there are synapses; connected neurons through several points. In biological systems, the changes of synaptic connections strength in response to impulse is where the learning process happened. ANN has adopted this biological analogy. Neuron and unit are the basic computation unit in an ANN that is connected by arranging them in different kinds of architectures and perceptron is the most basic architecture of the neural network that have a set of input nodes and an output node where the input units pass a set of inputs to output unit [25].

ANN is collection of highly interconnected processing elements or neurons that working together to solve specific problems such as related to pattern recognition or data classification through their ability to derive the meaning behind complicated or imprecise data. This is especially for extracting and detecting complex patterns and trends that are above the capability of humans or other computer techniques [26] and the result of using ANN in various field have showed many positive result with high probability to save time resources between input predictor and known output responses. They can track changes happen on signals over time with ability to continuously adapting to new data. Their ability to adapt continuously on new data allows them to track changes in a signal over time and can handle appropriately some problem that can't be solve by conventional statistical technique like adaptability to learn from arbitrary. Also ANN able to classify multi source data as they are non-parametric classifiers [27].

3.2. J48 Decision Tree

J48 algorithm is the improved version of the C4.5 or optimized implementation of the C4.5. The Decision tree will be the output of J48. A Decision tree is similar to tree structure where it has different nodes, such as root node, intermediate nodes and leaf node. Every node of the tree holds a decision and

the result is led by the decision as name is decision tree. Decision tree partitioned the data set input space in mutually exclusive areas with every area describing or elaborating its data points by having a label, a value or an action [23]. According to [28], decision tree mechanism is transparent and easy to follow a tree structure for the decision making. A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision-making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that characterizes the given data that leads to its being visited.

3.3. Support Vector Machine

Support Vector Machine is a machine learning technique based on Statistical Learning theory. SVM has been proved to be very effective in dealing with high-dimensional feature spaces, the most challenging problem of other machine learning techniques due to the so-called curse of dimensionality [29]. Support vector machine are basically binary classification algorithms. The basic support vector machine takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [30]. SVM has been constantly evolving to apply its excellent learning ability to many types Classification, information processing and so on [31].

SVM is one of machine learning methods, which based on statistical theory will automatically find methods utilizing hyperlink to improve precision, methods based on ontology and subject oriented methods with high dividing capacity. Classifier produced by these support vectors can maximize distances between categories. As a result, SVM gets high precision for classification. But if the training set is large, time consumed by SVM is too long [32]. When using SVM, first transforming data into high dimensional space may convert complex classification problem into simpler problem that can use linear discriminant function. Secondly, SVM provides the most useful information for classification [27].

4. Methodology

4.1. Research framework

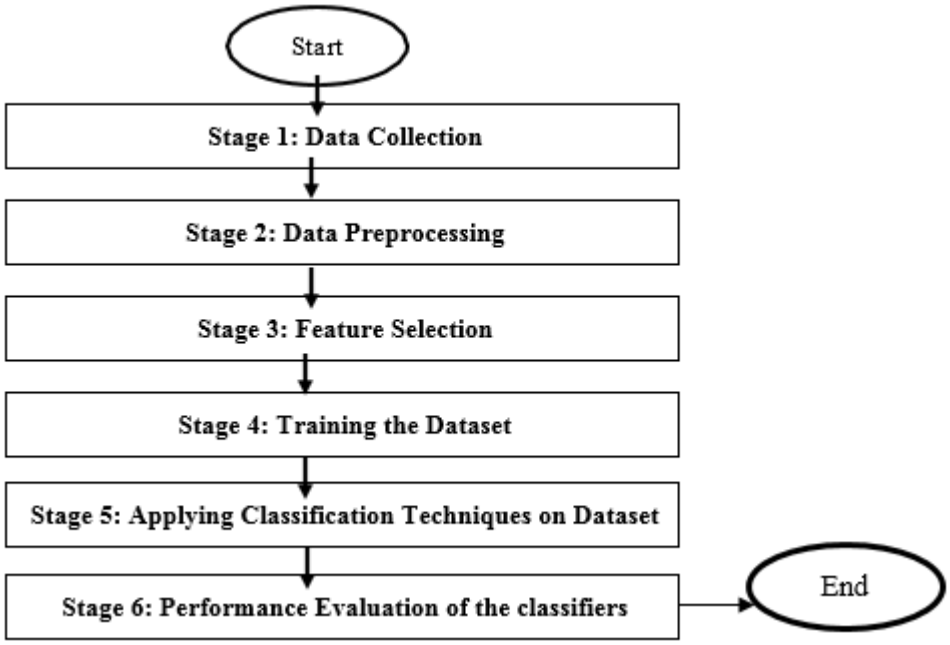

Figure 1. Research Framework

The process to evaluate the performance of classifier is composed of six steps: data collection, data pre-processing, feature selection, data training and applying the classifier techniques on the dataset which will be described in the following sections. **Figure 1** illustrates the research framework that consists of five main steps. The initial stage for evaluating classifier is to collect the data. The second step is data pre-processing which includes. The third, fourth and fifth step are the feature selection process followed with training the dataset and applying classification techniques on the dataset. The details of these stages will be elaborated in the following sections.

4.2. Data collection

The dataset used as shown in **Table 1** is from UCI repository. The researcher faced challenges about the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites have been disseminated these days, no reliable training dataset has been published publicly, may be because there is no agreement in literature on the definitive features that characterize phishing webpages, hence it is difficult to shape a dataset that covers all possible features. In this dataset, we shed light on the important features that have proved to be sound and effective in predicting phishing websites.

Table 1. Dataset Information

No	Name of Features	Rules
1.	Having_ip_Address	If the Domain Part has an IP Address → Phishing Otherwise→ Legitimate
2.	URL_Length	If URL length<54 → Legitimate else if URL length ≥54 and ≤75 → Suspicious otherwise→ Phishing
3.	Shortinig_Service	If TinyURL → Phishing Otherwise→ Legitimate
4.	Having_At_Symbol	If URL Having @ Symbol→ Phishing Otherwise→ Legitimate
5.	Double_slash_redirecting	IF The Position of the Last Occurrence of "//" in the URL > 7→ Phishing Otherwise→ Legitimate
6.	Prefix_suffix	If Domain Name Part Includes (-) Symbol → Phishing Otherwise → Legitimate
7.	Having_sub_Domain	If Dots in Domain Part=1 → Legitimate Else if Dots in Domain Part=2 → Suspicious Otherwise→ Phishing
8.	SSLfinal_State	If Use https and Issuer is Trusted &and Age of Certificate≥ 1 Years → Legitimate Else if Using https and Issuer Is Not Trusted → Suspicious Otherwise→ Phishing
9.	Domain_registeration _length	If Domains Expires on ≤ 1 years → Phishing Otherwise→ Legitimate
10.	Favicon	If Favicon Loaded from External Domain→ Phishing Otherwise→ Legitimate
11.	Port	If Port # is of the Preferred Status→ Phishing Otherwise→ Legitimate
12.	HTTPS_token	If Using HTTP Token in Domain Part of the URL→ Phishing Otherwise→ Legitimate

13. Request_URL	If % of Request URL <22% → Legitimate Else if %of Request URL≥22% and 61%→ Suspicious Otherwise→ Phishing
14. URL_of_Anchor	If % of URL Of Anchor < 31% → Legitimate Else if % of URL Of Anchor ≥ 31% And ≤ 67% → Suspicious Otherwise→ Phishing
15. Links_in_tags	If % of Links in “<Meta>”,”<Script>” and “<Link>” <17% → Legitimate Else if % of Links in “<Meta>”, “<Script>” and “<Link>” ≥ 17% And ≤ 81% → Suspicious Otherwise→ Phishing
16. Server from Handler (SFH)	If SFH is “about: blank” Or Is Empty → Phishing Else if SFH Refers to A Different Domain→ Suspicious Otherwise → Legitimate
17. Submitting_to_email	If Using “mail()” or “mailto:” Function to Submit User Information → Phishing Otherwise → Legitimate
18. Abnormal_URL	If the Host Name Is Not Included in URL → Phishing Otherwise→ Legitimate
19. Redirect	If number of Redirect Page ≤ 1 → Legitimate Else if number of Redirect Page ≥ 2 &And <4 → Suspicious Otherwise → Phishing
20. On_Mouseover	If onMouseOver Changes Status Bar → Phishing It Does't Change Status Bar → Legitimate
21. Right Click	If Right Click Disabled → Phishing Otherwise → Legitimate
22. popUpWidnow	If Pop-up Window Contains Text Fields → Phishing Otherwise → Legitimate
23. Iframe	If Using iframe→ Phishing Otherwise → Legitimate
24. Age_of_domain	If Age Of Domain ≥ 6 months → Legitimate Otherwise → Phishing
25. DNSRecord	If no DNS Record for The Domain → Phishing Otherwise → Legitimate
26. Web_traffic	If Website Rank < 100,000 → Legitimate Website Rank > 100,000 →Suspicious Otherwise → Phishing
27. Page_Rank	If PageRank < 0.2 → Phishing Otherwise → Legitimate
28. Google_Index	If Webpage Indexed by Google → Legitimate Otherwise → Phishing

29.	Links_poiniting_to_page	If number Of Link Pointing to The Webpage = 0 → Phishing Else if Of Link Pointing to The Webpage > 0 and ≤ 2 → Suspicious Otherwise → Legitimate
30.	Statistical_report	If Host Belongs to Top Phishing IPs or Top Phishing Domains → Phishing Otherwise → Legitimate

4.3. Data pre-processing

In the third phase, data need to be prepared for modeling tools. Therefore, features should be selected and extracted. For the purpose of classification, documents are represented by their features. A feature is simply a decimal value, a measure of a given aspect of a document. For instance, the frequency of a word in the document could be a feature. Note that features don't necessarily have a decimal part. They could be Boolean, integer, or some sort of label. But all of these can be represented as a decimal number. When considered as a whole, these features form a vector of decimal numbers. The length of the vector is equal to the number of features chosen to model the documents.

4.4. Feature selection

Feature selection is pretty important as it can reduce both the data and the computational complexity. It can also get more efficient and find out the useful feature subsets. The raw data collected is usually large, so it is desired to select a subset of data by creating feature vectors that feature subset selection is the process of identifying and removing much of the redundant and irrelevant information possible. This results in the reduction of dimensionality of the data and thereby makes the learning algorithms run in a faster and more efficient manner [33].

High dimensional data consists of features that can be irrelevant, misleading, or redundant which increase search space size resulting in difficulty to process data further thus not contributing to the learning process. Feature selection is the process of selecting best features among all the features that are useful to discriminate classes. Feature selection algorithm (FSA) is a computational model that is provoked by a certain definition of relevance [34].

For this research project, a single-attribute evaluator that doesn't evaluate a subset is used. It evaluates each attribute individually. This can help to eliminate irrelevant attributes, but it can't remove redundant attributes because it's only looking at individual attributes, one at a time. We applied AttributeSelectedClassifier using Ranker search method to get fair evaluation. The ranking search method doesn't really search; it just sorts them into rank order of the evaluation. It sorts attributes according to their evaluation and can specify the number of attributes to retain. The default is to retain them all, or to discard attributes whose evaluation falls below a certain threshold or can specify a certain set of attributes that need to ignore.

4.5. Training the dataset

Training Dataset is a dataset to be used in machine learning algorithm to train the model. Creating a train and test split of dataset is one method to quickly evaluate the performance of an algorithm on related problem. The training dataset is used to prepare a model, to train it. The chosen classification techniques which is ANN, J48 DT and SVM will be trained using the training data. In this research 75% of data will be used for training and the rest will be used for testing and evaluating the model as both processes can't be done with the same data to get the best reliable estimates of the models' performance. Comparing test vs. training performance help to avoid overfitting. If the model performs very well on the training data but poorly on the test data, then it's over fit.

4.6. Applying classification techniques

Applying classification techniques on dataset is where the process implementation of techniques is done on the selected dataset. The techniques will be applied on the dataset using WEKA. WEKA tool has widely used from Machine Learning researchers and proving its strength in classification as it provides many algorithms for these tasks that can be used on our datasets. Two cases will be tested to evaluate the classification techniques performance which is with and without feature selection. Three different classification techniques will be testes where the techniques were chosen based on their different nature of work in order to compare their efficiency in terms of accuracy in both cases with and without feature selection.

4.7. Performance evaluation

This section is how the results obtained by applying the classifier on dataset is compared and analyzed in order to evaluate their performance which classifier have the highest accuracy for web page classification. The experiments in this research are evaluated using the standard metrics of accuracy, precision, recall and f- measure for Web Classification. These were calculated using the predictive classification as in **Table 2**, known as Confusion Matrix.

Table 2. Predictive Classification

		PREDICTED	
		IRRELEVANT	RELEVANT
ACTUAL	IRRELEVANT	**TN**	**FP**
	RELEVANT	**FN**	**TP**

TN (True Negative) = Number of correct predictions that an instance is irrelevant
FP (False Positive) = Number of incorrect predictions that an instance is relevant
FN (False Negative) = Number of incorrect predictions that an instance is irrelevant
TP (True Positive) = Number of correct predictions that an instance is relevant

Precision – The proportion of the predicted relevant pages that were correct:

$$\text{Precision} = TP / (FP + TP) \quad (1)$$

Recall – The proportion of the relevant pages that were correctly identified

$$\text{Recall} = TP / (FN + TP) \quad (2)$$

F-Measure – Derives from precision and recall values:

$$\text{F-Measure} = (2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (3)$$

The F-Measure was used, because despite Precision and Recall being valid metrics in their own right, one can be optimized at the expense of the other. The F- Measure only produces a high result when Precision and Recall are both balanced, thus this is very significant.

5. Result and discussion

In this experiment, the classification task will be examined in order to classify instances and predict their category in web phishing problem; instances will be classified to legitimate or phishing website. Two cases have been tested to evaluate the classification performance; with and without feature selection. Different three types of classifiers have been tested; classifiers were chosen due to their different nature of work, to compare their efficiency in terms of accuracy in Phishing Website Prediction problem, in both cases with and without feature selection. Experiments were done using test set as testing mode.

5.1. Experiment with Full Features

Full features included all instances and category in web phishing were evaluated to get their percentage of accuracy. **Table 3** shows result of three types of classifiers from this full features experiment and **Figure 2** shows the same result in figure to make it more clear and easy to understand.

Table 3. Accuracy result for experiment with full features

Classifiers	Precision	Recall	F-Measure
Artificial Neural Network	0.963	0.963	0.963
J48 Decision Tree	0.953	0.953	0.953
SVM	0.937	0.937	0.937

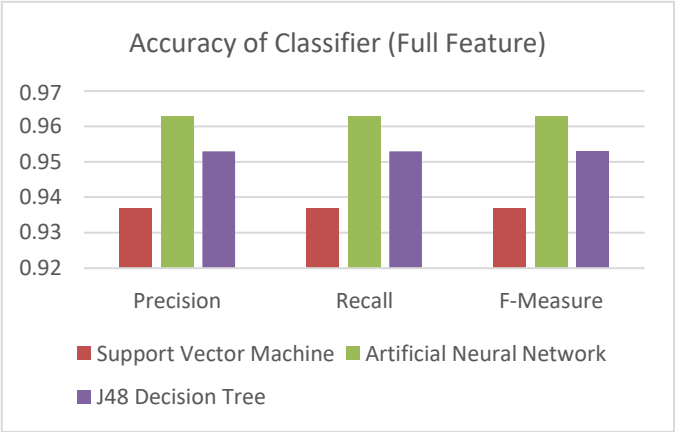

Figure 2. Graph of full features accuracy result

The above **Table 3** and **Figure 2** reveal that the classification process in the web page phishing data set, Artificial Neural Network classification holds highest F-Measure value (0.962), also followed by J48 classifier's F-Measure value which is 0.956 and Support Vector Machine is 0.934 with the lowest value.

As we observe, all classifiers accuracy without feature selection had better results than using feature selection as a preprocess step before classifying instances. The performance of the Artificial Neural Network classifier and J48 algorithms were almost similar on the data set, whereas the SVM classifier is slightly lower in accuracy.

5.2. Experiment with reduced features

Reduced features included some instances and some category in web phishing were evaluated to get their percentage of accuracy. **Table 4** shows result of three types of classifiers from this full features experiment and Figure 3 shows the same result in figure to make it more clear and easy to understand.

Table 4. Accuracy result for experiment with reduced features

Classifiers	Precision	Recall	F-Measure
Artificial Neural Network	0.962	0.962	0.962
J48 Decision Tree	0.956	0.956	0.956
SVM	0.934	0.934	0.934

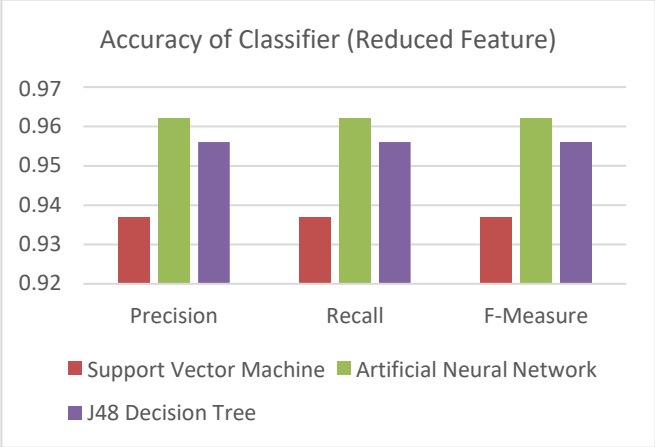

Figure 3. Graph of reduced features accuracy result

6. Conclusion

As a conclusion, we have met our objective which is to evaluate and investigate three selected classification algorithms based on Weka. Classification technique showed significant performance in Phishing Website Prediction. Three algorithms have been used and compared in terms of accuracy. The feature selection pre-processing was considered to observe the performance of classifiers with a minimal number of features. The obtained results showed that the classifiers are doing good without eliminating features in the tested dataset. But there is a tradeoff between the accuracy and the consumed time in the prediction process. Where if more accuracy was required, it's better to use the classification techniques without feature selection. And if the best performance was the target it's better to use the feature selection process in this dataset. From the experiment, it was found that the Artificial Neural Network Classifier had the highest accuracy and the Support Vector Machine had the least one. In the future, more classification techniques can be compared, with different measures, and more datasets can be used, in addition to the feature extraction from a number of phishing websites then many classification techniques can be applied for the prediction process.

Acknowledgement

A big appreciation to Universiti Malaysia Pahang for financial support under research grant RDU190310 to publish this research paper and a big thanks to all research team.

References

[1] "Internet Growth Statistics 1995 to 2019 - the Global Village Online," 2019. [Online]. Available: https://www.internetworldstats.com/emarketing.htm. [Accessed: 20- Mar-2019].
[2] "Number of internet users worldwide 2005-2018," 2019. [Online]. Available:

https://www.statista.com/statistics/273018/number-of- internet-users-worldwide/. [Accessed: 20-Mar-2019].
[3] P. Vinod and P. Prajapati, "Comparative Study of Web Page Classification Approaches," *Int. J. Comput. Appl.*, vol. 179, no. 45, pp. 6–9, 2018.
[4] S. S. Modi and S. B. Jagtap, "Web Content Filtration Using Different Web Mining Techniques in Educational System: An Overview," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 7, no. 3, pp. 135–139, 2017.
[5] T. Mahmoud, T. Abd-El-Hafeez, and D. El-Deen, "A Design of an Automatic Web Page Classification System," *Br. J. Appl. Sci. Technol.*, vol. 18, no. 6, pp. 1–14, 2017.
[6] B. Choi and Z. Yao, Web Page Classification. In: Chu W., Young Lin T. (eds) *Foundations and Advances in Data Mining*. Studies in Fuzziness and Soft Computing, vol 180. Springer, Berlin, Heidelberg.
[7] A. Siddiqui, M. Adnan, R. A. Siddiqui, and T. Mubeen, "A comparative study of web pages classification methods applied to health consumer web pages," *2015 2nd Int. Conf. Comput. Technol. Inf. Manag. ICCTIM 2015*, pp. 43–48, 2015.
[8] A. S. Patil and B. V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm," *Proc. Int. MultiConference Eng Comput. Sci.*, vol. I, 2012.
[9] R. K. Roul, "An Effective Approach for Web Document Classification using the Concept of Association Analysis of Data Mining," *Citeseer*, vol. 3, no. 10, pp. 483–491, 2012.
[10] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *IJCSI Int. J. Comput. Sci. Issues*, vol. 4, no. 1, pp. 16–23, 2009.
[11] Y. Yang, P. Liu, L. Ding, B. Shen, and W. Wang, "ServeNet: A Deep Neural Network for Web Service Classification," pp. 1–17, Jun. 2018.
[12] Z. Lu, "Web Page Classification Using Features from Titles and Snippets," no. May, 2015.
[13] R. R. and C. Aravindan, "An Effective and Discriminative Feature Learning for URL Based Web Page Classification," pp. 1374–1379, 2019.
[14] C. M. Chen, H. M. Lee, and Y. J. Chang, "Two novel feature selection approaches for web page classification," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 260–272, 2009.
[15] S. K. Dwivedi and C. Arya, "News web page classification using url content and structure attributes," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, no. October, pp. 317–322, 2017.
[16] E. Sarac and S. A. Ozel, "Web page classification using firefly optimization," in *2013 IEEE INISTA*, 2013, no. September, pp. 1–5.
[17] S. Moein, "Feature Extraction and Classification," *Med. Diagnosis Using Artif. Neural Networks*, pp. 159–169, 2014.
[18] K. He and C. Li, "Structure-Based Classification of Web Documents Using Support Vector Machine," 2016.
[19] P. Manchanda, S. Gupta, and K. K. Bhatia, "On The Automated Classification of Web Pages Using Artificial Neural Network," *IOSR J. Comput. Eng.*, vol. 4, no. 1, pp. 20–25, 2012.
[20] S. M. Kamruzzaman, "Web Page Categorization Using Artificial Neural Networks," *Training*, p. 4, 2010.
[21] X. Weimin, B. Hong, H. Weitong, and L. Yuchang, "Web page classification based on SVM," *Proc. World Congr. Intell. Control Autom.*, vol. 2, pp. 6111–6114, 2006.
[22] D. R. Gupta, "Comparison of classification algorithms to detect phishing web pages using feature selection and extraction," *Int. J. Res.*, vol. 4, no. 8, pp. 118–135, 2016.
[23] M. N. Amin and M. A. Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data," no. 3, 2015, pp. 55–61.
[24] S. Nandhini and V. Vasanthi, "Extraction of Features and Classification on Phishing Websites using Web Mining Techniques," vol. 5, no. 4, pp. 1215–1225, 2017.
[25] C. C.Aggarwal, Data Classification: Algorithms and Applications. CRC Press, 2015.

[26] V. A. Zilpe and M. Atique, "Web Usage Mining Using Neural Network Approach: A Critical Review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, no. 1, pp. 3073–3077, 2012.
[27] H. Bouali and J. Akaichi, "Comparative study of different classification techniques: Heart disease use case," *Proc. - 2014 13th Int. Conf. Mach. Learn. Appl. ICMLA 2014*, pp. 482–486, 2014.
[28] C. Nasa and S. Suman, "Evaluation of Different Classification Techniques for WEB Data," *Int. J. Comput. Appl.*, vol. 52, no. 9, pp. 34–40, 2012.
[29] W. . Awad, "Machine Learning Algorithms in Web Page Classification," *Int. J. Comput. Sci. Inf. Technol.*, vol. 4, no. 5, pp. 93–101, Oct. 2012.
[30] K. Wisaeng, "A Comparison of Different Classification Techniques for Bank Direct Marketing," *Int. J. Soft Comput. Eng.*, vol. 3, no. 4, pp. 116–119, 2013.
[31] W. Huang and H. You, "Web Page Classification Algorithm Based on Semi-Supervised Support Vector Machine," *Proc. 2018 2nd IEEE Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. IMCEC 2018*, no. Imcec, pp. 2144–2148, 2018.
[32] J. Wang, H. Cai, B. Xu, and L. Jiang, "CUCS: A web page classification algorithm for large training set," *Proc. - 2008 IFIP Int. Conf. Netw. Parallel Comput. NPC 2008*, pp. 440–445, 2008.
[33] T. Z. Phyu and N. N. Oo, "Performance Comparison of Feature Selection Methods," *MATEC Web Conf.*, vol. 42, p. 06002, Feb. 2016.
[34] S. Khalid, T. Khalil, and S. Nasreen, "A Survey Of Feature Selection And Feature Extraction Techniques in Machine Learning," pp. 372–378, 201.