International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Efficient Email Phishing Detection using Machine Learning

[*1] G Dayakar Reddy,[*2] Sneha Sreelata, [*3] D Shreya

[*1] Associate Professor & Vice Principal, Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India

[*2,3] Students, Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India

ARTICLEINFO

Article History:

Accepted: 10 May 2023
Published: 30 May 2023

Publication Issue
Volume 9, Issue 3
May-June-2023

Page Number
356-360

ABSTRACT

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 cyber security 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, Cyber security, Data Mining.

I. 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.

II. RELATED WORK

Traditionally, phishing detection research has focused on methods for automated phishing detection. This section presents related work covering different aspects of phishing detection. This section begins with a brief history of phishing and an overview of the most common phishing detection methods. Researchers have tackled this problem differently over time. Some researchers have focused on machine learning models, while others have focused on manual add-ins and natural language processing elements on email text.

The term “phishing” was coined by a then-teenager named Koceilah Rekouche. Rekouche developed the first phishing attack. With a small group of teenagers, Rekouche developed the AOHell software designed to steal the passwords of America Online (AOL) users. It was arguably the first phishing software, and it was used for stealing passwords and credit card information beginning in January 1995. AOHell's phishing system was made publicly available, its release leading to many other automated phishing systems over the years. Started by teenagers and adopted by several other amateurs, phishing activity spread from AOL to other networks. Slowly, professional criminals took notice of this phishing activity and got involved in phishing schemes. Although phishing started small, it became one of the major cyber security threats worldwide, leading to significant financial losses to individuals, corporations, and even governments. Phishing, which started as a very basic technology, soon became sophisticated methodical attacks. As organizations began building algorithms to identify phishing attempts, hackers continued to invent new ways to evade the detection. Phishing attackers have constantly developed new techniques to hide their phishing attacks like Smishing, Spear phishing, Malware phishing, and Malvertising. Humans are the weakest link in the phishing scheme as they can be easily manipulated for information or duped into clicking on malicious links via social engineering techniques.

III. PROPOSED SYSTEM

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].

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].

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].

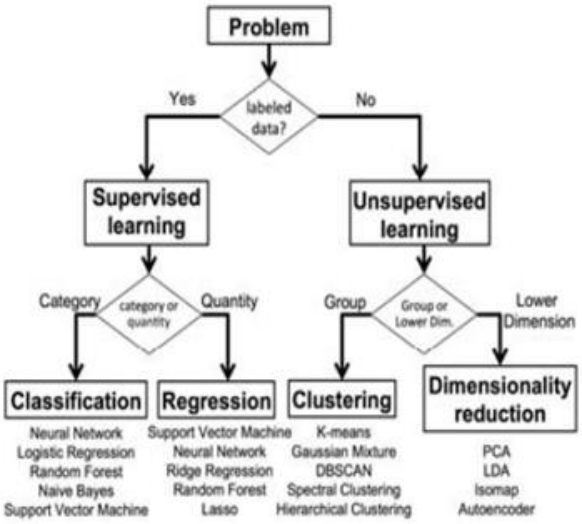

Fig. 1 Types of Machine Learning.

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 twothirds 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].

IV. RESULTS AND DISCUSSION

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. 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.

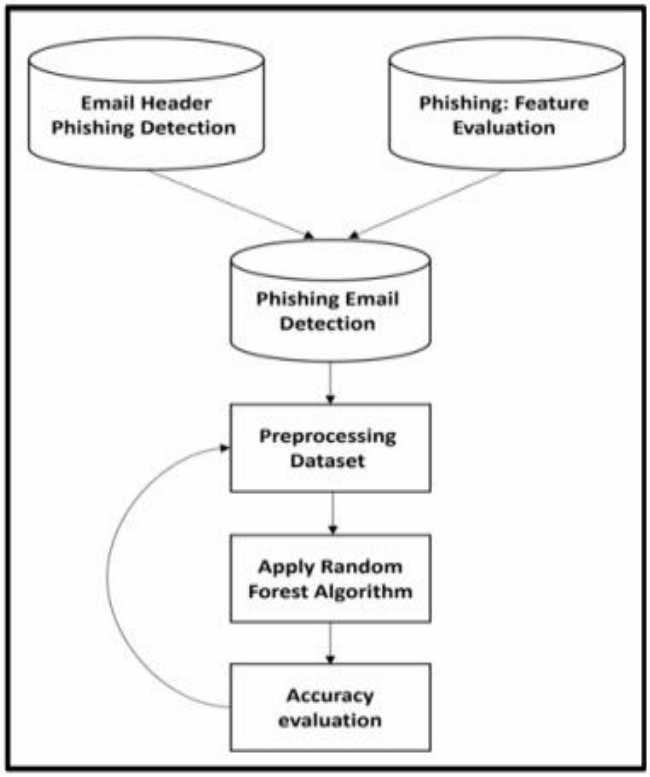

Fig. 2 Proposed Methodology

After collecting the data that investigates phishing emails from Combine between and it is well understood. There are 88489 instances of 126 attributes. 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.

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 Remove Misclassified, a filter that removes instances that are incorrectly classified.

Weka → filters → unsupervised → instances → Remove Misclassified.

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.

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.

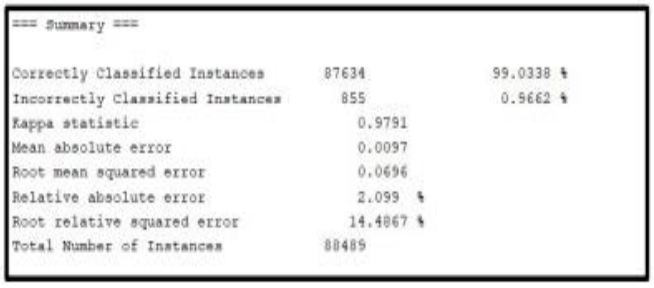

Fig. 3 Experiment Results

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 andfeature 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 %.

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 %**

Table 1 Random Forest Result

V. 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.

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.

VI. REFERENCES

[1]. M. Veale and R. Binns (2017) Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc 4(2):205395171774353. https://doi.org/10.1177/20539 51717743530

[2]. A. K. Dutta (2021) Detecting phishing websites using machine learning technique. PLoS ONE 16(10): e0258361. https://doi.org/10.1371/journal.pone.0258361

[3]. A. Akinyelu, and A. Adewumi (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. Journal of Applied Mathematics. 2014. 10.1155/2014/425731.

[4]. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “PhishNet: Predictive Block listing to Detect Phishing Attacks,” 2010 Proceedings IEEE INFOCOM, 2010, pp. 1-5, DOI: 10.1109/INFCOM.2010.5462216.

[5]. Detection of Phishing Websites Using Ensemble Machine Learning Approach Dharani M.,

Soumya Badkul, Kimaya Gharat, Amarsinh Vidhate, and Dhanashri Bhosale ITM Web Conf., 40 (2021) 03012, DOI:https://doi.org/10.1051/itmconf/20214003012

[6]. S. Abu-Nimeh, D. Nappa, X. Wang and S. Nair (2007) "Distributed Phishing Detection by Applying Variable Selection Using Bayesian Additive Regression Trees." 2009 IEEE International Conference on Communications, IEEE, 2009, pp. 1–5, https://doi.org/10.1109/ICC.2009.5198931.

[7]. N. Sanglerdsinlapachai and A Rungsawang. "Using Domain Top-Page Similarity Feature in Machine Learning-Based Web Phishing Detection." IEEE, 2010, pp. 187–190, https://doi.org/10.1109/WKDD.2010.108.

[8]. A. Alhogail and A. Alsabih (2021). Applying machine learning and natural language processing to detect phishing emails. Computers & Security, 110, 102414. https://doi.org/10.1016/j.cose.2021.102414

[9]. K. Haynes, H. Shirazi, and I. Ray (2021). Lightweight URL-based phishing detection using natural language processing transformers for mobile devices. Procedia Computer Science, 191, 127–134. https://doi.org/10.1016/j.procs.2021.07.040

[10]. V. Ramanathan and H. Wechsler (2012). phishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training. EURASIP Journal on Multimedia and Information Security, 2012(1), 1–1. https://doi.org/10.1186/1687-417X-2012-1

[11]. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri (2019). Machine learning-based phishing detection from URLs. Expert Systems with Applications, 117, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

[12]. A. Abbasi, D. Dobolyi, A. Vance, and F. M. Zahedi (2021). The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites. Information Systems Research, 32(2), 410–436. https://doi.org/10.1287/isre.2020.0973

[13]. A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.P. Niyigena (2020). An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. Electronics (Basel), 9(9), 1514. https://doi.org/10.3390/electronics9091514

[14]. C. Jones (2022, January 18). 50 phishing stats you should know in 2022. 50 Phishing Stats You Should Know In 2022. Retrieved January 27, 2022, from https://expertinsights.com/insights/50-phishing-stats-you-should-know/

[15]. A. Hannousse and S. Yahiouche (2021), "Web page phishing detection", Mendeley Data, V3, doi: 10.17632/c2gw7fy2j4.3

Cite this article as :

G Dayakar Reddy, Sneha Sreelata, D Shreya, "Efficient Email Phishing Detection using Machine Learning ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 3, pp.356-360, May-June-2023.