International Journal of Computers and Informatics (IJCI)
Vol. (3), No. (3)

IJCI

March 2024

المجلة الدولية للحاسبات والمعلوماتية

الإصدار (3)، العدد (3)

Exploiting the Capabilities of Classifiers to Examine a Website Defacement Data Set

Elrasheed Ismail Mohommoud Zayid
Dept. of Information Systems, College of Science & Arts-Alnamas, University of Bisha, Saudi Arabia
eazayid@ub.edu.sa , https://orcid.org/0000-0003-2375-6911

Ibrahim Isah
Dept. of Science and Lab. Technology, College of Science and Technology, Jigawa State Polytechnic Dutse, Nigeria

Nadir Abdelrahman Ahmed Farah
Dept. of Information Systems, College of Science & Arts-Alnamas, University of Bisha, Saudi Arabia

Yagoub Abbker Adam
Dept. of Computer Science, College of Computer Science and Information Tech., Jazan University, Saudi Arabia

Omar Abdullah Omar Alshehri
Educational Technology Dept., College of Education, University of Bisha, Saudi Arabia

Abstract

Website defacement is the illegal electronic act of changing a website. In this paper, the capabilities of robust machine learning classifiers are exploited to select the best input feature set for evaluation of a website's defacement risk. A defacement mining data set was obtained from Zone-H, a private organization, and a sample consisting of 93,644 data points was pre-processed and used for modelling purposes. Using multi-dimensional features as input, enormous modelling computations were carried out to determine the optimal outputs, in terms of performance. Reason and hackmode presented the highest contributions for the evaluation of website defacement, and were thus chosen as outputs. Various machine learning models were examined, and decision tree (DT), k-nearest neighbours (k-NN), and random forest (RF) were found to be the most powerful algorithms for prediction of the target model. The input variables 'domain', 'system', 'web_server', 'redefacement', 'type', 'def_grade', and 'reason/hackmode' were tested and used to shape the final model. Using the cross-validation (CV) technique, the key performance factors of the models were calculated and reported. After calculating the average scores for the hyperparameter metrics (i.e., max-depth, min-sample-leaf, weight, max-features, and CV), both targets were evaluated, and the learning algorithms were ranked as RF > DT > k-NN. The reason and hackmode variables were thoroughly analysed, and the average score accuracies for the reason and hackmode targets were 0.85 and 0.585, respectively. The results comprise a significant development, in terms of modelling and optimizing website defacement risk. This study successfully addresses key cybersecurity concerns, particularly website defacement.

Keywords: Website Defacement, Website Defacement Assessment, Classification Metrics, Website Hacktivism, Cyber Risks, Predict Cyber Threats.

1. Introduction

Top-ranking cybercrime references define website defacement as an illegal electronic attack (hack) of a webpage which changes the webpage’s appearance [1-3], including replacement of the site’s content with political, ideological, profane, or inappropriate content [4]. Defacement may be carried out on servers owned by the organization the attacker(s) have chosen [1, 5]. Previous studies [6-7] have outlined the common types of website defacement attacks, including unauthorized access, SQL injection, cross-site scripting (XSS), DNS hijacking, and malware infection [8, 9]. Standard methods to address website defacement follow the public wisdom that "prevention is better than a cure" [10]. A very effective way to secure a website is to be aware of attackers and data breaches, rather than detecting breaches and fixing them. Some methods limit the ability of offenders to upload files and their access to the server and controls within the organization, while others use a secure socket layer (SSL) certificate for confidential security across HTTPs and apply strong rules for securing login information, particularly usernames and passwords [11].

At present, most web pages are vulnerable to defacement and hacktivism [1, 12]. Mindful and continuously aware websites are often the only websites that are considered safe from destructive defacement threats. Defacement attacks can damage a company or organization's reputation, leading to the loss of trust and money. An affected website may be banned from search engine results, such as on Google. Strengthening the resilience of a website to prevent or become immune to defacement risks is the ultimate goal that each institution strives to achieve. Many techniques have been used to address website defacement [12-14]; however, prediction-based methods are preferred. These techniques are embedded with a powerful computation mechanism and can be used in a simple manner to realize the final design.

Many proposals utilizing robust machine learning algorithms to assess website defacement have been put forward. Nevertheless, an optimal method has not yet been designed [15]. A devastating website defacement attack motivated the exploitation of neural networks as a promising research method for addressing issues related to website defacement and cybercrime threats [16]. This analytical method is based on the process of deep learning, in order to mine the information contained inside the raw data points by learning from existing data. In this regard, various proposals have achieved advanced successes in website defacement classification and prediction [3, 6, 8, 17-21].

Our main idea was to collect a massive number of website defacement inputs, pre-process and filter these raw data, and examine several machine learning modelling scenarios to determine the best prediction model. Massive modelling permutations and combinations were assessed, and a single output/target CV technique was utilized to eliminate errors and determine the model's highest scores per target. These examined targets were chosen from the set of input features (i.e., domain, system, web_server, redefacement, type, def_grade, reason, and hackmode). Comparing the outputs for all targets, reason and hackmode presented the best results. In addition, reason and hackmode significantly contributed to measuring and assessing website defacement and yielded the lowest scores in terms of timing and classifier cost.

In this aspect, our model was designed to estimate two different outputs: Reason and hackmode. When predicting reason, hackmode was considered as an input feature; meanwhile, when predicting hackmode, reason was considered as an input feature. Therefore, our key input features included 'domain', 'system', 'web_server', 'redefacement', 'type', 'reason/hackmode', and 'def_grade'. However, the only dependent output variables were reason and hackmode. For computation, we employed three popular powerful machine learning prediction algorithms: Decision tree (DT), random forest (RF), and k-nearest neighbours (k-NN).

To simplify the computational process, the inputs and outputs were normalized to have a mean of zero and a variance of unity. Furthermore, sigmoid was chosen as an activation function, along with the Levenberg–Marquardt (LM) algorithm. Finally, the GridSearch CV technique was used to tune the hyperparameters of the DT, RF, and *k*-NN classifiers.

Many models were tested for each algorithm, and the averages are reported based on the optimal hyperparameter sets. These hyperparameter sets included factors such as CV, max-depth, min_sample_leaf, weights, maximum features, and neuron node neighbours. In summary, performance measures for the correlation coefficients, model timings, and average error rates were calculated. The attack reason(s) and hackmode type(s) were correctly developed. Furthermore, this study successfully addresses the most significant cybersecurity concerns particularly website defacement.

The remainder of this manuscript is organized as follows. In Section 2, the literature is reviewed. In Section 3, the proposed method is presented and the data set generation process is outlined. The results and discussion are provided in Section 4. Finally, this study is concluded in Section 5, followed by the references.

2. Related Works

Zone-H is an indispensable data source from which to retrieve information for the assessment of website defacement and hacktivism risks [5, 21]. In this regard, [22-26] have addressed attack types and profiling trends and presented a questionnaire which was completed by 119 active hackers. The paper in [22] examined hacker typologies by analysing their feedback and responses.

Many studies have validated these findings from an environmental criminology perspective [24,26,27]. In this regard, vital findings are compared in the discussion section of this paper.

In [27], factors such as the peak signal-to-noise ratio, cyclic redundancy check, secure hash algorithm, and structural similarity measure were calculated.

Excellent research on the detection of website defacement based on machine learning [28-34] has been published; for example, [34] performed an extensive experiment and obtained an overall accuracy of more than 99.26% and a false-positive rate of approximately 0.27%. The work in [34-35] focused on defacement heterogeneity, and [35-36] targeted reference group-motivated hackers and listed their motivations as being for fun, for a challenge, to be the best, for patriotism, for political reasons, and for revenge.

To better understand Islamic Jihadism, defacement features were obtained in [37-40]. Regression was the key factor used to validate the outcomes. Meanwhile, [41-43] utilized classification and case-based reasoning mathematics for the outputs.

The authors of [34] used website defacement and signature-based detection methods. The capabilities of machine learning approaches were examined, and classification metrics of 99.26% for accuracy and 0.26% for false positive rate were reported. Thus, the authors argued that only scalar assessment of machine learning classifiers is necessary; however, many graphic metrics can also be obtained using these methods.

The authors of [35] classified active hackers into mass/single hackers using a massive data set from Zone-H. The features used for modelling were defacement type, hacker type, operating system, hacker motivation, webpage type, site re-defacement, and method of attack. Concerning the analysis method, they utilized the Poisson distribution and a multinomial logistic regression classifier for prediction. It can be concluded that this study can provide information on criminological direction.

In [36], Jihadist features were characterized to differentiate them from features of common website attackers. This research presents excellent work regarding the assessment of hacktivism; however, a traditional and less computational analysis-based approach was followed. Utilizing a binary regression model, it was concluded that a Jihadist offensive was the least likely among other kinds of website defacement.

Summarizing website defacement development across Jihadist groups, [37] reported interesting results; for example, 20,000 websites have been attacked by these groups. Attack strategies have rapidly changed from cyberattacks to cyberterrorism, and the attacks have changed in terms of both number and sophistication. Moreover, the study introduced issues such as the Cyber Caliphate and Inspire Jihadist groups, as well as the age of digital natives (16–24). The CIA's World Factbook 2018-2019 [38], which reports a comprehensive world picture and classification, supports these facts concerning website defacement.

In [39], the classification and differences between traditional terrorism and cyberterrorism using Al Qaeda's network were described in detail, in order to understand how they exploit e-mail services to support Jihad e-mail.

Interesting research has been conducted to address the relationship between ideology and lethality [40]. A dataset was retrieved from the Global Terrorism Database, and the Global Jihadist Movement was determined to be the deadliest. Only logistic regression was used in this study, and models were assessed using variables and incident counting.

In [41], similarity measures were combined with clustering to assess website defacement. This method supports a case-based reasoning technique, and good results were achieved by inferring the attitudes of hackers to find evidence of website defacement. Such research demonstrates data-driven power as supporting the evaluation of website defacement.

After obtaining global website defacement and hacking data source files from Zone-H, website defacer samples from 114 nations were extracted and examined to explore/understand the relationship between the communication capabilities of the countries and the number of websites defaced [42]. For this purpose, routine activity factors and a methodological framework were used in this study. It was assumed that victimization is caused by a combination of the website attackers, the website itself, and a lack of secure systems with respect to time and within a hosting server. The analytical tools were typical machine learning correlation factors for the output variables. Finally, the study findings verified the hypothesis that website protection (guardianship) limits defacement numbers in a country-wise manner.

Previous studies [34-41] have focused on the usability of machine learning approaches to assess website defacement and hacktivism classifications. However, the approaches varied in terms of the algorithm(s) and data set used, the performance measures employed to evaluate the outcomes, and the study assumptions. These variations resulted in slightly different final comparisons; however, the machine learning methods were clarified, empowering the approach followed in this study.

The current study differs from those in the literature review, in that valuable modelling performance measures such as max_depth, min_sample_leaf, n_neighbors, accuracy, average error, and model evaluation time are considered, and the prediction performance of the models are validated with respect to three different algorithms. A large number of modelling permutations/computations are examined, in order to determine the optimal final models. Moreover, the mean prediction score is measured and the effect of increasing the size of n_neighbor on the mean accuracy is calculated for both reason and hackmode. Variation is also considered, through the use of max_depth with respect to the mean test score for the DT and RF algorithms.

3. Materials and Methods

Pre-processing is an indispensable process for any machine learning prediction method. The quality of the raw data set is measured to facilitate its multi-dimensional analysis. In this process, several measures are conducted to obtain a clean data set. These measures include 1) checking for correct/incorrect input(s), 2) examining the completeness of the data, 3) testing the consistency of the data (i.e., modified and invalid data), 4) updating the timelines, 5) assessing the believability of the data in terms of trustworthy inputs, and 5) examining the interpretability of the data (i.e., how easily the data can be understood).

The machine learning feature selection process is an effective means of predicting variables such as reason and hackmode. Many approaches can be employed to validate the selected features.

Figure 1 shows the details of the method applied in this manuscript. The 16 features were examined based on 7 features of the related algorithms; namely, 'Domain', 'system', 'web_server', 'hackmode', 'redefacement', 'type', and 'def_grade'. These features were sorted and selected to produce the final input feature set. The targets were determined as reason and hackmode, and the features were examined under five different powerful machine learning kernel algorithms; namely, the decision tree, random forest, *k*-nearest neighbours, SVR, and LR algorithms.

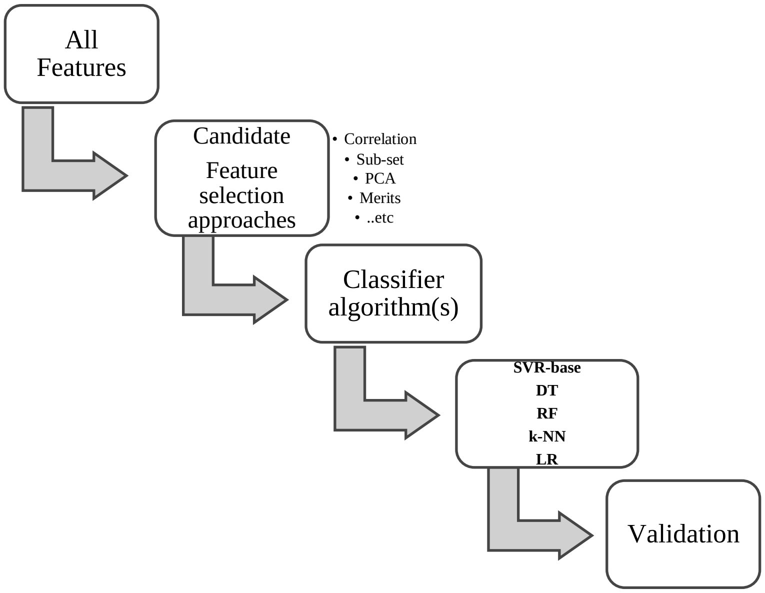

Figure (1): Prediction classifier flowchart

Five basic processes are shown in Figure 1. First, all of the input features are summed and passed into a feature selection phase, which accurately computes and filters the inputs using the correlations between the input parameters and the desired output. For this process, many techniques can be applied; however, correlation, PCA, and merits are the leading algorithms. Multilayer perceptrons and ANNs are very effective methods for the determination of outputs, which were introduced in [3, 8].

There are several active cybersecurity data science sources track hacking records. Very often, however, ZONE-H is considered the leading source of unrestricted and authenticated website defacement information. It provides an archive of defaced websites from all around the world [5]. The study population was obtained using a data set of terra-records offered by Zone-H for research purposes. The original data set package comprised 93,644 items with 16 dimensions and 16 features. After pre-processing, only 8 features relevant to the final data set were retained. The input and output variables and their statistics are presented in Table 1. Based on the 8 features with 80,382 items (rows), the final data set contained 9 columns.

Table (1): Descriptive statistics of the data set

Statistical Metric	**Data set Features**									
	Domain	**System**	**Web_Server**	**Reason**	**Hackmode**	**Type**	**Redefacement**	**Def_Grade**	**Domain**	**System**
Mean	2.8	0.85	0.67	1.26	3.93	0.72	0.12	0.24	2.8	0.85
Stdv.	2.89	1.72	1.13	2.25	5.14	0.45	0.33	0.43	2.89	1.72
Min.	0	0	0	0	0	0	0	0	0	0
25%	1	0	0	0	1	0	0	0	1	0
50%	1	0	0	0	2	1	0	0	1	0
75%	5	0	1	2	5	1	0	0	5	0
Max.	16	17	17	10	26	1	1	1	16	17
Count	80,382	80,382	80,382	80,382	80,382	80,382	80,382	80,382	80,382	80,382

Stdv: standard deviation; min. and max, minimum and maximum, respectively.

The equations below detail the mathematical relationships for X^2 (chi-square) [10]:

$$X^2 = \frac{\sum_{i=1}^{n}(Observed - Expected)^2}{Expected^2}, \qquad (1)$$

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - A^-)(b_i - B^-)}{(n-1)\sigma_A\sigma_B}, \qquad (2)$$

$$r_{A,B} = \frac{\sum_{i=1}^{n}((a_i b_i) - nA^- B^-)}{(n-1)\sigma_A\sigma_B}, \qquad (3)$$

where n is the number of tuples; A^- and B^- are the means of A and B, respectively; σ_A and σ_B are the standard deviations of A and B, respectively; and $\sum_{i=1}^{n}(a_i b_i)$ is the sum of the AB cross-product. From the equations above, the computations are performed as follows.

Figure (2): shows the chi-square correlation computation matrix, which measures the correlations between variables

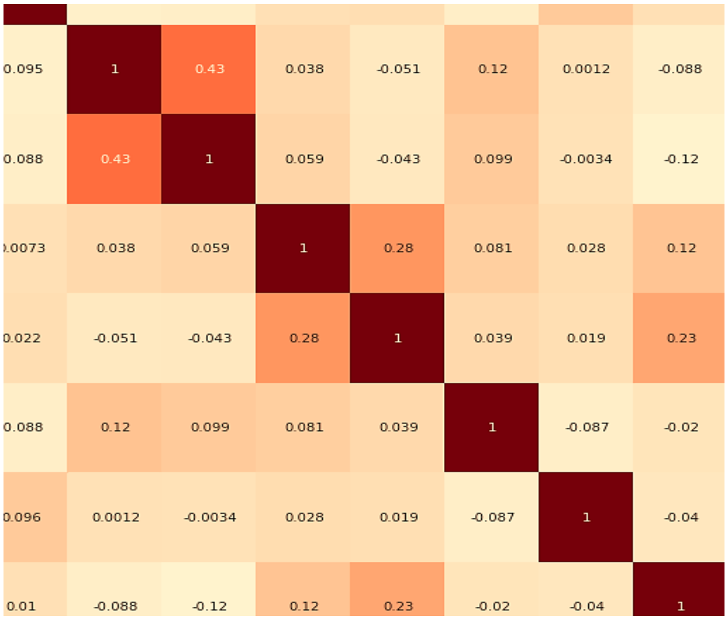

Figure (2): Correlation matrix

The following equations were used to calculate the performance measures [44-48]:

$$y_t^{'} = f(x_t, w), \tag{4}$$

where *w* is the ensemble of the synaptic weights and x_t are the input variables being fed into the network with errors.

$$E(w) = \frac{1}{2}\sum_{1}^{N}(y_t - y_t^{'}(x_t, w))^2 \quad . \tag{5}$$

Optimization is conducted as follows [44]:

$$\hat{w} = \arg\min E(w) = \arg\min(\frac{1}{2}\sum_{1}^{N}(y_t - y_t^{'}(x_t, w))^2) \quad . \tag{6}$$

Accuracy (Acc) represents the first measure used to assess the classification performance:

2.

1. $Acc = \frac{TP+TN}{TP+TN+FP+FN},$

where *P* and *N* indicate the number of positive and negative samples, respectively. The error rate (*ERR*) is

4.

3. $ERR = 1 - Acc,$

6.

5. $ERR = \frac{(FP+FN)}{(TP+TN+FP+FN)}\ .$

Python was considered to be the best programming language to develop our complex learning models for classification and prediction.

4. Results and Discussion

Table 2 lists the key features of the models used in this study. The used inputs included 'domain', 'system', 'web_server', 'redefacement', 'type', and 'def_grade'. The dependent output variables were chosen as reason for the first set and hackmode for the second set. In Table 2, the output parameters (i.e., reason/hackmode) are highlighted in grey.

These hyperparameters were permuted and tuned until the best scores were obtained in each case, as reported in the table. The most significant parameter for all three models was accurately adjusted using the above method and tuned to further improve the models. max_depth and min_sample_leaf were considered for DT and RF, while n_neighbors was considered for the case of k-NN. In addition, max_features was tuned for RF. The variation in these parameters was computed as a function of the average score, as reported in Figures 3–10.

As shown in Table 2, the CV values ranged from 14 to 16 for all algorithms. The hyperparameters of the random forest model were equal for both targets (i.e., reason and hackmode). For the k-NN algorithm, the sub-algorithm used was ball_tree, the leaf_size was set to 9, and the n_neighbors values were 18 and 7 for reason and hackmode, respectively. Different values were obtained for the leaf per algorithm(s) variable. The remaining parameters were primarily classifier-based values.

Table (2): Input/output features for each model

Model	Hyperparameters		Target	Hyperparameters		Target
	Parameter	Value		Parameter	Value	
Decision tree (DT)	CV[a]	16	Reason	CV	14	Hackmode
	max_depth	16		max_depth	14	
	min_sample_leaf	1		min_sample_leaf	1	
Random forest (RF)	CV	15	Reason	CV	16	Hackmode
	max_depth	16		max_depth	16	
	max_features	5		max_features	5	
	min_sample_leaf	3		min_sample_leaf	3	
k-nearest neighbours (k-NN)	CV	8	Reason	CV	14	Hackmode
	algorithm	ball_tree		Algorithm	ball_tree	
	n_neighbors	18		n_neighbors	7	
	leaf_size	9		leaf_size	9	
	weights	distance		weights	distance	

CV: cross-validation.

The CV training method obtained the highest training algorithm fit for all models, and its performance metrics for the three considered models are presented in Figures 3–10. The performance of the selected models was evaluated starting with the mean score with respect to cross-validation for the DT, RF, and k-NN algorithms.

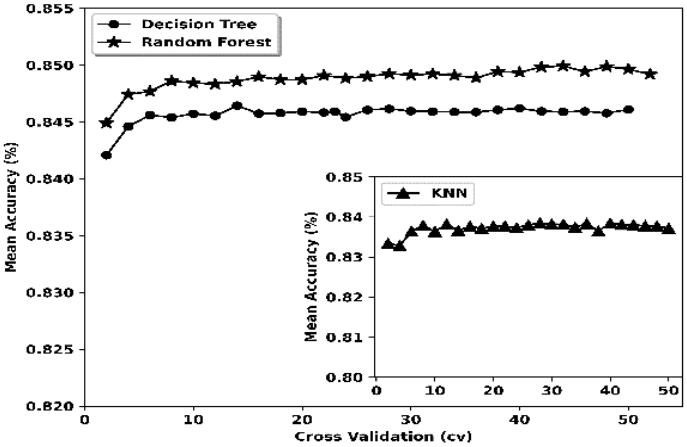

Figure (3): CV as a function of mean score for the reason target

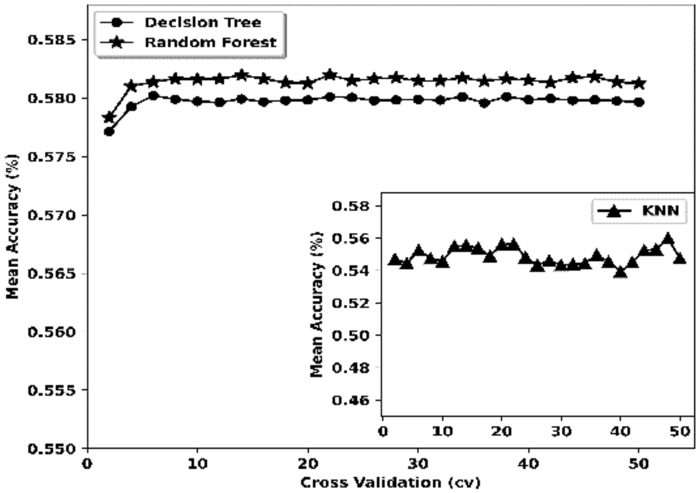

Figure (4): CV as a function of mean score for the hackmode target

The hyperparameters were then optimized, and only the most relevant parameters were tuned. Starting with max_depth, the optimized values for DT and RF were obtained, as shown in Figure 5 and Figure 6.

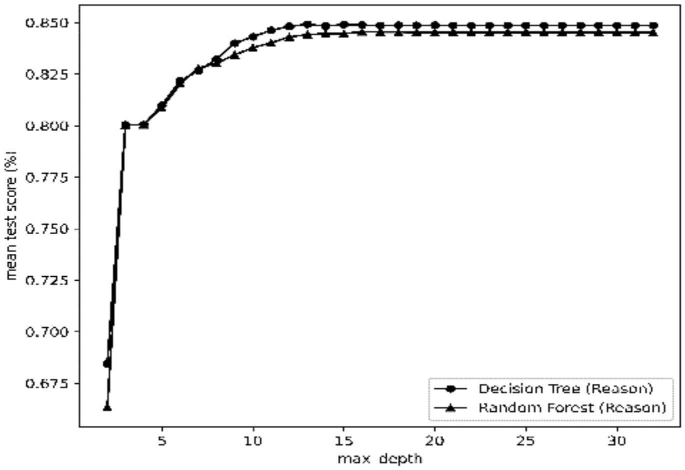

Figure (5): Variation in max_depth with respect to the mean test score for the reason target using DT and RF

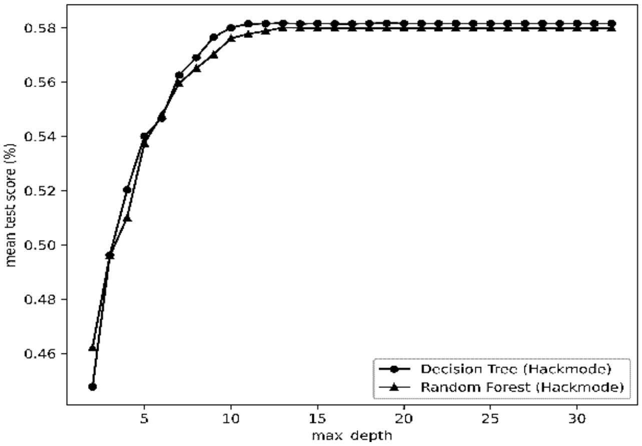

Figure (6): Variation in max_depth with respect to the mean test score for the hackmode target using DT and RF.

The minimum number of sample leaves (min_sample_leaf) was tuned after max_depth. The variations with respect to the mean score are presented in Figure 7 and Figure 8. It can be seen that the mean score decreased as the min_sample_leaf parameter increased for DT, and increased and then decreased continuously for RF.

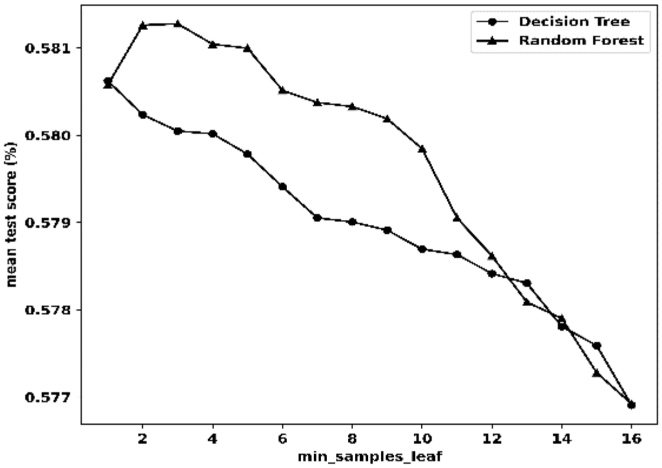

Figure (7): Variation in min_sample_leaf as a function of mean test score for DT and RF with hackmode as the target

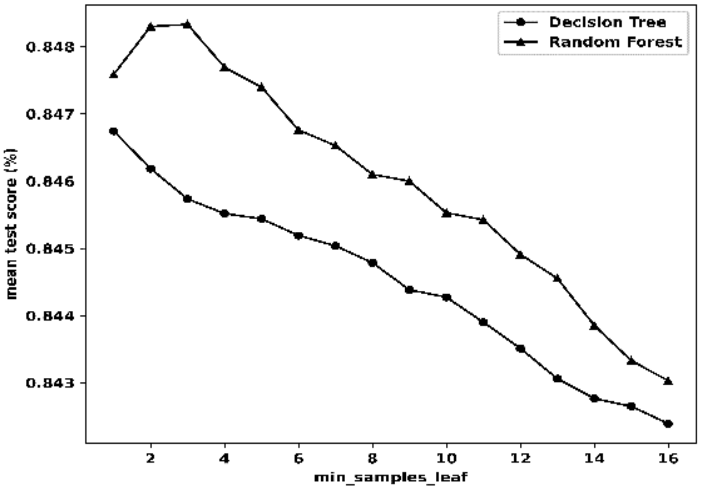

Figure (8): Variation in min_sample_leaf as a function of the mean test score for DT and RF with reason as the target

As presented in Figure 9 and Figure 10, the max_features and n_neighbors parameters were tuned for RF (reason and hackmode) and RF (reason) and k-NN (hackmode). The best values for each parameter were considered.

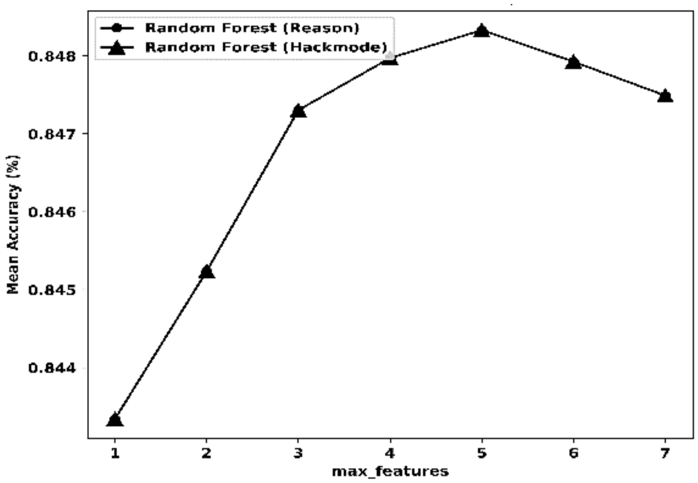

Figure (9): Variation in max_ features based on the mean score for RF with both reason and hackmode as targets

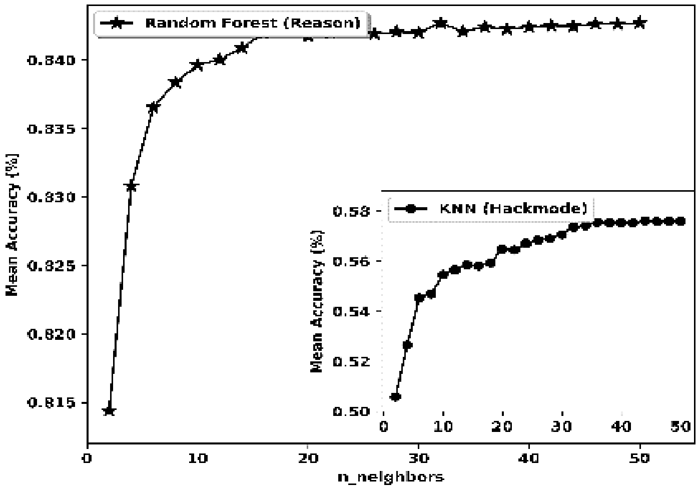

Figure (10): Effect of increasing the size of n_neighbors on the mean accuracy (reason and hackmode)

After model training, an accurate prediction of the target reason/hackmode could be made. The mean score of the model depended on the amount of the data set used as features for target prediction, and the mean score decreased with an increase in the number of prediction data sets. An accurate prediction could be made with 99% certainty in the case of a relatively small data set (with a typical size of one to five) for the DT, RF, and k-NN models. When predicting the reason for the hack, the mean score decreased exponentially with an increase in the number of prediction data sets.

When predicting hackmode as the target, a different (opposite) interpretation to that when predicting reason was obtained. In particular, the RF and DT models started with a mean score accuracy of 0.5 with a relatively small data set for prediction, and this score increased as the number of data sets (i.e., test sets) increased from 50 to 200. The mean accuracy of these two models for predicting hackmode ranged between 0.5 (for a relatively small test set) to 0.6 (for a relatively large test set). Finally, the k-NN model performed poorly when predicting hackmode for the data sets under consideration. The mean score decreased rapidly with an increasing number of test sets, as shown in Figure 11 and Figure 12.

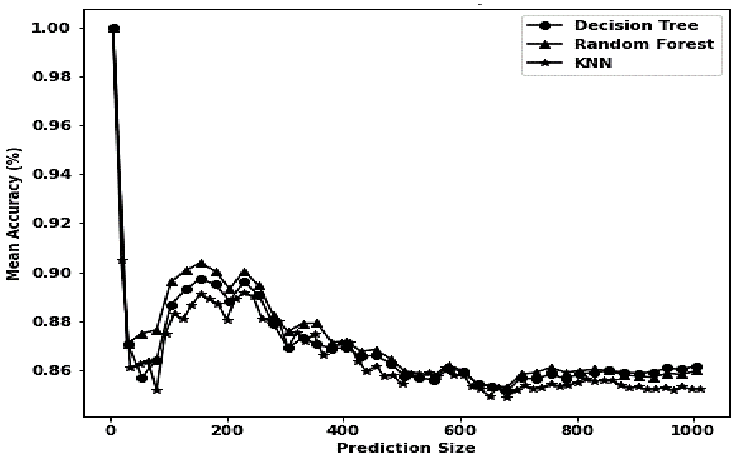

Figure (11): Effect of increasing the number of predictions on the mean score with reason as target

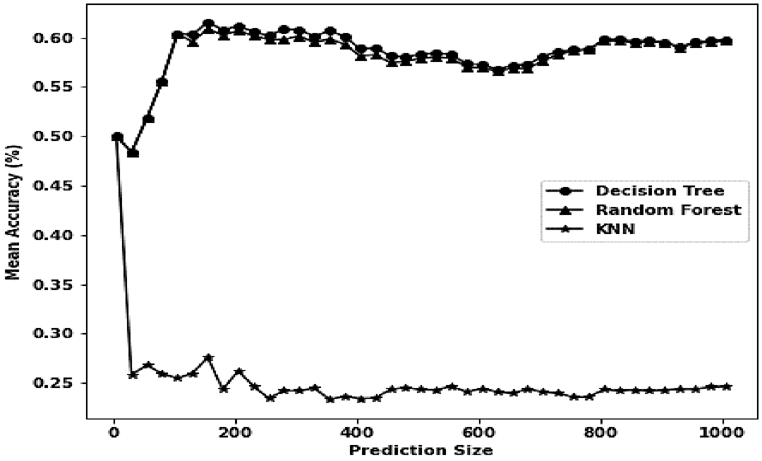

Figure (12): Effect of increasing the number of predictions on the mean score with hackmode as target

Model selection involves choosing the best performance score with minimal mean error(s). It also involves evaluating the average time for training, evaluating, and testing the model, among others. The best model for a given data set is the model with the highest performance and the shortest training and testing time. Table 3 shows the training and evaluation times for the three models, as well as the average errors.

Table (3): Model training and evaluation times and mean errors

Model	Training Time (s)	Evaluation Time (s)	Mean Error (%)	Target	Training Time (s)	Evaluation Time (s)	Mean Error (%)	Target
DT	0.342	0.018	0.1488	Reason	0.228	0.018	0.4131	Hackmode
RF	17.832	1.17	0.1479	Reason	14.712	2.13	0.4135	Hackmode
k-NN	39.678	68.34	0.156	Reason	37.86	73.692	0.4372	Hackmode

The analysis in this study demonstrated that DT, RF, and *k*-NN were the classifiers that are best suited to such a limited data set and, as such, can be used for website defacement prediction. The set of supervised learning classifiers covered in this study positively reflected the accuracy and superior performance of the methods, particularly in terms of the training and learning processing times, as well as the error and accuracy metrics, among others. As such, the algorithms employed for this method can be generalized and used in related website defacement assessment research. Moreover, we evaluated classification methods with respect to website defacement optimization factors. In this respect, the reported average scores for Acc, errors, times, and visual metrics supported the superior performance of our methodology, when compared with related approaches in the literature.

A deep analysis of the results and scores is provided in Table 3, demonstrating the high scores for reason as a key output classifier and the associated names, types, and ratios. Furthermore, the hackmode parameter reflects the necessity to warn our community against the serious security risks related to website defacement. Conducting such research increases the awareness of the community and cybersecurity-related individuals, as well as encouraging system administrators to secure their servers and applications against defacement risks due to the number of hacks, hacking types, hacking modes, and cases of website defacement that have been reported.

In summary, the findings of this study demonstrate the correct assumption that machine learning classifiers can be used to address many website defacement challenges, especially in terms of prediction and visualization. Furthermore, the website defacement input features—namely, reason and hackmode—are key factors to consider when reviewing any website defacement classifier.

Assessing and optimizing website defacement cybersecurity measurements can protect the community from malicious website attacks and encourage researchers to deepen their investigations. This study details several learning techniques, measures, and tools that could be used to analyse cybercrime risks and enable stakeholders to implement good cyber-domain prevention measures.

Table 4 lists the key factors that influence hackers and motivate them to hack websites, in the following order: To be the best defacer, for fun, for political reasons, as a challenge, for patriotism, and for revenge. Note that these were not the only reasons given in the study; they only represent the top six reasons. The ratio column lists the percentage who responded with each reason. It is clear that “being the best defacer” had the highest percentage (27.1%), while “revenge against a website” was the lowest-ranking reason (2.675%). Indeed, such findings motivate us to increase the amount of computing-based ethics in relevant education of our communities. In this regard, searching for the factors that influence and motivate defacers offers useful information. In addition, addressing the trends that have increased the risks of using information technology in an unethical manner may decrease website-related issues.

Table (4): Reasons and motivations.

Reason	Attack Reason Selected by Attacker	Ratio (%)
0	Heh...just for fun! ^{57046}	6.277%
1	Political reasons^{3619}	4.500%
2	I just want to be the best defacer^{21782}	27.098%
3	Revenge against the website^{2151}	2.675%
4	As a challenge^{2540}	3.156%
5	Patriotism^{2247}	2.795%

Table 5 details the methods of attack chosen by the attackers, with corresponding risk measurement percentages. This table outlines the hacking methods in descending order, listing only the fifteen most-frequently occurring methods. SQL injection was the highest-ranking hackmode level, with a risk of approximately 32%. Hackers usually exploit data-driven applications and may easily penetrate systems using SQL injection methods. These methods are used by attackers as direct-access vectors for websites, and have been extensively utilized to break into massive SQL databases.

Many proposals have addressed this problem; nevertheless, the number of SQL injection threats is growing. Overall, server intrusion had the second-highest attack risk (27.359%), as servers act as system incubators.

URL poisoning is one of the leading factors of website defacement, responsible for 1162 cases (1.5%).

“I Just want to be the best defacer” was the most selected attack reason by attackers (i.e., 27.1%), while “Revenge against the website” was the lowest (i.e., 2.68%).

From Table 5, the major website hacking methods were as follows: Server intrusion, SQL injection, DNS injection, and file inclusion.

Table (5): Major hacking methods

Hackmode	Attack method selected by notifier	Code#	Risks (%)
1	Overall Server intrusion**21992**	1	27.359
	¾ Web server intrusion**2032**	16	2.527
	¾ FTP server intrusion**480**	18	0.597
	¾ RPC server intrusion**353**	21	0.439
	¾ Telnet server intrusion**339**	15	0.421
	¾ SSH server intrusion **442**	20	0.549
	¾ Other server intrusion**18346**	1	22.823
2	SQL injection**25700**	0	31.972
3	DNS attacks**555**	12	0.69
4	File inclusion**12186**	2	15.16
5	brute force attack**1324**	3	1.647
6	configuration/admin.Mistake**1541**	4	1.917
7	known vulnerability(i.e. unpatchedsystem)**17960**	5	22.343
8	URL poisoning**1162**	6	1.445
9	Undisclosed (new) vulnerability**1292**	7	1.607
10	Other web application bug**4350**	8	5.411
11	Social engineering**762**	9	0.947
12	0t available**8007**	10	9.961
13	Cross-site scripting**401**	11	0.498
14	Remote admin. panel access through bruteforcing**385**	14	0.478
15	Shares misconfiguration**431**	22	0.536
16	Attack against the administrator/user (password stealing/sniffing)**140**	24	0.174

Table 6 reviews the major differences between the prediction models considered in the literature, indicating the significant contributions of our study in comparison with recent related studies.

Table (6): Comparison profile for website defacement prediction models using machine learning techniques.

Author(s), year(s)	Data set SRC, Size	# Objects	Duration, Type	Software tool	Algorithm(s)	Metrics	Purpose
Ours	Zone-H, 93,644 defacements	80,382 objects	2015–2016, standard	Python 3.10 (64-bit)	- DT - RF - k-NN	max_depth, min_sample_leaf, n_neighbors, timing, AVG error, and accuracy	Classifier prediction
Burruss et al., 2021	Zone-H, 1292 defacements 119 -> questionnaires	119 questionnaires	June–August 2017, 1062 -> standard 119 -> questionnaires	Stata v. 16.1, gsem command (StataCrop, 2017)	AIC = 657.653; BIC = 668.769, log-likelihood	- IRR% - SE	Classifier prediction
A. Moneva et al., 2022	Zone-H, 9,117,268 defacements	23.6% single attacks 76.4% mass attacks	2010–2017, standard	R-package 3.6.1 and R-Studio 1.2.5001	Statistical means	- Bar charts - Histograms - Log10 - Percentage % - Frequency	Regression assessment
Gurjwar R.K, Sahu D.R., and Tomar D.S., 2013	Monitoring 250 images MANIT, Bhopal (M.P.), India	100 webpages	**2013,** monitoring	CentOS Linux 5.9 C#.Net	CRC32, MD5, SHA 512, PSNR, and SSIM techniques	- Accuracy	Pre-processing/data cleaning
Hoang X. D. and Nguyen N. T., 2019	1200 English 217 Vietnamese 1200 defaced pages	50 attack signatures	2019, standard	Python Sklearn machine learning library	Multinominal naïve Bayes Random forest	- Accuracy - F1-score - Detection rate %	Conducting raw data collection
S.G.A. van de Weijer et al., 2021	Zone-H 2,745,311 defacements	66,553 hackers	2010–2017, standard	PL (i.e., C++/Java)	Logistic regression	Hacker's AVGs for: - Timing; - Length; - Frequency	Regression
Holt et al., 2021	Zone-H 2285172 defacements 2012–2016 @USA	29,035 attackers	2012–2016, standard	STATA 13 using cluster command	Routine activity theory (RAT) Binary logistic regression	-SE - b - # multicollinearity - Tolerance - Variance inflation	Classifier prediction
Mee Lan Han et al., 2019	Zone-H 212,093 defacements	k-hacker@DB randomly selected 100 hackers	2004–2019, standards	Data-driven and evidence-driven decision tools	CBR-based	- Similarity measure - Clustering	Data driven
Howell Jordan C. et al., 2019	13 M@Zone-H United States Central Intelligence Agency Freedom House Forum of Incident Response and Security Teams (FIRST.org) Kaspersky Lab	114 countries	2017, standard	Statistical analysis tools	Negative binominal regression	-IRR - SD, AVGs - b	Classifier prediction

Key: AIC and BIC denote the Akaike information criterion and Bayesian information criterion, respectively. AVG, average; b, binary regression; IRR, incident risk ratio (IRR%); B, change in the log of counts (b).

In particular, Table 6 demonstrates that our study gave the best results for the website prediction algorithms.

Regarding website defacement and hacktivism, the following points are valuable to consider.

When predicting website defacement, features such as domain, system/OS, webserver, reason, hackmode, type, defacement, state, and location are correlated and are key variables that can be used for prediction.

Based on their accuracy, the prediction algorithms can be ranked in the following order: DT > random forest > *k*-nearest neighbours.

The top five affected countries were ranked as follows: The USA$^{\mathbf{47.24\%}}$ > the UK$^{\mathbf{6.59\%}}$ > Germany$^{\mathbf{6.18\%}}$ > India$^{\mathbf{6.56\%}}$ > the Netherlands$^{\mathbf{5.15\%}}$. Additionally, this study revealed that all countries are routinely affected by defacement cyberattacks.

After mining the data set, our analysis revealed a lack of Islamic extremist and Jihadist defacement and hacktivism as well as a very low extremist contribution in terms of defacement (i.e., 0.00249%). This fact encouraged the authors to argue that: 1) Extremist Islamic groups may lack deep defacement coding skills or 2) powerful recently established guard systems limit their activity and bind their cybercrime aggressiveness, or both. Furthermore, the continuous disassembly of entire networks may have prevented them from training and acquiring superior IT skills.

Employing powerful machine learning methods to predict website defacement and hacktivism was our chosen approach to carry out relevant computations, and good results were achieved.

The outcomes of this study can be utilized by communities, institutions, organizations, governments, and individuals to promote immunity against defacement and hacking risks.

Concerning website cybercrimes, prevention is better than detection and avoidance.

Overall, server intrusion includes several sub-cybercrime types, including web server intrusion, FTP server intrusion, SSH server intrusion, RPC server intrusion, and Telnet server intrusion.

Massive embedded algorithms for prediction, including DT, RF, and *k*-NN, were found to be the most suitable algorithms for computing the outputs hackmode and reason.

5. Conclusion

In this study, a sample data set obtained from Zone-H was used to test the website defacement classification performance of machine learning supervised classifiers. The CV technique was used to avoid modelling errors and maintain the stability of the models. Differentiation was implemented for the output models, and measurements were made to choose the best targets. Reason and hackmode were the selected targets, due to their high output scores. Exploiting the capabilities of the machine learning classifiers, rigorous experiments were conducted to obtain the top-performing classifiers. The GridSearch CV technique was utilized to tune the hyperparameters of the three selected models—namely, DT, RF, and k-NN. The hyperparameters were permuted and tuned until the best scores were obtained in each case, as reported in the paper. The most significant parameters for all three models were accurately adjusted by tuning to further improve the models under the parameters obtained using the abovementioned method. The performance measurement factors max_depth and min_sample_leaf were considered in the cases of DT and RF, and n_neighbors was considered in the case of k-NN.

The study's limitations are summarized as follows: 1) Conducting research on a real data set is highly desirable. However, there is a lack of access to the right dataset, and obtaining a suitably large website defacement data set can be extremely expensive. 2) The limited sample size used in this study made it insufficient for mining/dataware, as the convergence criteria and the requirements for statistical metrics could not be achieved. As the performance of machine learning classifiers depends on the number of data points used, the use of big data will support the learning ability of classifiers, thus enhancing their prediction and accuracy. 3) Concerning visual classification measures, the medium outputs obtained in this study may be due to the nature of the data set used, as the scattering of data points affects the performance of such measures. 4) The probability distributions for each input/output target variable included some missing or noisy data points, which were manipulated during the pre-processing phase. This may have affected the ability to obtain better scores than those reported in the Results section.

The research presented in this study can be developed in several ways. For example, a deep mining exploration can be conducted in terms of clustering, outliers, and modelling using advanced neural network algorithms.

- **Availability of data**: The benchmark datasets generated and used in this research are available on request.
- **Informed Consent**: Not applicable.
- **Competing interests**: The authors declare that they have no competing interests.
- **Authors' Contributions**: All authors contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

6. References

[1] Romagna, M.; van den Hout, N. J (October 2017) Hacktivism and Website Defacement: Motivations, Capabilities and potential Threats. Proceedings of the 27th Virus Bulletin International Conference: 41–50. Retrieved 8 October 2017.

[2] Aslan, Çağrı Burak; Li, Shujun; Çelebi, Fatih V.; Tian, Hao (9 November 2020) The World of Defacers: Looking Through the Lens of Their Activities on Twitter. IEEE Access. 8: 204132–204143. doi:10.1109/ACCESS.2020.3037015.

[3] Hoang, Xuan Dau (2018) A Website Defacement Detection Method Based on Machine Learning Techniques. Proceedings of the Ninth International Symposium on Information and Communication Technology - SoICT 2018. Danang City, Viet Nam: ACM Press: 443–448. doi:10.1145/3287921.3287975. ISBN 978-1-4503-6539-0. S2CID 56403851

[4] Bartoli, A.; Davanzo, G.; Medvet, E (2010) A Framework for Large-Scale Detection of Web Site Defacements. ACM Trans. Internet Technol. 2010, 10, 10.

[5] Zone-H. (2022) News. www.zone-h.org/listingnews. Accessed (9th June 2021).

[6] Burruss, G. W., Howell, C. J., Maimon, D., & Wang, F (2021) Website defacer classification: A finite mixture model approach. Social Science Computer Review.

[7] Davanzo, G.; Medvet, E.; Bartoli, A (2011) Anomaly detection techniques for a web defacement monitoring service. J. Expert Syst. Appl. 38, 12521–12530.

[8] Banerjee, S., Swearingen, T., Shillair, R., Bauer, T. J., & Ross, A (2021) Using machine learning to examine cyberattack motivations on web defacement data. Social Science, Computer Review.

[9] Zhang, X., Tsang, A., Yue, W. T., & Chau, M (2015) The classification of hackers by knowledge exchange behaviors. Information Systems Frontiers, 17(6), 1239–1251.

[10] Maimon, David, Andrew Fukuda, Steve Hinton, Olga Babko-Malaya, and Rebecca Cathey (2017) On the Relevance of Social Media Platforms in Predicting the Volume and Patterns of Web Defacement Attacks. in 2017 IEEE International Conference on Big Data (Big Data), 4668-4673. IEEE.

[11] Andress, J., & Winterfeld, S (2013) Cyber warfare: Techniques, tactics and tools for security practitioners. Elsevier.

[12] Howell, C. J., Burruss, B. W., Maimon, D., & Sahani, S (2019) Website defacement and routine activities: Considering the importance of hackers' valuations of potential targets. Journal of Crime and Justice, 42, 536.

[13] Maggi, F., Balduzzi, M., Flores, R., Gu, L., & Ciancaglini, V (2018) Investigating web defacement campaigns at large. In Proceedings of the 2018 on asia conference on computer and communications security (pp. 443–456).

[14] Ooi, Kok Wei, Seung-Hyun Kim, Qiu-Hong Wang, and Kai Lung Hui (2012) Do Hackers Seek Variety? An Empirical Analysis of Website Defacements. AIS.

[15] Borgolte, K.; Kruegel, C.; Vigna, G. Meerkat (2015) Detecting Website Defacements through Image-based Object Recognition. In Proceedings of the 24th USENIX Security Symposium (USENIX Security), Washington, DC, USA, 12–14 August 2015.

[16] Yury Zhauniarovich, Issa Khalil, Ting Yu, Marc Dacier ((2018)) A Survey on Malicious Domains Detection through DNS Data Analysis, ACM Computing Survev, 1 (1), pp. 35.

[17] Rajesh Gupta, Sudeep Tanwar, Sudhanshu Tyagi, Neeraj Kumar (2020) Machine Learning Models for Secure Data Analytics: A taxonomy and threat model, Computer Communications, Volume 153, , pp. 406-440, https://doi.org/10.1016/j.comcom.2020.02.008.

[18] Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, Helge Janicke (2020) Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study, Journal of Information Security and Applications, Vol. 50, 2020, 102419, https://doi.org/10.1016/j.jisa.2019.102419.

[19] News, (2003)Web defacing contest stirs up conflict, Computer Fraud & Security, Vol. 2003, 8, 2003, pp.2-3, https://doi.org/10.1016/S1361-3723(03)08003-5.

[20] Samaneh Mahdavifar, Ali A. Ghorbani (2019) Application of deep learning to cybersecurity: A survey, Neurocomputing, Vol. 347, Pages 149-176, https://doi.org/10.1016/j.neucom.2019.02.056.Accessed December 23 2021

[21] Defacer.ID(2022) Available online: https://defacer.id (accessed on 10th April 2022).

[22] Burruss et al., (2021) Website defacer classification: a finite mixture model approach, Social Science Computer Review 1-13.

[23] Aslan, C¸. B., Li, S., C¸ elebi, F. V., & Tian, H (2020) The world of defacers: Looking through the lens of their activities on Twitter. IEEE Access, 8, 204132–204143.

[24] Fox, B. H., & Farrington, D. P (2015) An experimental evaluation on the utility of burglary profiles applied in active police investigations. Criminal Justice and Behavior, 42(2), 156–175.

[25] Braga, A. A., Turchan, B., Papachristos, A. V., & Hureau, D. M (2019) Hot spots policing of small geographic areas effects on crime. Campbell Systematic Reviews, 15(3). https://doi.org/10.1002/cl2.1046

[26] Bruinsma, G. J. N., & Johnson, S. D. (Eds.) (2018) The oxford handbook of environmental criminology. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190279707.001.0001.

[27] A. Moneva et al., (2022) Repeat victimization by website defacement: An empirical test of premises from an environmental criminology perspective, Computers in Human Behavior, 126 (2022), 106984.

[28] Gurjwar R.K, Sahu D.R., and Tomar D.S., (2013) An approach to reveal website defacement, International Journal of Computer Science and Information Security (IJCSIS), Vol. 11, No. 6, June 2013.

[29] Hoang, X.D (2018) A Website Defacement Detection Method based on Machine Learning. In Proceedings of the International Conference on Engineering Research and Applications (ICERA 2018), Thai-Nguyen, Vietnam, 1–2 December 2018.

[30] Banff Cyber Technologies (2022) Best Practices to Address the Issue of Web Defacement. Available online: https: //www.banffcyber.com/knowledge-base/articles/best-practices-address-issue-web-defacement/ (accessed on 26 April 2022).

[31] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi (2016) A review of data mining applications in crime, Statistical Analysis and Data Mining: 9e ASA Data Science Journal, vol. 9, no. 3, pp. 139–154,

[32] Y.-H. Tseng, Z.-P. Ho, K.-S. Yang, and C.-C. Chen (2012) Mining term networks from text collections for crime investigation, Expert Systems with Applications, vol. 39, no. 11, pp. 10082– 10090.

[33] A. Malathi and S. S. Baboo, (2011) An enhanced algorithm to predict a future crime using data mining, International Journal of Computer Applications, vol. 21, no. 1, 2011.

[34] Hoang X. D. and Nguyen N. T., (2019) Detecting website defacements based on machine learning techniques and attack signatures, Computers 2019, 8, 35.

[35] S.G.A. van de Weijer et al., (2021) Heterogeneity in trajectories of cybercriminals: a longitudinal analyses of web defacements, Computers in Human Behavior Reports, 4 (2021), 100113.

[36] Holt et al., (2021) Examining the characteristics that differentiate jihadi-associated cyberattacks using routine activities theory, Social Science Computer Review, pp.1-17.

[37] Berton, B., & Pawlak, P. (2015) Cyber jihadists and their web. European Union Institute for Security Studies.

[38] Central Intelligence Agency (2018) The CIA world factbook 2018. Skyhorse Publishing Inc.

[39] Heickero¨, R (2014) Cyber terrorism: Electronic jihad. Strategic Analysis, 38(4), 554–565.

[40] Carson, J. V., & Suppenbach, M. (2018) The Global Jihadist Movement: The most lethal ideology? Homicide Studies, 22(1), 8–44.

[41] Mee Lan Han et al., (2019) CBR-based decision support methodology for cybercrime investigation: focused on the data-driven website defacement analysis, Hindawi, Security and Communication Networks, Vol. 2019, (1901548), pp.21.

[42] Howell, Jordan C., George W. Burruss, David Maimon & Shradha Sahani (2019) Website defacement and routine activities: considering the importance of hackers' valuations of potential targets, Journal of Crime and Justice, 42, 2019, pp.536-550.

[43] Bernasco, W (2008) Them again?: Same-offender involvement in repeat and near repeat burglaries. European Journal of Criminology, 5(4), 411–431. https://doi.org/ 10.1177/1477370808095124

[44] E. ALPAYDIN (2010) Introduction to Machine Learning, 2nd ed., London: MIT press, 2010, pp. 67-97.

[45] V. N. Vapnik, (2000) The nature of statistical learning theory, 2nd ed., New York: Springer, 2000, pp. 112-235.

[46] V. CHERKASSKY and Y. MA, (2004) Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks, 17, 2004, pp.113–126.

[47] Holt, T. J., Leukfeldt, R., & van de Weijer, S (2020) An examination of motivation and routine activity theory to account for cyberattacks against Dutch websites. Criminal Justice and Behavior, 47(4), 487–505.

[48] Holt, T. J., Stonhouse, M., Freilich, J., & Chermak, S. M (2019) Examining ideologically motivated cyber-attacks performed by far-left groups. Terrorism and Political Violence, 33, 1–22.