International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 4565 - 4579

Associative Predictor with Optimal Rules for Predicting Phishing URL in Website via Oppositional Based Crow Optimization Model

Mr. M. Sathish Kumar[#1], Dr. B.Indrani[*2]
[#1]*Doctoral Research Scholar, Department of Computer Science, Directorate of Distance Education, Madurai Kamaraj University, Madurai, Tamilnadu, India.*
[2,*]*Assistant Professor, Department of Computer Science, Directorate of Distance Education, Madurai Kamaraj University, Madurai, Tamilnadu, India.*

Abstract

Phishing is an action of ambivalent web users to fake sites that can be utilized to take sensitive data from the web. In this paper, we proposed an optimal feature-based Association Rule (AR) to predict the phishing website. Improving the effectiveness of the work, we optimized the AR with optimal features by Opposition based Crow Search Optimization (OCS) model. The rules are deciphered to highlight the features that are increasingly pervasive in phishing URLs. Investigating the phishing Uniform Resource Locator (URL) website by two factors in predictor or classifier that is the greatest support value and confidence factor, in light of these two factors only determined the accuracy dimension of the proposed prediction process. When the rules are produced then the optimization model used to discover optimal AR for investigation, here closer to seventy-five rules is established. The optimal rules like URL length are low and several slashes are the minimum value, the accuracy is 0.922 acquired are deciphered to accentuate the features that are progressively pervasive in phishing URLs, at long last Associative Predictor (AP) for predicting the URL is Phishing or non-phishing or suspicious. From the implementation results, this proposed model analyzed by the confusion matrix, and this proposed work compared with existing predictors.

Keywords: *Phishing URL, Association Rules, Optimization, Features, Crow Search, Opposition, and predictor.*

1. Introduction

Phishing is a type of social crime called semantic attack and well perceived as online uniqueness fraud that bamboozles exploited people by guiding them to a false website seems authentic one [1]. The phishing website pages mainly have page designs, squares and text styles to impersonate genuine pages in a responsibility to impact web clients to get individual subtleties, for example, username and password [2]. The effect is the break of data security through the transaction of private information and the exploited people may eventually endure the loss of cash or different sorts [3]. Regardless of the wide field of phishing assault vectors, an ordinary reason for various vectors is the usage of connection misdirecting unfortunate casualties to phishing websites, this reality rouses us to manufacture a wide degree prediction framework by utilizing URLs features only [4, 5].

In the wake of having brand names used by Google to acquire authentic URLs is identified with these brand names. On account of entering this type of fake webpage [6], which is accepted to be the first website, computer clients can undoubtedly, give their sensitive data with no doubt [7]. The greater part of phishing prediction frameworks that work offline or online remains short to cover every one of the nuances expected to make the correct determination about potential phishing dangers by hailing them to genuine or malicious[8]. By ascertaining compelling estimations of each feature of the dataset, the optimal feature choice calculation chooses an optimal feature set. Among the procedure of optimal feature determination, much pointless and little effect features are interrupts [9]. To select captivating standards from the arrangement of conceivable principles, imperatives on various proportions of essentialness and

interest are utilized. Maintain and assurance is the proportions of a standard that repeat the quality and sureness of a standard, torecognizing the phishing identification [10].Various techniques are given by such huge numbers of analysts. Among them, Data Mining procedures are a standout amongst the most encouraging methods to recognize phishing movement [11,33].

Associative Classification is a grooming research method in data mining [32]. Therefore it is an interesting research topic that detecting phishing using associative classification [12]. Our work, we are used the association rule learning is a rule-based machine learning method for discovering relations between variables in the URL phishing database [13]. Comparing different Data Mining classification and association methods and techniques is also a goal of this investigation since there are only a few studies that compare different data mining techniques in predicting phishing websites [14,15]. The objective of the association rule (AR) Generation is used to discover associations among items in a set, by mining essential knowledge from the database [16].

Associative Classification is a preparing research strategy in data mining. Therefore it is a fascinating examination theme that distinguishing phishing utilizing associative classification [12]. In our work, we are utilized associative rule learning is a standard based AI technique for finding relations between factors in the URL phishing database [13]. Comparing various data mining groupings like association strategies and methods is likewise an objective of this examination since there are just a couple of concentrates that compares changed data mining procedures in predicting phishing websites [14,15]. The goal of the affiliation rule (AR) Generation is utilized to find the relationship among things in a set, by mining basic information from the database [16].

Contribution of this Research

In the existing system, numerous classifiers and feature selection processes are examined in various papers [17]. The prediction model means taking in the properties from different parts of URLs to assign the suspicion level of each portion [18, 19, and 20]. After the adjustment of the suspicion threshold of each bit, the framework would pick the most suspicious URL.

In this paper we are concentrated, on a real-world problem in site pages, that is phishing or noxious website pages (URLs) prediction with significant Optimal AR with an Associative classifier, and our proposed prediction approach compared with some ordinary classifiers and optimization. To get progressively successful prediction models, the extracted optimal features from our existing work Here will produce the AR with Oppositional based Crow Search (OCS) optimization system. The phishing URLs originated from PHISHTANK (dataset 1) and UCI data (dataset 2), a website utilized as a phishing URL store. Finally, the best rules-based cooperative classifier used to identify the given URL is malicious or begin. Theseresearch outcomes are validated by confusion matrix inform through some important metrics like precision, recall, F-Measure and accuracy values.

Organization of Paper

The rest of the paper is organized as follows: Section 1 discussed the background of phishing prediction models along with the prediction process and section 2 focused on the recent literature of phishing URL prediction papers. Section 3 discussed existing work then, the proposed Phishing URL prediction details explained in section 4. At finally conferred the implementation results and comparison of our work in section 5, at the end that is section 6 concluded our URL prediction with future scopes on this research are presented.

2. Literature Review

Build up a learning-based detection framework, training data must contain bunches of features, which are identified with phishing as well as legitimate website classes by Sahingoz, O. K et al. [17] (2018).

Enhancement techniques like Random Forest (RF) are used. Feature Extracted by Natural Language Processing (NLP) process. The main advantage of this article is Language independence, the Enormous Size of Phishing and Legitimate data, the Use of Feature-Rich classifiers, Finding of innovative Websites. The drawback as Checking these features need some additional time, accordingly, they may not be favored for ongoing identification. This examination prediction dependent on URL classification by utilizing data mining tool weka. The result of this examination is that the Random tree and Random forest both is the best classification calculation by Gupta, S., and Singhal, A. [18] (2017).

Parameters like Precision, Recall, F-Measure, accuracy, and sensitivity with pros is Random tree calculation is superior to anything RF as far as the time is taken to assemble model and time taken to test the model on training data. In 2016, S. Carolin Jeeva and Elijah Blessing Rajsingh [19] have proposed the Associative rule mining: apriori and predictive apriori to prediction process with Huge features that separate among legitimate and phishing URLs. Rules acquired are deciphered to underline the features that are increasingly pervasive in phishing URLs. Smart phishing detection & protection system and a combination of hybrid features by Association Rule Mining and Apriori Algorithm introduced by Tripathi, D et al. [20] (2018). Fraud detection and prevention in web advertising networks. Sample logs are increased the accuracy of the system is decreased due to the noise in the dataset by Adebowale, M. A et al. [21] (2019).

This model demonstrated the connection between some significant qualities like URL and Domain Identity, and Security and Encryption criteria in the last phishing detection rate by Aburrous, M et al. [22] (2010). sufficient to identify new phishing sites. An efficient security alert mechanism makes use of a classification model. well-known 11 best classification features attained 94.75% accuracy. Then valuable enough to rapid 97.50% security alerts as phishing URL correctly by LIEW, S et al. [23] (2019). Applying pattern recognition abilities of machine learning to phishing detection areas, we can accomplish critical execution upgrades by Pradeepthi, K. V, et al [24].The detection level of phishing URLs with a high classification rate.

In our current work, we have chosen the UCI database for the URL prediction model, so at first, a few features are separated from the database and then are fed to the optimization model. The motivation behind the optimization model gets optimal features of URL phishing prediction by Discrete Bat Algorithm (DBA), at last,the Decision tree classifier used to distinguish the given website is non-phishing, phishing or suspicious. The principle downside of existing work is minimum confusion matrix consequences of the URL prediction model, at long last for further improvements we will think about those optimal features from DBA procedure.

3. Prediction of Phishing URLs: A Novel approach

The most direct route for a phisher to cheat individuals is to make the phishing website page like their objective. Phisher endeavors to fraudulently procure approved users' confidential or sensitive accreditations by imitating electronic correspondences from any association, shopping website or any space. For achieving better feature selection results at a faster rate, optimization algorithms like Ant Colony Optimization (ACO), particle swarm optimization (PSO) algorithm, genetic algorithm (GA), simulated annealing, etc., are used [34-37]. Our proposed model depends on the perfect features phishing URL prediction process. For this proposed model, Opposition based Crow Search (OCS) utilized for optimal AR selection with AP, It's to anticipate the testing URL is phishing, non-phishing or suspicious. From this model analyzes the web get to log which tends to the exercises performed by the end customers, it is utilized for recognizing a fraud sequence of repeated web URL.

3.1 General Process and Information Extraction

Extricated features about the URL of the pages and composed feature matrix. Probably most of the features are in the category of texture and lexical features. Here over 30 features are utilized, to identify the expect phishing URLs, the optimization (Discrete Bat Algorithm) we need to locate the optimal features. Because of the heuristics, fourteen features were characterized and are exposed to association rule mining to successfully decide the legitimate and phished URL.

Data preprocessing is to get a reasonable organization from the gathered datasets due to conflictingly, deficiency, and certain practices lacking among the principle features of genuine data. A standout amongst the most significant qualities of this strategy, the extraction of such features isn't time expend and counteract the risk and inertness of the page loading.

3.2 DBA based best Features

By utilizing the technique of DBA the best feature is considered for the phishing URL prediction model, a suspicious URL is now and again loaded up with confounded numerical characters to decrease its understandability, so we in like manner determine the extent of numerical characters as a component. Here chose the optimal features are absolutely nine, it's clarified in underneath segment [27, 28 and 29].

URL Anchor (F1): Like the URL feature, however here the links inside the website page may point to a domain different from the domain typed in the URL address bar.

Request URL (F2): A page generally comprises content and a few articles, for example, pictures and recordings. Commonly, these items are loaded into the website page from a similar server of the site page. On the off chance that the articles are loaded from a domain other than the one composed in the URL address bar, the site page is possibly suspicious.

Server from Handler (F3): When the client submitted data; the page will exchange the data to a server with the goal that it can process it. Ordinarily, the data is processed from a similar area where the website page is being loaded.

URL Length (F4): Phishers shroud the suspicious piece of the URL to divert data's put together by clients or divert the uploaded page to a suspicious domain. Logically, there is no standard dependable length that differentiates between phishing URLs and non-phishing ones.

Prefix/Suffix (F5): Phishers endeavor to trick clients by reshaping the suspicious URL so it looks non-phishing. One procedure utilized is adding a prefix or addition to the non-phishing URL. In this manner, the client may not see any distinction.

IP Address (F6): Utilizing an IP address in the domain name of the URL is a pointer somebody is attempting to get to the individual data. This trap includes links that may start with an IP address that most organizations don't regularly utilize anymore. In the frequency analysis directed before, 20% of the information contains the "IP" address and every one of them is related tophishing sites.

Sub Domain (F7): Another procedure utilized by phishers to scam clients is by adding a subdomain to the URL so clients may trust they are managing a genuine site.

Special Characters (F8):The hostname in the URL of the legitimate and phished dataset is researched for understanding the nearness of special characters in both the data sets.

Number of Slashes (F9):Here considers the number of slashes in URLs as an element of identification of phishing and analyzes the number of slashes in legitimate and phishing URLs.

This model heuristics to remove optimal features from the URL and are exposed to association rule mining to decide the legitimate and phished URL. The features that recognize phishing websites from legitimate ones, yet they all fizzle ordering exact standards to extricate the features to association rules to characterize a website as either phishing or legitimate [30-31].

3.3 Association Rule (AR) Generation

This AR procedure to discover the phishing URLs by realized optimal feature's, Generally association rule mining working by confidence and support value, Moreover here two significant procedure are utilized that is frequent itemset and rule generation process. In our work we will upgrade the generated rules by utilizing the OCS [26] behavior utilized; this behavior effortlessly gets the optimal riles to distinguish the phishing URLs, this process shows in figure 1. Here AR for the generation of a frequent itemset in the middle of request, answer, time and reference.

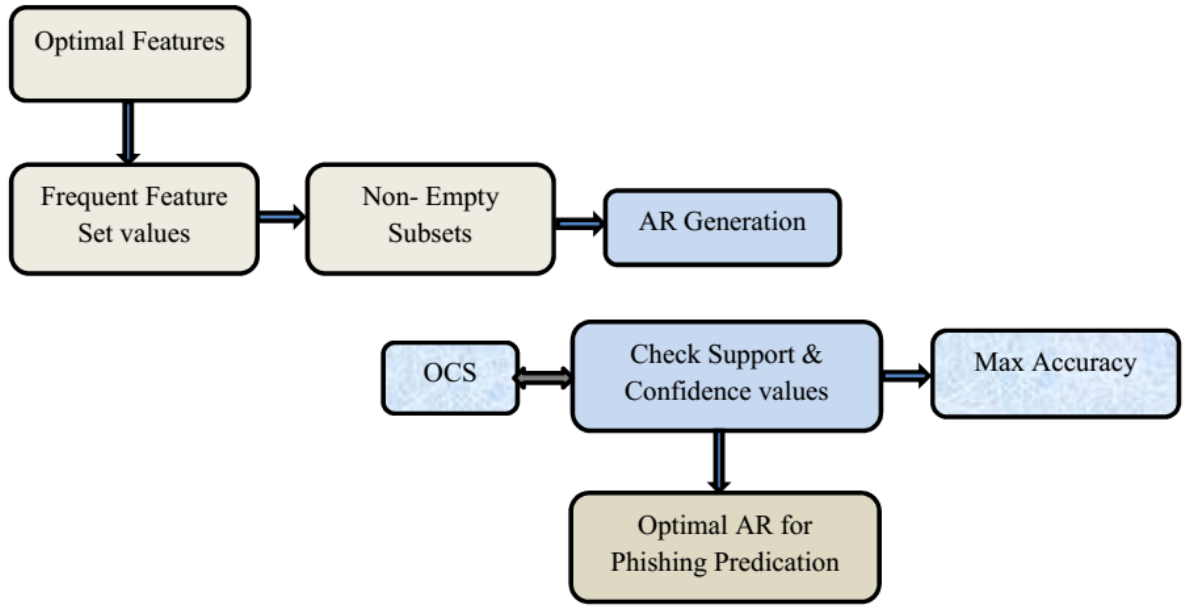


Fig 1: Model for Generating Association Rules

(i) Rule Generation

The rule for confidence alone however prescient apriori deliberates the confidence and support together in positioning the principles [25]. Confidence alone however prescient apriori thinks the confidence and support together in positioning the guidelines; Based on this just determined the accuracy.

(ii) Support value

Support and confidence are the proportions of the rule that reproduce the quality and assurance of a rule. Condition (1) exploits the accuracy of AR on hidden data. Apriori grades the rules as for confidence alone however prescient apriori ponders the confidence and support together in ranking the rules.

$$\text{Support(Feature)} = \text{Number of features(A \& B)} * \text{Appears together} \Big/ \text{N} \quad (1)$$

Ascertaining the exactness of the URL Phishing forecast, the support measure is significant. This measure characterizes the rate or division of records or passages in the list of features contains A & B to the absolute number of records or sections. A URL package along with support value beneath this minimum support value is typically removed.

(iii) Confidence Value

Confidence (Conf) is a measure that characterizes the rate or division of records or passages in the dataset that contains A and B to the total number of records or sections that contain just X. Utilizing the equation, this confidence value can be resolved.

$$Conf = Support(A \& B) \Big/ Supp(A) = Pro(A \cap B) \Big/ Pro(A) \quad (2)$$

To create association guidelines, our item set made at last just as the least confidence esteem is used.

3.4 Optimal AR Selection

Optimizing the guidelines of the AR-URL phishing prediction the generated optimal features based on the principles are produced, and after that, we will utilize OCS utilized. On the off chance that the desired value is more noteworthy than the threshold value, we expel the principles from the input dataset and the remaining item set or standards are stored in the new feature sets. When a rule is built then the majority of the training data occasions secured by it are expelled.

3.4.1 Crow Search (CS) Behavior

URL phishing prediction the optimal features have been gathered utilizing a web-based scripting device and the examination will fundamentally concentrate on prediction rate, the numbers of principles incited and the time taken to develop the prescient models. Crows have been known to watch other different birds, see where different birds conceal their food and take it once the owner leaves [26]. On the off chance that a crow has submitted burglary, it will play it safe, for example, moving concealing places to abstain from being a future injured individual. Here some fundamental condition is pursued optimal rule selection procedure.

- Features live in the form of a flock
- Features memorize the position of their hiding places.
- Features follow each other to do thievery.
- Features protect their caches from being pilfered by a probability.

(i) Objective function: Phishing Detection

Phishing detection, the target work is accuracy; it's calculated by utilizing the support and confidence esteems. The support value for every standard, if the support value is superior to threshold values the patters or evacuated generally store the database. It's communicated by

$$Optimal\ \ feature\ Accuracy = \{ Support, Conf \} \quad (3)$$

The strong rule produced by the predictive with accuracy level 100% has been considered for further investigation and different guidelines are disposed of. On the off chance that unfit to get the greatest accuracy implies the Crow that is feature solutions are refreshing by OCS behavior.

(ii) Oppositional Process of Crow Search (OCS)

Opposition process, the solution produced oppositional by real solutions, here three arrangements of factors are comprised of this objective, opposition learning parameter,and search space, of getting optimal standards in URL phishing detection. It's portrayed by

$$O_i^* = M_i + H_i - O_i \quad (4)$$

Here O_i^* represents the opposition solution that means new solutions for analysis, $M_i and H_i$ represent the existing search space as per the opposition based learning model. From this process, the actual and opposition feature set, represented by

$$F = \begin{pmatrix} F_1^1 F_2^1 .....F_d^1 \\ F_1^2 F_2^2 .....F_d^2 \\ ..... \\ F_1^N F_2^N .....F_d^N \end{pmatrix} OF = \begin{pmatrix} O_1^1 F_2^1 .....O_d^1 \\ O_1^2 F_2^2 .....O_d^2 \\ ..... \\ O_1^N F_2^N .....O_d^N \end{pmatrix} \quad (5)$$

Where $F_i^{\backslash}$ shows the feature which is taken in this study as F_1, F_2, F_3, F_d^N ; OF Represents the opposition feature

(iii) New Feature updating procedure

The memory of each crow is initialized. Since at the initial iteration, the crows have no encounters, it is expected that they have shrouded their foods at their underlying positions, the new position having two significant conditions that are depicted in this segment and this crow behavior shows in figure 2.

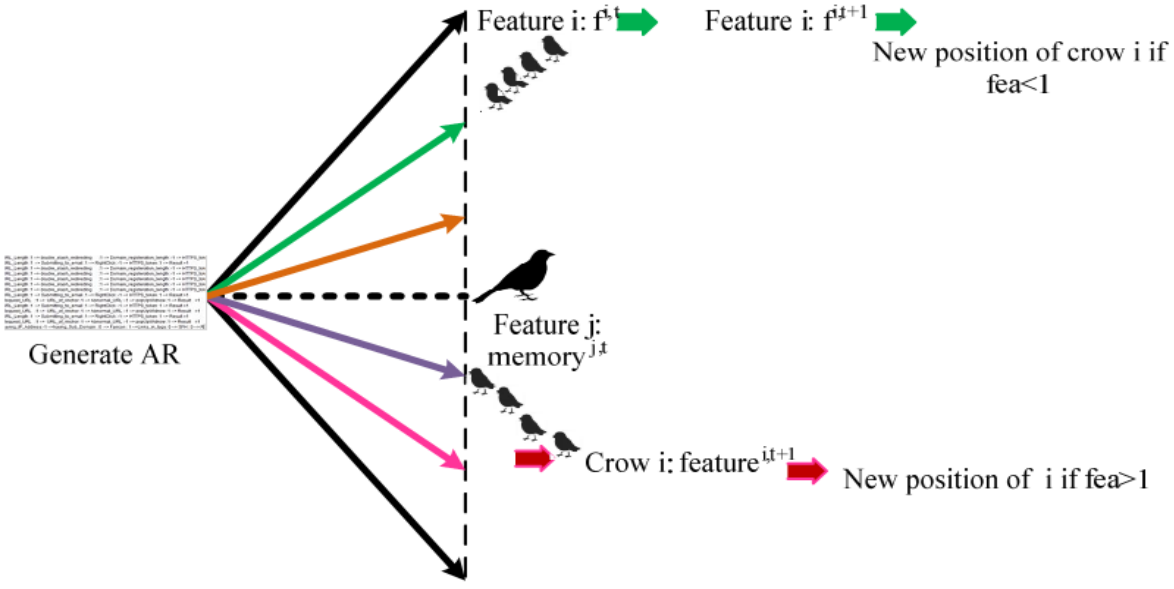


Fig 2: Crow behavior

Criteria 1: Update the Feature Memory

Update the new memory of the feature selection process, by the fitness calculation the new position of a crow is superior to fitness function value of the memorized position; the crow refreshes its memory by the new position.

$$M^{i,iter+1} = \begin{cases} Obj^{i,iter} & f(obj^{j,iter+1}) > obj(m^{j,iter}) \\ m^{i,iter} & otherwise \end{cases} \quad (6)$$

Here m^i represents the memory of the crow; that isold memory value of search solutions and the optimal objective (Obj) feature value. At that point, the best position that is memorized by the members of the crow flock is reported as the optimum solution.

Criteria 2: Generate New Position

The position of the i^{th} crow individual at the t^{th} iteration in the search space is represented by $\text{objective(feature)}$ which is, in fact, a realistic array of the decision variables. For example, crow "feature" likewise, its position and memory.

$$F^{i,iter+1} = \begin{cases} F^{i,iter} + randM_i \times FL^{i,iter} \times (m^{j,iter} - F_N{}^{i,iter}) if \ \ rand_j \geq A.P^{j,iter} \\ F_R \qquad otherwise \end{cases} \quad (7)$$

From equation (7) $rand$ indicates a random number of crows and the range is between [0, 1], $FL^{i,iter}$ represents flight length of a crow, $F_N{}^{iter}$ symbolizes the position of a crow, $m^{j,iter}$ denotes the memory location of j^{th} crow and $A.P^{j,iter}$ indicates the awareness probability of crow j at iteration.

(iv) At End

From this OCS system, the AR is enhanced with most maximum accuracy (support and confidence values). At the point, when the termination criterion is met, the best position of the memory as far as the objective function value, Moreover this refreshing system is rehashed to improve optimal qualities.

3.4.2 OCS based Optimal Rules for Phishing Model

The best unique rules produced from OCS and prescient apriori for different size of the phishing input dataset is considered, shows in figure 3. We have around a hundred standards in the optimization process, in the wake of applying the OCS approach [26] the best guidelines closer to seventy-five, here referenced some example optimal rules of phishing URL prediction process.

R1: If (Request URL is genuine) & (URL length is large IP address is genuine), the accuracy is 0.82222.
R2: If (URL anchor is malicious) & (URL length is short) & (subdomain is high) the accuracy is 0.784.
R3: If (the special characters) in subdomain the confidence is high and accuracy is 0.922
R4: If (Number of slashes is the maximum value) & (the URL length is low) and (unique code of URL as a minimum), the accuracy is 0.86.
R5: If (Server from handles is malicious) & (the request URL is unique) & (the length of URL is low) the accuracy is 0.755
R6: If (layer security is high) & (URL is genuine) & (IP address is malicious) the accuracy is 0.697
R7: If Server from the handler is genuine and length is medium, the prefix-suffix of URL is malicious the accuracy is b0.69
R8: If (URL length is low) & (Number of slashes is the minimum value), the accuracy is 0.922.
R9: If (a URL subdomain is malicious) & (other features are good), the accuracy is 62.85.
R10: If (URL anchor is genuine) & (URL is genuine) & (Length of URL is medium) the accuracy is 0.962.

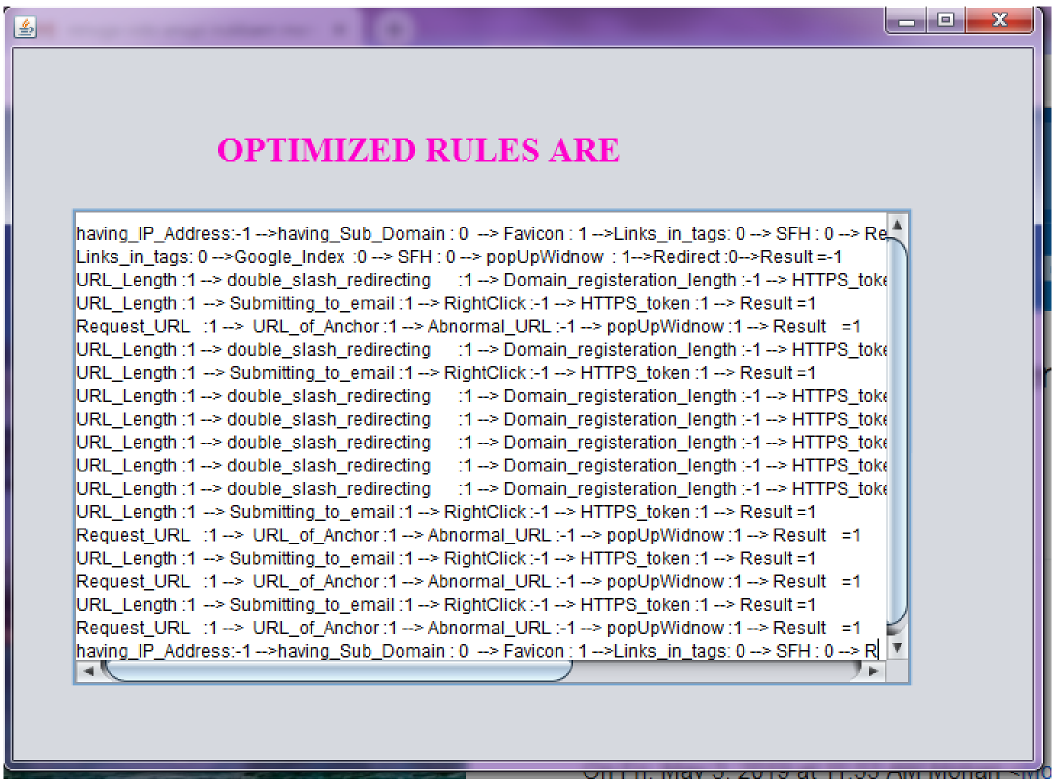


Fig 3: Optimal AR

In light of previously mentioned optimal rules, the phishing URL is anticipated. Uncover the proficiency and viability of methodologies that produce models with rules in fighting to phish. Rule mining is utilized to investigate the concealed connection between the attributes. This model has been utilized for recognizing them as often as frequently happening features in phishing URLs.

3.5 Optimal Rule-based Associative Prediction

Classification is forecast that dependent on AR, which both mirrors the application qualities of expectation. The optimal rules are Attributes that can be discrete and can be persistent, for a consistent attribute value, discrete. Mined every single optimal rule to anticipate the given URL is non-phishing, phishing or suspicious, this entire process shows in figure 4.

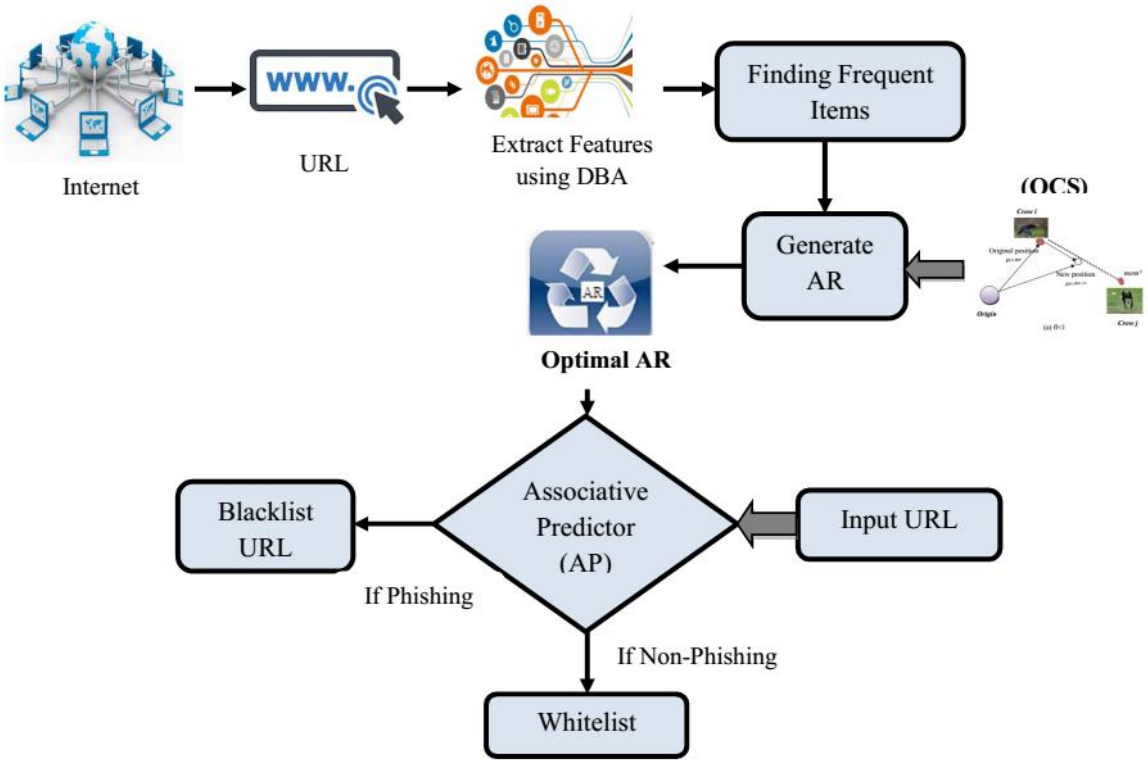


Fig 4: Overall View of Proposed URL phishing Detection

The probability is to appraise the predicted error rate is to utilize another, extra set of test tests on the off chance that they are accessible, or to utilize the cross-approval systems. This AP mines unique guidelines the principles mined by associative are considered for further procedure. The rule generated by the training dataset is checked for an example data set.

4. Results Analysis

Phishing URL forecast models implemented in the stage of Java language with Net beans 8.0.1, and 4GB RAM i3 processor in Windows 10 (64-bit). The phishing URLs originated from PHISHTANK (dataset 1) and UCI data (dataset 2), a website utilized as a phishing URL store. An extensiveconfusion matrix is a Precision, recall, F-measure for predictor and Optimal guidelines with, accuracy (support and confidence) values. Our proposed model contrasted with the other regular method resembles Decision Tree (DT), DT with Discrete Bat Algorithm (DBA), AR.

In the phishing URL forecast process, the confusion matrix is discussed in table 1. Here two datasets are utilized, so we are discussed True positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) with execution measures. This showed that our proposed URL phishing detection has the greatest TP and FP value. Proposed associative classifier against Phishing URLs, and the rate of right classification with 94% accuracy and non-phishing class additionally dissected.

Table 1: Confusion Matrix results for Proposed Model

Dataset	TP	TN	FP	FN	Precision	Recall	F-Measure	Accuracy
PHISHTANK	2568	1524	1058	306	0.94	0.938	0.924	0.9655
UCI Archive	567	324	425	37	0.925	0.89	0.94	0.936

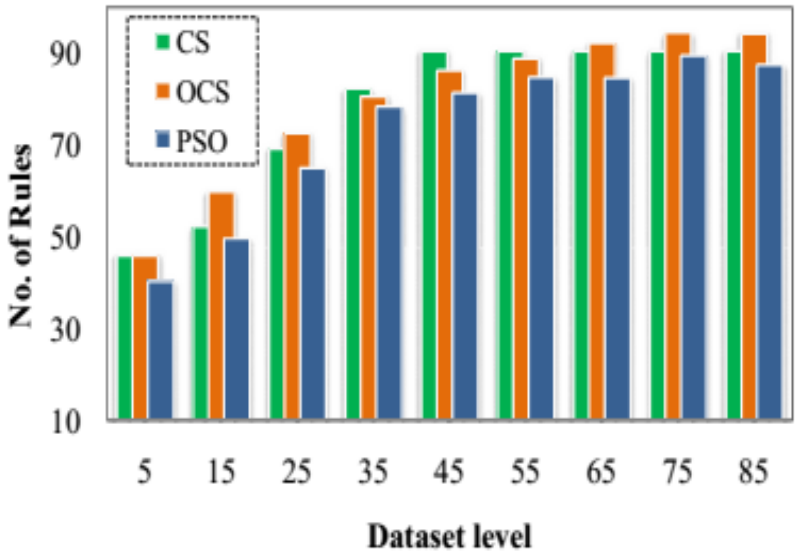


Fig 5: Effective optimal rule selection

The graphical portrayal of optimal rules with the accuracy level and dataset Vs optimal AR appears in Figures 5 and 6. This URL forecast, mines optimal rules the standards mined by OCS with the greatest accuracy.The optimal rules in CS, OCS, and Particle Swarm Optimization (PSO) are appeared in figure 5, as far as changing the dataset level that is optimal feature is 75 in OCS and the accuracy is 94.22%, Moreover, the CS getting some feature that is 46 out of hundred rates with 90.44% accuracy. The test data set contains dots in the hostname of the URL. This test exhibits that our model offers more accuracy than these strategies in identifying phishing websites. The optimal AR by OCS time expended for the indicator to distinguish single URL as phishing or not is researched by processing the time taken from extricating the features to test and giving the last results.

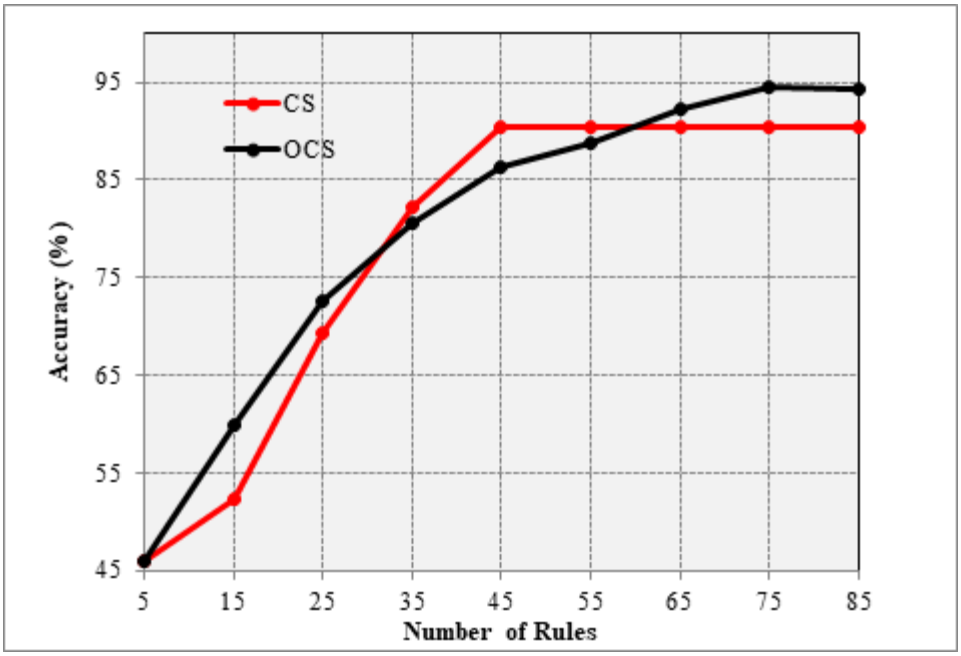


Fig 6: Accuracy analysis for Optimal AR

Table 2: Optimal Features and AR results

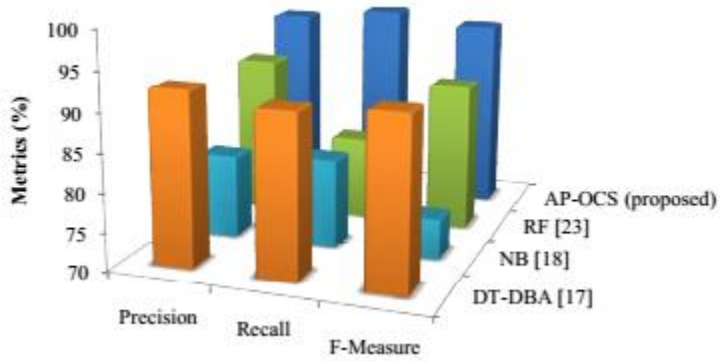
No. of optimal Features			*No. of optimal AR*				
Technique		***Accuracy***	***Technique***		***Support***	***Confidence***	***Accuracy***
BA	15	84.55	CS	46	0.148	0.29	90.44
DBA	9	92.25	OCS	75	0.22	0.32	94.42

Table 3: Optimal features with optimal AR model, Accuracy analysis

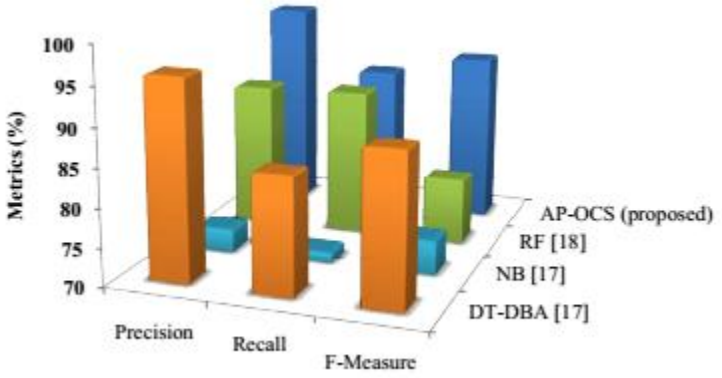
Methods	***PHISHTANK***				***UCI Archive***			
	Precision	***Recall***	***F-Measure***	***Similarity Index***	***Precision***	***Recall***	***F-Measure***	***Similarity Index***
DT-DBA	96	85.22	89.58	89.55	92.55	91.22	92.22	93.14
AP	76.89	78.48	76.55	69.55	82.78	86.55	79.22	84.55
AP-CS	86.22	89.51	76.55	82.52	89.58	75.85	86.22	82.58
AP-OCS (Proposed)	97.89	89.55	92.44	93.22	94.85	96.22	94.88	92.52

Table 2 and 3 demonstrates the diverse predictor like DT-DBA, AR, Associative Predictor (AP)-Crow search (CS) and our proposed methodology AP-OCS with all execution measures. Tentatively, the AP framework can create the features vector that is 75 with a high accuracy level. The exactness of optimal feature selection stands 92.25 in DBA, it's contrasted with BA the deviation is almost 10 to 12%, its minimum deviation only. For choosing the optimal rules in AR generation the support and confidence rules are significant, here the novel methodology has support and confidence (0.22, 0.32) by the accuracy as 94.42%. Investigations demonstrated that as confidence expands, the accuracy would build first at that point decline. This might be because when confidence is excessively low; numerous pointless rules will

be created, which will exasperate the classifier. Furthermore, a few assessments likewise bring up that with similar corpora and feature vectors, the associative algorithm additionally performs better classification results. At last table 3 demonstrate the forecast outcomes with Similarity index values, the most extreme value is 95.52in AP-OCS, its better, contrasted with others.



(a)



(b)

Fig 7: Comparative analysis of URL phishing Detection: (a) dataset I- PHISHTANK (b) dataset II- UCI

Comparative analysis of various predictors appears in figure 7 (a) and (b), by the confusion matrix measures precision, recall, and F- measure values. From the figure, precision means the extent of existent dynamic cases that were anticipated effectively as reality. Precision is really what machine learning. For dataset 1 (PHISHTANK) the precision is 94.85, its greatest contrasted with others. The general normal training exactness result for optimal features was recorded as in the experiment utilizing approval with normal time. The best relative execution result accomplished in training. For Dataset 2 (UCI) better URL phishing prediction results in figure 6 (b), our outcome for phishing detection recorded a little improvement of 10.5% contrasted with different predictors. If any methodology utilizes just a URL based feature, it yields a high false-negative rate, which wrongly decided the phishing websites.

5. Conclusion

In this paper, the optimal features based ideal AR of the Phishing URL are dissected with AP. The outcomes got from rule mining have highlighted the valuable features accessible in the phished URL. The list of features of our phishing detection approach altogether relies upon the URL and source code of the

website by OCS by optimal features. The execution results demonstrate the most extreme prediction rate with precision, recall and F measure in both datasets that is 96.37, 92.885 and 93.66. Also our proposed ideal AR process additionally the better support and confidence value of URL detection with the greatest rate is 94.22%. The explanation behind gettingthegreatest accuracy in OCS is the highest confidence value that is chosen by the client according to his experience and judgment. Notwithstanding, trust that utilizing optimal features with optimal AR, for example, those introduced here can fundamentally help with identifying this class of phishing website. In future work, we will utilizea deep learning model with the number of developing technology that could incredibly help with phishing detection utilizing some frequent pattern mining models.

Reference

1. Jain, A.K. and Gupta, B.B., 2019. A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, 10(5), pp.2015-2028.
2. Cheng, H., Yan, X., Han, J. and Philip, S.Y., 2008, April. Direct discriminative pattern mining for effective classification. In 2008 IEEE 24th International Conference on Data Engineering (pp. 169-178). IEEE.
3. Aburrous, M., Hossain, M.A., Dahal, K. and Thabtah, F., 2010. Intelligent phishing detection system for e-banking using fuzzy data mining. Expert systems with applications, 37(12), pp.7913-7921.
4. Khonji, M., Jones, A. and Iraqi, Y., 2011, February. A novel Phishing classification based on URL features. In 2011 IEEE GCC conference and exhibition (GCC), pp. 221-224.
5. Abutair, H.Y. and Belghith, A., 2017. A multi-agent case-based reasoning architecture for phishing detection. Procedia Computer Science, 110, pp.492-497.
6. Li, Y., Yang, Z., Chen, X., Yuan, H. and Liu, W., 2019. A stacking model using URL and HTML features for phishing webpage detection. Future Generation Computer Systems, 94, pp.27-39.
7. Chiew, K.L., Tan, C.L., Wong, K., Yong, K.S. and Tiong, W.K., 2019. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences, 484, pp.153-166.
8. Ding, Y., Luktarhan, N., Li, K. and Slamu, W., 2019. A Keyword-based Combination Approach for Detecting Phishing Webpages. Computers & Security.
9. Basnet, R.B. and Sung, A.H., 2012, December. Mining web to detect phishing URLs. In 2012 11th International Conference on Machine Learning and Applications, Vol. 1, pp. 568-573.
10. Inuwa-Dutse, I., Liptrott, M. and Korkontzelos, I., 2018. Detection of spam-posting accounts on Twitter. Neurocomputing, 315, pp.496-511.
11. Aburrous, M., Hossain, M.A., Dahal, K. and Thabtah, F., 2010. Intelligent phishing detection system for e-banking using fuzzy data mining. Expert systems with applications, 37(12), pp.7913-7921.
12. Sananse, B.E. and Sarode, T.K., 2015. Phishing URL detection: a machine learning and web mining-based approach. International Journal of Computer Applications, Vol. 123, No.13, pp.1-8.
13. Feroz, M.N. and Mengel, S., 2015, June. Phishing URL detection using URL ranking. In *2015* ieee international congress on big data, IEEE, pp. 635-638.
14. Tahir, M.A.U.H., Asghar, S., Zafar, A. and Gillani, S., 2016, December. A hybrid model to detect Phishing-Sites using supervised learning algorithms. In 2016 International Conference on Computational Science and Computational Intelligence (CSCI),pp. 1126-1133.
15. Agrawal, P. and Mangal, D., 2013,A Novel Approach for Phishing URLs Detection.Journal of Science and Research,pp.1117-1122.
16. Daeef, A.Y., Ahmad, R.B. And Yacob, Y., 2016. Lexical Based Method For Phishing Urls Detection. Journal of Theoretical & Applied Information Technology, Vol. 88, No.3.pp.1-10.
17. Sahingoz, O.K., Buber, E., Demir, O. and Diri, B., 2019. Machine learning based phishing detection from URLs. Expert Systems with Applications, Vol. 117, pp.345-357.
18. Gupta, S. and Singhal, A., 2018. Dynamic Classification Mining Techniques for Predicting Phishing URL. In Soft Computing: Theories and Applications, pp. 537-546. Springer, Singapore.

19. Jeeva, S.C. and Rajsingh, E.B., 2016. Intelligent phishing url detection using association rule mining. Human-centric Computing and Information Sciences, Vol.6,No.1, pp.1-10.
20. Tripathi, D., Nigam, B. and Edla, D.R., 2017. A novel web fraud detection technique using association rule mining. Procedia computer science, 115, pp.274-281.
21. Adebowale, M.A., Lwin, K.T., Sánchez, E. and Hossain, M.A., 2018. Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Systems with Applications.
22. Aburrous, M., Hossain, M.A., Dahal, K. and Thabtah, F., 2010, April. Predicting phishing websites using classification mining techniques with experimental case studies. In 2010 Seventh International Conference on Information Technology: New Generations pp. 176-181.
23. Liew, S.W., Sani, N.F.M., Abdullah, M.T., Yaakob, R. and Sharum, M.Y., 2019. An effective security alert mechanism for real-time phishing tweet detection on Twitter. Computers & Security, 83, pp.201-207.
24. Pradeepthi, K.V. and Kannan, A., 2014, December. Performance study of classification techniques for phishing url detection. In 2014 Sixth International Conference on Advanced Computing (ICoAC) (pp. 135-139). IEEE.
25. Agrawal, A., Thakar, U., Soni, R. and Chaurasia, B.K., 2011, September. Efficiency enhanced association rule mining technique. In International Conference on Parallel Distributed Computing Technologies and Applications (pp. 375-384). Springer, Berlin, Heidelberg.
26. Askarzadeh, A., 2016. A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Computers & Structures, 169, pp.1-12.
27. Zouina, M. and Outtaj, B., 2017. A novel lightweight URL phishing detection system using SVM and similarity index. Human-centric Computing and Information Sciences, 7(1), p.17.
28. Ramanathan, V. and Wechsler, H., 2012. phishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training. EURASIP Journal on Information Security, Vol.2012, No. 1, pp.1-10.
29. Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A. and Liang, Z., 2019. Phishing page detection via learning classifiers from page layout feature. EURASIP Journal on Wireless Communications and Networking, 2019, Vol.1, pp.1-43.
30. Iuga, C., Nurse, J.R. and Erola, A., 2016. Baiting the hook: factors impacting susceptibility to phishing attacks. Human-centric Computing and Information Sciences, Vol.6,No.1, pp.1-8.
31. Jain, A.K. and Gupta, B.B., 2016. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security, pp.1-9.
32. Zenkert, J., Klahold, A., & Fathi, M. (2018). Knowledge discovery in multidimensional knowledge representation framework. Iran Journal of Computer Science, 1(4), 199-216.
33. Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M. R., Razeghi, F., Mazaheri, S., Zamani-Harghalani, Y., & Moosavi-Tayebi, R. (2020). New hybrid method for feature selection and classification using meta-heuristic algorithm in credit risk assessment. Iran Journal of Computer Science, 3(1), 1-11.
34. Joshua Samuel Raj, S. Jeya Shobana, Irina Valeryevna Pustokhina, Denis Alexandrovich Pustokhin, Deepak Gupta, K. Shankar, "Optimal Feature Selection based Medical Image Classification using Deep Learning Model in Internet of Medical Things", IEEE Access, March 2020. In press: https://doi.org/10.1109/ACCESS.2020.2981337
35. Sankhwar, S., Gupta, D., Ramya, K. C., Rani, S. S., Shankar, K., & Lakshmanaprabu, S. K. (2020). Improved grey wolf optimization-based feature subset selection with fuzzy neural classifier for financial crisis prediction. Soft Computing, 24(1), 101-110.
36. Elhoseny, M., Shankar, K., & Uthayakumar, J. Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease, Nature Scientific Reports, July 2019. Press. DOI: https://doi. org/10.1038/s41598-019-46074-2.
37. Elhoseny, M., Bian, G. B., Lakshmanaprabu, S. K., Shankar, K., Singh, A. K., & Wu, W. (2019). Effective features to classify ovarian cancer data in internet of medical things. Computer Networks, 159, 147-156.