Zhongguo Kuangye Daxue Xuebao
Journal of China University of Mining and Technology

2024 | Vol 29 | Issue 2 | Page 45-52
Journal Homepage: https://zkdx.ch/
DOI: 10.1654/zkdx.2024.29.2-4

Phishing Detection Using Hybrid Machine learning Techniques

Rasha Gaffer M. Helali*
Assistant professor, Faculty of Computing and Information Technology, University of Bisha, Saudi Arabia
*Corresponding author

Abstract

Cyber security has become a crucial component of the new digital age with more than 820 million users of internet in year 2023 and social media users are expected to reach 82.3% from the total number of internet users by 2024. According to these figures, security systems are required to shield the public from phishing scams, which have a negative impact not only on financial resources but also on people's mental health by making them fearful to use the internet or surf. This drives efforts to find effective solutions for the issue. The swift alterations in phishing attack patterns necessitate constant improvement of existing phishing detection systems in order to effectively counter new and upcoming phishing attempts.

This research aims to identify common characteristics displayed by phishing websites and create a model to identify them. The dataset was used to train a number of models, including the Random Forest Classifier, Artificial Neural Networks, and Principal component Analysis. Feature selection and clustering technique were also integrated to detect unknown attacks. The dataset was collected from Kaggle and contains information of 549,346 entries. RF attained the highest accuracy of 94%.

Keywords

Cyber security, Phishing attacks, Machine learning, Neural Network, PCA

1. Introduction

With the significant growth of internet usage, people increasingly share their personal information online. As a result, an enormous amount of personal information and financial transactions become vulnerable to cybercriminals [1].Phishing is a type of digital assault, which adversely affects people where the client is coordinated to counterfeit sites and hoodwinked to screen their touchy and private data which integrates watchwords of records, monetary data, ATM pin-card data, etc. This growth can lead to the theft of users' private information for malicious purposes. Phishing is one technique that can cause users to be redirected to sites with malicious content and steal all of their information [2].

PHISHING is understood to be a criminal attack on obtaining personal information, such as passwords and payment card information, through web pages or e-mails [3]. Webpage creators can easily make fake pages which are virtually identical to the original ones, so people can easily fall victim to them. An alarming sign is the availability of guides about how to make fake web pages directly on the internet, e.g. [4]. At the same time, online payments are increasingly being used and many other activities are being moved to the internet. For example, the transaction value of digital payments is expected to show a growth rate of 17.0 percent between 2020 and 2024 [4]. The number of internet users has grown 1,187 percent since 2000 and there are 4,648,228,067 internet users at this moment, which is 59.6 percent of the world population [4]. According to Proof point annual report of phishing attacks, Phishing attacks on the rise on 2022 comparing to 2021 [5].

The report also stated that there is a gap in awareness and training to face phishing attack in organizations. Fewer than 60% of organizations deliver organization-wide training. Nearly 30% focus their efforts strictly on specific departments and roles, and another 15% are only concerned about specific individuals. Fewer than 50% of organizations formally cover email-based phishing in their training programmers, and just 43% cover ransom ware. In comparison, more than 80% of organizations experienced at least one successful phishing attack in 2021, and nearly 70% dealt with at least one ransom ware infection. (See more on topics later in this section [5]. The following diagram shows the percentage of phishing attack in 2021 according to APWG annual report [5].

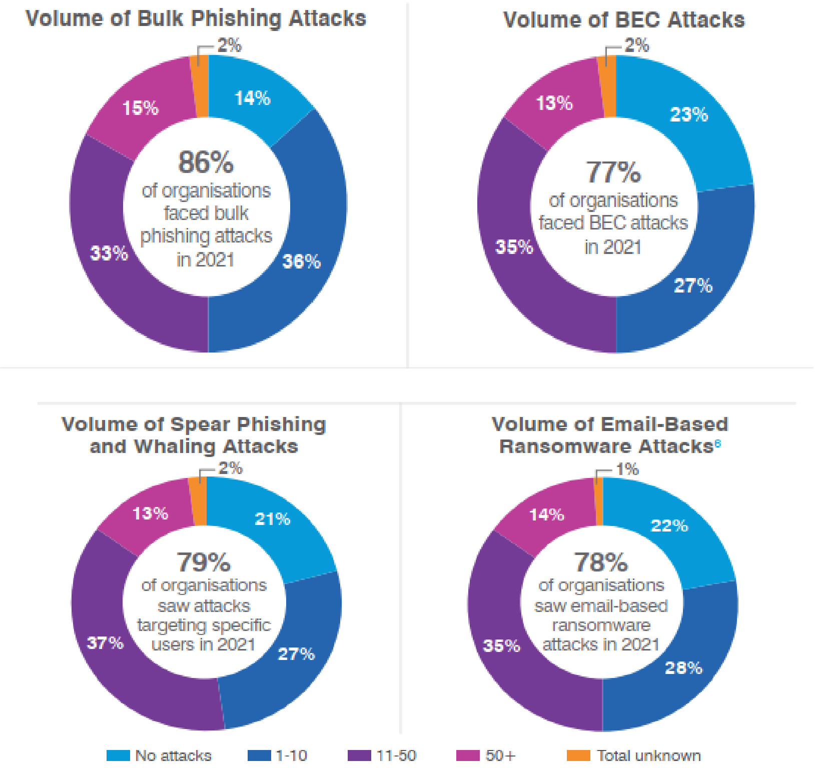


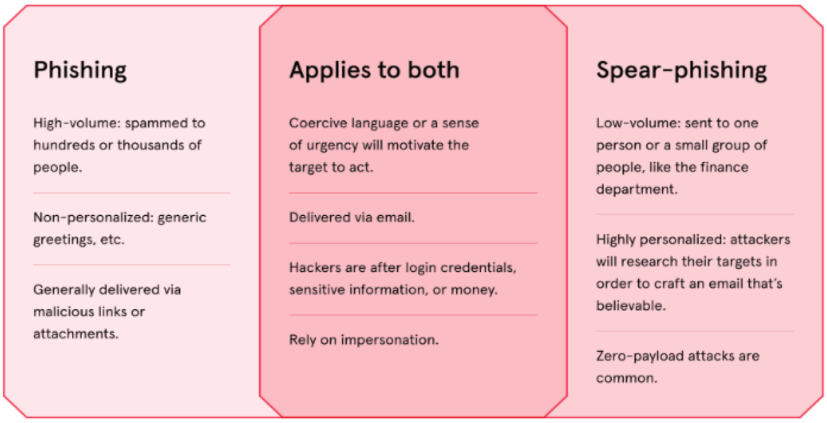
Fig. 1 Results from APWG annual report [5]

2. Method for Phishing

The term "Phishing" which was also called brand spoofing, was first time in 1996 when the hackers created randomized credit card numbers using an algorithm to steal users' passwords from America Online (AOL) [1]. The following is phishing techniques:

2.1 Spear Phishing

Mass emails are sent to as many people as possible, spear phishing is a much more targeted attack in which the hacker knows which specific individual or organization they are after. They do research on the target in order to make the attack more personalized and increase the likelihood of the target falling into their trap.



Phishing

High-volume: spammed to hundreds or thousands of people.

Non-personalized: generic greetings, etc.

Generally delivered via malicious links or attachments.

Applies to both

Coercive language or a sense of urgency will motivate the target to act.

Delivered via email.

Hackers are after login credentials, sensitive information, or money.

Rely on impersonation.

Spear-phishing

Low-volume: sent to one person or a small group of people, like the finance department.

Highly personalized: attackers will research their targets in order to craft an email that's believable.

Zero-payload attacks are common.

Fig. 2 Differences between phishing and spare phishing [6]

2.2 Session Hijacking

The phisher exploits the web session control mechanism to steal information from the user. The phisher can use a sniffer to intercept relevant information so that he or she can access the Web server illegally.

2.3 Email/Spam

The same email is sent to millions of users with a request to fill in personal details.

2.4 Content Injection

Is the technique where the phisher changes a part of the content on the page of a reliable website?

2.5 Vishing (Voice Phishing)

In phone phishing, the phisher makes phone calls to the user and asks the user to dial a number. The purpose is to get personal information of the bank account through the phone. Phone phishing is mostly done with a fake caller ID.

2.6 Smishing (SMS Phishing)

Phishing conducted via Short Message Service (SMS), a telephone-based text messaging service. A smishing text, for example, attempts to entice a victim into revealing personal information via a link that leads to a phishing website.

2.7 Malware

Phishing scams involving malware require it to be run on the user's computer. The malware is usually attached to the email sent to the user by the phishers. Once you click on the link, the malware will start functioning. Sometimes, the malware may also be attached to downloadable files.

2.8 Ransom ware

Ransom ware denies access to a device or files until a ransom has been paid. Ransom ware for PC's is malware that gets installed on a user's workstation using a social engineering attack where the user gets tricked in clicking on a link, opening an attachment, or clicking on maladvertising.

The remainder of this paper is organized as follows: Section 2 describes the related work and provides an overview of existing work. Section 3 describes the methodology used in this study in addition to the proposed method. The evaluation metrics are discussed in Section 4. Finally, conclusion is presented.

3. Related work

Much research moves to use machine learning techniques for detecting phishing attacks effectively and few tried to mix between different techniques with machine learning to improve detection accuracy. The following section lists some of researches done in the field.

Fatima Salahdine et. al. Proposed a phishing attack detection technique based on machine learning. They collected and analyzed more than 4000 phishing emails targeting the email service of the North Dakota University [3]. Sundara p. et. al. Examined the association of Machine Literacy routes in identifying phishing assaults and records their advantages and drawbacks [4].

Ala M. et. al. Proposed a detection model using machine learning techniques by splitting the dataset to train the detection model and validating the results using the test data, to capture inherent characteristics of the email text, and other features to be classified as phishing or non-phishing using three different data sets, After making a comparison between them, they obtained that the most number of features used the most accurate and efficient results achieved. The best ML algorithm accuracy was 0.88, 1.00, and 0.97 consecutively for boosted decision tree on the applied data sets [7].

Trivikram M. and NirNissima presented the first fully automated malicious email detection framework using deep ensemble learning to analyze all email segments (body, header, and attachments); this eliminates the need for human expert intervention for feature engineering the proposed framework's results surpass state-of-the-art malicious email detection methods, including human expert feature-based machine learning models by a TPR of 5% [8].

Ameya Chawla analyzed some common attributes shown by phishing websites and develops a model to detect these websites. Various models where trained on the dataset like Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K Nearest Neighbors, Artificial Neural Networks and Max Vote Classifier of Random Forest, Artificial Neural Networks and K Nearest Neighbors. Highest accuracy was achieved by Max Vote Classifier of Random Forest (max depth 16), Decision Tree (max depth 18) and Artificial Neural Network of 97.73% [9]. Wend-B et.al focused on determined the importance of the features by using cross validation as well as the correlation between features. They found that the Logistic Regression classifier had better accuracy for the best accuracy [2]. In [10] Mohammad N. et. al presented a study about phishing attack using machine learning techniques. They developed a model to detect the phishing attacks using machine learning (ML) algorithms like random forest (RF), decision tree (DT) and PCA.

Authors [11] developed a detection approach for classifying malicious and normal webpages. The outcome of this study indicated that the value of true positive was higher rather than the false positive rate. In other study [12], authors proposed a Convolutional Neural Network (CNN) to detect a phishing URL. In this study, researchers employed a sequential pattern to capture the URL information. It achieved an accuracy of 98.58%, 95.46%, and 95.22%, respectively on benchmark datasets.

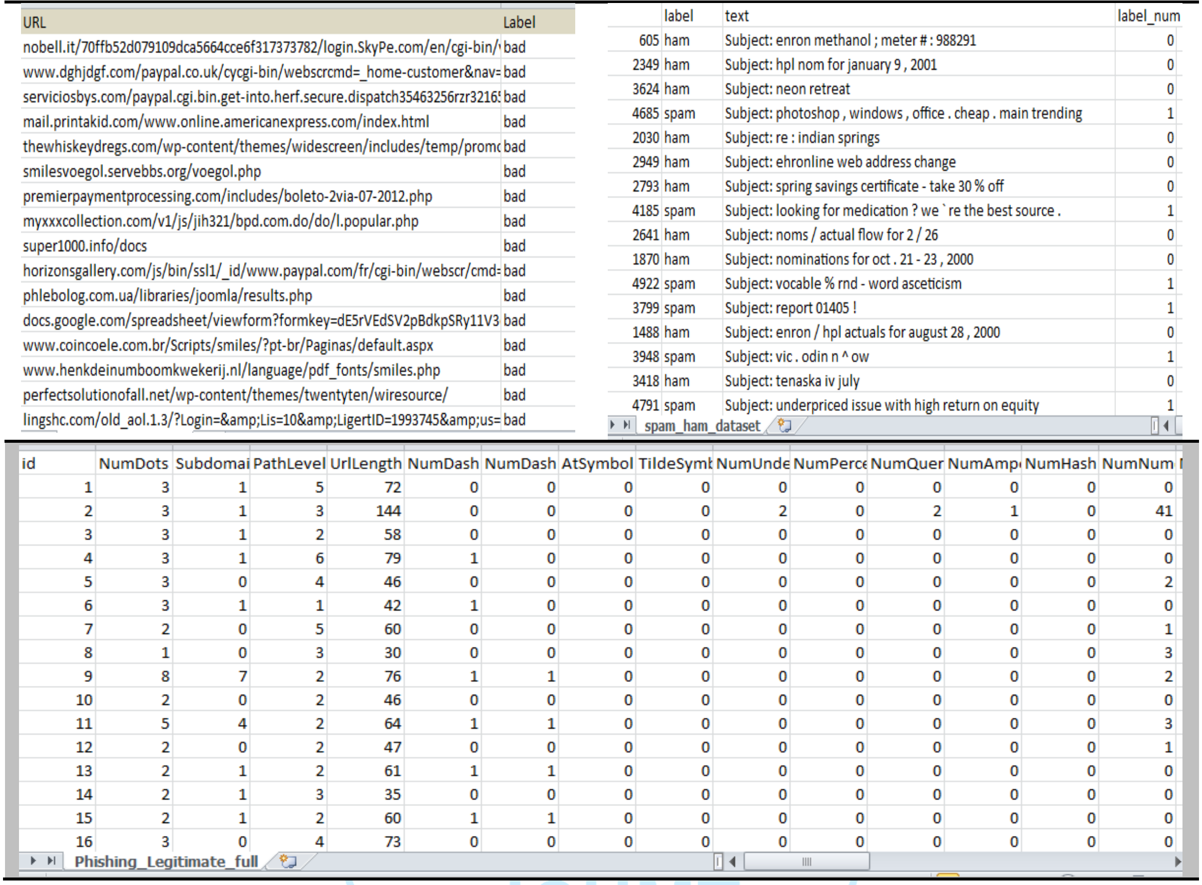
Recently, R. Jayaraj et. al. Presents a study of Intrusion detection based on phishing detection with machine learning. They used hybrid ensemble feature selection techniques [15]. Van Geest et. al. Combined multiple models to enhance both the robustness and effectiveness of phishing detection. They introduce an innovative methodology for simulating bypass attacks on single-analysis base models. Their experiments demonstrate that the proposed hybrid framework outperforms individual models [16]. Another study done by R. Alazaidah, they considered two objectives. The first is to identify the best classifier that can detect phishing among twenty-four different classifiers that represent six learning strategies. The second objective aims to identify the best feature selection method for websites phishing datasets. Their results revealed the superiority of Random Forest, Filtered Classifier, and J-48 classifiers in detecting phishing websites [17].

4. Methodology

4.1 Dataset Description and Representation

The dataset was collected from the public and well-known repository Kaggle.com [13]. Datasets for both phishing URL detection and phishing Email were utilized for the study. The first dataset was called the phishing site predict dataset, and it contained 549,346 entries with two columns. The website links (URLs) are represented as the first attribute, while countries are represented as the second attribute. The website's labels were classified into good or bad. Another dataset serves in the same domain including 50 attributes and total of 10000 entries for phishing attack labeled as normal and phishing. The third dataset used for email detection using email subject. Table 1 denotes a small sample of the selected datasets.

Table 1 Sample of selected datasets [14]



URL	Label
nobell.it/70ffb52d079109dca5664cce6f317373782/login.SkyPe.com/en/cgi-bin/	bad
www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrcmd=_home-customer&nav=	bad
serviciosbys.com/paypal.cgi.bin.get-into.herf.secure.dispatch35463256rzr3216	bad
mail.printakid.com/www.online.americanexpress.com/index.html	bad
thewhiskeydregs.com/wp-content/themes/widescreen/includes/temp/promc	bad
smilesvoegol.servebbs.org/voegol.php	bad
premierpaymentprocessing.com/includes/boleto-2via-07-2012.php	bad
myxxxcollection.com/v1/js/jih321/bpd.com.do/do/l.popular.php	bad
super1000.info/docs	bad
horizonsgallery.com/js/bin/ssl1/_id/www.paypal.com/fr/cgi-bin/webscr/cmd=	bad
phlebolog.com.ua/libraries/joomla/results.php	bad
docs.google.com/spreadsheet/viewform?formkey=dE5rVEdSV2pBdkpSRy11V3	bad
www.coincoele.com.br/Scripts/smiles/?pt-br/Paginas/default.aspx	bad
www.henkdeinumboomkwekerij.nl/language/pdf_fonts/smiles.php	bad
perfectsolutionofall.net/wp-content/themes/twentyten/wiresource/	bad
lingshc.com/old_aol.1.3/?Login=&amp;Lis=10&amp;LigertID=1993745&amp;us=	bad

	label	text	label_num
605	ham	Subject: enron methanol ; meter # : 988291	0
2349	ham	Subject: hpl nom for january 9 , 2001	0
3624	ham	Subject: neon retreat	0
4685	spam	Subject: photoshop , windows , office . cheap . main trending	1
2030	ham	Subject: re : indian springs	0
2949	ham	Subject: ehronline web address change	0
2793	ham	Subject: spring savings certificate - take 30 % off	0
4185	spam	Subject: looking for medication ? we ` re the best source .	1
2641	ham	Subject: noms / actual flow for 2 / 26	0
1870	ham	Subject: nominations for oct . 21 - 23 , 2000	0
4922	spam	Subject: vocable % rnd - word asceticism	1
3799	spam	Subject: report 01405 !	1
1488	ham	Subject: enron / hpl actuals for august 28 , 2000	0
3948	spam	Subject: vic . odin n ^ ow	1
3418	ham	Subject: tenaska iv july	0
4791	spam	Subject: underpriced issue with high return on equity	1

id	NumDots	Subdomai	PathLevel	UrlLength	NumDash	NumDash	AtSymbol	TildeSymt	NumUnde	NumPerce	NumQuer	NumAmp	NumHash	NumNum
1	3	1	5	72	0	0	0	0	0	0	0	0	0	0
2	3	1	3	144	0	0	0	0	2	0	2	1	0	41
3	3	1	2	58	0	0	0	0	0	0	0	0	0	0
4	3	1	6	79	1	0	0	0	0	0	0	0	0	0
5	3	0	4	46	0	0	0	0	0	0	0	0	0	2
6	3	1	1	42	1	0	0	0	0	0	0	0	0	0
7	2	0	5	60	0	0	0	0	0	0	0	0	0	1
8	1	0	3	30	0	0	0	0	0	0	0	0	0	3
9	8	7	2	76	1	1	0	0	0	0	0	0	0	2
10	2	0	2	46	0	0	0	0	0	0	0	0	0	0
11	5	4	2	64	1	1	0	0	0	0	0	0	0	3
12	2	0	2	47	0	0	0	0	0	0	0	0	0	1
13	2	1	2	61	1	1	0	0	0	0	0	0	0	0
14	2	1	3	35	0	0	0	0	0	0	0	0	0	0
15	2	1	2	60	1	1	0	0	0	0	0	0	0	0
16	3	0	4	73	0	0	0	0	0	0	0	0	0	0

4.2 Add dataset class

As covered in the previous section, there is a still gap in phishing detection techniques in prior research. Most research focuses on using one machine learning techniques for detecting phishing emails, while few studies attempted to mix multiple mining algorithms to simplify multi-feature approach.

In this research, the proposed frame work addresses the phishing problem and attempts to use multiple ML techniques to improve the accuracy of proposed model. Both classification and clustering techniques in addition to feature selection algorithm used to fulfill the required goal figure 3 below showed the proposed model stages.

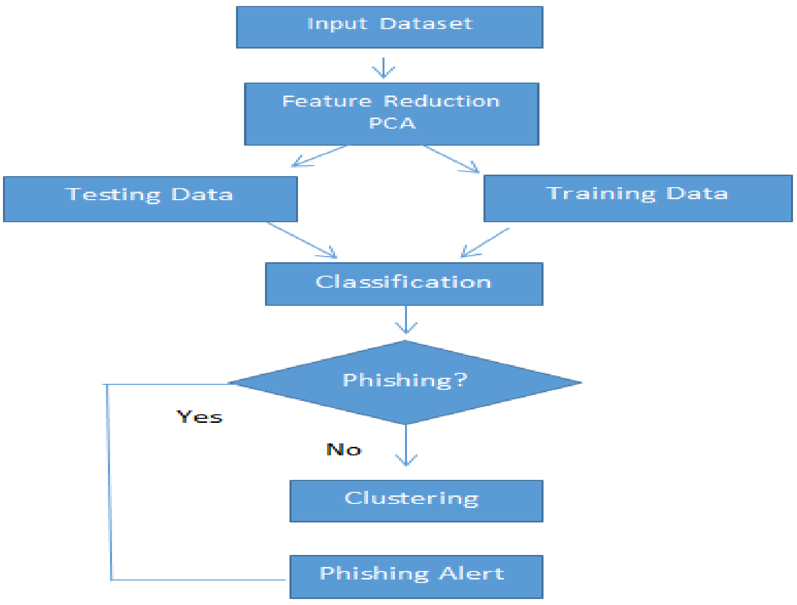
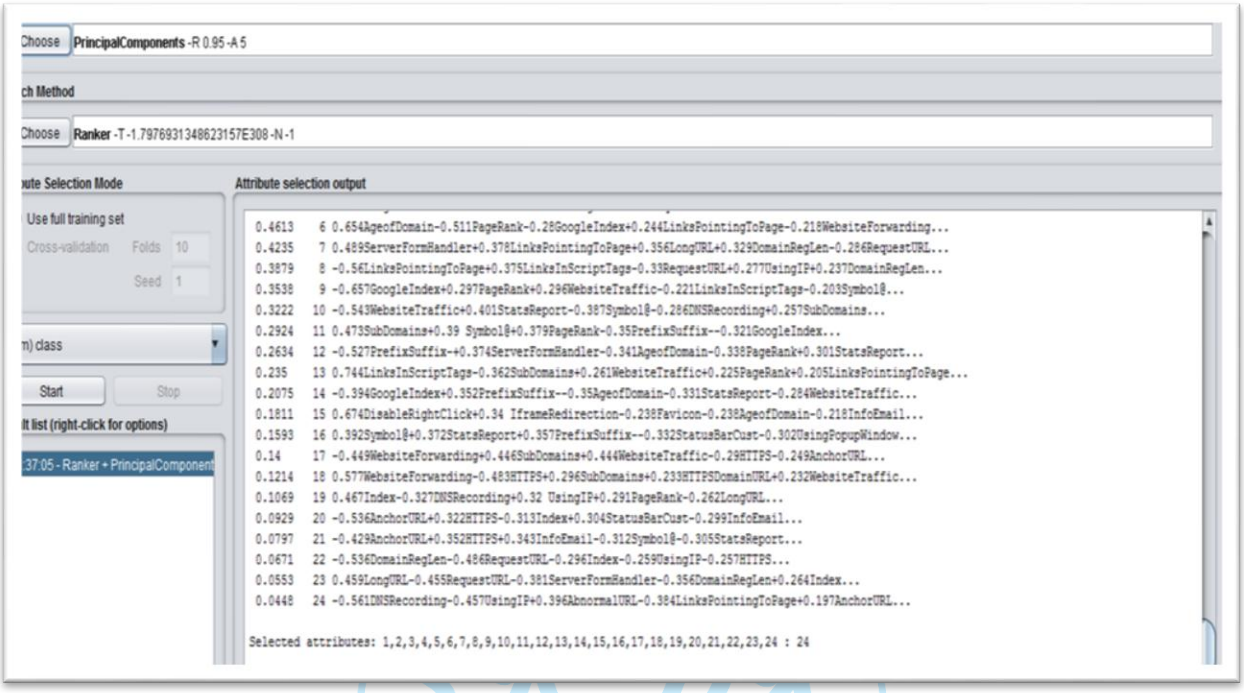


Fig. 3 The proposed Model

4.3 Using feature reduction Technique

PCA is an unsupervised machine learning algorithm that attempts to reduce the dimensionality (number of features) of used dataset. This is done by finding a new set of features called *components*, which are composites of the original features that are uncorrelated with one another. PCA is one of the algorithms based on feature selection, also known as feature extraction. Feature selection works by selecting the right attributes and removing non-essential attributes to get good analysis results [14]. PCA used for reduces the dataset form 50 attribute to 18 attribute. The reduction done in three stages the first stage the algorithm reduces the features to 31 and the second reduces to 24 then the third stage reduces the features to 18 Feature but the model accuracy remains the same. Figure 4 below shows the PCA reduction result.



PrincipalComponents -R 0.95 -A 5

Ranker -T -1.7976931348623157E308 -N -1

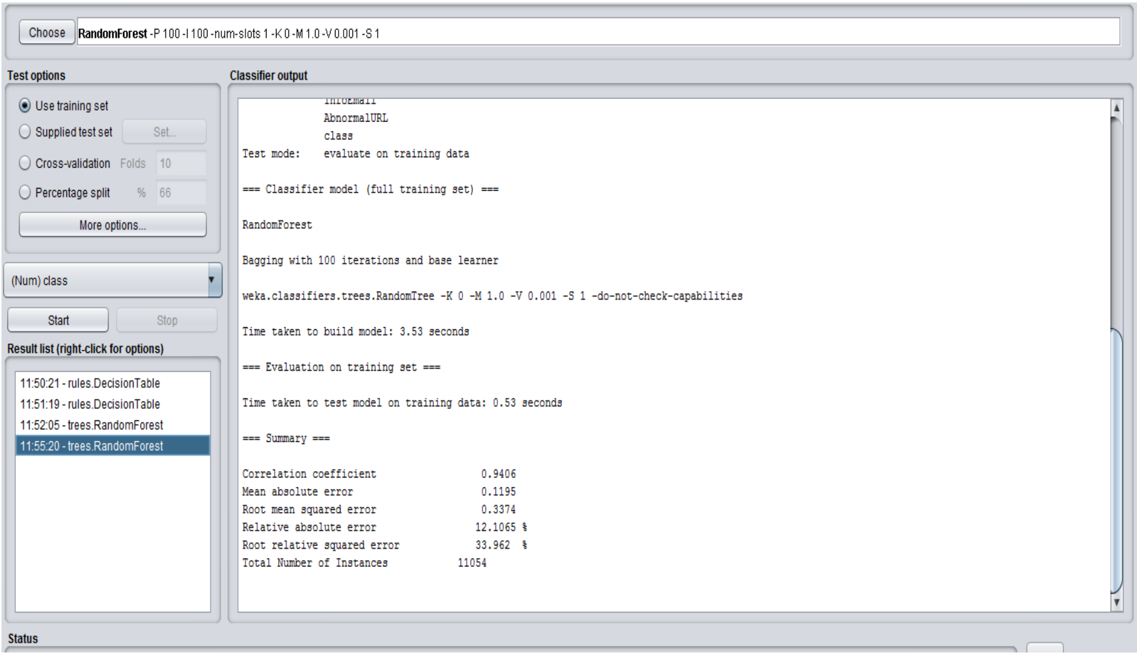
```
0.4613    6 0.654AgeofDomain-0.511PageRank-0.28GoogleIndex+0.244LinksPointingToPage-0.218WebsiteForwarding...
0.4235    7 0.489ServerFormHandler+0.378LinksPointingToPage+0.356LongURL+0.329DomainRegLen-0.286RequestURL...
0.3879    8 -0.56LinksPointingToPage+0.375LinksInScriptTags-0.33RequestURL+0.277UsingIP+0.237DomainRegLen...
0.3538    9 -0.657GoogleIndex+0.297PageRank+0.296WebsiteTraffic-0.221LinksInScriptTags-0.203Symbol@...
0.3222   10 -0.543WebsiteTraffic+0.401StatsReport-0.387Symbol@-0.286DNSRecording+0.257SubDomains...
0.2924   11 0.473SubDomains+0.39 Symbol@+0.379PageRank-0.35PrefixSuffix--0.321GoogleIndex...
0.2634   12 -0.527PrefixSuffix-+0.374ServerFormHandler-0.341AgeofDomain-0.338PageRank+0.301StatsReport...
0.235    13 0.744LinksInScriptTags-0.362SubDomains+0.261WebsiteTraffic+0.225PageRank+0.205LinksPointingToPage...
0.2075   14 -0.394GoogleIndex+0.352PrefixSuffix--0.35AgeofDomain-0.331StatsReport-0.284WebsiteTraffic...
0.1811   15 0.674DisableRightClick+0.34 IframeRedirection-0.238Favicon-0.238AgeofDomain-0.218InfoEmail...
0.1593   16 0.392Symbol@+0.372StatsReport+0.357PrefixSuffix--0.332StatusBarCust-0.302UsingPopupWindow...
0.14     17 -0.449WebsiteForwarding+0.446SubDomains+0.444WebsiteTraffic-0.29HTTPS-0.249AnchorURL...
0.1214   18 0.577WebsiteForwarding-0.483HTTPS+0.296SubDomains+0.233HTTPSDomainURL+0.232WebsiteTraffic...
0.1069   19 0.467Index-0.327DNSRecording+0.32 UsingIP+0.291PageRank-0.262LongURL...
0.0929   20 -0.536AnchorURL+0.322HTTPS-0.313Index+0.304StatusBarCust-0.299InfoEmail...
0.0797   21 -0.429AnchorURL+0.352HTTPS+0.343InfoEmail-0.312Symbol@-0.305StatsReport...
0.0671   22 -0.536DomainRegLen-0.486RequestURL-0.296Index-0.259UsingIP-0.257HTTPS...
0.0553   23 0.459LongURL-0.455RequestURL-0.381ServerFormHandler-0.356DomainRegLen+0.264Index...
0.0448   24 -0.561DNSRecording-0.457UsingIP+0.396AbnormalURL-0.384LinksPointingToPage+0.197AnchorURL...

Selected attributes: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24 : 24
```

Choose
ch Method
Choose
ute Selection Mode
Attribute selection output
Use full training set
Cross-validation Folds 10
Seed 1
m) class
Start
Stop
lt list (right-click for options)
:37:05 - Ranker + PrincipalComponent

Fig. 4 PCA Reduction result

4.4 Using classification and clustering techniques

In this stage the selected features are further input to classification algorithm after splitting to training and testing sets. For classification two algorithms are voted Random forest and neural networks based on the review done in previous studies. Random forest is an ensemble learning classification and regression method suitable for handling problems involving grouping of data into classes [18]. The algorithm was developed by Breiman and Cutler. RF method has also been used to solve similar problem in the literature, such as in [19, 20]. Neural network is a mathematical model inspired by biological neural networks and one of famous techniques used for classifying phishing data [21]. This technique was also used in previous literature such as [22, 23]. Figure 5 and 6 shows classification results respectively.



RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) class

```
             InfoEmail
             AbnormalURL
             class
Test mode:   evaluate on training data

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 3.53 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.53 seconds

=== Summary ===

Correlation coefficient                  0.9406
Mean absolute error                      0.1195
Root mean squared error                  0.3374
Relative absolute error                 12.1065 %
Root relative squared error             33.962  %
Total Number of Instances            11054
```

11:50:21 - rules.DecisionTable
11:51:19 - rules.DecisionTable
11:52:05 - trees.RandomForest
11:55:20 - trees.RandomForest

Fig. 5 RF Classification result using weka miner

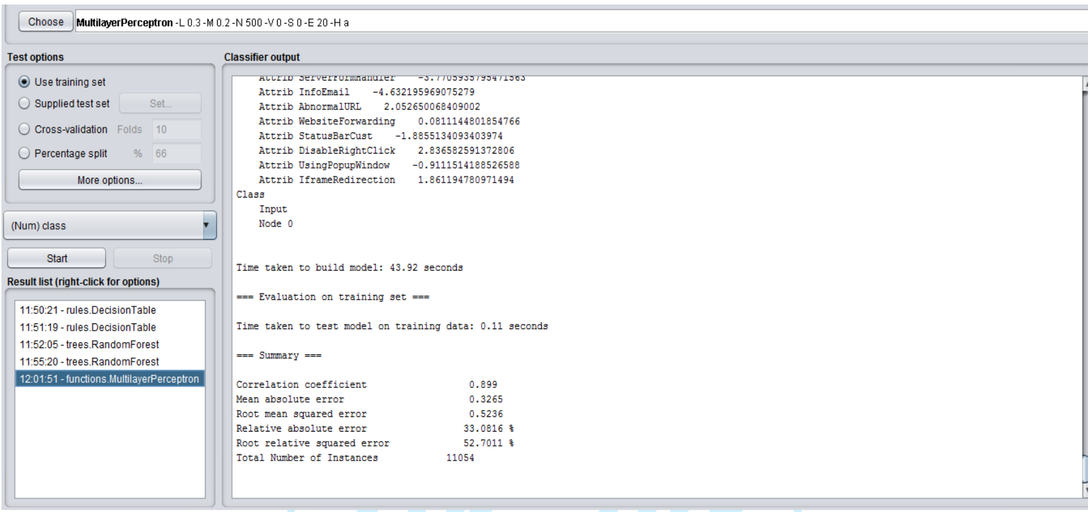
The following tables (2and3) show the statistical result of using RF classification algorithm including TP, FP, and precision in addition to Confusion Matrix.

Table 2 Detailed statistical Results for RF algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.929	0.038	0.950	0.929	0.939	Phishing
	0.962	0.071	0.944	0.962	0.953	Normal
Weighted Avg.	0.947	0.057	0.947	0.947	0.947	

Table 3 Confusion Matrix for RF Algorithm

	Phishing	Normal
Phishing	1547	119
Normal	81	2025



Choose MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Test options
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) class

Result list (right-click for options)
11:50:21 - rules.DecisionTable
11:51:19 - rules.DecisionTable
11:52:05 - trees.RandomForest
11:55:20 - trees.RandomForest
12:01:51 - functions.MultilayerPerceptron

Classifier output

```
    Attrib ServerFormHandler    -3.7705935795471563
    Attrib InfoEmail    -4.632195969075279
    Attrib AbnormalURL    2.052650068409002
    Attrib WebsiteForwarding    0.0811144801854766
    Attrib StatusBarCust    -1.8855134093403974
    Attrib DisableRightClick    2.836582591372806
    Attrib UsingPopupWindow    -0.9111514188526588
    Attrib IframeRedirection    1.861194780971494
Class
    Input
    Node 0


Time taken to build model: 43.92 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.11 seconds

=== Summary ===

Correlation coefficient                  0.899
Mean absolute error                      0.3265
Root mean squared error                  0.5236
Relative absolute error                 33.0816 %
Root relative squared error             52.7011 %
Total Number of Instances            11054
```

Fig. 6 Neural Network classification result

Fig. 6 shows the reached results of using neural network for classification with Detailed statistical results (Below table 4).

Table 4 Detailed statistical Results

	TP Rate	FP Rate	Precision	Recall	Class
	0.714	0.201	0.730	0.714	Phishing
	0.799	0.286	0.786	0.799	Normal
Weighted Avg	0.762	0.762	0.762	0.762	

Table 5 Confusion Matrix for Neural network

	Phishing	Normal
Phishing	100	40
Normal	37	147

From the result RF give more accuracy over neural network. RF accuracy reaches 94% and neural network give 89%. Time taken to build RF model: 1.9 seconds in contrast to 43 seconds for NN. Based on this result the model will use RF for classification stage.

Data spited to training and testing sets before fed to classifier the training percent will be 66% to 34% testing. The output from this step contains 2 classes “phish” and “normal” in addition to misclassified records. The number of correctly classified records are3572 compared to 200 incorrect classified records. Normal classified records including phishing records that are misclassified as normal were further input to clustering algorithm to identify newly suspicious URLs that doesn’t discovered in the previous stage. Clustering technique classify data based on similarity between its attributes. K-means algorithm is one of the simplest and popular unsupervised machine learning algorithms used for clustering purpose and previous research addressed the use of this algorithms in [24,25]. Based on previous research findings K-means used as the second stage in the proposed model. The result from clustering algorithm is two main groups of URLs as shown in the Table below.

Table 6 Clustering Results

Clusters	Instances
Phishing Cluster	140
Normal cluster	184

Time taken to build the clustering model is 0.01 seconds that considered as relatively low. Using the same algorithm with 3 clustering group's settings produce number of 17 records marked as not known or suspicious which is added to phishing group in case of two clusters. So, the user could revise the suspicious records whether it is phishing or normal. The reached results confirm the results of using another clustering algorithm expectation–maximization (EM) algorithm that concluding there is 17 records classified as unknown mean that it requires additional inspection. EM algorithm was used previously by Indu S. and Rajni Jindal in [25], for detecting intrusive transactions in databases. Clustering results can be used for checking the newly undiscovered kinds of phishing attack.

5. Results and Discussion

Statistics computation such as True Positive, False Positive, True Negative, False Negative, Precision, F-score, Accuracy, and Recall were computed for each model. These statistics were used as a standard for comparison, and the best performing model picked based on these measurements. Based on the listed results we find that using three stages model (feature reduction, classification and clustering) can enhance detection performance for phishing detection model.

6. Conclusion

Phishing is a crucial threat to individual's data nowadays. Detection of phishing sites is actually a tiresome task, as the outcome phishers are actually quickly enhancing. To overcome the problem, researchers and specialists dealt with lots of methods and techniques, however it led to reduced prices of detection. The main contribution of the study is to use three mining stages to improve detection process. The area is an ongoing research area much research are still required to enhance not only detection but also detection performance.

Acknowledge

The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

References

1. Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.
2. Zongo, W. B. S., Kabore, B., & Vaghela, R. S. (2023, January). Phishing URLs Detection Using Machine Learning. In Advancements in Smart Computing and Information Security: First International Conference, ASCIS 2022, Rajkot, India, November 24–26, 2022, Revised Selected Papers, Part II (pp. 159-167). Cham: Springer Nature Switzerland.
3. Salahdine, F., El Mrabet, Z., & Kaabouch, N. (2021, December). Phishing Attacks Detection A Machine Learning-Based Approach. In 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 0250-0255). IEEE.
4. Sundara Pandiyan S, Prabha Selvaraj, Vijay Kumar Burugari, Julian Benadit P, Kanmani P,Phishing attack detection using Machine Learning , Measurement: Sensors, Volume 24, ,2022 ,100476 ISSN 2665-9174.
5. Proof point, 2022 State of the Phish, https://www.proofpoint.com/us/resources/threat-reports/state-of-phish , visit date 30 january 2023
6. https://www.tessian.com/blog/phishing-vs-spear-phishing/
7. Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A., & Elsoud, E. A. (2022). An intelligent cyber security phishing detection system using deep learning techniques. Cluster Computing, 25(6), 3819-3828.
8. Muralidharan, T., & Nissim, N. (2023). Improving malicious email detection through novel designated deep-learning architectures utilizing entire email. Neural Networks, 157, 257-279.
9. Chawla, A. (2022). Phishing website analysis and detection using Machine Learning. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 10-16.
10. Alam, M. N., Sarma, D., Lima, F. F., Saha, I., & Hossain, S. (2020, August). Phishing attacks detection using machine learning approach. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 1173-1179). IEEE.
11. Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 2019 Jun 1;83:246–67.
12. Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020 Sep;9(9):1514.
13. Kaggle.com, P.S.U.A.O. Available online: https://www.kaggle.com/taruntiwarihp/phishing-site-urls (accessed on 8 march 2023).
14. Deyanara Tuapattinaya, Antoni Wibowo , Phishing Website Detection using Neural Network and PCA based on Feature Selection, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020.
15. Jayaraj, R., Pushpalatha, A., Sangeetha, K., Kamaleshwar, T., Shree, S. U., & Damodaran, D. (2024). Intrusion detection based on phishing detection with machine learning. Measurement: Sensors, 31, 101003.

16. van Geest, R. J., Cascavilla, G., Hulstijn, J., & Zannone, N. (2024). The applicability of a hybrid framework for automated phishing detection. Computers & Security, 139, 103736.
17. Alazaidah, R., Al-Shaikh, A., AL-Mousa, M. R., Khafajah, H., Samara, G., Alzyoud, M., ... & Almatarneh, S. (2024). Website phishing detection using machine learning techniques. Journal of Statistics Applications & Probability, 13(1), 119-129
18. Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014
19. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in Proceedings of the 16th International World Wide Web Conference (WWW '07), pp. 649–656, Alberta, Canada, May 2007.
20. C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in Proceedings of the 17th Annual Network & Distributed System Security Symposium (NDSS '10), The Internet Society, San Diego, Calif, USA, 2010.
21. Zhang, N., & Yuan, Y. (2012). Phishing detection using neural network. CS229 lecture notes, 34
22. Wang, W., Zhang, F., Luo, X., & Zhang, S. (2019). PDRCNN: Precise phishing detection with recurrent convolutional neural networks. Security and Communication Networks, 2019, 1-15.
23. Wanawe, K., Awasare, S., & Puri, N. V. (2014). An efficient approach to detecting phishing a web using k-means and naïve-bayes algorithms. International Journal of Research in Advent Technology, 2(3), 106-111.
24. Mhaske-Dhamdhere, V., & Vanjale, S. (2018). A novel approach for phishing emails real time classification using k-means algorithm. International Journal of Engineering and Technology, 7, 96-100.
25. Singh, I., & Jindal, R. (2021). Expectation maximization clustering and sequential pattern mining based approach for detecting intrusive transactions in databases. Multimedia Tools and Applications, 80(18), 27649-27681.