Phishing Email Detection Using Machine Learning Techniques

Meaad Alammar[1†] and Maria Altaib Badawi[2††].

College of Science, Department of Computer Science and Information in Majmaah University, Az Zulfi 15971, SA

Summary
Email phishing has become very prevalent especially now that most of our dealings have become technical. The victim receives a message that looks as if it was sent from a known party and the attack is carried out through a fake cookie that includes a phishing program or through links connected to fake websites, in both cases the goal is to install malicious software on the user's device or direct him to a fake website. Today it is difficult to deploy robust cybersecurity solutions without relying heavily on machine learning algorithms. This research seeks to detect phishing emails using high-accuracy machine learning techniques. using the WEKA tool with data preprocessing we create a proposed methodology to detect emails phishing. outperformed random forest algorithm on Naïve Bayes algorithms by accuracy of 99.03 %.

Keywords:
WEKA, Random Forest, Phishing Email, Cybersecurity, Data Mining,

1. Introduction

In the digital revolution, many people are doing their daily work by relying on the services provided by various Internet sites, such as online shopping, financial transactions, and many more, in the hope of saving effort and time. But what is wrong with these services is the disclosure of the user's personal information, various account numbers, and passwords, which formed an environment that attracted cybercriminals who excelled in inventing methods and methods of fraud and phishing to get what they want without the user feeling anything. Phishing emails are usually of poor style, however, cybercriminal groups use the same techniques as professional marketers to find out the most effective types of messages. With all this development but humans may overlook these attacks, we seek to make data protection without human intervention by using machine learning. We can simplify the concept of machine learning as one of the branches of artificial intelligence based on programming computers in all their forms; To be able to perform the tasks and carry out the commands assigned to them based on the data available to it and its analysis with the limitation or complete absence of human intervention in directing it. It is worth noting that the machine in this case must rely on analyzing the data entered into it in advance to meet the commands and tasks required of it.

2. Theoretical Consideration

2.1 Phishing

Cybercrime is a widespread occurrence in the realm of technology, and it can happen to anyone at any time. Cybercrime is a type of criminal activity that targets computers and networks. A thief who we know is a criminal steals data documents, money, and confidential private information. But consider who does these same things in the virtual world, which we have dubbed PHISHER. And the phisher's work is known as PHISHING [1]. Phishing is a dangerous type of social engineering that aims to trick people into disclosing personal or confidential information. Despite frequent warnings and methods to teach users how to recognize phishing communications, phishing is a common practice and profitable industry [2].

2.2 Email Phishing

When a recipient clicks on a malicious file or link in an email sent by a cybercriminal, malware is installed. In the past, cybercriminals utilized broad-based spamming techniques to spread their virus, but contemporary ransomware efforts have been more focused and smart. Criminals may also employ precursor malware to infiltrate a victim's email account, allowing the cybercriminal to utilize the victim's email account to propagate the infection further [3]. Phishing emails are a type of targeted email assault in which social engineers persuade recipients to take specified actions, such as clicking on a harmful link, opening a malicious attachment, or visiting a website and entering personal information [4].

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.38

2.3 Machine Learning

Training and testing are the two phases of machine learning. They execute mathematical computations over the training dataset and learn the behavior of traffic over time during the training phase [5]. The term "machine learning" refers to a process in which computers analyze current data and learn new skills and information from it. Machine learning systems employ algorithms to search for patterns in datasets that may include structured data, unstructured textual data, numeric data, or even rich media such as audio files, photos, and videos [6].

2.4 Types of Machine Learning

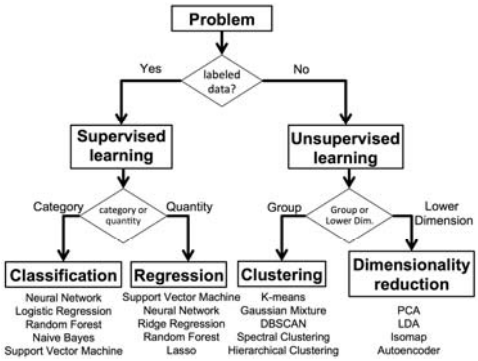

Fig. 1 Types of Machine Learning [7].

2.5 Random Forest Algorithm

Random Forests use random bootstrapped samples of the training data to create several decision trees. RF, unlike other classifiers, does not produce overfitting or necessitate a lengthy training period. The nodes are divided using the best split variable from a subset of m randomly selected variables, and each tree is formed using a subset that differs from the original training data, containing around two-thirds of the cases [8]. One of the key advantages of a random forest technique is that it can fit nonlinearities and interactions [9]. It can handle huge datasets with a lot of dimensionalities. It improves the model's accuracy and eliminates the problem of overfitting [10].

3. Methodology for Result Implantation

3.1 WEKA

WEKA is open-source software that is free to use. It's written in Java and may operate on any Java-enabled platform, including Linux, Mac OS X, and Windows [11]. WEKA is a collection of data mining-related machine learning algorithms. The methods are immediately applied to a dataset. Data pre-processing, classification, clustering, regression, and feature selection and visualization are data-mining operations WEKA provides [12].

3.2 Proposed Methodology

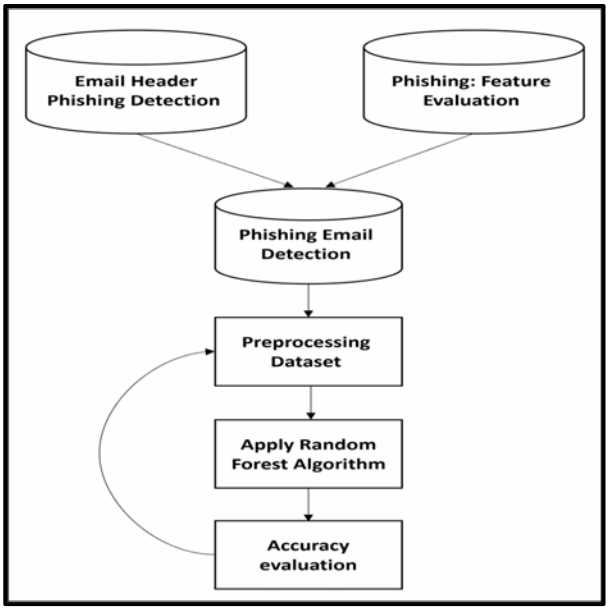

Fig. 2 Proposed Methodology

3.3 Dataset

After collecting the data that investigates phishing emails from Combine between [13] and [14] it is well understood. There are 88489 instances of 126 attributes. The detailed information about Dataset is given below in table 1.

Table 1 Dataset

Sr. No	Feature Name	Type
1	Hops	Nominal
2	MissingSubject	Nominal
3	MissingTo	Nominal
4	MissingContentType	Nominal
5	MissingMime-version	Nominal
6	MissingX-mailer	Nominal

7	MissingContent-transfer-encoding	Nominal
8	MissingX-mimeole	Nominal
9	MissingX-priority	Nominal
10	MissingList-id	Nominal
11	MissingLines	Nominal
12	MissingX-virus-scanned	Nominal
13	MissingStatus	Nominal
14	MissingContent-length	Nominal
15	MissingPrecedence	Nominal
16	MissingDelivered-to	Nominal
17	MissingList-unsubscribe	Nominal
18	MissingList-subscribe	Nominal
19	MissingList-post	Nominal
20	MissingList-help	Nominal
21	MissingX-msmail-priority	Nominal
22	MissingX-Spam-status	Nominal
23	MissingSender	Nominal
24	MissingErrors-to	Nominal
25	MissingX-beenthere	Nominal
26	MissingList-archive	Nominal
27	MissingReply-to	Nominal
28	MissingX-mailman-version	Nominal
29	MissingX-miltered	Nominal
30	MissingX-uuid	Nominal
31	MissingX-virus-status	Nominal
32	MissingX-spam-level	Nominal
33	MissingX-spam-checker-version	Nominal
34	MissingReferences	Nominal
35	MissingIn-reply-to	Nominal
36	MissingUser-agent	Nominal
37	MissingThread-index	Nominal
38	MissingCC	Nominal
39	MissingReceived-spf	Nominal
40	MissingX-original-to	Nominal

41	MissingContent-disposition	Nominal
42	MissingMailing-list	Nominal
43	MissingX-spam-check-by	Nominal
44	MissingDomainkey-signature	Nominal
45	MissingImportance	Nominal
46	MissingX-mailing-list	Nominal
47	Content-encoding-val	Nominal
48	ReceivedStrForged	Nominal
49	StrContent-encodingEmpty	Nominal
50	StrFromQuestion	Nominal
51	StrFromChevron	Nominal
52	StrToChevron	Nominal
53	StrToEmpty	Nominal
54	StrMessage-IDDollar	Nominal
55	StrReturn-pathBounce	Nominal
56	StrContent-typeTextHtml	Nominal
57	StrPrecedenceList	Nominal
58	LengthFrom	Nominal
59	NumRecipientsTo	Nominal
60	NumRecipientsCc	Nominal
61	NumberReplies	Nominal
62	TimeZone	Nominal
63	X-priority	Nominal
64	ContentLength	Nominal
65	Lines	Nominal
66	DayOfWeek	Nominal
67	DateCompDateReceived	Nominal
68	SpanTime	Nominal
69	ConseqNumRecievedIsOne	Nominal
70	ConseqRecievedGood	Nominal
71	ConseqRecievedBad	Nominal
72	ConseqRecievedDate	Nominal
73	EmailMatchFronReply-to	Nominal
74	DomainValMessage-id	Nominal
75	DomainMatchMessage-idFrom	Nominal
76	DomainMatchFromReturn-path	Nominal

77	DomainMatchMessage-idReturn-path	Nominal
78	DomainMatchMessage-idSender	Nominal
79	DomainMatchMessage-idReply-to	Nominal
80	DomainMatchReturn-pathReply-to	Nominal
81	DomainMatchReply-toTo	Nominal
82	DomainMatchToIn-reply-to	Nominal
83	DomainMatchErrors-toMessage-id	Nominal
84	DomainMatchErrors-toFrom	Nominal
85	DomainMatchErrors-toSender	Nominal
86	DomainMatchErrors-toReply-to	Nominal
87	DomainMatchSenderFrom	Nominal
88	DomainMatchReferencesReply-to	Nominal
89	DomainMatchReferencesIn-reply-to	Nominal
90	DomainMatchReferencesTo	Nominal
91	DomainMatchFromReply-to	Nominal
92	DomainMatchToFrom	Nominal
93	DomainMatchToMessage-id	Nominal
94	DomainMatchToReceived	Nominal
95	Label	Nominal
96	HavingIPAddress	Nominal
97	URLLength	Nominal
98	ShortiningServices	Nominal
99	HavingAtSymbol	Nominal
100	DoubleSlashRedirecting	Nominal
101	PrefixSuffix	Nominal
102	HavingSubDomain	Nominal
103	SSLfinalState	Nominal
104	DomainRegisterationLength	Nominal
105	Favicon	Nominal
106	Port	Nominal
107	HTTPSToken	Nominal
108	RequestURL	Nominal
109	URLOfAnchor	Nominal
110	LinksInTags	Nominal
111	SFH	Nominal
112	SubmittingToEmail	Nominal
113	AbnormalURL	Nominal
114	Redirect	Nominal
115	OnMouseover	Nominal
116	RightClick	Nominal
117	PopUpWindow	Nominal
118	Iframe	Nominal
119	AgeOfDomain	Nominal
120	DNSRecord	Nominal
121	WebTraffic	Nominal
122	PageRank	Nominal
123	GoogleIndex	Nominal
124	LinksPointingToPage	Nominal
125	StaticalReport	Nominal
126	Result	Nominal

3.4 Data Preprocessing

The data obtained from the field comprises a number of undesirable elements that lead to incorrect analysis. The data could, for example, contain null fields or columns that are irrelevant to the current study, and so on. As a result, the data must be preprocessed to satisfy the needs of the analysis you're performing.

3.5 Applying Filters

Filters help with data preparation and sometimes lead to better classification. We will increase Email Phishing Detection accuracy to high by applying filters to our raw data. The filter we used is RemoveMisclassified, a filter that removes instances that are incorrectly classified.

Weka → filters → unsupervised → instances → RemoveMisclassified.

3.6 Applying Algorithm

Random Forest is a multiple learning classifier that works by building a large number of decision trees during training, This classifier aids in the correction of overfitting in decision trees during training [15].

3.7 Experiment Results

The dataset output was rated using Phishing Email Detection Accuracy as 0 for phishing features and 1 for legitimate features. After using Random Forest Algorithm in the WEKA tool to detect emails that contain Phishing. Figure 3 shows the result of the accuracy of using the algorithm that was reached 99.03%, this means that we have 88489 cases, 87634 correctly detected and 855 mis detected.

```
=== Summary ===

Correctly Classified Instances       87634               99.0338 %
Incorrectly Classified Instances       855                0.9662 %
Kappa statistic                          0.9791
Mean absolute error                      0.0097
Root mean squared error                  0.0696
Relative absolute error                  2.099  %
Root relative squared error             14.4867 %
Total Number of Instances            88489
```

Fig. 3 Experiment Results

4. Statical Analysis And Evaluation

4.1 Metrics

According to the study in [16] it was found that the best standards for measuring accuracy are the following:

(1) $Precision = \frac{TP}{TP+FP}$,

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

(3) $F - Measure = \frac{2*Precision*Recall}{Precision+Recall}$

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

The symbols are used as follows:

1- True Positive (TP): Number of phishing emails detected correctly.
2- False Negative (FN): Number of phishing emails detected as legitimate emails.
3- False Positive (FP): Number of legitimate emails detected as phishing emails.
4- True Negative (TN): Number of legitimate emails detected as legitimate emails.

4.2 Experimental Results

The proposed methodology was implemented to search and extract the results. In the first experiment the results of collecting data sets without processing. In the second experiment after data processing and feature selection. In the third experiment after applying the Remove misclassified filter. Table 2, the results of the random forest algorithm and raises the accuracy to 99.03 %.

Table 2 Random Forest Result

No.	Precision	Recall	F-Measure	Accuracy
Exp. 1	0.914	0.905	0.886	90.53 %
Exp. 2	0.990	0.990	0.990	98.97 %
Exp. 3	0.991	0.990	0.990	**99.03 %**

4.3 Comparing Algorithms

Figure 4 shows the result of the accuracy of using the Naive Bayes algorithm that was reached 90.13 %, this means that we have 88489 cases, 79762 correctly detected and 8727 mis detected.

```
=== Summary ===

Correctly Classified Instances       79762               90.1378 %
Incorrectly Classified Instances      8727                9.8622 %
Kappa statistic                          0.4913
Mean absolute error                      0.1073
Root mean squared error                  0.3128
Relative absolute error                 41.9079 %
Root relative squared error             87.4264 %
Total Number of Instances            88489
```

Fig. 4 Naïve Bayes Result

The results for the Random Forest algorithm were compared with the Naive Bayes algorithm and the following results are shown in Table 3. This means that the random forest algorithm is outperformed by it.

Table 3 Naïve Bayes Result

No.	Precision	Recall	F-Measure	Accuracy
Exp. 1	0.852	0.852	0.852	85.17%
Exp. 2	0.856	0.856	0.856	85.60 %
Exp. 3	0.903	0.901	0.883	**90.13 %**

5. Conclusion

An email phishing attack occurs when someone tries to trick you into sharing your personal information online. Phishing emails have become a common problem. We can present and process a data set to become highly accurate in detecting phishing emails through a random forest machine learning algorithm using the WEKA tool. In this work, the accuracy of the phishing email detection model was examed based on two datasets from Header anomaly detection and Phishing. In WEKA tool uses classifiers algorithms. Finally, a comparison was made between the two algorithms. outperformed the Random Forest algorithm on Naïve Bayes algorithms. The study concluded that the selection of efficient features influences the accuracy of the task of phishing emails classification. Therefore, the highest accuracy of 99.03% was obtained when we used a Random Forest classifier based on the set from the extracted features.

Future work focuses on: Spreading sufficient awareness to detect phishing e-mail and increasing the security of companies or institutions to their users by reducing the risk of threats using highly accurate machine learning algorithms. We hope that the algorithm will be used in real life by all segments of society, allowing them to benefit from it and raise their awareness of the dangers that individuals face in society.

Acknowledgments

First and foremost praise is to Allah. his constant grace and mercy were with us during life and throughout this project duration. we are extremely grateful to our parents for their love and continued support in preparing for my future. we would like to thank the Department of Computer Science & information, Majmaah University for their constant support, guidance, and encouragement. We appreciate the discussions, suggestions, criticism, and support of our colleagues, and friends, We would also like to thank them for all the aspects that facilitated the smooth work of my project.
Finally, we owe everything to our family who at every point of our personal and academic life, supported and motivated us, and longed to see this accomplishment come true.

References

[1] Soni, P., Pawar, M., & Goyal, S. (2019). A Survey on Detection and Defense from Phishing.

[2] Ferreira, A., & Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. International Journal of Human-Computer Studies, 125, 19-31.

[3] "Internet Crime Report 2020" (PDF). FBI Internet Crime Complaint Centre. U.S. Federal Bureau of Investigation. Retrieved February 10, 2022.

[4] Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., & Ebner, N. C. (2019). Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. ACM Transactions on Computer-Human Interaction (TOCHI), 26(5), 1-28.

[5] Mishra, P., Varadharajan, V., Tupakula, U., & Pilli, E. S. (2018). A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials, 21(1), 686-728.

[6] Tsymbal, O. (2022, January 20). 5 Essential Machine Learning Algorithms For Business Applications. MobiDev. Retrieved February 9, 2022, from https://mobidev.biz/blog/5-essential-machine-learning-techniques

[7] Concept of Machine Learning — Python Numerical Methods. (2021). Book. Retrieved February10,2022,from https://pythonnumericalmethods.berkeley.edu/notebooks/chapter25.01-Concept-of Machine-Learning.html

[8] Boonprong, S., Cao, C., Chen, W., Ni, X., Xu, M., & Acharya, B.K. (2018). The Classification of Noise-Afflicted Remotely Sensed Data Using Three Machine-Learning Techniques: Effect of Different Levels and Types of Noise on Accuracy. ISPRS Int. J. Geo Inf., 7, 274.

[9] Grigorescu, A., Maer-Matei, M. M., Mocanu, C., & Zamfir, A. M. (2020). Key drivers and skills needed for innovative companies focused on sustainability. Sustainability, 12(1), 102.

[10] Machine Learning Random Forest Algorithm - Javatpoint. (2021). Www.Javatpoint.Com. Retrieved February 17, 2022, from https://www.javatpoint.com/machine-learning-random-forest-algorithm

[11] Brownlee, J. (2019). Machine learning mastery with Weka. Ebook. Edition, 1(4).

[12] Fong, S., Biuk-Aghai, R. P., & Millham, R. C. (2018, February). Swarm search methods in weka for data mining. In Proceedings of the 2018 10th international conference on machine learning and computing (pp. 122-127).

[13] Tan, Choon Lin (2018), "Phishing Dataset for Machine Learning: Feature Evaluation", Mendeley Data, V1, doi: 10.17632/h3cgnj8hft.1

[14] K. (2021, Aug 3). GitHub - kregg34/EmailHeaderAnomalyDetection: Using machine learning and features extracted from email headers to detect anomalies (i.e., spam, phishing) in email datasets. [Dataset]. https://github.com/kregg34/EmailHeaderAnomalyDetection

[15] Pandey, A. K., & Rajpoot, D. S. (2016, December). A comparative study of classification techniques by utilizing

WEKA. In 2016 International Conference on Signal Processing and Communication (ICSC) (pp. 219-224). IEEE.
[16] Toolan, F., & Carthy, J. (2010, October). Feature selection for spam and phishing detection. In 2010 eCrime Researchers Summit (pp. 1-12). IEEE.

Meaad Alammar department of computer and Information science bachelor's degree from Majmaah University, KSA in 2022.

Maria Altaib Badawi Assistant Professor, Department of Computer and Information Science, Majmaah University, KSA. Several scientific papers have been published in the field of cyber security, which is one of the most important areas that the world needs now.