ANALYSIS OF MACHINE LEARNING TECHNIQUES ON URL PHISHING DATASET

M.A.H. Ahmad[1], R. Yusof[1], N. A. Zakaria[1] and N.S. Ismail[2]

[1]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

[2]College of Computing, Informatics and Media,
Universiti Teknologi Mara Kuala Terengganu, 21080 Kuala Terengganu, Terengganu, Malaysia.

Corresponding Author's Email: [1]robiah@utem.edu.my

Article History: Received 6 March 2023; Revised 21 April 2023; Accepted 25 May 2023

ABSTRACT: The Internet has become a vital part of daily life, as do almost all online social and financial practices. However, rising phishing sites today faced significant threats because of their appallingly imperceptible danger. Phishing is an online fraudulent act that uses social engineering and technical subterfuge to trick internet users and capture their sensitive data or critical information online. There is a lack of knowledge on implementing a suitable classification technique on machine-learning tools for analyzing phishing URLs or Websites. This research aims to identify the best classification technique using the orange tool on these three datasets and implement the phishing URL analysis methodology comprising six phases. Based on the result, Decision Tree is the best classification technique for identifying the URL phishing attack. It has obtained the highest accuracy result of 88.30% and 70.70% in Dataset 2 and Dataset 3, respectively. In the future, more classification techniques or machine learning tools with different performances are explored to analyze Phishing URLs or Websites for better results.

KEYWORDS: *Phishing, Machine-learning, Orange Tool, URL, classification technique*

1.0 INTRODUCTION

Phishing is a web-related scheme focused on cyber attackers' illegal activity that can persuade users to reveal their login passwords, pin numbers, credit card information, and personal data. The attackers may gather the financial credentials information to gain access and participate in fraudulent activities. The malicious Uniform Resource Locator (URL) plays a significant role in phishing [1]. Cybercriminals send malicious URLs to victims inside a post using different channels such as private text messages, emails, websites, and banners, as well as on forums. The URLs appeared to be valid source. Users can use URL metadata to identify the presence of phishing in the URL, as described by [2]. Thus, classification techniques can accurately predict the phishing URL website [3]. Moreover, [4] proposed using classification techniques to implement a precise, intelligent phishing website detection framework. However, [5] stated that the problem arises in selecting the best classifier for phishing websites due to a lack of knowledge on implementing classification techniques for analyzing phishing URLs or Websites. Hence, this research is to identify the best classification technique for identifying phishing URLs or Websites. The rest of the paper is structured as follows. Section 2 discusses the related work on phishing, the taxonomy of phishing, machine learning tool, classification techniques, datasets, and evaluation metrics. Section 3 presents the methodology used in this research. Section 4 discusses the result and discussion of the study, and finally, Section 5 concludes and summarizes the future directions of this work.

2.0 RELATED WORK

2.1 What is Phishing?

Phishing is a technique of social engineering in which a malicious attacker impersonates a trusted third party to trick the user into revealing sensitive data [6]. It is also known as the fraudulent effort to acquire confidential information such as usernames, passwords, and credit card data by disguising yourself in an online message as a trusted person [1].

Phishing is a significant threat to all internet users, and it is hard to track or protect against because it does not seem malicious [7, 19, 20].

Thus the protection of personal credentials is at risk. Phishing can be seen as one of the oldest and simplest ways to steal people's information and is used to obtain a wide variety of personal data. It also has a reasonably simple approach by emailing a victim, which can lure him to a site that steals his information.

2.2 Taxonomy of Phishing

There are eight categories of Phishing: *spear phishing, search engine phishing, pharming, email phishing, smishing, vishing, whaling, and watering hole phishing*, as depicted in Figure 1.

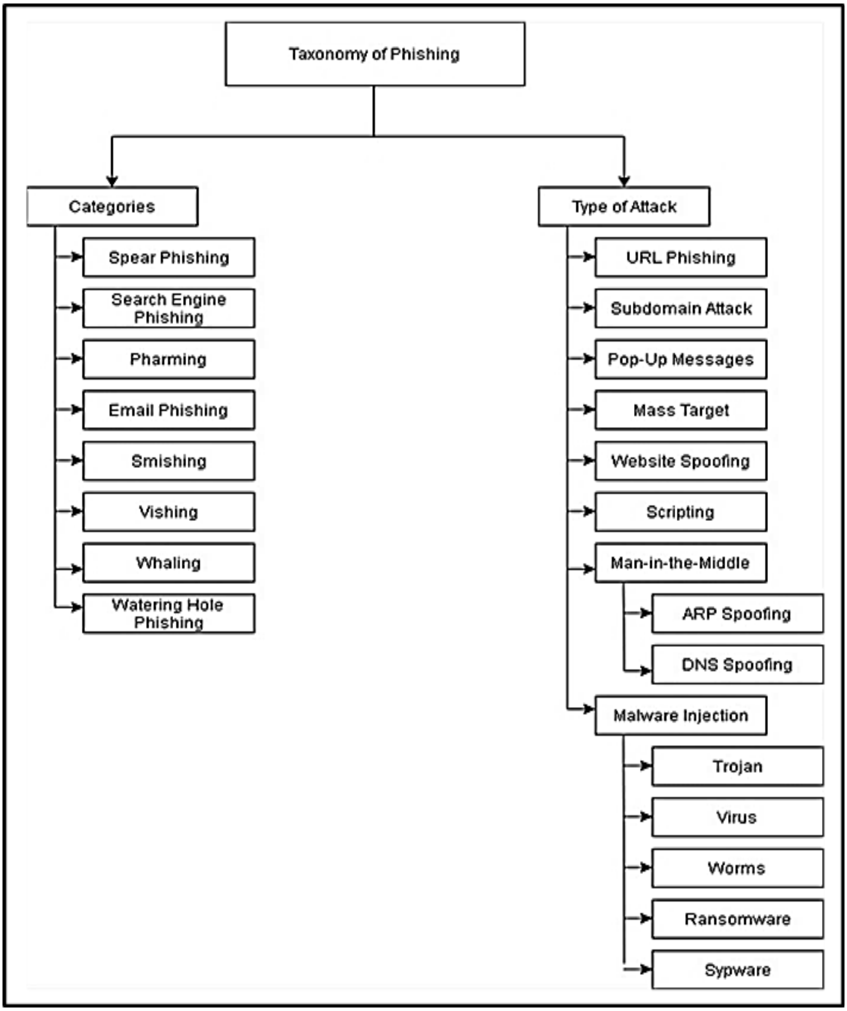


Figure 1: Category of Phishing

The elaboration of each category is shown in Table 1.

Table 1: Description of Phishing Category

Category	Description
Spear Phishing	It is a preliminary stage of an advanced persistent threat (APT) attack to create a point of entry into an organization [9]. It is an email or electronic communication fraud targeted at a specific person, organization, or company. Cybercriminals often intend to steal data for malicious purposes, and cyber criminals may often wish to install malware on a targeted user's computer.
Search Engine Phishing	It is a reasonably new form of phishing attack which does not bother fraudsters to send a targeted email. Search Engine Phishing builds its platform by selling inexpensive goods or spectacular offers and getting them indexed by legitimate search engines. Online shoppers can notice these pages on a standard Google result page, and we cannot identify the differences. A platform aimed at phishing the search engine allows users to pass on their personal information. They may ask to register for a National Insurance number, or it may depend on purchasing a bank account number. They may use the data to rob, hijack identity, or destroy reputation.
Pharming	Pharming leverages malicious code such as viruses, worms, trojans, and spyware to execute sophisticated attacks, including alteration of the host server, DNS cache poisoning, and so on, which will not be known to the user [9]. For example, an attempt by a hacker to change or exploit a server's DNS settings to redirect to a fake or copy of the original site hosted somewhere else when entering the address of a legitimate website.
Email Phishing	It is a game of numbers. An attacker sending thousands of fake messages, even if only a tiny percentage of the recipients fall for the scam, may net important information and money. More e-mails are received daily to make Web users assume that the same e-mail is genuine and comes from trusted institutions [10].
Smishing	Smishing is a malware attack achieved by sending a fake message intended to steal smartphone users' credentials [11]. Smishing attack is becoming popular nowadays due to mobile users' massive growth. Since its aim is for financial benefits, the smishing message is very harmful.
Vishing	Vishing uses IP-based voice messaging technologies (mainly Voice over Internet Protocol or VoIP) to socially engineer the intended target to supply personal, financial, or other confidential information for financial reward purposes. Since the advent of the telephone, landline telephony systems have been used to convince others to conduct unintentional acts [12].
Whaling	Whaling is not so distinct from spear phishing, but the target category is more precise and limited to this phishing attack. Whaling attacks are also more aggressive, threatening senior executives. While the ultimate aim of whaling is the same as any other form of the phishing attack, the strategy appears to be even more subtle. Tricks such as false links and malicious URLs are not helpful because perpetrators attempt to mimic senior staff.
Watering Hole Phishing	A watering hole attack is a security vulnerability in which the attacker tries to compromise a particular group of end-users by infecting websites known to be accessed by community members. The purpose is to infect the machine of a targeted user and gain access to the network at the victim's place of employment. Watering hole attacks, which focus on legal, popular websites, are a derivative of pivotal attacks.

This research shall focus on spear phishing, pharming, and watering hole phishing as they are closely related to phishing URLs or websites. There are also eight attack types, as shown in Fig. 1, and this research shall only focus on URL Phishing. This attack is a malicious Uniform Resource Locator (URL), which plays a significant role in phishing. Cybercriminals send malicious URLs to victims inside a post using different channels such as private text messages, emails, websites, and banners, as well as on forums [1]. The URLs appear as valid sources. As the internet proliferates, users shift their preference from traditional shopping to electronic commerce. By implementing the internet's anonymous structure, attackers set out new techniques, such as phishing, to deceive victims using false websites to gather sensitive information, such as account IDs, usernames, and passwords [8]

2.3 Machine Learning Tool

There are three machine learning types: supervised, unsupervised, and reinforcement [9]. This research used supervised learning and implemented Orange as the machine learning tool. Orange is an open-source data visualization, machine learning, and data mining toolkit. It features a visual programming front-end for explorative data analysis and interactive data visualization. This machine-learning tool can implement several classification techniques, which will be further discussed in the following sub-section.

2.4 Classification Techniques

This research shall implement a support vector machine, random forest, decision tree, and naive Bayes.

i. Support Vector Machine

Support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. It is a typical family of linear classifiers using a space-linear hypothesis [13]. It is trained by learning algorithms to refine the principle of practice learning bias derived from statistical learning theory.

ii. Random Forest

Random Forest (RF) is one of the most popular algorithms [14]. It is a collection of decision trees that produces better prediction accuracy and a research tool used to build multiple decision trees. The final decision will be based on most trees and random forests. An RF reduces the variance of a single decision tree leading to better predictions of new data. In other words, the RF is a mixture of tree predictors, such that each tree depends on the values of the random variable sampled independently and with the same distribution for all trees in the forest.

iii. Decision Tree

Decision trees are considered one of the most common methods for classifier representation [15]. Researchers from diverse disciplines, such as statistics, machine learning, pattern recognition, and data mining, discussed the issue of developing a decision tree from available data. A decision tree is a classifier that expresses itself as a recursive instance space partition.

The decision tree consists of nodes forming a rooted tree, meaning that it is a directed tree with a node called "root," which has no incoming edges. Decision tree inducers are algorithms that create a decision tree automatically from a given dataset. The goal is usually to find the optimal decision tree by minimizing the generalization error. Other target functions, however, may also be specified, such as reducing the number of nodes or minimizing the mean depth.

iv. Naïve Bayes

The naive Bayes algorithm is a primary probabilistic classifier for calculating a series of probabilities to measure the frequency and combination of values in a given data set [16]. Bayes theorem uses the algorithm which claims that all the attributes depend on the value of the class variables.

The concept of conditional independence is uncommon in real-world implementations. Hence, the definition is Naive, but the algorithm performs well and learns quickly in various supervised classification problems. A naive Bayesian classifier is based on the Bayes theorem and the theorem of absolute probability. The Naive Bayes approach assesses the likelihood of each function independently, regardless of any correlation, and estimates the probability based on the Bayes theorem.

2.5 Dataset

This research uses three different datasets: Dataset 1 and Dataset 2, obtained from [17], and Dataset 3, obtained from Mendeley Phishing

Dataset for Machine Learning. The attributes and URL of the website in each dataset are summarized in Table 2.

Table 2: Information on Phishing Dataset

Dataset	Attributes and websites
Dataset 1 (D1)	30 attributes and 1 target attributes. It consists of 2456 entries of phishing as well as non-phishing Websites.
	URL: http://archive.ics.uci.edu/ml/datasets/Phishing+Websites
Dataset 2 (D2)	1353 Websites with 10 attributes. These Websites are classified into 3 categories: Phishing, non-phishing and suspicious
	URL: http://archive.ics.uci.edu/ml/datasets/Website+Phishing [
Dataset 3 (D3)	It has total of 10000 different URLs where 5000 URLs are phishing and rest 5000 URLs are the information of legitimate webpages.
	URL: https://data.mendeley.com/datasets/h3cgnj8hft/l

2.6 Evaluation Metrics

The performance of the classification algorithm is assessed using four metrics, including *Accuracy*, *Precision*, *Recall*, and *F-Measure*. The measures of all these metrics depend on a variety of factors, including *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), and *False Negative* (FN).

A *confusion matrix* is used to display and summarize the results of a classification algorithm. A sample of the *confusion matrix* is shown in Table 3. A *Correctly Classified Instance* is a combination of 00 and 11, while an *Incorrectly Classified Instance* is a combination of 01 and 10.

Table 3: Sample of Confusion Matrix

	Predicted	
Actual	0	1
0	2689 (TP)	25 (FN)
1	85 (FP)	201 (TN)

In Table 3, the TP in the actual and predicted class will combine with TN in the actual and predicted class to generate a *Correctly Classified Instance*. In this case, TP of 2689 will integrate with TN of 201 to generate 2890 of *Correctly Classified Instance*. While FN of 25 will combine with FP of 85 to generate 110 *Incorrectly Classified Instances*. Thus, we can determine *Accuracy*, *Precision*, *Recall*, and *F-Measure* using this *confusion matrix* as shown in equation (1) – (4).

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F\text{−}measure = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

The higher Precision, Recall, F-Measure, and Accuracy value indicated the classifier's precision in classifying the instances. All the measurement parameters value will be used to determine the best algorithm obtained in the machine learning tool.

3.0 METHODOLOGY

The analysis methodology for phishing URLs consists of six phases as shown in Figure 2. The phases are finding tools and dataset, selecting tool and dataset, installing the tool, information collection, information analysis, and documenting the result [18].

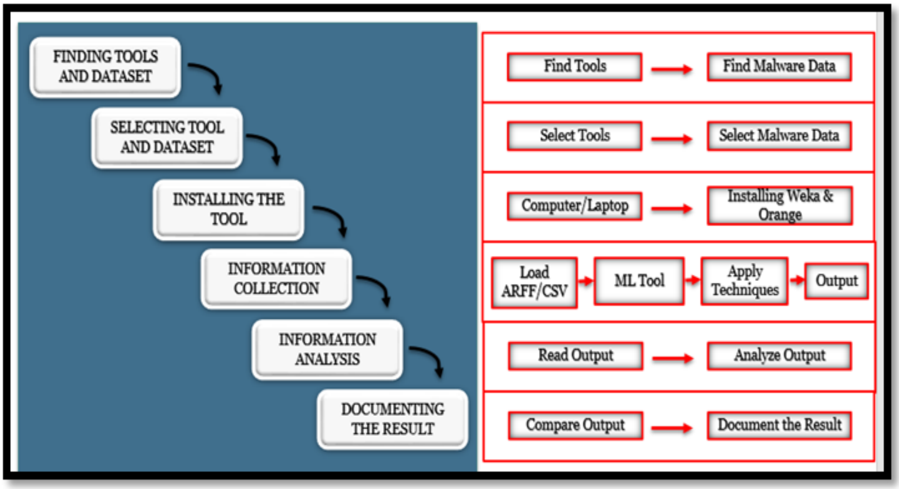


Figure 2: Phishing URL analysis methodologies

- **Phase 1: Finding Tools and Dataset**

During this initial phase, appropriate machine learning tools and malware datasets need to be identified.

- **Phase 2: Selecting Tool and Dataset**

In this phase, the machine learning tool selected is Orange. The selected malware data is a phishing URL obtained from UCI Machine Learning Repository and Mendeley Dataset for Machine Learning.

- **Phase 3: Installing the Tool**

In this phase, the hardware used is laptops or computers, and Orange will be installed in windows 10.

- **Phase 4: Information Collection**

Different classification techniques or algorithms and measurement attributes are applied in Orange, and the results are collected.

- **Phase 5: Information Analysis**

The information obtained will be analyzed to identify the best classification techniques or algorithms for identifying Phishing URLs.

- **Phase 6: Documenting the Result**

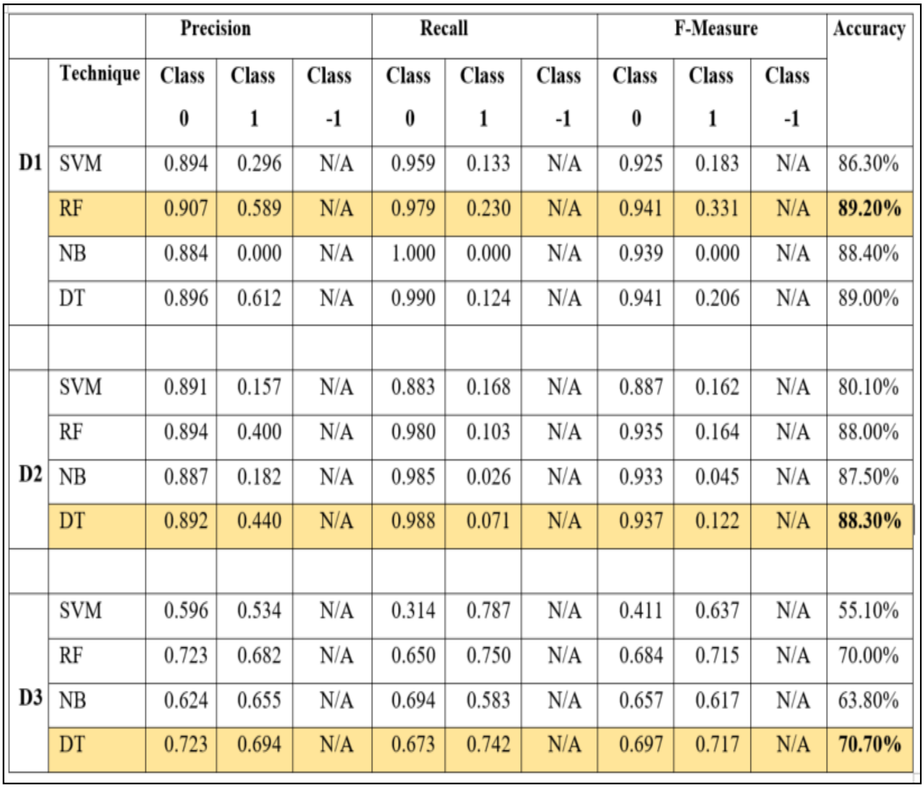
The output obtained in the previous phase will be documented after comparing it to identify the best classification technique or algorithm.

4.0 RESULT AND DISCUSSION

The result depicted in Table 4 shows the Precision, Recall, F-Measure, and Accuracy obtained using different datasets, namely Dataset 1 (D1), Dataset 2 (D2), and Dataset 3 (D3), and different classification techniques, namely SVM, RF, NB and DT. Each Precision, Recall, and F-Measure has class 0 for non-phishing, class 1 for phishing, and class -1 for suspicious.

Referring to Table 4, in D1, RF has the highest accuracy, 89.20%, compared to other techniques. However, in D2 and D3, the highest accuracy is using DT, which is 88.30% and 70.70%, respectively. The value of Precision, Recall, and F-Measure obtained in Table 4 is used to calculate the weighted average.

Table 4: Result of Classification Techniques Implemented using Orange Tool



		Precision			Recall			F-Measure			Accuracy
	Technique	Class 0	Class 1	Class -1	Class 0	Class 1	Class -1	Class 0	Class 1	Class -1	
D1	SVM	0.894	0.296	N/A	0.959	0.133	N/A	0.925	0.183	N/A	86.30%
	RF	0.907	0.589	N/A	0.979	0.230	N/A	0.941	0.331	N/A	**89.20%**
	NB	0.884	0.000	N/A	1.000	0.000	N/A	0.939	0.000	N/A	88.40%
	DT	0.896	0.612	N/A	0.990	0.124	N/A	0.941	0.206	N/A	89.00%
D2	SVM	0.891	0.157	N/A	0.883	0.168	N/A	0.887	0.162	N/A	80.10%
	RF	0.894	0.400	N/A	0.980	0.103	N/A	0.935	0.164	N/A	88.00%
	NB	0.887	0.182	N/A	0.985	0.026	N/A	0.933	0.045	N/A	87.50%
	DT	0.892	0.440	N/A	0.988	0.071	N/A	0.937	0.122	N/A	**88.30%**
D3	SVM	0.596	0.534	N/A	0.314	0.787	N/A	0.411	0.637	N/A	55.10%
	RF	0.723	0.682	N/A	0.650	0.750	N/A	0.684	0.715	N/A	70.00%
	NB	0.624	0.655	N/A	0.694	0.583	N/A	0.657	0.617	N/A	63.80%
	DT	0.723	0.694	N/A	0.673	0.742	N/A	0.697	0.717	N/A	**70.70%**

The weighted average shown in Table 5 is calculated by adding the value of class 0 and class 1 of each parameter found in Table 4. In Table 4, the value of Precision and Recall for RF in D1 is high and above 0.5, which is suitable for this technique.

Similar high values above 0.5 are found in Precision and Recall for the DT technique in D2 and D3. For the value of F-Measure in D1 and D2, the RF technique has the highest value, 0.636 and 0.550, respectively, followed by the DT technique in D3, which is 0.707. The value of the F-measure that falls between 0.0 and 1.0 indicates that the technique is suitable for implementation on the datasets.

Table 5: Result of Weighted Average for Precision, Recall and F-Measure

Dataset	Algorithms	Precision	Recall	F-Measure	Accuracy
D1	SVM	0.595	0.546	0.554	86.30%
	RF	**0.748**	**0.605**	**0.636**	**89.20%**
	NB	0.442	0.500	0.470	88.40%
	DT	0.754	0.557	0.574	89.00%
D2	SVM	0.524	0.526	0.525	80.10%
	RF	0.647	0.542	**0.550**	87.80%
	NB	0.535	0.506	0.489	87.50%
	DT	**0.666**	**0.530**	**0.530**	**88.30%**
D3	SVM	0.565	0.551	0.524	55.10%
	RF	0.703	0.700	0.700	70.00%
	NB	0.640	0.639	0.637	63.80%
	DT	**0.709**	**0.708**	**0.707**	**70.70%**

A histogram graph of the accuracy result in Figure 3 uses the data obtained in Table 4.

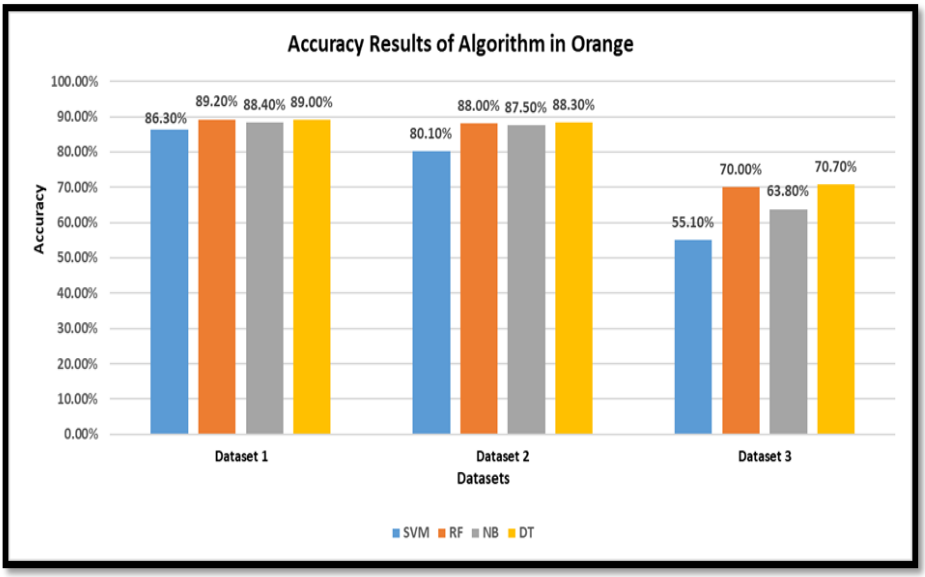


Figure 3: Graph of Accuracy Results of Various Classification Techniques using Orange Tool

In Figure 3, the accuracy result shows that the RF technique in Dataset 1 has gained the highest accuracy of 89.20% compared to other classification techniques. Meanwhile, in Dataset 2 and Dataset 3, the DT technique obtained the highest accuracy of 88.30% and 70.70%, respectively. From the accuracy result obtained in Table 4, Table 5 and Figure 3, the DT technique is the best classification technique to implement in identifying the URL phishing attack as it has obtained the highest accuracy result in D2 and D3.

5.0 CONCLUSION

In conclusion, based on the result, the best classification technique to identify URL phishing attacks using the Orange tool is Decision Tree (DT) classification technique. Thus, this research allows the researcher to apply classification techniques to selected machine learning tools when analyzing URL phishing datasets. In the future, more machine learning tools and algorithms will be implemented on this dataset to identify the best classification technique for machine learning tools.

ACKNOWLEDGMENTS

The authors would like to thank the Research Group of Information Security, Forensics and Computer Networking (InsForNet), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

[1] Rahman, S. S. M. M., Islam, T., and Jabiullah, M. I., "PhishStack: Evaluation of stacked generalization in phishing URLs detection", *Procedia Computer Science*, vol.167, pp. 2410-2418, 2020.

[2] Jagadeesan, S., Chaturvedi, A., and Kumar, S., "URL phishing analysis using random forest", *International Journal of Pure and Applied Mathematics*, vol. 118, no. 20, pp. 4159-4163, 2018.

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

[4] Subasi, A., and Kremic, E., "Comparison of adaboost with multiboosting for phishing website detection", *Procedia Computer Science*, vol. 168, pp. 272-278, 2020.

[5] Amiri, I. S., Akanbi, O. A., and Fazeldehkordi, E. "A machine-learning approach to phishing detection and defense", *Syngress*, 2014

[6] Lawson, P., Pearson, C. J., Crowson, A., and Mayhorn, C. B. "Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy", *Applied ergonomics*, vol. 86, 103084, 2020.

[7] Vayansky, I., and Kumar, S., "Phishing–challenges and solutions. Computer Fraud & Security", vol. 2018, no. 1, pp. 15-20, 2018

[8] Sahingoz, O. K., Buber, E., Demir, O., and Diri, B., "Machine learning based phishing detection from URLs", *Expert Systems with Applications*, vol. 117, pp. 345-357, 2019.

[9] Dubey, R., and Banu, P., "Analysis of supervised and unsupervised technique for authentication dataset", *International Journal of Engineering & Technology*, vol.7, no. 4, pp. 2867-2873, 2018.

[10] Purkait, S., "Phishing counter measures and their effectiveness–literature review", *Information Management & Computer Security*, vol. 20, no. 5, pp. 382-420, 2012.

[11] Jain, A. K., and Gupta, B. B., "Feature based approach for detection of smishing messages in the mobile environment", *Journal of Information Technology Research (JITR)*, vol. 12, no. 2, pp.17-35, 2019.

[12] Ollmann, G. "The vishing guide. IBM Global Technology Services", pp. 1-16, 2007.

[13] Choubey, S. M. V., Pandey, S., and Shukla, J. "An Efficient Approach of Support Vector Machine for Runoff Forecasting", *International Journal of Computer Applications*, vol. 92, 2014.

[14] Chumachenko, K. "Machine learning methods for malware detection and classification", 2017

[15] Aksu, D., Turgut, Z., Üstebay, S., and Aydin, M. A., "Phishing analysis of websites using classification techniques", In International Telecommunications Conference: Proceedings of the ITelCon 2017, Istanbul. Springer Singapore. pp. 251-258, 2019.

[16] Patil, T. R. Mrs. SS Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data", *International Journal of Computer Science And Applications*, vol. 6, no. 2, pp.1-6, 2013.

[17] Dua, D. and Graff, C. *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science., 2019.

[18] Yusof, R., Adnan, N. S., Jalil, N. A., and Abdullah, R. S. "Analysis of data mining tools for android malware detection", *Journal of Advanced Computing Technology and Application (JACTA)*, 1(2), 21-24, 2019.

[19] Alkhalil Z, Hewage C, Nawaf L and Khan I, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy", *Front. Comput. Sci.* vol. 3, no. 563060, 2021.

[20] Vaishnavi Bhavsar, Aditya Kadlak and Shabnam Sharma, "Study on Phishing Attacks", *International Journal of Computer Applications*, vol. 182, no. 33, pp. 27-29, 2018.