International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2390563

Machine Learning-Based Detection of Phishing in COVID-19 Theme-Related Emails and Web Links

Usman Ali[1], Dr. Isma Farah Siddiqui[2]

[1]Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro, Pakistan

[2]Asscoiate Professor, Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro, Pakistan

ARTICLEINFO

Article History:

Accepted: 15 Sep 2023

Published: 13 Oct 2023

Publication Issue

Volume 9, Issue 5

September-October-2023

Page Number

276-285

ABSTRACT

During the COVID-19 epidemic phishing dodges increased in frequency mostly the links provided current updates about COVID-19 hence it became easy to trick the victims. Many research studies suggest several solutions to prevent those attacks but still phishing assaults upsurge. There is no only way to perform phishing attacks through web links attackers also perform attacks through electronic mail. This study aims to propose an Effective Model using Ensemble Classifiers to predict phishing using COVID-19-themed emails and Web Links. Our study comprises two types of Datasets. Dataset 1 for web links and Dataset 2 for email. Dataset 1 contains a textual dataset while Dataset 2 contains images that were downloaded from different sources. We select ensemble classifiers including, Random Forest (RF), Ada Boost, Bagging, ExtraTree (ET), and Gradient Boosting (GB). During the analysis, we observed that Dataset 1 achieves the highest accuracy rate as compared to Dataset 2 which is 88.91%. The ET classifier performs with an accuracy rate of 88.91%, a precision rate of 89%, a recall rate of 89%, and an f1 score of 89% which is better as compared to other classifiers over both datasets. Interesting concepts were found during the study.

Keywords : Component, Formatting, Style, Styling, Insert

I. INTRODUCTION

Due to infectious illness, COVID-19 culminated in a pandemic that put millions of lives at risk spread globally. Advances in technologies made it possible to spread critical information about the virus and its impacts through various sources including social media platforms i.e. FB, Twitter, YouTube healthcare websites, blogs, etc, Many attackers took advantage during the lockdown period by offering jobs, and funds, and sending fake test results. COVID-19-related domains have seen a substantial surge in

Copyright © 2023 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)** which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

popularity because of people's curiosity in determining the threat's scope and identifying protective measures [1][2]. The attacker attempts to get sensitive data from the victim through email, text, messages, or websites is known as Phishing and it is a kind of social engineering attack where attackers use email to steal data for example, bank credentials, health reports, and home address [3] [4]. Electronic mail is one of the ways to send and receives message, documents, videos, files, etc. [5]. The combination of text, special characters, and numeric, hexadecimal codes is known as Web links. Web links connect several pages of a website and communicate over the internet. Due to its complex structure, it is easy to perform hacking. [6].

In the era of AI and the capabilities of the AI field, it is informal to detect those attacks. Machine learning (ML) and Deep learning (DL) techniques proved to be efficient techniques. Due to their capabilities over analytical methods and advanced algorithms can detect phishing and real [7][8].

There a is lack of horizontal research on the Ensemble model for the detection of Phishing on COVID-19 emails and web links. The objective of this study is to Create ensemble classifiers to predict phishing using COVID-19-themed emails and Web Links. In this study, our contribution is listed below:

- We experimented with two types of Datasets. Dataset 1 for web links that contain only URLs and were collected from GitHub sources. Dataset 2 for email contains only images related to COVID-19 and was collected from a Google search.
- We applied image processing techniques to convert images into text datasets.
- We selected features from both datasets.
- We train 5- ensemble classifiers including GB, RF, ET, Adaboost, and Bagging.
- The performance was evaluated with different measuring matrices such as accuracy, precision, recall, f1 score, and confusion matrix.
- We present a comparative study among ensemble classifiers and datasets.

II. RELATED WORK

The authors performed a survey about social engineering which is the basic mechanism through which phishing attacks can be performed to hack the target internet users' vulnerabilities, which is critical in phishing instances [18].

In [19] python-based command line solution was proposed to detect spam emails by using deep learning and techniques of NLP. The LSTM model was used to detect text-based content while the MLP model was selected to detect numeric-based content. It was revealed that the LSTM model achieved the highest accuracy rate which is 99%.

Many of the attackers took advantage of COVID-19 contagion by deceiving people with fake email as well as web links to get necessary personal information, bank details, contact numbers, etc. In ref [20] the authors performed a systematic mapping study to analyze phishing attacks performed during COVID-19 and which types of phishing methods were used.

This paper uses batch and online learning methods to categorize the domain names related to Covid-19. For categorizing the domain names lexical features were used for detection. Different ML classifiers were applied, and their performance was evaluated over both learning methods. [21].

In this work [23] the phishing email detector was proposed. The detector identifies unique 26 features of email content concentrating on word counts, stop word counts, punctuation counts, and uniqueness. The detector achieved 80% of performance for phishing and 95% for ham emails.

In this work [24] the authors present a mechanism of phishing detection by executing 3-stages: DNS blacklist, using web sycophant a heuristic-based detection, and investigation. The mechanism achieved the best accuracy rate of 95.18%, 85.45%, and 78.89% for NN, SVM, and RF classifiers, respectively.

In ref [25] the author focuses on detecting and predicting phishing websites using machine learning

classifiers and ensemble-based techniques including bagging, adaboost, ET, GB, voting, and XGB over two various datasets. According to their results, ET achieves the highest accuracy rate which is 98.59%

III. METHODOLOGY

This study aims to propose an ensemble model using Classifiers to predict phishing using COVID-19-themed emails and Web Links. We experimented on two types of datasets one for web links and the other for email content and those datasets were downloaded from different sources. Our study was performed in the following steps:

Step-1: Web Links

We collected web links from GitHub sources that are only related to Covid-19. After pre-processing and analysis 17 features were selected. Algorithm-1 below shows the working of web links.

Algorithm-1: Phishing web Links detection model

Input: Phishing + legitimate URL related to Covid-19

Output: Phishing and Legitimate detection = X

Begin

1. loading CSV file
2. Features including (url_length, counting of special characters, etc.) would be extracted.
3. Splitting training and testing with ratios of 70% & 30% respectively
4. Making prediction using Ensembled Classifiers
5. Evaluate performance using Confusion Matrix and Accuracy
6. Return X

End

Step-2: Email Themed

We collected email images related to COVID-19 from the Google search engine. At first, by applied image processing techniques to convert images into text so that we have email content. We selected 5 various features to detect phishing and spam emails. algorithm-2 below depicts the working of email content detection.

Algorithm-2: Phishing Email Detection

Input: Phishing + legitimate email images related to Covid-19

Output: Phishing and Legitimate detection = X

Begin

1. Loading the raw images dataset
2. Using image preprocessing techniques translating raw images into CSV files
3. Features such as (ratio of Chars and special Chars, Body Richness, Frequencies of Unique Words, etc.) would be extracted.
4. Splitting training and testing with ratios of 70% & and 30% respectively
5. With Ensembled classifiers making predictions
6. Evaluate performance using the Confusion Matrix and accuracy.
7. Return X

Finish

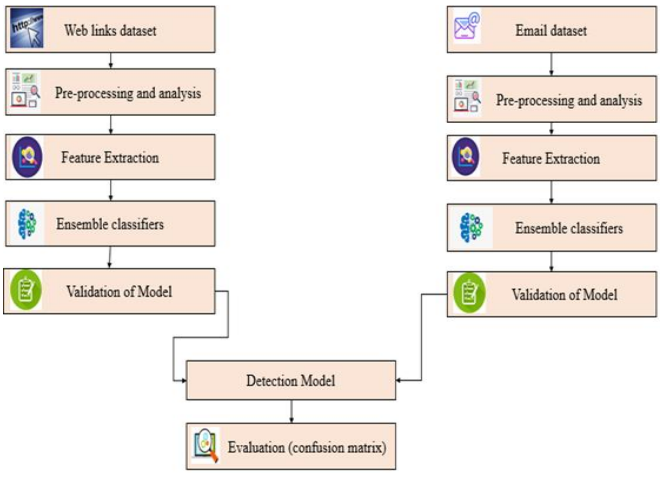

Fig. 1 : System Diagram

Fig. 1: shows the working flow of our study. We collected two types of datasets from different sources. One dataset relates to weblinks and other email images. We applied diverse pre-processing methods (handling missing data with imputation and removing duplicates). We chose diverse features based on the nature of each dataset. We trained the ensembled classifiers (Bagging, Adaboost, ET, GB, RF), and then evaluated them with measuring matrices for example, confusion matrix, accuracy, precision, recall, and f1 score.

IV. RESULT AND DISCUSSION

The purpose of this study is to ensemble classifiers to predict Phishing using COVID-19 web links and Email themes. Our study was performed in two parts:

1. Web Link Themed
2. Email Themed

Part-1: Web links

TABLE I. DATASET DESCRIPTION

Dataset	Source	Phishing	Legitimate	features
Dataset 1	Github	916	916	1832
Dataset 2	Google	197	199	396

Table I: shows the dataset's description. Dataset 1 for Web Links and Dataset 2 for emails. Dataset 1 was collected from GitHub and Dataset 2 was collected from the Search engine i.e., Google. There was a total of 1832 samples after preprocessing there was 916 phishing dataset and 916 was the ham dataset in dataset 1while there was a total of 396 email images after pre-processing, we have 197 Phishing data and 199 legitimate data in dataset 2.

TABLE II. DESCRIPTION OF FEATURES FOR DATASET 1

Ser	Features	Datatypes
1.	www	Bool
2.	url_length	Number
3.	No_of_digits	Number
4.	No_of_letters	Number
5.	Is_shortening_service	Bool
6.	Hash_url_region	Number_hash
7.	Hash_root_domain	Number_hash
8.	Is_https	Bool
9.	Count_@	Number
10.	Count_*	Number
11.	count_?	Number
12.	count_-	Number
13.	count_=	Number
14.	count_#	Number
15.	count_%	Number
16.	count_+	Number
17.	count_//	Number

Table II. describes the features that were selected after pre-processing and their datatypes whether it is Boolean or a number. We have selected 17 features for our model to detect phishing and no-phishing web links.

- **Experimental Evaluation**

We experimented on Windows 10 @ 1.80GHZ using Python programming. This study aims to create ensemble classifiers to detect phishing weblinks and email content only related to COVID-19. We split training and testing at a ratio of 70 and 30% respectively. We use the following measuring matrices:

- **Accuracy**

Percentage of all observations accurately and classified by the model. Accuracy refers to the percentage of all observations accurately classified by the model can be calculated using eq. (1) below:

$$Accuracy = \frac{TP+TN}{T} \quad (1)$$

Where TP denotes true positives, TN denotes true negatives and T denotes the total number of samples.

- **Precision**

The ratio of true positive divided by the sum of true positive and false positive can be calculated using below eq. (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Here
TP = True positives.
FP = False positives.

- **Recall**

The ratio of true positive and summation of true positive and false negative is calculated using a formula:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Where TP shows the true positive and FN shows the false negative.

- **F1 score**

The harmonic means of precision and recall.

$$F1\ score = 2 * \frac{Precision*Recall}{Precision + Recall} \quad (4)$$

TABLE III: PERFORMANCE OF ENSEMBLE CLASSIFIERS

Ensemble classifiers	Accuracy	Precision	Recall	F1 score
AdaBoost	87.64%	88%	88%	88%
Bagging	86%	86%	86%	86%
ET	88.91%	89%	89%	89%
GB	87.64%	88%	88%	88%
RF	88%	88%	88%	88%

Table III: examines the accuracy, average precision, average recall, and F1 score of five ensemble classifiers to see how well they perform. The Extra Tree classifier's maximum accuracy, precision, recall, and f1 scores are 88.91%, 89%, 89%, and 89%, respectively. The Bagging classifier has the lowest accuracy, precision, recall, and f1 score rates of all the classifiers are 86%, 86%, 86%, and 86% respectively.

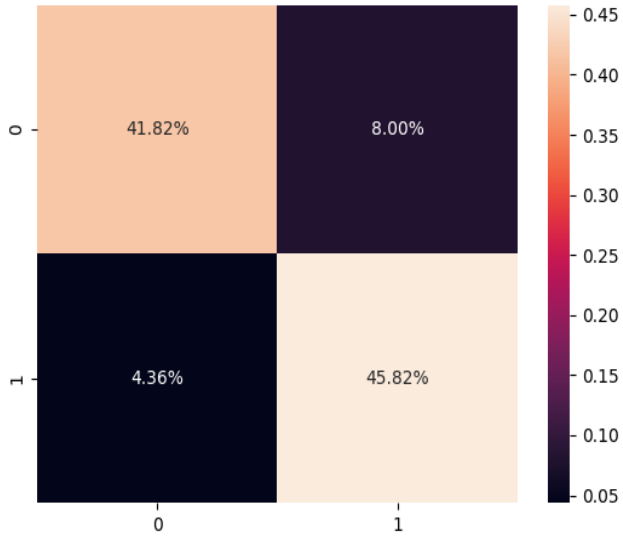

Fig. 2 : Confusion matrix for Adaboost

Fig. 2: illustrates the confusion matrix for the AadaBoost classifier. The classifier accurately predicted that is 87.64% while the classifier wrongly predicted the 12.36% samples.

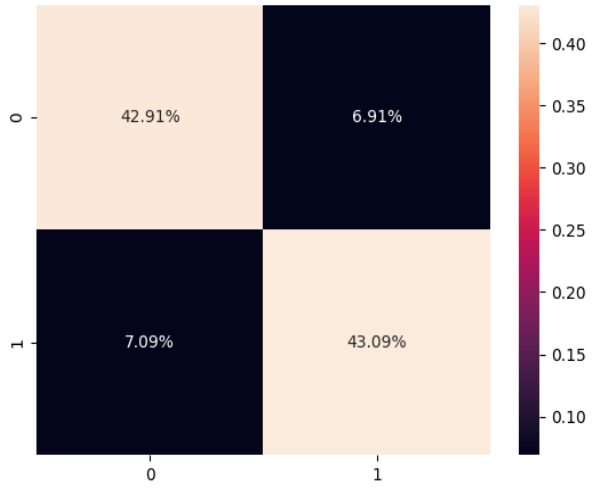

Fig. 3: confusion matrix for Bagging

Fig. 3: signifies the confusion matrix for the Bagging classifier. Among the 100% samples in total, the classifier correctly forecasted that is 86% while the classifier wrongly predicted the 14% samples.

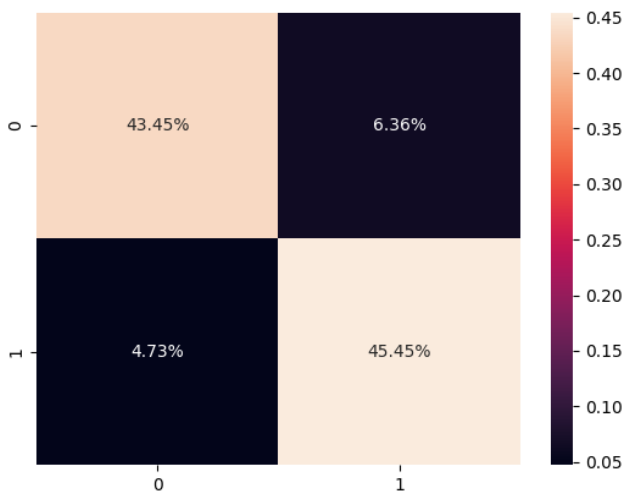

Fig. 4: Confusion matrix for ET

Fig. 4: illustrates the confusion matrix for the ET classifier. The classifier precisely predicted that is 88.91% while the classifier wrongly predicted the 11.09% samples.

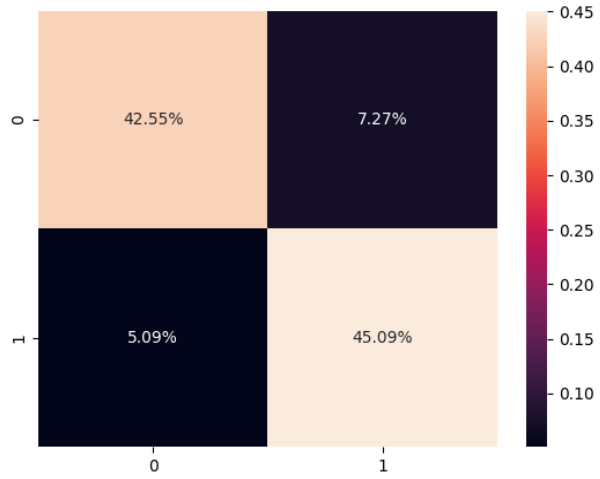

Fig. 5: Confusion Matrix for GB

Fig. 5: represents the confusion matrix for GB. There are 100% samples in total. Out of which the classifier perfectly predicted that is 87.64% while the classifier incorrectly predicted the 12.36% samples.

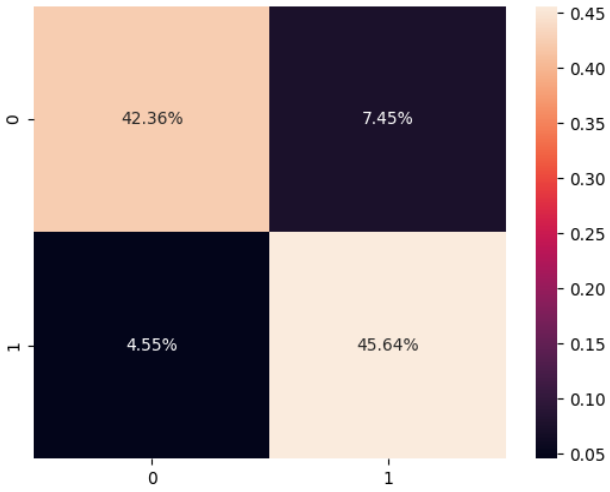

Fig. 6: confusion matrix for RF

Fig. 6: depicts the confusion matrix for the RF classifier. There are 100% samples in total. 88% of samples were accurately predicted by the classifier only 12% of samples were wrongly predicted.

Part 2: Email themed:

in our study, we deal only with email content to detect phishing emails and features selected as shown in below table IV.

TABLE IV. FEATURES SELECTED FOR DATASET 2

Ser	Features	Datatype	Description
1	Occurrence of Special Character	Continuous	Total of Special characters
2	Regularity of Characters	Continuous	Total of only characters other than special characters
3	Ratio of longest Sentence	Continuous	Count of characters of longest sentence including spaces / Count of characters of email characters including spaces
4	Word richness	Continuous	Count of Unique lemmatized words / total tokenized words in an Email
5	body Richness	Continuous	No of Words/ No of characters

TABLE V: PERFORMANCE EVALUATION OF EMAIL CONTENT

Ensemble classifiers	Accuracy	Precision	Recall	F1 score
Adaboost	79.83%	80%	80%	80%
Bagging	78.99%	79%	79%	79%
ET	83.19%	83%	83%	83%
GB	81.51%	82%	81%	81%
RF	83.19%	83%	83%	83%

Table V: describes the performance of 5- Ensemble classifiers. The highest performance rate was recorded and obtained by the RF classifier which are 83.19%, 83%, 83%, and 83% respectively. It can be observed that the lowest performance rate was achieved by the Bagging classifier which are 78.99%, 79%, 79%, and 79% respectively.

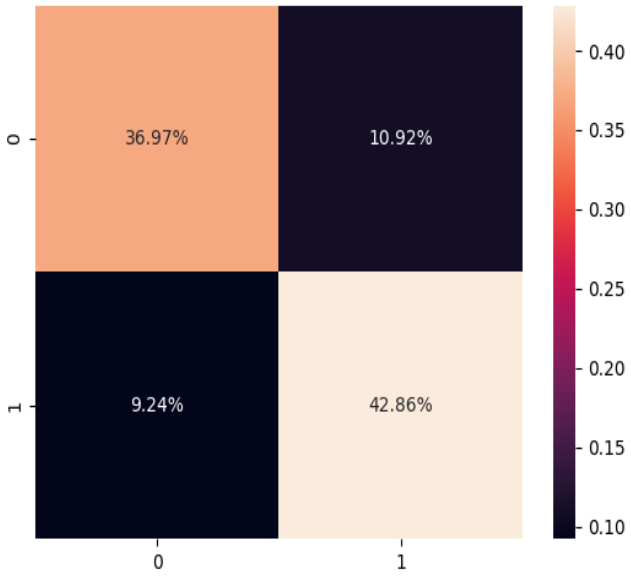

Fig.7: Performance of Adaboost

Fig.7: portrays the confusion matrix for the AadaBoost classifier. In total, we have 119 samples out of which the classifier is exactly foreseen which is 79.83% while the classifier wrongly predicted 20.16% of samples.

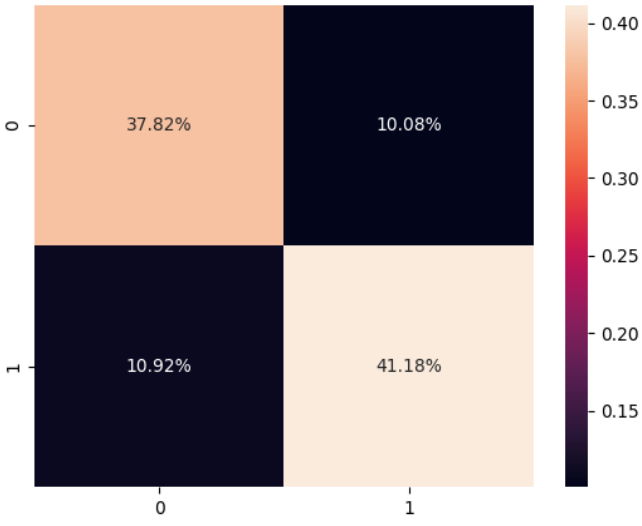

Fig. 8: performance of Bagging

Fig.8: represents the confusion matrix for the Bagging classifier. The classifier accurately predicted that is 79% of 100% samples and wrongly predicted 21% of samples.

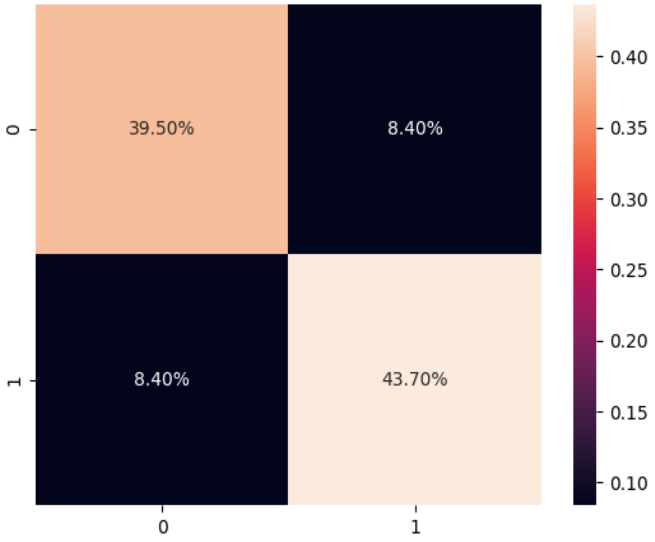

Fig. 9: performance evaluation of ET

Fig. 9: depicts the confusion matrix for the ExtraTree classifier. There are 119 samples in total. Out of which the classifier precisely predicted that is 83.2% while the classifier wrongly predicted the 16.8% samples.

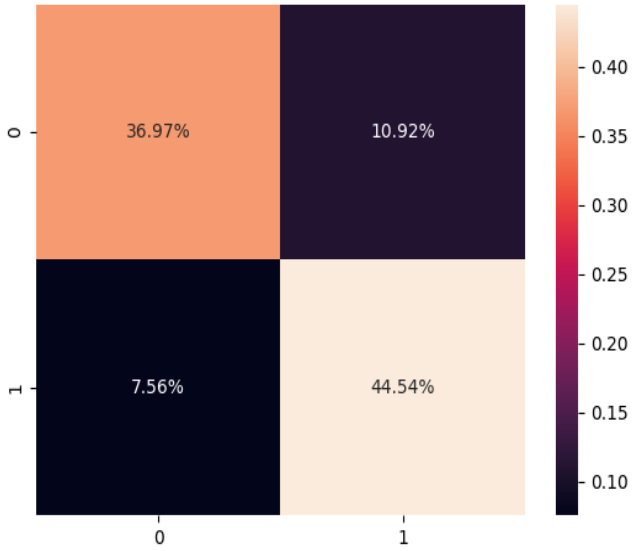

Fig.10: performance of GB

Fig. 10: represents the confusion matrix for the GB classifier. There are 119 samples in total. 81.51% of samples were correctly predicted by the classifier while 18.49% of samples were wrongly predicted.

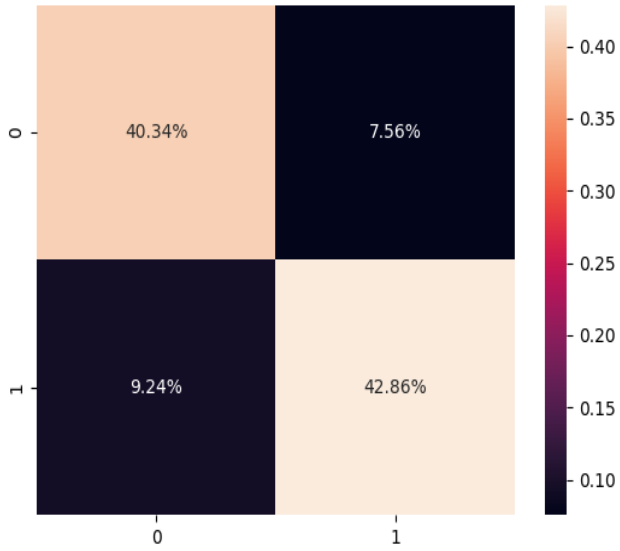

Fig. 11: performance evaluation of RF

Fig. 11: illustrates the confusion matrix for the RF classifier. The classifier correctly predicted 83.2% of the total samples that is 119 while the classifier wrongly predicted 16.8% of samples.

- **Comparative Study**

The objective of this study is to provide a comparative study among the ensembled classifiers and datasets so that we can conclude which dataset the model performs better and on which ensembled classifiers the model performs better.

TABLE VI : COMPARISON OF ENSEMBLED CLASSIFIERS OVER BOTH DATASETS

Ensemble classifiers	Accuracy of Dataset2	Accuracy of Dataset 1
AdaBoost	79.83%	87.64%
Bagging	78.99%	86%
ET	83.19%	88.91%
GB	81.51%	87.64%
RF	83.19%	88%

Table 4.6: describes the performance of 5- Ensemble classifiers in terms of Accuracy. The ET classifier achieved the highest accuracy rate and performed well over both datasets.

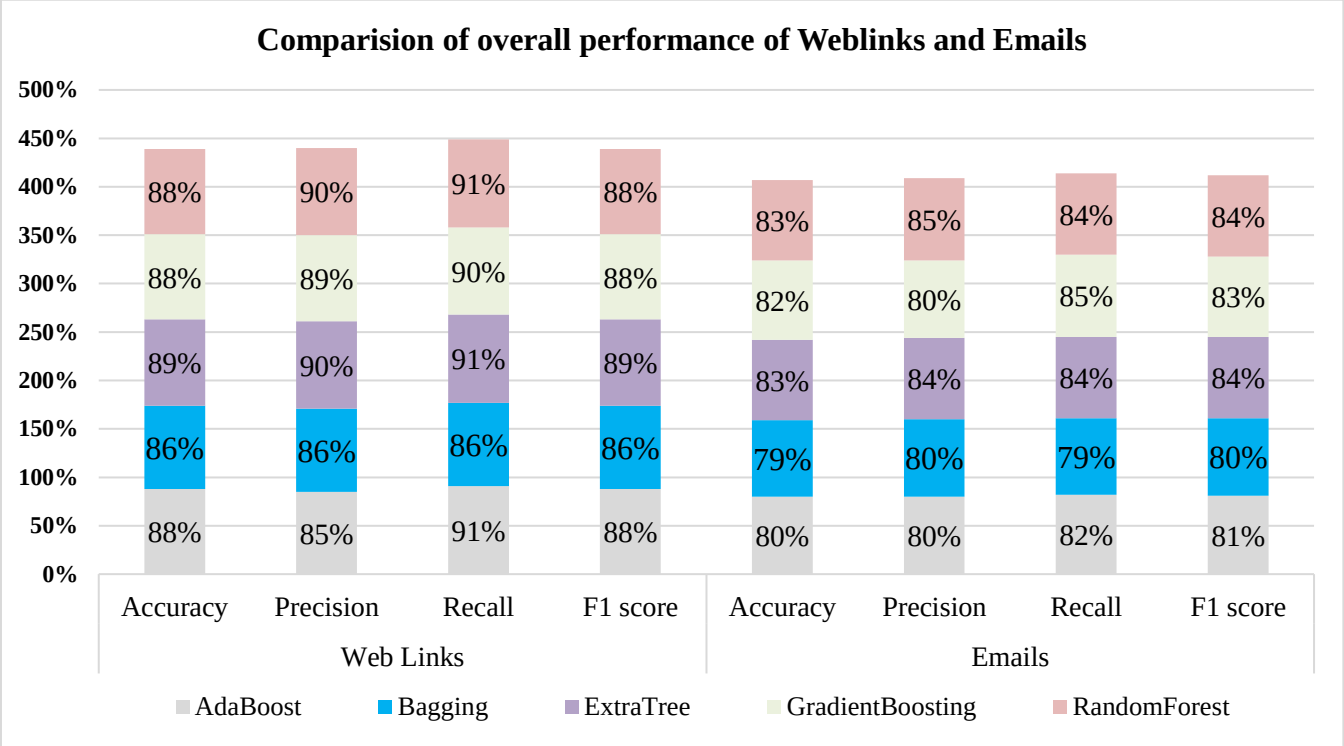

Fig. 13 : Overall Performance of ensembled classifier over both datasets

Fig. 13: depicts the overall performance of the ML algorithms over weblinks and Email. We can conclude that Weblinks achieve promising results for the detection of phishing attacks as compared to emails. Among the classifiers, ExtraTree and Random Forest perform good.

V. CONCLUSION

The threat of phishing URLs in cyber security can steal sensitive information from users. URLs can be sent via email. Phishers design phishing URLs in various ways to bypass detection techniques. Therefore, the objective of this study is to create an Ensemble model for the detection of phishing URLs and emails. Two types of datasets were selected Dataset 1 contains a description of the URLs and was downloaded from GitHub. Dataset 2 contains email images downloaded from Google search engines. We applied 5 ensemble classifiers (AdaBoost, ET, GB RF, and Bagging) were chosen. It was observed that among ensembled classifiers, ET performed well on both Datasets. It can be concluded that with web links we have higher chances of Finding phishing attacks as compared to email content.

In the future, we will explore other fields where phishing attacks are performed and will utilize deep learning techniques to detect the attacks.

VI. REFERENCES

[1]. N. A. Afandi, I. Rahmi, and A. Hamid, "Covid-19 Phishing Detection Based on Hyperlink Using K-Nearest Neighbor (KNN) Algorithm," Applied Information Technology And Computer Science, vol. 2, no. 2, pp. 287–301, 2021, doi: 10.30880/aitcs.2021.02.02.020.

[2]. A. F. Al-Otaibi and E. S. Alsuwat, "A study on social engineering attacks: phishing attack." [Online]. Available: www.ijramr.com

[3]. P. Sharma, B. Dash, and M. F. Ansari, "Anti-Phishing Techniques – A Review of Cyber Defense Mechanisms," IJARCCE, vol. 11, no. 7, Jul. 2022, doi: 10.17148/ijarcce.2022.11728.

[4]. S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey," in Procedia CIRP, Elsevier B.V., 2021, pp. 19–28. doi: 10.1016/j.procs.2021.05.077.

[5]. A. Aljofey et al., "An effective detection approach for phishing websites using URL and HTML features," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-10841-5.

[6]. J. Ispahany and R. Islam, "Detecting Malicious Urls of COVID-19 Pandemic Using ML Techniques."

[7]. Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Frontiers in Computer Science, vol. 3. Frontiers Media S.A., Mar. 09, 2021. doi: 10.3389/fcomp.2021.563060.

[8]. V. Gomes, J. Reis, and B. Alturas, "Social Engineering and the Dangers of Phishing," in 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, Jun. 2020, pp. 1–7. doi: 10.23919/CISTI49556.2020.9140445.

[9]. M. Dewis and T. Viana, "Phish Responder: A Hybrid Machine Learning Approach to Detect Phishing and Spam Emails," Applied System Innovation, vol. 5, no. 4, Aug. 2022, doi: 10.3390/asi5040073.

[10]. A. F. Al-Qahtani and S. Cresci, "The COVID-19 scamdemic: A survey of phishing attacks and their countermeasures during COVID-19," IET Information Security, vol. 16, no. 5. John Wiley and Sons Inc, pp. 324–345, Sep. 01, 2022. doi: 10.1049/ise2.12073

[11]. P. K. Mvula, P. Branco, G. V. Jourdan, and H. L. Viktor, "COVID-19 malicious domain names classification[Formula presented]," Expert Syst Appl, vol. 204, Oct. 2022, doi: 10.1016/j.eswa.2022.117553.

[12]. G. Egozi and R. Verma, "Phishing email detection using robust NLP techniques," in IEEE International Conference on Data Mining Workshops, ICDMW, IEEE Computer Society, Feb. 2019, pp. 7–12. doi: 10.1109/ICDMW.2018.00009.

[13]. G. Mohamed, J. Visumathi, M. Mahdal, J. Anand, and M. Elangovan, "An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approach," Processes, vol. 10, no. 7, Jul. 2022, doi: 10.3390/pr10071356.

[14]. A. Awasthi and N. Goel, "Phishing website prediction using base and ensemble classifier techniques with cross-validation," Cybersecurity, vol. 5, no. 1, Dec. 2022, doi: 10.1186/s42400-022-00126-9.

[15]. A. L. S. Saabith, M. Fareez, and T. Vinothraj, "Python current trend applications-an overview PYTHON CURRENT TREND APPLICATIONS-AN OVERVIEW POPULAR WEB DEVELOPMENT FRAMEWORKS IN PYTHON," International Journal of Advance Engineering and Research Development, vol. 6, no. 10, 2019, [Online]. Available: https://www.researchgate.net/publication/34456995

Cite this article as :

Usman Ali, Dr. Isma Farah Siddiqui, "Machine Learning-Based Detection of Phishing in COVID-19 Theme-Related Emails and Web Links ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 5, pp.276-285, September-October-2023. Available at doi :
https://doi.org/10.32628/CSEIT2390563
Journal URL : https://ijsrcseit.com/CSEIT2390563