Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 11, Issue 2, pp: (45-59), Month: May - August 2024, Available at: www.noveltyjournals.com

# Challenges of Data Collection and Preprocessing for Phishing Email Detection

Obianuju N. Mbadiwe[1], Obi C. Nwokonkwo[1], Anthony I. Otuonye[1], Charles O. Ikerionwu[1], Chukwuemeka Etus[1]

[1]Department of Information Technology, Federal University of Technology, Owerri, Imo State, Nigeria.

**DOI:** https://doi.org/10.5281/zenodo.12651047

**Published Date: 04-July-2024**

***Abstract:*** **Phishing remains a pervasive threat in the realm of cybersecurity, necessitating effective detection mechanisms to safeguard individuals and organizations from malicious attacks. However, the successful implementation of phishing email detection systems hinges upon the availability of high-quality datasets and meticulous preprocessing techniques. This paper, through a comprehensive systematic review of literature delves into the challenges encountered in data collection and preprocessing for phishing email detection. It aims at shedding light on the complexities involved in obtaining reliable datasets and refining raw data for analysis. A careful analysis of facts presented in this study reveals such challenges as complexities in feature extraction, data imbalance, and ethical concerns present substantial obstacles in acquiring dependable and high-quality datasets for phishing detection systems. Creating standardized protocols for gathering and preprocessing data, while advocating for transparency and accountability in research methodologies, are suggested potential solutions for addressing these challenges which can guarantee enhanced the robustness and efficacy in phishing email detection systems.**

***Keywords:*** **Phishing datasets, Data collection, Data preprocessing, Phishing email, Data imbalance, Machine learning.**

## I. INTRODUCTION

Detecting phishing emails has become an increasingly vital task in the realm of cybersecurity, given the rising sophistication of cyber threats and the significant risks posed to individuals, organizations, and even entire economies. Phishing attacks, which aim to deceive recipients into divulging sensitive information such as login credentials or financial details, continue to evolve in complexity and prevalence, making them a formidable challenge for both users and security systems alike. Machine learning methods is increasingly being adopted to address this escalating threat, as they present an opportunity to analyze extensive datasets and detect patterns indicative of malicious behavior [1].

The effectiveness of phishing email detection systems heavily relies on the quality and quantity of data used for training and evaluation. However, the collection and preprocessing of such data present numerous challenges that can significantly impact the performance and reliability of detection algorithms. The ever-changing landscape of phishing attacks, coupled with the immense influx of emails, poses challenges in curating a comprehensive and current dataset suitable for training machine learning models [2]. This paper investigates the multifaceted obstacles encountered in the process of gathering and preparing datasets specifically tailored for phishing email detection.

One of the primary challenges lies in acquiring authentic phishing emails for analysis. Phishing emails are often dispersed sporadically across various platforms and communication channels, making systematic collection a daunting task. Furthermore, the dynamic nature of phishing campaigns necessitates continuous updates to the dataset to reflect emerging tactics and trends. As highlighted by research conducted by [3], the lack of standardized repositories for phishing email datasets exacerbates this issue, hindering reproducibility and comparability across studies.

Moreover, the preprocessing stage introduces additional complexities due to the inherent variability and obfuscation techniques employed by attackers. Cleaning and structuring raw email data while preserving relevant features pose significant computational and methodological challenges. For instance, techniques such as feature extraction and dimensionality reduction, as discussed by [4], are essential for enhancing the efficiency and effectiveness of detection algorithms but require careful consideration of trade-offs between information loss and computational overhead.

In light of these challenges, this paper aims to provide a comprehensive overview of the hurdles faced in collecting and preprocessing data for phishing email detection. By identifying key obstacles and proposing potential solutions, this research contributes to the advancement of cybersecurity practices and the development of more robust detection mechanisms. Through a synthesis of existing literature and empirical insights, we endeavor to offer actionable recommendations for researchers and practitioners striving to enhance the efficacy of phishing email detection systems in an ever-evolving threat landscape.

## II. BACKGROUND AND RELATED LITERATURE

Section 2.1 of this chapter provides background on Data Collection and Preprocessing Challenges in Phishing Email Detection. Section 2.2 follows with a review of related work.

### A. Data Collection and Preprocessing Challenges

Collecting and preprocessing relevant and representative data is a major determinant in developing effective and reliable detection mechanism for phishing email and in the fight to secure our cyberspace.

#### i.Data Collection Challenges

One primary data collection challenge lies in the dynamic nature of phishing attacks, characterized by rapidly evolving tactics and strategies employed by cybercriminals. Phishing campaigns often vary in their objectives, targeting, and methodologies, making it challenging to capture a comprehensive dataset that encompasses the full spectrum of malicious behaviors. As highlighted by [5], the lack of standardized repositories for phishing email datasets exacerbates this issue, hindering the establishment of a common benchmark for evaluation and comparison across different detection systems.

The sheer volume of emails being sent daily adds another layer of complexity to the data collection process. Identifying and isolating phishing emails from legitimate ones amidst the deluge of messages flooding inboxes requires robust filtering mechanisms and automated tools [6]. However, indiscriminate sampling may inadvertently bias the dataset, leading to skewed representations of phishing patterns and reducing the efficacy of detection algorithms. The importance of carefully balancing sample diversity with data integrity to ensure the robustness and generalizability of machine learning models trained on such datasets cannot be over emphasized.

Furthermore, acquiring authentic phishing emails for analysis poses significant ethical and legal considerations. While some datasets are publicly available, obtaining consent from individuals or organizations whose data may be included in these emails is essential to ensure compliance with privacy regulations. Additionally, ensuring the confidentiality and security of sensitive information contained within phishing emails is paramount to prevent inadvertent exposure or misuse. The work by [7] underscores the importance of ethical guidelines and protocols for data collection in cybersecurity research to mitigate potential risks and safeguard the privacy rights of individuals.

A significant obstacle lies in the inherent imbalance and bias within existing phishing email datasets. The overwhelming majority of email traffic consists of legitimate correspondences, complicating efforts to assemble a balanced dataset that accurately reflects the prevalence of phishing emails [2]. Also, the data often exhibits biases, such as the disproportionate representation of certain types of phishing attacks, which can significantly undermine the performance of machine learning models and their ability to generalize to novel threats [8].

Addressing these challenges in data collection is paramount for improving the efficacy of machine learning models in detecting and mitigating the risks associated with phishing attacks. In the forthcoming sections of this paper, we will delve into innovative methodologies and potential remedies to tackle these obstacles, with the aim of bolstering the resilience and dependability of phishing email detection systems.

**ii.Preprocessing Techniques for Email Data Sets**

Preprocessing plays a crucial role in preparing email data sets for effective analysis and modeling in phishing email detection systems. By cleaning, transforming, and enhancing raw email data, preprocessing techniques aim to improve the quality, usability, and performance of machine learning models. Here, we are looking at various preprocessing techniques commonly employed in the context of email data sets, along with their implications and contributions to the field.

Text Cleaning and Normalization: Text cleaning involves removing irrelevant characters, symbols, and formatting artifacts from email bodies to enhance readability and facilitate subsequent analysis. Techniques such as removing HTML tags, punctuation, and special characters help to standardize the text and reduce noise in the dataset. Additionally, text normalization processes, including stemming and lemmatization, ensure consistency in word forms and reduce the dimensionality of the feature space, thereby improving model efficiency [9].

Tokenization and Feature Extraction: Tokenization involves breaking down email text into individual tokens or words, enabling the extraction of meaningful features for analysis. Feature extraction techniques such as bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF) transform tokenized text into numerical representations, capturing the importance of words in distinguishing between phishing and legitimate emails. These techniques facilitate the creation of feature vectors that serve as input for machine learning algorithms, enabling effective classification and detection [10].

Email Header Parsing: Email headers contain metadata such as sender information, timestamps, and routing details, which can provide valuable contextual information for phishing email detection. Preprocessing techniques for email headers involve parsing and extracting relevant fields, such as the sender's domain, IP address, and email client information. Analyzing header attributes can help identify suspicious senders, detect anomalies in email routing, and enhance the accuracy of phishing detection algorithms [11].

Feature Engineering and Selection: Feature engineering involves creating new features from existing ones or transforming raw data into more informative representations. In the context of email data sets, feature engineering techniques may include extracting domain-specific features such as URL presence, attachment types, and language identifiers. Moreover, feature selection methods such as information gain and chi-squared tests help identify the most discriminative features for distinguishing between phishing and legitimate emails, reducing model complexity and improving generalization [12].

Handling Imbalanced Data: Imbalanced data sets, where one class (e.g., phishing emails) significantly outnumbers the other (e.g., legitimate emails), pose challenges for machine learning algorithms, leading to biased predictions and reduced performance. Preprocessing techniques for handling imbalanced data include oversampling, under-sampling, and synthetic data generation. These techniques aim to rebalance the class distribution in the dataset, ensuring that the model learns from representative examples of both classes and improves its ability to detect rare events [13].

**iii.Ethical Considerations in Data Collection**

The collection of data for phishing detection research raises several ethical considerations that must be carefully addressed to ensure the protection of individuals' privacy rights and adherence to ethical standards. As we take a critical look at key ethical concerns surrounding data collection in the context of phishing detection and highlight the importance of implementing appropriate safeguards, the following will be considered;

a) *Informed Consent:* Obtaining informed consent from individuals whose email data is used for research purposes is paramount to uphold ethical principles. Transparency in data collection processes fosters trust and respect for participants' autonomy [14].

b) *Data Anonymization and Privacy Preservation:* Researchers should anonymize or use pseudonyms for email data by removing or encrypting personally identifiable information (PII) to prevent the identification of individuals [15].

c) *Minimization of Harm:* Phishing email datasets may contain malicious content or links that could pose risks to individuals' cybersecurity and privacy. Therefore, researchers should exercise caution when handling and analyzing such data to prevent inadvertent exposure to harmful content [16].

d) *Fair and Responsible Use of Data:* Researchers should refrain from using collected data for purposes other than those explicitly stated in informed consent agreements and should obtain appropriate approvals for any secondary uses of data [16].

e) *Community Engagement and Stakeholder Involvement:* Engaging with the community and involving relevant stakeholders, such as cybersecurity experts, legal professionals, and privacy advocates, can provide valuable insights and perspectives on ethical considerations in data collection [2].

f) *Bias and Fairness:* Researchers should strive to mitigate biases by employing representative sampling techniques, ensuring diversity in the dataset, and conducting bias audits to identify and address potential sources of bias [17].

g) *Responsible Disclosure and Transparency:* Researchers should consider disclosing vulnerabilities or weaknesses identified during the data collection process to relevant stakeholders, such as email service providers or cybersecurity organizations, to facilitate timely mitigation and remediation efforts [16].

h) *Cultural Sensitivity and Contextual Awareness:* Researchers should seek input from local communities and stakeholders to ensure that data collection practices align with cultural expectations and values [18].

i) *Accountability and Governance:* Establishing clear accountability mechanisms and governance structures is essential to oversee and regulate data collection activities effectively [19].

j) *Long-Term Impact and Sustainability:* There is the need to strive to minimize the environmental footprint of data collection activities, such as reducing energy consumption and waste generation associated with data storage and processing, to promote environmental sustainability and responsible stewardship of resources [6].

**iv.Mitigating Bias in Training Data for Phishing Detection**

Biased training data can lead to skewed model predictions, reinforcing existing disparities and hindering the effectiveness of detection and prediction systems. Addressing bias in training data is essential for developing accurate and reliable phishing detection models. We will explore strategies and techniques for mitigating bias in training data for phishing detection, along with their implications.

a) *Diverse Data Collection:* By capturing a wide range of phishing tactics, targets, and contexts, researchers can minimize the risk of bias stemming from underrepresentation or overrepresentation of certain types of attacks. [2].

b) *Bias-Aware Sampling Techniques:* Techniques such as oversampling of minority classes, under-sampling of majority classes, and synthetic data generation help address disparities in class representation and ensure that the model learns from diverse examples of phishing and legitimate emails [20].

c) *Feature Engineering and Selection:* Feature selection methods such as information gain, mutual information, and feature importance analysis help identify the most discriminative features while mitigating the influence of biased or noisy features [21].

d) *Fairness-aware Algorithms:* Fairness-aware algorithms explicitly incorporate fairness considerations into the model training process to mitigate bias and promote equitable outcomes [22].

e) *Continuous Monitoring and Evaluation:* Continuous monitoring and evaluation of model performance are essential for identifying and mitigating bias throughout the model lifecycle [23].

f) *Cross-Validation and Holdout Sets:* Employing cross-validation techniques and holdout sets can help assess the generalization performance of phishing detection models while mitigating bias. By systematically validating model performance across diverse subsets of the data, researchers can identify and address biases that may arise from specific training-validation splits [24].

g) *Ensemble Learning and Model Diversity:* Ensemble learning techniques, such as bagging, boosting, and stacking, combine multiple base models to improve predictive performance and mitigate bias [25].

h) *Transfer Learning and Domain Adaptation:* By leveraging pre-trained models or representations learned from related tasks or domains, researchers can mitigate biases in training data and improve model generalization [26].

i) *Community Engagement and Participatory Research:* By soliciting input from diverse perspectives and incorporating domain knowledge from stakeholders, researchers can mitigate biases, address community concerns, and enhance the social impact and acceptance of phishing detection technologies [27].

### B. Related Works

In their review paper, [28] suggested that the choice of datasets utilized by researchers significantly influences the credibility of their models when testing and training. Additionally, while the focus of the papers may be on phishing email detection, some researchers utilize malware and spam emails for training and testing purposes. This introduces complexities as personal data sources may be included in the research, leading to potential privacy concerns. The disclosure of information regarding these sources varies, with some being kept private by the authors and others being made public. According to the authors, various approaches rely on ground truth datasets obtained from diverse cyber intelligence sources, each employing different testing and evaluation methodologies. As these sources target different types of phishing activities, there is a notable contrast between evaluations relying on one dataset compared to another. Consequently, there is ongoing debate regarding the necessity of publicly available reference datasets to classify different evaluation approaches. Such reference datasets could serve as benchmarks for contrasting the efficacy of different approaches and facilitate systematic improvements in the field. Ultimately, having access to a standardized reference dataset could streamline the process for analysts to enhance phishing detection methodologies in a more systematic manner. In a separate review paper, [29] noted that certain researchers commonly gather spam email data using spam-bots, which are automated applications designed to scour email addresses across the Internet. [30] analyzed 35 widely recognized cyber datasets, categorizing them into seven distinct groups. These categories encompass Internet traffic-based, network traffic-based, Intranet traffic-based, electrical network-based, virtual private network-based, android apps-based, IoT traffic-based, and Internet-connected device-based datasets. It's important to highlight that dealing with data imbalance is a prevalent challenge in the detection of phishing emails, necessitating suitable strategies for handling it. Therefore, addressing the issue of data imbalance involves ensuring the effective categorization of minority classes [31]. [32] proposed system outlined in their paper involves the efficient extraction of data from web log data. This process utilizes web usage mining techniques to extract features indicative of user behavior. Additionally, the system incorporates URL analysis for feature extraction to detect phishing website addresses. The infiltration of bots continues to pose significant challenges to data integrity. Moreover, researchers encounter ethical dilemmas regarding the inadvertent misuse of research funds when they unintentionally pay for bot responses during data sourcing [33]. Upon closer scrutiny of several criminology articles, it became evident that only a small number of researchers explicitly discussed the issue of informed consent or raised other ethical considerations linked to the data collection process. This lack of attention may stem from uncertainty regarding whether data obtained through automated software originates from human subjects. Researchers intending to utilize automated data collection tools must meticulously assess the privacy implications to make informed judgments about privacy in the online environments they aim to investigate. To guide such decisions, some scholars advocate that assumptions concerning privacy should align with the norms prevalent within the community under investigation [34]. In order to address the impact of data imbalance on model performance and improve the detection accuracy of phishing emails, [25] introduces two novel algorithms incorporating under-sampling techniques: The Fisher–Markov-based phishing ensemble detection (FMPED) method and the Fisher–Markov–Markov-based phishing ensemble detection (FMMPED) method. These algorithms initially eliminate benign emails in overlapping regions, then under-sample the remaining benign emails, and subsequently merge the retained benign emails with phishing emails to form a new training dataset. Ensemble learning algorithms are employed for training and classification purposes. Experimental findings indicate that the proposed algorithms surpass the performance of alternative machine learning and deep learning algorithms. Notably, they achieve an F1-score of 0.9945, an accuracy of 0.9945, an AUC of 0.9828, and a G-mean of 0.9827. It is important to point out that normalizing feature values before implementing low variance filtering helps prevent unnecessary bias stemming from irregularities in the data. Also, from a statistical standpoint, eliminating randomization via bagging offers an advantage in terms of bias reduction [35]. In accordance with [36] findings, employing k-fold cross-validation is typically recommended to address the challenges posed by imbalanced datasets. This method involves iteratively changing validation and training data samples, offering a robust approach to handle data imbalances effectively.

## III. RESEARCH OBJECTIVE AND METHODOLOGY

The objective of this work are to offer a complete overview of the challenges encountered while collecting and preparing data for phishing email detection. It seeks to identify major problems and provide feasible solutions.

### A. Methodology

The systematic literature review is a research procedure that adheres to a set of guidelines, and this study employs the approach proposed by [37]. The review technique entails developing research questions, defining a list of electronic resources to investigate, data collecting and data analysis, and recommendations. This study will begin by developing research

questions and then listing the databases utilized for email phishing detection. The technique also comprises searching the databases for keywords, applying inclusion-exclusion criteria, interpreting the results, and drawing conclusions.

**i.The Research Questions**

Q1: What are the data collection challenges in phishing email detection and what is the most common challenge?

Q2: What are the data sources used by the researchers to collect phishing/legitimate emails and which is the most commonly used?

Q3: Which preprocessing techniques are most commonly used for preprocessing email phishing datasets?

Q4: To what extent can bias be mitigated in datasets for email phishing detection?

**ii.The Relevant Documents for the Review**

The databases used to deliver relevant results in this paper are selected based on the keywords. Some of the databases explored for this review include;

IEEE Explore, Elsevier, ACM Digital Library, Springer, and others.

**iii.Sources of Review Papers**

*a)* Journals

*b)* Conference Proceedings.

*c)* Researchers' theses.

*d)* Published Reports

*e)* Review Articles

*f)* Webpages

*g)* Books Chapters

**iv.Essential keywords for the Study**

The database search was conducted between February and April 2024. No criteria were set for the publication date. The database search yielded 48 papers. This is because there are fewer research work on email phishing when compared to some other forms of phishing especially website phishing. A keyword search carried out on a group of paper downloaded from online sources combined in a Microsoft Word document revealed some useful information about the papers as illustrated the word cloud in Fig.1 below. The Word Cloud approach displays how closely connected the articles are based on the topical linkage theme. Typically, word clouds are used for summarizing text content. In a word cloud, the larger and bolder the term, the more frequently and significantly it appears.


**FIG. 1: WORD CLOUD FOR THE KEYWORDS OF THE SELECTED RESEARCH ITEMS.**

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 11, Issue 2, pp: (45-59), Month: May - August 2024, Available at: www.noveltyjournals.com

**v.Inclusion and exclusion criteria**

The inclusion–exclusion criteria were used at three levels. Unrelated papers are eliminated after each stage or level. The initial search focused on papers from the fields of computer science and engineering. However, because the term ‘‘data” is interdisciplinary, articles from other fields were made the list, but such types of papers were excluded from the study. Only English-language papers were eligible for inclusion. The systematic review included research publications published over a period of ten years, between January 2014 and February 2023. The same research papers from multiple libraries are discarded. After the initial exclusion, 115 papers were selected but later 53 articles were selected and included in the literature based on the selected keywords.

**vi.Quality evaluation**

Certain criteria were used to determine the quality of the papers included in the literature and these were;

*a)* Papers with clearly stated objective(s) of study.

*b)* Papers with well-defined context and experimental design of study.

*c)* Papers with adequately documented research process.

*d)* Papers that have the main findings fully stated.

*e)* Papers whose conclusions are clearly relatable to the aim of the study.

## IV. RESULTS

This section mainly presents results of systematic review of email phishing datasets and responses to the research questions. Most of the answers to the research question are synthesized form TABLE I.

**TABLE I: RESULTS OF SYSTEMATIC REVIEW OF EMAIL PHISHING DATASETS**

| S/N | Reference | Dataset | Source | Data Collection Challenges | Preprocessing Technique | Bias Mitigation Technique | Remarks |
|---|---|---|---|---|---|---|---|
| **1.** | [38] | This version of the DMOZ dataset comprises 1,562,978 website URLs spanning 15 categories. | https://github.com/UTCID/DMOZ-Privacy-Policy-Corpus-CODASPY21 | Duplicate data and lack of maintenance. identifying true privacy policies among the candidates. | Text cleaning, normalization. | - | The most difficult aspect of constructing the corpus lies in identifying true privacy policies among the candidates. These policies are often lengthy and complex, making them difficult for their intended audience to understand. Consequently, users rarely invest the time and effort to read them thoroughly so compliance is affected. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **2.** | [39] | Data set comprises 8266 instances, 47 features - 4150 phishing and 4116 ham E-mail types. | khonji's anti-phishing website, http://khonji.org/phishing_studies.html | Not updated. | Text cleaning, normalization. | Feature engineering -Gain Ratio (GR) feature ranking. | Balanced dataset. Has no missing values. Manually curated.<br><br>An ensemble of C4.5 and CART achieved 99.11% accuracy suggesting that a balanced datasets could achieve better results than unbalanced datasets. |
| **3.** | [40] | ACM IWSPA 2018<br>Subtask A: only body emails, 8913 legitimate and 1087 phishing<br>Subtask B: Header + body emails, 7781 legitimate and 496 phishing. | Legitimate emails were gathered from different sources like collections from WikiLeaks archives, some emails from the Enron Dataset6 and also SpamAssassin7. The phishing emails in the dataset were collected from the Information technology departments of various universities. emails from the popular Nazario's phishing corpora8 were included as well. Some of the emails were synthetic emails created by the organizers. | Unbalanced Dataset. Challenges due to the highly diverse nature of emails. | Normalization and Email Header Parsing. | Various Teams used different methods. | Unbalanced dataset. Collecting datasets and preprocessing them for both subtasks proved to be notably challenging, with particular emphasis on the Header Subtask. This aspect necessitated impeccably clean and comprehensive headers, a requirement that numerous datasets failed to meet.<br><br>The No-header Subtask achieved the highest F1 score of 83.54%, while the Header Subtask achieved the highest F1 score of 96.8%, indicating that the header may contain a wealth of important features for detecting phishing emails. |
| **4.** | [41] | Generated dataset comprising 30,000 samples, evenly distributed between phishing and legitimate webpages, each accounting for 50 percent of the total. | Different sources which included PhishTank, Alexa, DMOZ, and BOTW. | Not Stated or implied. | Manual inspection. | Not stated. | An attempt at benchmarking phishing datasets - developed a substantial standard offline dataset that is accessible for download, universally applicable, and thorough in scope. This principle could also be applied to email spoofing. |
| **5.** | [42] | Dataset consisting of 659,673 emails having 613,048 legitimate and 46,525 phishing emails. | Live emails actively received by WestPac and their customers during the year 2007. | semi-automatic classification of live emails. Unbalanced dataset. | Normalization. | 10-fold cross-validation. | Unbalanced dataset. Decision tree produced the highest accuracy of over 99%. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **6.** | [43] | Manually collected emails numbering 7315 emails. Also another dataset having 6047 emails (4951 legitimate and 1096 phishing). | PhishingCorpus and SpamAssassin corpus respectively. | Time taken for the manual collection and classification. Unbalanced dataset. | Text cleaning. | Patch training. | Unbalanced dataset from SpamAssassin. The mean accuracy recorded is 98.6% with Neural Network & Reinforcement Learning. |
| **7.** | [44] | Dataset comprises of 8266 emails, 47 features having phishing and ham. 4156 legitimate, and 4110 phishing emails. | From black Hat DC 2009 and Jose Nazario Phishing corpus. http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus | Not stated or implied. | Feature Engineering and Selection. | Not stated. | There is no missing value in this dataset. Utilizing this dataset, the Bayesian network classification model achieves the highest test accuracy of 99.32% when combined with CART in an ensemble. |
| **8.** | [45] | PhishBench Benchmark Dataset: comprises 10,500 genuine and 10,500 phishing emails, featuring unaltered headers originating from various sources. | The legitimate email samples: 6,779 from Wikileaks archives, 718 from Hacking Team, 3,098 from DNC, 1,066 from GI files, 1,120 from Sony, 678 from National Socialist Movements, 88 from Citizens Commission On Human Rights and 11 from Plum emails. Also 2,046 from Enron dataset and 1,675 from SpamAssassin.<br><br>The phishing email sample: 9,481 from the Nazario 7, and Nazario 2015 - 2017 phishing email datasets, also included are 1,019 spam emails from SpamAssassin. | Observed as sifting datasets from diverse sources. | Text Cleaning and Normalization. | Ensuring a diverse dataset source. | A balanced and comprehensive phishing benchmarking dataset. PhishBench offers a convenient platform for the research community to run their models and evaluate their progress against that of others who have utilized widely shared email datasets. This still needs to be periodically updated to retain its relevance. Achieved the highest accuracy of 99% with Deep Learning. |
| **9.** | [46] | 12 legitimate and 12 phishing emails. | A meticulously and systematically assembled collection of both phishing and legitimate emails. | No challenge. Dataset was manually created for experimental purpose. | Text Cleaning, Tokenization and Normalization. | No Bias. | Balanced dataset. Highest accuracy of 91.6% with Random Forest classifier. |
| **10.** | [47] | Two datasets from the landing page of APWG and phishing e-mails reported by users to APWG; i. from September 2008 - November 2009 | Anti-Phishing Working Group (APWG) | Manual feature analysis is time consuming and some emails presented in different language. | Not stated. | Not stated. | Phishing e-mails have evolved substantially over time since phishers have embraced new strategies like delivering marketing e-mails as well as emotionally targeting |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | ii. from January 2014 - April 2014<br><br>320 unique phishing domains in 2008 dataset and 1,893 in 2014 dataset. | | | | | people into clicking phishing URLs thereby increasing the number of victims.<br><br>Also collecting datasets from recent emails will improve accuracy. |
| **11.** | [48] | Dataset comprised of 8266 emails. 4133 phishing emails and 4133 legitimate emails. | The apache SpamAssassin public corpus. URL: http://spamassassi.apache.org/publiccorpus | Publicly existing dataset, challenge not stated. | Text cleaning and feature extraction technique. | A hybrid of information gain and genetic algorithm in feature engineering. | Balanced dataset from a publicly available email corpus; achieved 98.9% accuracy rate via a hybrid approach. |
| **12.** | [49] | 4000 emails of which 2000 are legitimate and 2000 are phishing. | Emails targeted at the University of North Dakota's email service. | Redundant emails, duplicate emails. | Standard text cleaning. | Balancing the datasets. | A balanced dataset. Achieved accuracy of 92.9% with ANN algorithm. |
| **13.** | [50] | Dataset containing of 500 authentic emails and 500 phishing emails. | The phishing emails are from https://monkey.org/~jose/phishing/ While the legitimate emails come from CSDMC2010. | One of the sources https://monkey.org/~jose/phishing/ Reported as unsafe. | Text cleaning. | Balancing datasets from different sources. | Balanced dataset. Accuracy of 95.0% with SVM. |
| **14.** | [51] | Dataset having 300 ham emails and 300 spam emails. | Not clearly stated. | Manual collection from a bankrupt company may involve permission issues. | Feature Extraction (TF-IDF). | Balancing datasets from different sources. | Balanced dataset. It is inappropriate for researchers not to state their data sources because this tends to lower trust in the study. |
| **15.** | [52] | This dataset has 37,055 email samples (17,902 phishing and 19,153 legitimate). | phishing data from Millersmile and legitimate emails from the Enron corpus. | Identifying elements that reflect the effectiveness of phishing and quantifying them within the phishing email sample. | Text Cleaning, Normalization and Tokenization. | 10-fold cross-validation. | Balanced dataset. 96.52% accuracy with support vector machine model. |
| **16.** | [53] | Dataset consisting of 10,606 emails (4,150 legitimate and 6,456 phishing emails.) | SpamAssassin project, and PhishingCorpus. | Phishload has raw web-based coding structures. Some of the e-mails may contain extensive HTML structural codes. Evaluating this dataset to determine if such is present can be time consuming. Emails may be corrupt. | Text Cleaning and Normalization. | Combining datasets from various sources through multiple iterations. | Balanced dataset. The ANN model achieved the best accuracy of 98.39%. |

**A. Q1: The data collection challenges in phishing email detection and the most common challenge.**

Information on TABLE II were extracted from TABLE I. Please note that some authors presented more than one challenge while some authors did not state any challenge.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 11, Issue 2, pp: (45-59), Month: May - August 2024, Available at: www.noveltyjournals.com

**TABLE II: PHISHING EMAILS DATA COLLECTION CHALLENGES**

| | Duplicate Data | Unbalanced Dataset | Non-Update of Dataset | Sorting High Volume of Emails | Email Language Issue | Permission Issue | Corrupt Site/Email |
|---|---|---|---|---|---|---|---|
| **Number of examined papers with these challenges** | 2 | 3 | 1 | 7 | 1 | 1 | 2 |

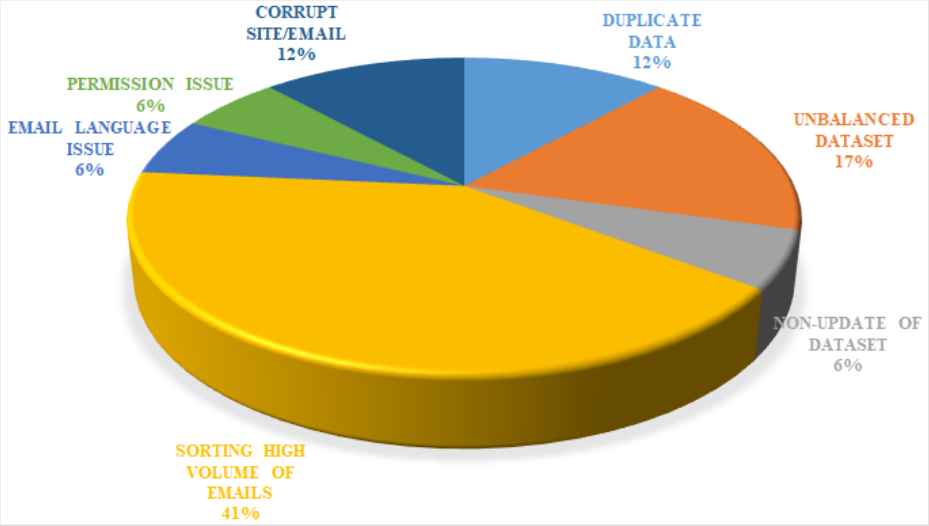


**FIG. 2: PERCENTAGE OF EXAMINED PAPERS WITH DIFFERENT CHALLENGES**

FIG. 2 has answered Research Question 1. Observe from FIG. 2 that sorting of high volume of email Corpus is the most common challenge in collecting email phishing datasets.

**B. Q2: Data sources used by the researchers to detect phishing/legitimate emails and the most commonly used.**

We extracted the information for FIG. 3 from TABLE I. Please note that some authors presented more than one data source while some authors did not state their data source. FIG. 3 below shows that the most common data sources are non-public sources followed by SpamAssassin.

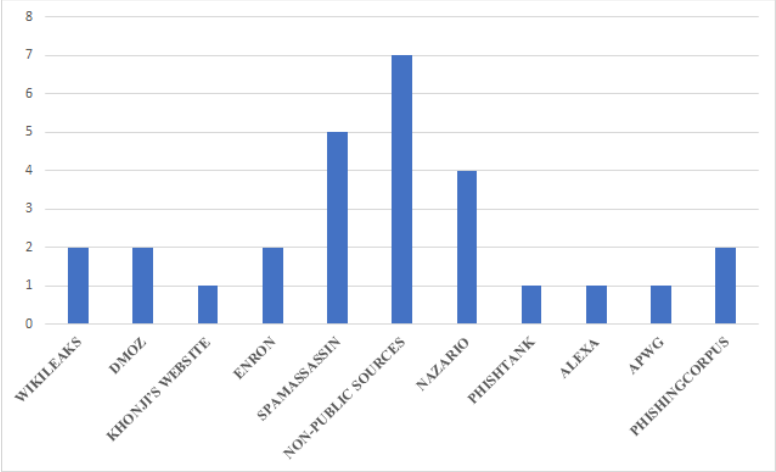


**FIG. 3: COMMON SOURCES OF EMAIL DATASETS**

### C. Q3: Preprocessing techniques that are most commonly used for preprocessing email phishing datasets.

Observation from the Preprocessing Technique Column of TABLE I suggests that Text Cleaning and Normalization are the most commonly used data preprocessing techniques for email phishing datasets.

### D. Q4: To what extent can bias be mitigated in datasets for email phishing detection?

Section II A number 4 outlines various techniques for bias mitigation in datasets and the "bias mitigation technique" column of TABLE I clearly illustrates that authors have practically applied some of these techniques to mitigate bias in email phishing datasets. Also observing from the last column of TABLE I, it can be inferred that bias mitigation in phishing email datasets corrects the anomaly in unbalanced datasets such that the accuracy levels of models built with both balanced and unbalanced seem to be at par.

## V. CONCLUSION AND RECOMMENDATIONS

In conclusion, the challenges associated with collecting and preprocessing data for phishing email detection are multifaceted and require careful consideration. Sorting high email volumes is the highest, though not insurmountable challenge and at such feature extraction complexities demand innovative approaches to effectively capture relevant information from raw email data. Data imbalance poses a significant hurdle, necessitating the development of techniques such as under-sampling and ensemble learning to mitigate its impact on model performance. Ethical considerations, including issues surrounding informed consent and data privacy, must be prioritized to uphold ethical standards and maintain trust with participants.

Moving forward, researchers and practitioners in the field of phishing email detection should collaborate to address these challenges collectively. Establishing standardized protocols for data collection and preprocessing, along with promoting transparency and accountability in research practices, can foster advancements in the field. Additionally, ongoing efforts to enhance dataset quality, mitigate biases, and prioritize privacy concerns are crucial for the development of robust and reliable phishing detection systems. By embracing interdisciplinary collaboration and leveraging emerging technologies, the cybersecurity community can effectively combat the evolving threat landscape posed by phishing attacks.

## REFERENCES

[1] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Mirco, "On the Effectiveness of Machine Learning and Deep Learning Algorithms for Cyber Security," in Arc2018 10th International Conference on Cyber Conflict. Tallinn, Estonia., T. Minárik, R. Jakschis, and L. Lindström, Eds., Tallinn, 2018, pp. 371–390. doi: 10.23919/CYCON. 2018.8405026.

[2] V. Zeng, S. Baki, A. El Aassal, R. Verma, L. Felipe, and T. De Moraes, "Diverse Datasets and a Customizable Benchmarking Framework for Phishing," IWSPA '20, March 18, 2020, New Orleans, LA, USA, no. Section 3. pp. 35–41, 2020. doi: 10.1145/3375708.3380313.

[3] H. Al-Hamadi et al., "A Novel Protocol for Security of Location Based Services in Multi-agent Systems," Wirel. Pers. Commun., vol. 108, pp. 1841–1868, 2019.

[4] S. A. Alsenan, I. M. Al-Turaiki, and A. M. Hafez, "Auto-KPCA: A Two-Step Hybrid Feature Extraction Technique for Quantitative Structure-Activity Relationship Modeling," IEEE Access, vol. 9, pp. 2466–2477, 2021, doi: 10.1109/ACCESS.2020.3047375.

[5] Y. Al-Hammadi, A. H. Al-Bayatti, M. N. Al-Kabi, and N. B. Anuar, "A review on phishing email datasets," in Proceedings of 6th International Conference on Cyber Security and Privacy in Communication Networks (Cyber Security). Patna, India., 2020, pp. 111–116.

[6] R. M. Verma, V. Zeng, and H. Faridi, "Poster: Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets," Proc. ACM Conf. Comput. Commun. Secur., pp. 2605–2607, 2019, doi: 10.1145/3319535.3363267.

[7] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, "Ethical issues in research using datasets of illicit origin," Proc. ACM SIGCOMM Internet Meas. Conf. IMC, vol. Part F1319, pp. 445–462, 2017, doi: 10.1145/ 3131365.3131389.

[8] H. Zuhair and A. Selamat, "Phishing classification models: issues and perspectives," Int. J. Digit. Enterp. Technol., vol. 1, no. 3, p. 219, 2019, doi: 10.1504/ijdet.2019.10019065.

[9] N. Muthukrishnan, F. Maleki, K. Ovens, C. Reinhold, B. Forghani, and R. Forghani, "Brief History of Artificial Intelligence," Neuroimaging Clin. N. Am., vol. 30, no. 4, pp. 393–399, Nov. 2020, doi: 10.1016/J.NIC.2020.07.004.

[10] N. Majumdar, S. Shukla, and A. Bhatnagar, "Survey on applications of internet of things using machine learning," Proc. 9th Int. Conf. Cloud Comput. Data Sci. Eng. Conflu. 2019, no. January, pp. 562–566, 2019, doi: 10.1109/CONFLUENCE.2019.8776951.

[11] M. D. Behera et al., "Species-level classification and mapping of a mangrove forest using random forest—utilisation of aviris-ng and sentinel data," Remote Sens., vol. 13, no. 11, 2021, doi: 10.3390/rs13112027.

[12] V. S. Mohan, J. R. Naveen, R. Vinayakumar, and K. P. Soman, "A.R.E.S: Automatic rogue email spotter," in R. Verma, A. Das (eds.): Proceedings of the 1st AntiPhishing Shared Pilot at 4th ACM International Workshop on Security and Privacy Analytics (IWSPA 2018), Tempe, Arizona, USA., 2018, pp. 57–63. [Online]. Available: http://ceur-ws.org

[13] A. Jain et al., "Overview and Importance of Data Quality for Machine Learning Tasks," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York; NY; United States, 2020, pp. 3561–3562. doi: 10.1145/3394486.3406477.

[14] P. Finn and M. Jakobsson, "Designing ethical phishing experiments," IEEE Technol. Soc. Mag., vol. 26, no. 1, pp. 46–58, 2007, doi: 10.1109/MTAS.2007.335565.

[15] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," Telecommun. Syst., vol. 76, no. 1, pp. 139–154, 2021, doi: 10.1007/s11235-020-00733-2.

[16] G. Yu, W. Fan, W. Huang, and J. An, "An Explainable Method of Phishing Emails Generation and Its Application in Machine Learning," in Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020. Chongqing, China., 2020, pp. 1279–1283. doi: 10.1109/ITNEC48623.2020.9085171.

[17] K. Crawford et al., "AI Now 2019 Report," New York, 2019.

[18] A. Mislove and C. Wilson, "A Practitioner's Guide to Ethical Web Data Collection," in The Oxford Handbook of Networked Communication, 2020, pp. 538–558. doi: 10.1093/oxfordhb/9780190460518.013.27.

[19] L. Floridi, J. Cowls, T. C. King, and M. Taddeo, "How to Design AI for Social Good: Seven Essential Factors," Sci. Eng. Ethics, vol. 26, no. 3, pp. 1771–1796, 2020, doi: 10.1007/s11948-020-00213-5.

[20] S. M. H. Mirsadeghi, H. Bahsi, R. Vaarandi, and W. Inoubli, "Learning From Few SDN Cyber-Attacks: Addressing The Class Imbalance Problem in Machine Learning-based Intrusion Detection in Software-Defined Networking," IEEE Access, vol. PP, p. 1, 2023, doi: 10.1109/ACCESS.2023.3341755.

[21] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique," in Proceedings - 2020 1st International Conference of Smart Systems and Emerging Technologies, SMART-TECH 2020. Riyadh, Saudi Arabia, 2020, pp. 43–46. doi: 10.1109/SMART-TECH49988.2020.00026.

[22] H. Yang, Z. Liu, Z. Zhang, C. Zhuang, and X. Chen, "Towards Robust Fairness-aware Recommendation," in Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023. Singapore., 2023, pp. 211–222. doi: 10.1145/3604915.3608784.

[23] A. Chouldechova, "Fair Prediction with Disparate Impact - A Study of Bias in Recidivism Prediction Instruments," Big Data, vol. 5, no. 2, pp. 153–163, 2017, doi: https://doi.org/10.1089/big.2016.0047.

[24] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," in Proceedings - 6th International Advanced Computing Conference, IACC 2016. Bhimavaram, India., 2016, pp. 78–83. doi: 10.1109/IACC.2016.25.

[25] Q. Qi, Z. Wang, Y. Xu, Y. Fang, and C. Wang, "Enhancing Phishing Email Detection through Ensemble Learning and Undersampling," Appl. Sci., vol. 13, no. 15, pp. 1–19, 2023, doi: 10.3390/app13158756.

[26] W. Xu, J. He, and S. Yanfeng, "Transfer Learning and Deep Domain Adaptation," in IntechOpen, 2016, p. 13. [Online]. Available: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics

[27] V. K. Ahuja, H. K. Rosser, A. Grover, and M. Hale, "Phish Finders - Improving Cybersecurity Training Tools Using Citizen Science," in ICIS 2022 Proceedings. Copenhagen, Denmark., 2022. [Online]. Available: https://aisel.aisnet.org/icis2022/security/security/1

[28] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," IEEE Access, vol. 10, pp. 65703–65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

[29] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges," Secur. Commun. Networks, vol. 2022, 2022, doi: 10.1155/2022/1862888.

[30] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," J. Inf. Secur. Appl., vol. 50, p. 102419, Feb. 2020, doi: 10.1016/J.JISA.2019.102419.

[31] J. Doshi, K. Parmar, R. Sanghavi, and N. Shekokar, "A comprehensive dual-layer architecture for phishing and spam email detection," Comput. Secur., vol. 133, p. 103378, Oct. 2023, doi: 10.1016/J.COSE.2023.103378.

[32] A. J. Obaid, K. K. Ibrahim, A. S. Abdulbaqi, and S. M. Nejrs, "An adaptive approach for internet phishing detection based on log data," Period. Eng. Nat. Sci., vol. 9, no. 4, pp. 622–631, 2021, doi: 10.21533/pen.v9i4.2398.

[33] M. Griffin et al., "Ensuring survey research data integrity in the era of internet bots," Qual. Quant., vol. 56, no. 4, pp. 2841–2852, 2022, doi: 10.1007/s11135-021-01252-1.

[34] R. Brewer, B. Westlake, O. Arauza, and T. Hart, "The Ethics of Web Crawling and Web Scraping in Cybercrime Research: Navigating Issues of Consent, Privacy, and Other Potential Harms Associated with Automated Data Collection," in Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches, 2021, pp. 435–456. doi: 10.1007/978-3-030-74837-1_22.

[35] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: review and approaches," Artif. Intell. Rev., vol. 53, no. 7, pp. 5019–5081, 2020, doi: 10.1007/s10462-020-09814-9.

[36] I. Ul Hassan, R. H. Ali, Z. Ul Abideen, T. A. Khan, and R. Kouatly, "Significance of Machine Learning for Detection of Malicious Websites on an Unbalanced Dataset," Digital, vol. 2, no. 4, pp. 501–519, 2022, doi: 10.3390/digital2040027.

[37] B. Kitchenham and S. M. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," ResearchGate, no. October, 2021.

[38] R. Nokhbeh Zaeem and K. S. Barber, "A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ," CODASPY 2021 - Proc. 11th ACM Conf. Data Appl. Secur. Priv., pp. 143–148, 2021, doi: 10.1145/3422337.3447827.

[39] H. S. Hota, A. K. Shrivas, and R. Hota, "An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique," Procedia Comput. Sci., vol. 132, no. January 2018, pp. 900–907, 2018, doi: 10.1016/j.procs.2018.05.103.

[40] A. El-Aassal, L. Moraes, S. Baki, A. Das, and R. Verma, "Evaluating Performance with New Metrics for Unbalanced Datasets," in R. Verma, A. Das (eds.): Proceedings of the 1st AntiPhishing Shared Pilot at 4th ACM International Workshop on Security and Privacy Analytics (IWSPA 2018), Tempe, Arizona, USA., 2018, pp. 21–24. [Online]. Available: http://ceur-ws.org

[41] K. L. Chiew, E. H. Chang, C. L. Tan, J. Abdullah, and K. S. C. Yong, "Building Standard Offline Anti-phishing Dataset for Benchmarking," Int. J. Eng. Technol., vol. 7, no. 4.31, pp. 7–14, 2018, doi: 10.14419/ijet.v7i4.31.23333.

[42] L. Ma, B. Ofoghi, P. Watters, and S. Brown, "Detecting phishing emails using hybrid features," UIC-ATC 2009 - Symp. Work. Ubiquitous, Auton. Trust. Comput. Conjunction with UIC'09 ATC'09 Conf., pp. 493–497, 2009, doi: 10.1109/UIC-ATC.2009.103.

[43] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," Decis. Support Syst., vol. 107, pp. 88–102, 2018, doi: 10.1016/j.dss.2018.01.001.

[44] N. Vaishnaw and S. R. Tandan, "Development of Anti-Phishing Model for Classification of Phishing E-mail," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 39–45, 2015, doi: 10.17148/IJARCCE.2015.4610.

[45] V. Zeng, S. Baki, A. El Aassal, R. Verma, L. F. T. De Moraes, and A. Das, "Diverse datasets and a customizable benchmarking framework for phishing," IWSPA 2020 - Proc. 6th Int. Work. Secur. Priv. Anal., no. Section 3, pp. 35–41, 2020, doi: 10.1145/3375708.3380313.

[46] L. F. Gutierrez, F. Abri, M. Armstrong, A. S. Namin, and K. S. Jones, "Email Embeddings for Phishing Detection," in Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, 2020, pp. 2087–2092. doi: 10.1109/BigData50022.2020.9377821.

[47] S. Gupta and P. Kumaraguru, "Emerging phishing trends and effectiveness of the anti-phishing landing page," in eCrime Researchers Summit, eCrime, 2014, pp. 36–47. doi: 10.1109/ECRIME.2014.6963163.

[48] Y. M. Mansour and M. A. Alenizi, "Enhanced Classification Method for Phishing Emails Detection," J. Inf. Secur. Cybercrimes Res., vol. 3, no. 1, pp. 58–63, 2020, doi: 10.26735/ygmy6142.

[49] F. Salahdine, Z. El Mrabet, and N. Kaabouch, "Phishing Attacks Detection A Machine Learning-Based Approach," in 2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2023. New York USA., 2021, pp. 0250–0255. doi: doi: 10.1109/UEMCON53757.2021.9666627.

[50] Z. Yang, C. Qiao, W. Kan, and J. Qiu, "Phishing Email Detection Based on Hybrid Features," in IOP Conference Series: Earth and Environmental Science, 2019. doi: 10.1088/1755-1315/252/4/042051.

[51] R. Hassanpour, E. Dogdu, R. Choupani, O. Goker, and N. Nazli, "Phishing e-mail detection by using deep learning algorithms," no. October 2022, pp. 1–1, 2018, doi: 10.1145/3190645.3190719.

[52] R. Valecha, P. Mandaokar, and H. Raghav Rao, "Phishing Email Detection Using Persuasion Cues," IEEE Trans. Dependable Secur. Comput., vol. 19, no. 2, pp. 747–756, 2022, doi: 10.1109/TDSC.2021.3118931.

[53] T. Chin, K. Xiong, and C. Hu, "Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking," IEEE Access, vol. 6, pp. 42513–42531, 2018, doi: 10.1109/ACCESS.2018.2837889.