Why Phishing Emails Escape Detection: A Closer Look at the Failure Points

Arifa I. Champa Md Fazle Rabbi Minhaz F. Zibran
Department of Computer Science, Idaho State University
Pocatello, ID, USA
{arifaislamchampa, mdfazlerabbi, minhazzibran}@isu.edu

2024 12th International Symposium on Digital Forensics and Security (ISDFS) | 979-8-3503-3036-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISDFS60797.2024.10527344

***Abstract*—This research uncovers why phishing emails often escape machine learning (ML) detection algorithms. For training and testing ML algorithms in detecting phishing emails, we produce and publicly release *11 curated datasets* consisting of 217,470 emails categorized and labeled as phishing and legitimate emails. Then, we perform a quantitative analysis to assess the effectiveness of five ML algorithms and confirm the suitability of our curated datasets. Through an in-depth analysis of misclassified emails, we identify patterns indicating when ML fails to detect phishing emails. These findings inform the design and development of better phishing email filtering systems while our datasets will allow further studies in this direction.**

***Index Terms*—Phishing email, data curation, machine learning, quantitative analysis, qualitative analysis, detection**

I. Introduction

Email is the most widely used communication method for both professional and personal purposes, making it a primary target for phishing attacks. In these attacks, cybercriminals create deceptive emails that appear genuine but are actually intended to trick users into giving away sensitive data. The scammers keep improving their tactics, and even individuals with advanced education are falling for these scams. Statistics show that 91% of hacking attempts start with a phishing email, and 3.4 billion of these deceptive emails are sent every day [1]. This highlights the importance of devising more effective and robust methods to detect these phishing attempts.

This work addresses the expanding problem of phishing email attacks, which are increasing in both frequency and complexity. In 2023, there are nearly 6 billion security breaches, with phishing attacks occurring approximately every 11 seconds [2]. Despite technological advancements, phishing emails continue to pose a significant threat, as cybercriminals persistently develop new scams. For instance, Gmail's filters detect millions of phishing emails, a considerable portion of which are previously unseen and entirely new [3]. Among these phishing emails, 67% have no subject line. However, when they do, common subjects include "fax delivery report" (9%) and "business proposal request" (6%). These statistics motivate us to conduct a more thorough examination of the characteristics of phishing emails, leading us to pursue a qualitative study focused on incorrectly predicted emails.

The rise of phishing attacks in the digital environment has become a growing concern, leading to extensive research efforts [4]–[9]. One major challenge in this field is the lack of high-quality, diverse, and meticulously curated datasets [10]. This inspires us to investigate the suitability of potential email repositories or datasets for the purpose of phishing email detection. We observe that existing email repositories and datasets are not readily usable for machine learning (ML) algorithms. This is due to various issues such as encryption, HTML formatting, non-English languages, and even empty email bodies. These repositories/datasets require thorough curation and preprocessing before they can be effectively utilized in ML-based approaches. In addressing the aforementioned difficulties, this paper makes the following main contributions.

- We produce and publicly release 11 phishing email *datasets* [11], [12] containing 217,470 emails sourced from nine repositories, ensuring their readiness for ML algorithm utilization. Then to verify the suitability of these datasets, we operate five prominent ML algorithms on the datasets and conduct a *quantitative* analysis.
- We conduct an in-depth *qualitative* analysis of the misclassified emails to identify distinctive features in phishing emails that mislead ML algorithms. The insights derived from the thorough qualitative analysis is the primary contribution of this work, which informs devising new techniques for capturing phishing emails.

As portrayed in Figure 1, the procedural steps of this work include data curation (Phase-1), quantitative analysis (Phase-2), and qualitative analysis (Phase-3) described respectively in Section II-B, Section III-A, and Section IV-A. Section V discusses the limitations of this work and Section VII concludes this paper with some future research directions.

II. Curation and Creation of Datasets

A. Sources of Email Collections

We utilize mainly two repositories - the Nazario Phishing Corpus (2015-2022) [13] and the Nigerian Fraud repository (1998-2007) [14] - both of which contain collections of raw phishing emails. Additionally, we incorporate curated datasets from our prior research, including the Enron corpus, the TREC Public corpora from 2005-2007 (TREC-05, TREC-06, and TREC-07), and the CEAS 2008 Challenge Lab Evaluation Corpus (CEAS-08) [15].

B. Processing for Curation

All the selected repositories require varying levels of processing for ML applications. At first, we decode/decrypt the encoded/encrypted emails using the corresponding decoding

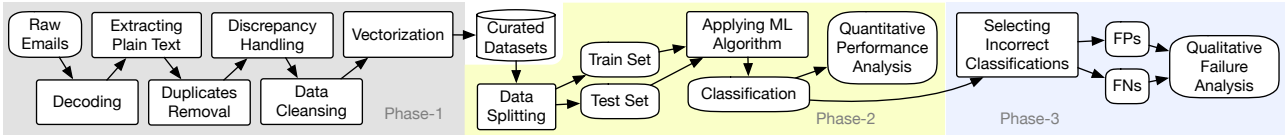

Fig. 1. Procedural steps in different phases of our work

techniques. We also manage different character sets to ensure accurate translation into readable text. We use the Python `email` library [16] for this decoding process.

We observe that the repositories have a mix of HTML-formatted and plain text emails. To retain consistency, we convert HTML-formatted emails into plain text by removing formatting tags and replacing consecutive newlines with single spaces. We then identify duplicate emails based on their identical 'Body' content and remove duplicates by keeping only one email from each group. Moreover, to handle discrepancies, we discard emails with empty bodies and focus solely on those written in English from all the selected sources.

Then, we perform data cleansing by eliminating common stop words such as 'and', 'the', and 'is' from each email to enhance dataset quality and reduce noise. Finally, to prepare the datasets for applying ML algorithms, we convert email data into numerical vectors using the most widely used statistical method known as term frequency-inverse document frequency (TF-IDF) [7], [17], [18]. Additionally, we transform the 'Urls' attribute into a binary feature, with a value of 1 indicating the presence of url(s) and 0 indicating their absence.

C. Resultant Curated Datasets

Since the Nazario Corpus and Nigerian Fraud repository contain only phishing emails, we need to incorporate legitimate emails as well to properly train ML algorithms. We randomly select 600 legitimate emails for the Nazario Corpus and 300 legitimate emails for the Nigerian Fraud repository from each of our previously curated datasets, which include Assassin, TREC-05, TREC-06, TREC-07, and CEAS-08 [15]. We then combine these legitimate emails with the curated Nazario and Nigerian Fraud datasets, creating two new datasets named Nazario-5 and Nigerian-5. Table I provides a summary of all 11 datasets we have produced and released. The first two rows identify each dataset and the release year of the original source repositories.

For each dataset, the third through fifth rows represent the number of email instances processed during decoding, duplicate removal, and discrepancy handling. The subsequent rows in Table I offer an overview of each dataset, including details such as the total number of emails, the number of legitimate and scam/phishing emails, the Legit:Scam ratio, and available features. For instance, in the Nazario dataset, we have 1,939 emails after decoding. Then, we get a total of 1,565 curated phishing emails after removing 373 duplicates and addressing discrepancies in one email. This results in a legit:scam ratio of 0:100. The Nazario dataset, the same as Assassin, TREC-05, TREC-06, TREC-07, and CEAS-08, includes six email features: 'Sender,' 'Receiver,' 'Date,' 'Subject,' 'Body,' and 'Urls.' In contrast, the Ling and Enron datasets contain only two features: 'Subject' and 'Body.'

As seen in Table I, the curated datasets Nazario-5, Nigerian-5, Enron, TREC-07, and CEAS-08 demonstrate a relatively balanced distribution of phishing and legitimate emails. In contrast, the Ling dataset is notably the most imbalanced, with a Legit:Scam ratio of 84:16. Similarly, the TREC-06 and Assassin datasets also display imbalances. In phishing email detection, real-world scenarios often involve imbalanced datasets, challenging ML algorithms. Our curated datasets vary in size, including both balanced and imbalanced ones, to challenge ML algorithms and similar analyses.

III. Quantitative Analysis

To validate the suitability of our curated datasets, we operate ML algorithms on the datasets and quantitatively measure their performances. We exclude the Nazario and Nigerian datasets because they exclusively contain phishing emails, lacking legitimate ones necessary for binary classification. Instead, we work with the extended versions, Nazario-5 and Nigerian-5, which incorporate legitimate emails, thus focusing on phishing email detection across nine curated datasets.

A. Procedure

We use five ML algorithms, Support Vector Machine (SVM), Random Forest (RF), Extra Tree (ET), XGBoost (XGB), and AdaBoost (ADB), known for their effectiveness in phishing email detection on previous research [7], [17]–[19]. Further details on these ML algorithms can be found elsewhere [20]. We employ 10-fold cross-validation using 'StratifiedKFold' to ensure that the distribution of classes in both the training and testing datasets closely reflects the distribution in the complete dataset.

For each ML algorithm on every dataset, we record the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). FN indicates phishing emails incorrectly classified as legitimate, FP represents legitimate emails misclassified as phishing, TP occurs when phishing emails are correctly labeled, and TN is when legitimate emails are correctly identified. Algorithm performance is assessed using metrics including accuracy, recall, precision, and F-score [7]. To apply ML algorithms to the curated datasets, we use all features of each curated dataset. We then train the models with default hyperparameters on the training subsets and assess their performance on the testing subsets.

B. Findings

We have assessed the performance of five ML algorithms using seven datasets in our previous work [15]. For each of those datasets, Table II displays the accuracy, precision, recall, and F-score of the best-performing ML algorithm.

TABLE I
SUMMARY OF 11 DATASETS THAT WE HAVE CURATED AND RELEASED

Dataset	Ling	Enron	Assassin	TREC-05	TREC-06	TREC-07	CEAS-08	Nigerian Fraud	Nazario	Nigerian-5	Nazario-5
Release Date	2000	2006	2002-2006	2005	2006	2007	2008	1998-2007	2015-2022	1998-2008	2005-2022
Decoded	0	30,494	6,047	92,188	37,786	75,417	137,701	3,970	1,939	-	-
Duplicates	34	724	220	29,500	20,079	19,026	70,100	633	373	-	-
Discrepancies	0	3	18	3,413	254	3	1,217	6	1	-	-
Total Emails	2,859	29,767	5,809	55,414	16,416	53,757	39,154	3,331	1,565	6,331	3,065
Legitimate Emails	2,401	15,791	4,091	32,329	12,411	24,358	17,312	0	0	3,000	1,500
Scam Emails	458	13,976	1,718	23,085	4,005	29,399	21,842	3,331	1,565	3,331	1,565
Legit:Scam	84:16	53:47	70:30	58:42	76:24	45:55	44:56	0:100	0:100	47:53	49:51
Features	Subject, Body	Sender, Receiver, Date, Subject, Body, Urls									

TABLE II
BEST-PERFORMING ML ALGORITHMS ON SEVEN CURATED DATASETS

Dataset	Best ML Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
Ling	XGB	98.95	98.96	98.95	98.93
Enron	ET	98.69	98.69	98.69	98.69
Assassin	ADB	98.79	98.79	98.79	98.79
TREC-05	ET	99.12	99.12	99.12	99.12
TREC-06	XGB	97.99	97.99	97.99	97.97
TREC-07	ET	99.85	99.85	99.85	99.85
CEAS-08	ET	99.69	99.69	99.69	99.69

In this work, we evaluate the performance of these same five ML algorithms on the extended curated datasets, Nigerian-5 and Nazario-5, and present the results in Table III. Here, the cells highlighted in green represent the best metric values achieved by the ML algorithms, while the cells highlighted in red indicate the worst metric values. For the Nazario-5 dataset, ADB achieves a flawless 100% score across all evaluation metrics. In the case of the Nigerian-5 dataset, both ET and ADB achieve a perfect score of 100% across all evaluation metrics, while the other algorithms also score above 99%. This emphasizes the exceptional quality of our curated datasets.

TABLE III
PERFORMANCE OF ML ALGORITHMS ON EXTENDED CURATED DATASETS

Dataset	Metric	SVM	RF	ET	XGB	ADB
Nigerian-5	Accuracy(%)	99.84	99.53	100.00	99.84	100.00
	Precision(%)	99.84	99.53	100.00	99.84	100.00
	Recall(%)	99.84	99.53	100.00	99.84	100.00
	F-score(%)	99.84	99.53	100.00	99.84	100.00
Nazario-5	Accuracy(%)	97.06	99.35	99.67	99.67	100.00
	Precision(%)	97.23	99.36	99.68	99.68	100.00
	Recall(%)	97.06	99.35	99.67	99.67	100.00
	F-score(%)	97.06	99.35	99.67	99.67	100.00

As observed in Table III and Table II, ET outperforms all other algorithms when handling large balanced datasets, especially those with nearly 30K instances or more, such as Enron, CEAS-08, TREC-07, and TREC-08. However, for relatively smaller datasets, whether they are balanced (Nazario-5 and Nigerian-5) or imbalanced (Ling, Assassin, and TREC-06), boosting algorithms namely, ADB and XGB show the best performance in detecting phishing emails.

IV. QUALITATIVE ANALYSIS

Although the ML algorithms exhibit high accuracy in the quantitative assessment, which validates the suitability of the curated datasets, we proceed with a qualitative investigation of the failure points that lead the ML algorithms to produce those few incorrect classifications.

A. Approach

At first, from the false classifications made by the best-performing ML algorithm, we analyze up to a maximum of 20 FPs and 20 FNs from each curated dataset, as displayed in Table IV. For each selected false prediction, we examine any complexities or pattern present within the features that may have led to incorrect classification. After identifying the failure points, we categorize similar discrepancies and complexities into subcategories, which collectively contribute to the formulation of broader predicament categories.

B. Findings

Table IV displays the ratio of false predictions (F) generated by the best-performing ML algorithm, along with the number of false positives (FPs) and false negatives (FNs) chosen for the qualitative analysis in each curated dataset. Nazario-5 and Nigerian-5, both with no false predictions, are not presented in this table. We observe that the more balanced datasets exhibit little to no false predictions, while the imbalanced datasets have a comparatively higher rate of false predictions. Despite being fairly balanced, Enron registers the second highest false prediction rate of 1.31%, while TREC-06 has the highest false prediction rate of 2.01%.

TABLE IV
NUMBER OF FP AND FN INSTANCES QUALITATIVELY STUDIED

	Ling	Enron	Assassin	TREC-05	TREC-06	TREC-07	CEAS-08
Best ML algorithm	XGB	ET	ADB	ET	XGB	ET	ET
F (%)	1.05	1.31	1.21	0.92	2.01	0.15	0.31
FN	3	13	4	20	20	1	8
FP	0	20	3	18	7	7	4

After analyzing the false predictions, we categorize the identified predicaments into nine categories. Table V presents the number of FPs and FNs contributing to the overall number of false predictions (F) within each category that affects phishing email detection.

(P1) Subject Line Characteristics: This category includes cases such as the subject line being irrelevant, casual, or non-standard, absent, and using 're' in the subject line, which contribute to 54 FNs. Whereas, subject line being short and vague, informal or sentimental, and clickbaity, tempting, or suspicious, and missing subject line contributed to 29 FPs.
For instance, the subject line below is both tempting and suspicious, which might have caused this genuine email to be marked as a phishing attempt.

`"get a $25 certificate just for responding to this e - mail"`(Email # 1669, Enron)

TABLE V
PREDICAMENTS MISLEADING ML IN PHISHING EMAIL DETECTION

Predicaments	FN	FP	F
P1: Subject Line Characteristics	54	29	83
P2: Communication Legitimacy Indicators	77	0	77
P3: Content Relevance and Coherence	43	33	76
P4: Content Formatting and Presentation	44	30	74
P5: Suspicion and Security Triggers	0	64	64
P6: Contact and Response Directives	32	26	58
P7: Tone and Linguistic Patterns	24	19	43
P8: Deceptive Techniques	43	0	43
P9: Sender Authenticity Indicators	42	0	42

Similarly, the relevance of subject line to the email body content has confused the classifier to misclassify the following example as legitimate email.

"Subject: Network+Bootcamp on Feb 28-Mar 2, 2005
Body: invites you to attend.

Network+ Bootcamp
February 28-March 2, 2005
March 14-16, 2005
M - W (9am - 5pm)
Q1 Promo: P 12,200
Regular Fee: P 19,900
Course Description

Your success in the IT ..."(Email # 13378, TREC-05)

(P2) Communication Legitimacy Indicators: The absence of obvious grammar or spelling errors, mention of external distribution, given specific reference code, copyright statement and affiliation notices, and reference to legislation can give an email a sense of legitimacy, which can lead to misclassification. Moreover, the inclusion of technical or academic terminology, legal and financial terms, privacy assurance, and reputable references also play a role in 77 FN predictions.

"Involved in Fundraising?
Here is the answer ... All marketing expenses
are met by Amex, and there is no breaching of
Privacy ..." (Email # 11025, TREC-06)

For instance, the privacy assurance mentioned in the preceding example contribute to the failure of the ML algorithm to correctly identify this phishing email.

(P3) Content Relevance and Coherence: The aspects including personal messages referring to past conversations, combination of both legitimate content as well as phishing elements, blend of fictional content along with scientific terms, mimicry of typical automated system messages, and discussion of spam-related issues contributed to misclassification by ML algorithms. Moreover, composition of various phrases and sentences, nonsensical content or word salad, lack of context, gibberish text in email also contributed to a total of 43 FNs.

"Subject:mail system error - returned mail
Body: this message was undeliverable due to
the following reason : the user (s) account
....."(Email # 10070, Enron)

The email in the above example mimics typical automated system messages, thereby confusing the ML classifiers and causing them to identify this phishing email as a legitimate one. Additionally, emails with a lack of personalization, detailed information, contextual information, or mentioning a privacy policy contributed to 33 FPs.

"unsubscribe
Thierry VOINIER
CNRS-LMA-IM
31, Chemin Joseph Aiguier
13402 Marseille, France
Tel : 04 91 16 44 73
Fax : 04 91 22 08 75
E-mail : voinier ... (Email # 4640, TREC-06)

To illustrate, the lack of contextual and detailed information led the algorithm mistakenly classify the preceding legitimate email as a phishing email.

(P4) Content Formatting and Presentation: The genuine-looking or informative content or offer, legitimate-looking structure, job advertisement or newsletter format, promoting investments or stocks, adult content lead to 44 FNs. For example, the below genuine-looking invitation to a lunch party confused the ML algorithm, misidentifying it as legitimate.

"networking lunch
chesapeake , bay golf club in north east
wednesday , november 17 th
11 : 15 am

small companies ..."(Email # 14905, Enron)

On the contrary, emails having fragmented content, multiple embedded email addresses, lack of human-readable content, gibberish content, unusual or non-standard formatting, long list of quotes without much context, sparse content, and mentions of HTML encoding give rise to 30 FPs. For example, the unusual and context-lacking content in the following email led to the misclassification of this genuine email.

"Lots of jobs at EEI.
http://www.eei.org/careers/openings.htm#sdea

Sue Mara
164 Springdale Way
Emerald Hills, CA 94062
Cell: (415) 902-4108
Home: (650) 369-8268"(Email # 33659, TREC-05)

(P5) Suspicion and Security Triggers: Various email aspects can trigger suspicion and security concerns such as the presence of clickable text, URLs, suspicious or sensitive keywords, unknown file formats, repetitive phrasing, explicit sharing of credentials, X-Authentication-Warnings, complex metadata, random numbers at the end, promotional content, offers of services, monetary rewards, gifts, prizes, promises, benefits lead to misidentification of phishing emails by ML algorithms. Moreover, characteristics like insufficient information, requests for personal information, mentions of credit card or financial or business transactions, technical boundaries and encodings, and requests for action contributed to 64 FPs. To illustrate, the presence of clickable text and the presence of URL in the above email content, misguide ML algorithm to classify it as phishing email.

"Bicycling's Maintenance Repair Guide
Dear Spie,
Inside this revised popular guide, you...
satisfied. Click here!
Sincerely, ... (Email # 14614, TREC-07)

(P6) Contact and Response Directives: The several factors of an email related to contact and response instructions namely the inclusion of footer, an alternate email, contact details,

contact address, unsubscribe or opt-out information play a part to 32 FNs. For instance, the following email has a contact details provided that misdirect the algorithm to predict it as legitimate.

"FOR IMMEDIATE RELEASE
PROS Revenue Management Announces Record Third Quarter and YTD 2001
November 20, 2001 ...
Contact:
Candy Haase - VP Marketing
713-335-5253 / 713-335-8144 - fax
chaase@prosRM.com" (Email # 40349, TREC-05)

However, the inclusion of controversial or negative information, contact information or address, subscribe links, and mismatched date contributed to 26 FPs. For instance, inclusion of subscription link in the following email mislead the ML algorithm to identify this legitimate email as phishing.

"**Ashfield Online (C) 2002 Aric McKeown**
Comic ..... To subscribe to a new list or to resubscribe to this list if you are unsubscribed, go to
http://www.keenspot.com/subscribe.html "
(Email # 4049, Assassin)

(P7) Tone and Linguistic Patterns: A business-like or professional tone, emotional appeal, casual or personal tone, use of unusual or nonstandard language, incorrect or hard to decode contents in emails contribute to 24 FNs. For example, the following email has a business-like or professional tone that guide the algorithm mistakenly classify this phishing emails as legitimate.

"Charles Schwab & Co., Inc.
Email Alert
Mutual Fund Viewpoint(TM)
==========

Dear American Century Shareholder,
As a valued ..."(Email # 116937, TREC-05)

Conversely, emotionally charged content, informal tone, instructional tone, pressure, urgency, and flattery has been found responsible to misguide the classifier to 19 FPs. On the other hand, emails with emotionally charged content, like the following email, contributed to FPs.

"Remember me ? I'm the one who loved you so much that I gave my very best at raising you. and now I'm missing ..."(Email # 1647, Enron)

(P8) Deceptive Techniques: The absence of URLs or attachments, absence of typical phishing phrases or keywords, no sense of urgency, coercion, or threat, no request for sensitive information or action, and obfuscation of suspicious URLs are used in phishing emails. These deceptive techniques contribute to 43 FNs. For instance, the absence of URL in the following email causes the algorithm to fail to correctly identify this phishing email even though it has other phishing indicators.

"ATTENTION,
I have been waiting for you ... your money but ... deposited the $1.500.000.00 Million USA Dollars ..." (Email # 32353, CEAS-08)

(P9) Sender Authenticity Indicators: This category deals with various indicators that aid in giving a false impression of authenticity of the sender of an email. These indicators include the email address matching the expected sender, random and generic sender, and consistency in domain. These characteristics contributed to 42 FNs. To illustrate, the consistency in the sender domain in the following example lead the ML algorithm mistakenly classify this phishing email as legitimate.

"Innovation Thinktank<editors@innovationthinktank.org>" (Email # 35851, TREC-05)

The qualitative analysis of FPs and FNs highlights the various aspects, from the subject line characteristics to the lack of information and personalization, contribute to the misclassification of emails. Phishing emails often mimic legitimate emails in terms of tone, language use, and content, making it even more challenging for algorithms to accurately classify them. Moreover, legitimate emails sometimes contain characteristics commonly associated with phishing emails, leading to FPs. It is recommended that future work in phishing email detection should focus on developing more sophisticated algorithms that can better understand the nuances of language, tone, and content, and incorporate a broader range of indicators to accurately classify phishing emails.

V. Threats to Validity

Our curated datasets include emails written in English only. We simply consider the presence or absence of URLs. A thorough examination of the complete URL links, including their domain, structure, and other characteristics, may reveal signs of phishing attempts.

In our work, we utilize five well-known ML algorithms on our curated datasets, none of which are deep learning models. It is noteworthy that these ML models have consistently scored 98% or higher accuracy across our curated datasets. Moreover, our primary objective has been to create high-quality datasets and understand the issues that mislead ML algorithms. However, we do have plans to investigate the potential of deep learning algorithms in future research.

VI. Related Work

ML and NLP techniques have been applied to several research domains [21]–[23], including security components in software systems [24], [25]. Likewise, recent efforts to detect phishing emails have used ML techniques such as SVM, RF, Decision Tree, AdaBoost, XGBoost, K-Nearest Neighbors [6], [19], [26] and NLP techniques [4]. Some studies have focused on single algorithms [5], [6] while others have evaluated multiple ones [7], [19] and analyzed email content [27]–[29]. Many have tried to enhance the detection of phishing emails by raising suspicion [8], understanding IT experts' recognition methods [9], investigating attack tactics and motives [29].

Phishing emails typically exhibit a significant volume disparity compared to legitimate emails, and thus a 1:1 ratio no longer considered the norm by many researchers [30]. Das et al. [10] investigated the complexities of email phishing and spear-phishing, highlighting the importance of comprehensive and high-quality curated datasets. In our research, we have carefully compiled and made available a total of 11 phishing email datasets, having various sizes and distributions.

Jakobsson et al. [31] examined trust indicators in emails and web pages, exploring how phishing content can appear authentic while legitimate content may seem suspicious to users. Kikerpill and Siibak [32] analyzed only 42 phishing emails, concentrating on content and tactics, centered on psychological manipulation. Ferreira et al. [33] developed a method to detect persuasive elements in phishing emails, with a specific focus on email subject lines. In contrast, our objective has been to identify points of failure of ML algorithms by analyzing all features in a total of 128 false predictions, encompassing both legitimate and phishing emails.

In our earlier work [7], we applied six ML algorithms on only two datasets for phishing email detection. In a follow-up work [15], we curated seven datasets and conducted feature importance analysis by applying five ML algorithms to the datasets. In this work, we have extended our curated datasets to make it have a total 11 datasets, and we carry out an in-depth qualitative analysis of the reasons for misclassifications of five ML algorithms we have applied to the datasets.

VII. Conclusion

In this paper, we present a comprehensive analysis of phishing email tactics and the effectiveness of ML algorithms in detecting them. Our approach involves producing 11 curated datasets [11] followed by quantitative validation with performance assessment of five ML algorithms and then an in-depth analysis of causes for misclassifications. All of the five ML algorithms operated on our curated datasets achieve very high accuracy, precision, recall, and F-score. This indicates that our curated datasets are suitable for the application of ML in phishing email detection or similar purposes.

Our qualitative analysis reveals that scammers are crafting emails that closely resemble authentic communication. They frequently mimic the tone, language, and format of legitimate emails, making it increasingly difficult to identify them solely based on content. Moreover, the relevance of the subject line with email body content, use of technical or academic terminology, privacy assurances, imitation of standard automated system messages, obfuscation of suspicious URLs, and lack of typical phishing phrases or keywords contribute to the inaccurate detection of phishing emails.

The findings from both the quantitative and the qualitative analyses highlight the importance of refining ML algorithms along with understanding how scammers manipulate human psychology. In the future, we will work on addressing the limitations of our current work. Additionally, we will investigate the effectiveness of our curated datasets when applied to deep learning algorithms.

References

[1] MimeCast, *How to Stop Phishing Attacks (Whitepaper)*. https://www.mimecast.com/resources/white-papers/how-to-stop-phishing-attacks/, Verified: Sept 2023.

[2] "The latest 2023 phishing statistics (updated august 2023)," 2023.

[3] N. James, "81 phishing attack statistics 2023: The ultimate insight," Verified: Sept 2023.

[4] R. Al-Yozbaky and M. Alanezi, "Detection and analyzing phishing emails using nlp techniques," in *HORA*, 2023, pp. 1–6.

[5] W. Pan, J. Li, L. Gao, L. Yue, Y. Yang, L. Deng, and C. Deng, "Semantic graph neural network: a conversion from spam email classification to graph classification," *Scientific Programming*, vol. 2022, pp. 1–8, 2022.

[6] R. Ibrahim, M. Argungu, and I. Mungadi, "Development of an ensemble classification model based on hybrid filter-wrapper feature selection for email phishing detection," *International Journal of Computer and Systems Engineering*, vol. 17, no. 9, pp. 519–523, 2023.

[7] M. Rabbi, A. Champa, and M. Zibran, "Phishy? detecting phishing emails using ml and nlp," in *SERA*, 2023, pp. 77–83.

[8] M. Harris and D. House, "Invoking suspicion for improved accuracy in phishing email identification," 2023.

[9] R. Wash, "How experts detect phishing scam emails," *PACM HCI*, vol. 4, pp. 1–28, 2020.

[10] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, "Sok: a comprehensive reexamination of phishing research from the security perspective," *Commun Surv Tutor*, vol. 22, no. 1, pp. 671–708, 2019.

[11] *11 Phishing Email Datasets*. https://doi.org/10.6084/m9.figshare.25437178.v1, 2024.

[12] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Curated datasets and feature analysis for phishing email detection with machine learning," in *ISDFS*, 2024, pp. 1–6 (to appear).

[13] J. Nazario, *The online phishing corpus*. http://monkey.org/jose/wiki/doku.php, Verified: August 2023.

[14] D. Radev, *CLAIR collection of fraud email, ACL Data and Code Repository*. https://aclweb.org/aclwiki, 2008.

[15] A. Champa, M. Rabbi, and M. Zibran, "Curated datasets and feature analysis for phishing email detection with machine learning," in *ICMI*, 2024, pp. 1–7 (to appear).

[16] Python. (2023) email — an email and mime handling package. [Online]. Available: https://docs.python.org/3.8/library/email.html

[17] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A new english/arabic parallel corpus for phishing emails," *TALLIP*, vol. 22, pp. 1–17, 2023.

[18] M. Islam, M. Al Amin, M. Islam, M. Mahbub, M. Showrov, and C. Kaushal, "Spam-detection with comparative analysis and spamming words extractions," in *ICRITO*, 2021, pp. 1–9.

[19] Y. Murti and P. Naveen, "Machine learning algorithms for phishing email detection," *JLISS*, vol. 10, no. 2, pp. 249–261, 2023.

[20] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2022.

[21] S. M. Mahedy Hasan, M. F. Rabbi, A. I. Champa, and M. A. Zaman, "An effective diabetes prediction system using machine learning techniques," in *ICAICT*, 2020, pp. 23–28.

[22] A. I. Champa, M. F. Rabbi, S. M. Mahedy Hasan, A. Zaman, and M. H. Kabir, "Tree-based classifier for hyperspectral image classification via hybrid technique of feature reduction," in *ICICT4SD*, 2021, pp. 115–119.

[23] A. I. Champa, M. F. Rabbi, M. F. Zibran, and M. R. Islam, "Insights into female contributions in open-source projects," in *MSR*, 2023, pp. 357–361.

[24] A. Champa, M. Rabbi, F. Eishita, and M. Zibran, "Are we aware? an empirical study on the privacy and security awareness of smartphone sensors," in *SERA*, 2023, pp. 1–8.

[25] M. F. Rabbi, A. I. Champa, C. Nachuma, and M. F. Zibran, "Sbom generation tools under microscope: A focus on the npm ecosystem," in *ACM SAC*, 2024, pp. 1–9 (to appear).

[26] P. Dinesh, M. Mukesh, B. Navaneethan, R. Sabeenian, M. Paramasivam, and A. Manjunathan, "Identification of phishing attacks using machine learning algorithm," in *E3S Web of Conf.*, vol. 399, 2023, p. 04010.

[27] Y. Teo, "Phishing attack detection using machine learning techniques," Ph.D. dissertation, TAR UMT, 2023.

[28] B. Ampel, Y. Gao, J. Hu, S. Samtani, and H. Chen, "Benchmarking the robustness of phishing email detection systems," 2023.

[29] D. Bera, O. Ogbanufe, and D. Kim, "Towards a thematic dimensional framework of online fraud: An exploration of fraudulent email attack tactics and intentions," *Decision Support Systems*, p. 113977, 2023.

[30] M. Zareapoor and K. Seeja, "Feature extraction or feature selection for text classification: A case study on phishing email detection," *IJIEEB*, vol. 7, no. 2, p. 60, 2015.

[31] M. Jakobsson, A. Tsow, A. Shah, E. Blevis, and Y.-K. Lim, "What instills trust? a qualitative study of phishing," in *Financial Cryptography and Data Security*, 2007, pp. 356–361.

[32] K. Kikerpill and A. Siibak, "Living in a spamster's paradise: Deceit and threats in phishing emails," *Masaryk UJL Tech.*, vol. 13, p. 45.

[33] A. Ferreira and S. Teles, "Persuasion: How phishing emails can influence users and bypass security measures," *Intl. J. Human Comp. Interaction*, vol. 125, pp. 19–31, 2019.