A Comparative Study of Machine Learning Techniques for Phishing Website Detection

Mohammad Farhan Khan, Rohit Kumar Tiwari, Sushil Kumar Saroj, and Tripti Tripathi

1 Introduction

The Internet has become one of the integral parts of our life in recent years due to the availability of various services in online mode like online banking, social media, entertainment, etc. These online services have caused an exponential increase in Internet users, which in turn has given an opportunity to cyber criminals for cyber fraud causing huge financial loss to users every year. Cyber criminals use various techniques to harm the users' system or steal their sensitive information like username, password, email ID, and other credentials by deceiving the users as a trustworthy entity [1]. Phishing is one of the techniques of cybercrime where the attacker presents himself as a trustworthy entity to the users to collect sensitive information through email or websites. In a phishing website, the attacker creates a fake website by cloning the legitimate website and sends an email to target users to update their information. Once a user goes through the email and clicks on the URL of the website, it redirects the target users to a fake website where the users enter their credential to update the information. The attacker gets the credentials of the users and uses them for financial or any other type of fraud.

Phishing website attack has become one of the main challenges in cyberspace. It is causing huge financial loss to users every year with the increase of Internet users. According to the RSA quarterly fraud report for the period of 1st of January to 31st of March 2018, phishing is responsible for 48 percent of all cyberattacks [2]. The report says that Canada, United States, India, and Brazil are the most victim countries of the phishing attack. Figure 1 shows the statistics of phishing attacks of various countries in the above period.

M. F. Khan · R. K. Tiwari (✉) · S. K. Saroj · T. Tripathi
Department of Computer Science & Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, Uttar Pradesh, India

S. Pandey et al. (eds.), *Role of Data-Intensive Distributed Computing Systems in Designing Data Solutions*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-031-15542-0_6

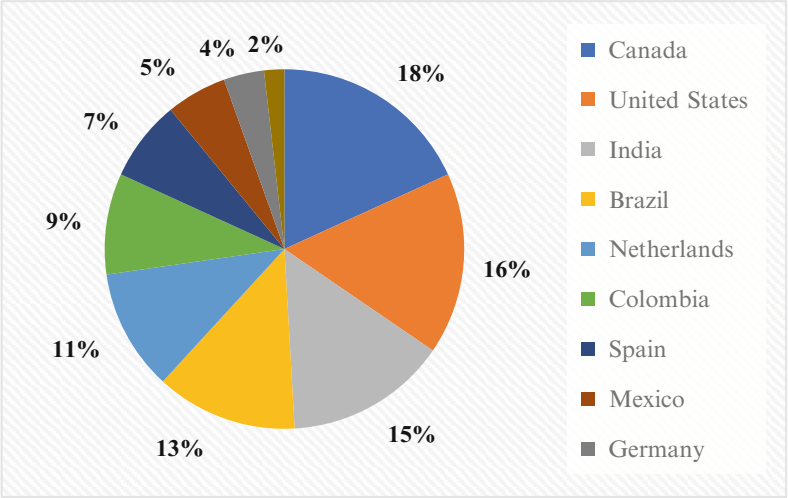

Fig. 1 Percentage of phishing attack in different countries during January to March 2018

Due to the increasing number of cyber frauds through phishing websites, it has become necessary to develop some techniques to prevent it. Many techniques have been proposed to detect phishing websites with the help of black and whitelisted databases like PhishTank. In this approach, whenever a user visits a website, its URL is checked in the database. If the URL is present in the blacklist label, then the website is phishing. However, these techniques are not sufficient to detect phishing websites as new phishing websites are created every minute and their addition to the database takes time. Therefore, there is a need for an intelligent phishing website detection system that can detect phishing websites automatically.

In this paper, we have developed intelligent phishing detection systems based on machine learning techniques that use structural features of the website to detect whether a website is phishing or not. The features used for training the detector are based on the structure of the webpage and URL. We have used various machine learning algorithms to train a detector using a standard dataset consisting of phishy and non-phishy websites. We have also compared them in terms of various metrics to show the usefulness of the machine learning algorithm for phishing website detection.

The remaining part of this paper is organized as follows: Section 2 discusses the categorization of phishing website detection techniques with a detailed review of them. Section 3 discusses the various structural features, which are being used to train the phishing website detector. Section 4 explains the working of the proposed method of machine learning-based phishing detector followed by result and discussion in Sect. 5. At last, the conclusion is presented in Sect. 6.

2 Literature Review

Phishing website detection is one of the major challenges in cybersecurity. There are various methods that exist in the literature to detect phishing websites. Figure 2 presents the taxonomy of phishing website detection systems. In user awareness-based techniques, a user recognizes a phishing website from its experience or knowledge whereas software detection-based techniques use automated techniques to detect phishing websites. Vision and web page structure-based techniques use website design and structure to detect phishing. The vision-based technique detects a phishing website based on the visual comparison of a legitimate and illegitimate websites. They use interest point detector techniques of computer vision to locate a phishing website. The web page structure-based technique detects a phishing website using a structure of web page like referencing, HTML structure, URL, etc. Various authors have proposed phishing website detectors based on the above categorizations. Some of them have been discussed below.

Rao et al. [3] have proposed an approach to detect phishing websites based on machine learning techniques. They trained eight machine learning algorithms to detect phishy websites out of which random forest-based technique was more accurate. Sönmez et al. [4] also proposed a machine learning-based phishing website detector. They initially extracted features of the website and used them to train the classifier. They used support vector machine, naive Bayes, and extreme learning machine (ELM) as machine learning techniques out of which ELM has the highest accuracy. Sharmin et al. [5] have proposed a supervised learning technique in which they discussed the problem of spam detection in social media platforms. They worked on the comment section of YouTube to filter out spam comments. They tried to solve the phishing problem by applying various methods. Ensemble classifier has the best response among them.

Altaher et al. [6] have proposed a hybrid algorithm by combining KNN and SVM methods. They first applied the KNN method to remove noisy data followed by

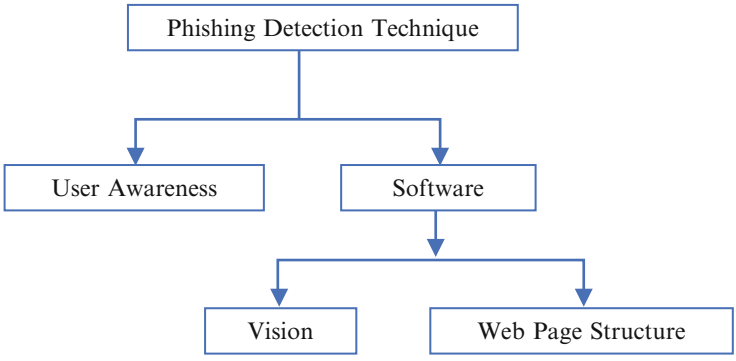

Fig. 2 Taxonomy of phishing detection techniques

the SVM method that is used to classify the phishy website. This hybrid approach has 90.04% accuracy. Karnik et al. [7] proposed the phishy website detector using the SVM algorithm to solve the phishing attack problem. They used features like webpage contents, DNS information, link structure, textual properties, and network traffic of webpage to train the detector. The proposed approach has 95% accuracy. Abunadi et al. [8] presented a review of features used in machine learning-based techniques to detect phishy websites. They also added some new features, which are helpful in phishing websites detection and experimentally show that new features are more helpful in phishy website detection.

James et al. [9] proposed machine learning based to prevent phishing attacks. They used three more features like host properties, page importance properties, and lexical features to train the detector. Xiang et al. [10] proposed an extension of CANTINA called CANTINA+ to detect the phishy website. In CANTINA +, they added new features with previous features to get better accuracy. The proposed system filters website without entering the login form in the first step to reduce the false positive rate. The proposed approach utilizes 15 attributes like URL, HTML document object model, search engines, other services to train SVM to identify phishing websites. The true positive rate of CANTINA+ is 92%, and the false-positive rate is 0.4%. Aburrous et al. [11] presented a method using data mining techniques to search and identify the Internet banking system to prevent phishing attacks in the banking system. They used 27 phishing characteristics divide into six categories like source code and JavaScript, protection and encryption, URL and domain identity, content and page style, the web address bar, and social human factors to train the detector. Wenyin et al. [12] proposed a method to detect phishy websites in two stages. The first stage detects the keywords and suspicious URLs on the local email server. After detection of the URLs or suspicious keywords in the email, the second module compares the layout, block-level, and style equality for the suspicious webpage to detect a phishy website.

3 Phishing Websites Features

The phishing website follows some patterns related to web page structure and URL. There are many features to identify phishing websites. These patterns and features are used to categorize the websites as legitimate or illegitimate websites. Figure 3 shows the classification of phishing websites features.

3.1 Address Bar Features

Address bar features correspond to features obtained from the URL or address of a website. There are various patterns in the address of a website that indicate the website is phishy or not. Figure 4 shows an example of the address of a website

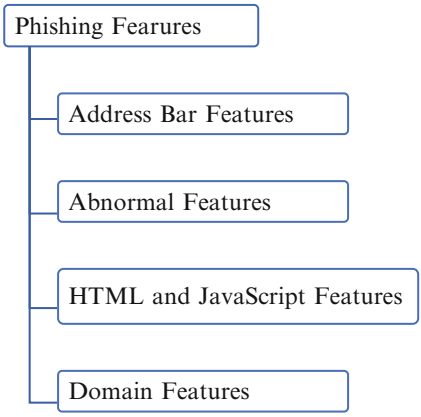

Fig. 3 Phishing websites features

Fig. 4 URL and its components

with its various parts. A typical URL consists of a domain, a protocol, and a file path. Some important address bar features used for detecting phishing websites are discussed below.

- **IP address:** An IP address uniquely identifies a computer on the Internet. The websites are hosted on the server, and they are accessed through URL. Sometimes, the URL consists of an IP address that normally does not exist. So, if the website consists of a URL, then users perceive that it is phishy. The rule used to identify a phishy website is given in Eq. 1:

$$\text{Rule} : \text{If} \begin{cases} \text{IP address exits in URl} \rightarrow \text{Phishy} \\ \text{Otherwise} \rightarrow \text{Feature} = \text{Legitimate} \end{cases} \tag{1}$$

- **URL Length:** URL length also indicates whether a website is phishing or not. The rule for classifying a website as phishing, legitimate, or suspicious is given in Eq. 2:

$$\text{Rule} : \text{If} \begin{cases} \text{URL length} < 54 \rightarrow \text{Legitimate} \\ \text{URl length} \geq 54 \text{ and} \leq 75 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishy} \end{cases} \tag{2}$$

- **URL with @ symbol**: If a URL consists of an @ symbol, it is categorized as a phishing URL otherwise legitimate. The rule is shown in Eq. 3:

$$\text{Rule} : \text{If} \begin{cases} \text{URl has@Symbol} \rightarrow \text{Phishy} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{3}$$

- **Prefix or suffix:** To make users perceive that they are working with a legitimate website, phishers use suffixes or prefixes separated by "−"in the area name. Therefore, the users think that they are working with a valid webpage with a domain name. The rule used to categorize a website based on suffix and prefix is given in Eq. 4:

$$\text{Rule} : \text{If} \begin{cases} \text{Domain has}\,' - '\,\text{symbol} \rightarrow \text{Phishy} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{4}$$

- **Subdomain and multisubdomains:** A website is classified as legitimate, suspicious, or phishy based on the number of subdomains in its URL. The rule used to classify it is shown in Eq. 5:

$$\text{Rule} : \text{If} \begin{cases} \text{Dots in the domain part} < 3 \rightarrow \text{Legitimate} \\ \text{Else if dots in domain part} = 3 \rightarrow \text{Supicious} \\ \text{Otherwise} \rightarrow \text{Phishy} \end{cases} \tag{5}$$

3.2 Abnormal Features

Abnormal features are the information or features obtained from a website. We have examined different abnormal features. Some important abnormal features used for identifying phishing websites are presented below.

- **Request URL:** Websites use different application programming interfaces (API) to access some resources. The API is identified by URL and request data. The website accessing API with the same domain name is legitimate. The ratio of request URL with same domain name to another domain name is used to identify if a website is phishy or not. If the web page has a ratio of request URL less than 22%, then it is legitimate. If request URLs are between 22% and 61%, then it is suspicious; otherwise, the website is a phishing website. The rule to detect phishing websites is given in Eq. 6:

$$\text{Rule} : \text{If} \begin{cases} \text{Request URL\%} < 22\% \rightarrow \text{Legitimate} \\ \text{Request URL\%} \geq 22\% \text{and} < 61\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishy} \end{cases} \tag{6}$$

- **Anchor tag URL:** A website is classified as phishy based on the URL of the anchor tag. An anchor is defined by HTML element <a > tag. A website can be classified as phishy based on the percentage of URLs pointing to another domain. The rule used to classify a website as phishy is given in Eq. 7:

$$\text{Rule : If} \begin{cases} \text{Anchor URL\%} < 31\% \rightarrow \text{Legitimate} \\ \text{Anchor URL\%} \geq 31\% \text{and} \leq 67\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishy} \end{cases} \tag{7}$$

3.3 HTML and JavaScript Features

These features are extracted from the HTML and JavaScript files of the websites. The various features that can be extracted from HTML and JavaScript files of a website to classify a website phishy or non-phishy are discussed below.

- **Redirect page:** Page redirection is a situation where when we click on a URL to reach page x but it redirects to page y. The rule used to classify a website as phishy is given in Eq. 8:

$$\text{Rule : If} \begin{cases} \text{page redirect} \leq 1 \rightarrow \text{Legitimate} \\ \text{page redirect} > 1 \text{ and} < 4 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishy} \end{cases} \tag{8}$$

- **Hide link:** JavaScript supports various functions to make a website responsive based on user mouse click. One of the functions provided by it is *onMouseOver()*, which hides the text when the mouse hovers over it. Phishers use this feature to hide phishy links. The rule used to categorize a website phishy is given in Eq. 9:

$$\text{Rule : If} \begin{cases} \text{status bar change on nmouseover} \rightarrow \text{Phishy} \\ \text{no status bar change} \rightarrow \text{Supicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{9}$$

- **Right click disable:** JavaScript is used by phishers to block the right clicks on a webpage, which makes users unable to view source code and helps them to do cyber fraud. The rule used to classify a website as phishing or legitimate based on this feature is given in Eq. 10:

$$\text{Rule : If} \begin{cases} \text{disabled right click} \rightarrow \text{Phishy} \\ \text{alert on right click} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{10}$$

3.4 Domain Features

Domain features are the information or features extracted from the domain of a website. We have discussed various domain-based features following based on which we can categorize a website as phishy or legitimate

- **Domain age:** The domain age of a website is obtained from the WHOI database. The domain age information helps us to identify a phishing website. If a website is very old, it indicates that it is valid and not created for phishing purpose. The rule used to classify a website phishy on domain age is given in Eq. 11:

$$\text{Rule} : \text{If} \begin{cases} \text{domain age} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{otherwise} \rightarrow \quad \text{Phishy} \end{cases} \tag{11}$$

- **DNS record:** A website is categorized as phishing or not based on its domain name system (DNS) record. The rule used to classify it is given in Eq. 12:

$$\text{Rule} : \text{If} \begin{cases} \text{no DNS record} \quad \rightarrow \text{Phishing} \\ \text{Otherwise} \quad \rightarrow \quad \text{Legitimate} \end{cases} \tag{12}$$

4 Proposed Method

The proposed method used to detect a website is phishy or non-phishy is shown in Fig. 5. It aggregates the different website features and provides them as input to a trained machine learning classifier to classify it as phishy or non-phishy. The classifier is trained on a standard dataset. We have used six machine learning algorithms such as k-nearest neighbor (KNN), logistic model tree, support vector machine, naive Bayes, multilayer perceptron, and decision tree machine learning algorithms to classify a website as legitimate or non-legitimate. The details of the machine learning algorithm are discussed below.

- **K-nearest neighbor's algorithm**: KNN is a simple and most commonly used algorithm. It is a type of supervised learning method that classifies a new website based on similarity measures. It uses distance measures to find the distance of the new website from the phishy and non-phishy websites available in the dataset, and based on a similarity measure, it predicts whether the website is phishy or not. The distance measure generally used in KNN is Euclidean; however, hamming distance is also used in some cases.
- **Logistic model tree**: The logistic model is a supervised learning classification model built by combining logistic regression and decision tree. It uses the concept of both decision tree and logistic regression tree. The decision tree classifies the problem as a tree where logistic regression generates the result as a discrete

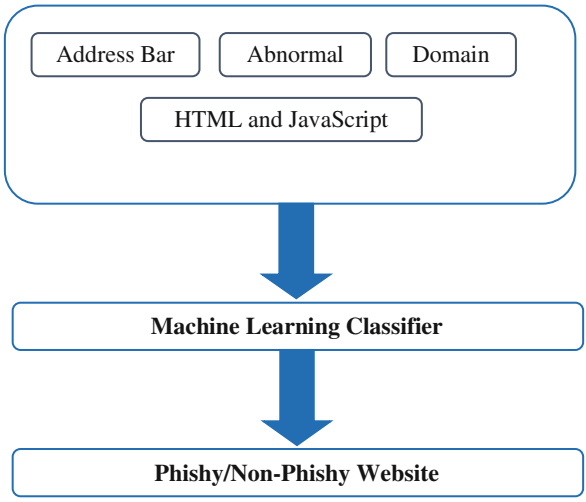

Fig. 5 Flowchart of proposed method

value such as yes or no, 0 or 1, true or false, or high or low. So, we can say that the logistic model tree works on combining two methods into a model tree and generates a tree with nodes containing a logistic regression function.

- **Support vector machine**: Support vector machine (SVM) is one of the most effective machine learning classifier that is used in various fields such as face recognition, cancer classification, and many more. It is a supervised classification method that separates data using a hyperplane where a hyperplane acts like a decision boundary between the various classes. It is a representation of training examples as points in space such that the points of different categories are separated by a gap as wide as possible. It can also perform nonlinear classification and work well with large datasets.
- **Naive Bayes**: Naïve Bayes is a simple probabilistic classifier based on the Bays theorem with an assumption of independence among training cases. It assumes that the quantity of interest is governed by probability distributions and the optimal decision can be made by reasoning about these probabilities together with observed data. Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability of a hypothesis.
- **Multilayer layer perceptron**: A multilayer perceptron is a perceptron with multiple layers. It is a type of feed-forward artificial neural network. It has an input layer, output layer, and hidden layer with perceptron. The perceptron consists of weights, the summation processor, and an activation function. A perceptron takes a weighted sum of inputs and outputs a single value. From the input layer, input signals are taken, and all the computations are performed at the hidden layers, and the final output is reflected on the output layer. If the predicted output is the same as the desired output, then the performance is considered satisfactory, and no changes to the weight are made. However, if the output does not match the desired output, then the weights are changed to reduce the error.

- **Decision tree**: A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions. It is a decision support tool that arranges datasets in the form of a tree-like structure. It is also called the tree-like model, in which each internal node represents the test attribute and all the branches represent the outcome of the test. It built the decision tree based on the training examples using computing entropy and information gain of samples. Once the decision tree is constructed, a new sample is classified into a category based on the decision rule of each node of the decision tree.

5 Results and Discussions

We compare the different machine learning algorithms for phishing website detection using a publicly available dataset available at UCI Machine Learning Repository collected by organizations Phish Tank, MillerSmiles, and Google [13]. The dataset consists of a total of 11,055 entries of phishy and non-phishy website features; out of which 4898 are non-phishy, and the rest are phishy websites. There are total 30 features of each website, which is used for classification purpose. Some of them are IP address, URL length, right click, etc., which are discussed in Sect. 3.

We have used Windows 8 operating system and Weka tool to train and compare the accuracy of machine learning algorithms like multilayer perceptron, support vector machine, decision tree, logic model tree, random forest, and k-nearest neighbor machine learning algorithm for phishing website detection. We used tenfold cross-validation during training to remove the biases. The confusion matrix obtained during the training and validation phase for different machine learning algorithms is shown in Fig. 6. It can be observed from Fig. 6 that random forest has the highest value of true positive while k-nearest neighbor has the highest value of false negative.

We further compared the accuracy of different machine learning algorithms used for phishing website detection in the proposed approach. The accuracy of different algorithms is shown in Fig. 7. It can be observed from it that the random forest is efficient in terms of accuracy to detect phishing websites. The accuracy of random forest is 97.20%. K-nearest neighbor has an accuracy of 97.2% while logistic model tree and multilayer perceptron both have an accuracy of 96.9%. The decision tree has an accuracy of 95.9% while the support vector machine has the least accuracy of 94%.

6 Conclusions

Phishing is one of the important challenges for today's era in cybersecurity. The cases of phishing are growing exponentially and causing many cyber frauds, which result in the loss of money of business organizations or individuals. In this paper,

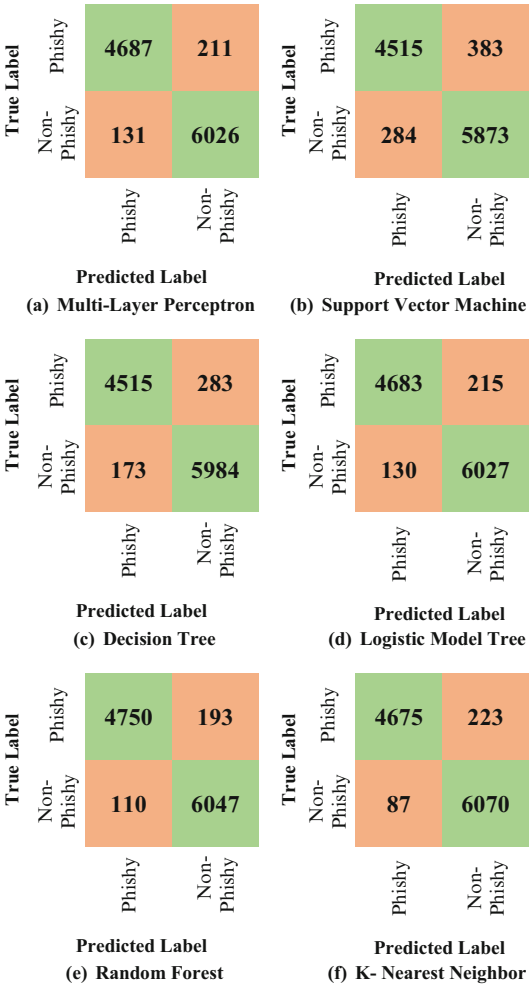

Fig. 6 Confusion matrix for phishing website detection using different machine learning algorithm

we have proposed a machine learning-based approach to detect phishing websites. We used various website features to train a classifier to detect a phishy website. We used six machine learning like the random forest, multilayer perceptron, naïve Bayes, support vector machine, decision tree, and logistic model tree algorithms to train the classifier. It was found that the random forest is the most efficient algorithm to detect phishing websites while other methods detect phishing websites with less accuracy. A comparison of machine learning methods to detect a phishy website in terms of accuracy is given at last.

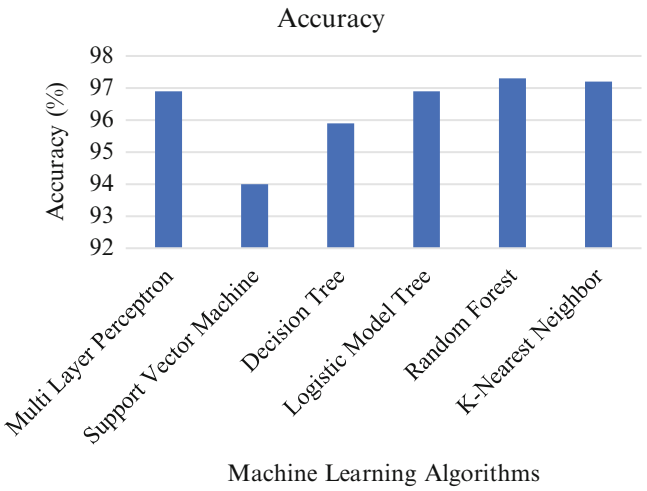

Fig. 7 Accuracy of various machine learning algorithms for phishing detection

References

1. Singh, P., Maravi, Y. P., Sharma, S. (2015, February). Phishing websites detection through supervised learning networks. In *2015 international conference on computing and communications technologies (ICCCT)* (pp. 61–65). IEEE.
2. HeidiBleau. *RSA quarterly fraud report: Q4 2018*. URL: https://www.rsa.com/en-us/offers/rsa-fraud-report-q4-2018. Accessed 15 Sept 2021.
3. Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications, 31*(8), 3851–3873.
4. Sönmez, Y., Tuncer, T., Gökal, H., & Avcı, E. (2018, March). Phishing web sites features classification based on extreme learning machine. In *2018 6th international symposium on digital forensic and security (ISDFS)* (pp. 1–5). IEEE.
5. Sharmin, S., & Zaman, Z. (2017, December). Spam detection in social media employing machine learning tool for text mining. In *2017 13th international conference on signal-image Technology & Internet-Based Systems (SITIS)* (pp. 137–142). IEEE.
6. Altaher, A. (2017). Phishing websites classification using hybrid svm and knn approach. *International Journal of Advanced Computer Science and Applications, 8*(6), 90–95.
7. Karnik, R., & Bhandari, G. M. (2016). Support vector machine-based malware and phishing website detection. *International Journal of Computer Applications in Technology, 3*(5), 295–300.
8. Abunadi, A., Akanbi, O., & Zainal, A. (2013, December). Feature extraction process: A phishing detection approach. In *2013 13th international conference on Intellient systems design and applications* (pp. 331-335). IEEE.
9. James, J., Sandhya, L., & Thomas, C. (2013, December). Detection of phishing URLs using machine learning techniques. In *2013 international conference on control communication and computing (ICCC)* (pp. 304–309). IEEE.
10. Xiang, Y. A. N. G., Li, Y. A. N., Bo, Y. A. N. G., & LI, Y. F. (2017). Phishing website detection using C4. 5 decision tree. *DEStech Transactions on Computer Science and Engineering*.
11. Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert Systems with Applications, 37*(12), 7913–7921.

12. Fu, A. Y., Wenyin, L., & Deng, X. (2006). Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). *IEEE Transactions on Dependable and Secure Computing, 3*(4), 301–311.
13. Mohammad, R., Thabtah, F., & McCluskey, T. L. (2015). *Phishing websites dataset*.