frontiers | Frontiers in Computer Science

TYPE Data Report
PUBLISHED 06 March 2024
DOI 10.3389/fcomp.2024.1308634

OPEN ACCESS

EDITED BY
Saqib Saeed,
Imam Abdulrahman Bin Faisal University,
Saudi Arabia

REVIEWED BY
Lidia Fotia,
University of Salerno, Italy
Mohammad Jabed Morshed Chowdhury,
La Trobe University, Australia

*CORRESPONDENCE
Maruf Ahmed Tamal
✉ tamal25-103@diu.edu.bd

RECEIVED 06 October 2023
ACCEPTED 23 February 2024
PUBLISHED 06 March 2024

CITATION
Tamal MA, Islam MK, Bhuiyan T and Sattar A (2024) Dataset of suspicious phishing URL detection. *Front. Comput. Sci.* 6:1308634. doi: 10.3389/fcomp.2024.1308634

Dataset of suspicious phishing URL detection

Maruf Ahmed Tamal[1]*, Md Kabirul Islam[2], Touhid Bhuiyan[1] and Abdus Sattar[1]

[1]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh, [2]Faculty of Graduate Studies, Daffodil International University, Dhaka, Bangladesh

KEYWORDS

phishing detection dataset, phishing URLs, legitimate URLs, cybersecurity, classification

1 Introduction

The contemporary world is witnessing a transformative shift driven by technological advancement. As of October 2023, there were 5.3 billion Internet users globally, comprising 65.7 percent of the world's population (Internet and Social Media Users in the World 2023 | Statista, 2023). This exponential growth of the Internet has brought about significant transformations in traditional systems and people's daily lives (Hoehe and Thibaut, 2020). However, alongside this progress suspicious online activities have also increased alarmingly, especially phishing has taken a terrifying shape. It is a form of cyber-enabled crime, uses social engineering and technical subterfuge to deceive individuals into divulging confidential information (Ejaz et al., 2023). Unlike other cybercrimes with consistent victim profiles and known attacker motives, phishing attacks are characterized by their diverse targets, motivations, and goals.

To combat phishing attacks, two types of approaches are commonly adopted: (1) preventive approach (Daengsi et al., 2021; Quinkert et al., 2021; Alahmari et al., 2022), and (2) detective approach (Chiew et al., 2015; Rao and Pais, 2017; Aljofey et al., 2022). While phishing preventive approaches focus on educating individuals to raise user awareness against phishing attacks, detective approaches leverage technical measures like list-based, rule-based, similarity-based, and machine learning (ML)-based methods. However, among all the approaches, ML-based approaches have been extensively utilized by scholars and security experts globally. Considering phishing detection as a binary classification problem, both supervised (Nagaraj et al., 2018; Sahingoz et al., 2019; Zamir et al., 2020) and deep learning algorithms (Dhanavanthini and Chakkravarthy, 2023) have been employed to differentiate phishing sites from legitimate ones. However, none of the approaches performs as a "bullet of silver" against phishing (Gupta et al., 2016). The dynamic and sophisticated nature of phishing attacks has made phishing detection a pressing challenge for both end-users and security experts. Phishers continuously evolve their tactics, seeking new and creative ways to bypass existing anti-phishing tools. Consequently, phishing has become one of the most organized and challenging cybercrimes of the 21st century. As reported by the Anti-Phishing Working Group, 1270883 unique phishing attacks took place in the 3rd quarter of 2022, which was the worst APWG had ever recorded (APWG | Phishing Activity Trends Reports, 2022). This rising tendency underscores the limitations of current anti-phishing methods, particularly their inability to detect zero-hour attacks and their lack of robustness. Unfortunately, existing resources and countermeasures are demonstrably inadequate in detecting and preventing these attacks. One of the most significant challenges hindering the development of robust and effective ML-based phishing detection systems is the lack of a comprehensive and up-to-date labeled training dataset (Catal et al., 2022; Salloum et al., 2022; Zieni et al., 2023). As ML models rely heavily on labeled data to learn the distinguishing characteristics of phishing attacks, this scarcity of labeled data significantly hinders the development of data-driven approaches for designing effective anti-phishing tools.

To address this gap, this article introduces a new, large-scale labeled dataset specifically designed for URL-based phishing detection. This dataset comprises 247,950 instances, meticulously categorized into 128,541 phishing URLs and 119,409 legitimate URLs (see full specification in Table 1). Instead of content-based aspects like text, message, DOM, CSS, logos, etc., this dataset solely focuses on intra-URL features. This strategic choice leverages the fact that many phishing red flags are readily apparent within the URL itself, encompassing typosquatting, unusual extensions, subdomains mimicking legitimate brands, and excessive parameters. So, URLs can reveal patterns and anomalies indicative of phishing attempts. To extract the most discriminatory features from URLs, we employed the Optimal Feature Vectorization Algorithm (OFVA). This rigorous approach yielded 42 optimal intra-URL features. These features demonstrate high efficacy in classifying phishing URLs, contributing significantly to the advancement of data-driven anti-phishing techniques. The availability of this extensive dataset is expected to assist security experts, practitioners, and researchers in developing more sophisticated, resilient, and effective solutions for combating phishing attacks.

2 Value of the data

- The scarcity of large labeled data has been a significant challenge in developing robust and effective anti-phishing tools. To this end, this dataset can address this gap by providing a large number of labeled instances, consisting of both phishing and legitimate URLs.
- The dataset can be used for phishing URL detection using supervised machine learning and deep learning algorithms.
- This dataset can benefit various stakeholders, especially security experts, practitioners, and researchers in the cybersecurity domain by enabling them to stay up-to-date on evolving phishing attacks, advance anti-phishing research, and design sophisticated data-driven anti-phishing solutions for combating phishing attacks.
- The dataset can be utilized to gain insights and develop experiments in phishing detection, including training machine learning models, analyzing intra-URL feature significance and relevance, improving classification performance, developing tailored feature engineering techniques, and exploring model generalization to new phishing attack patterns.

3 Experimental design, materials, and methods

In the process of preparing the phishing detection dataset, we considered three key phases depicted in Figure 1.

3.1 Dataset acquisition

In the first phase, raw unstructured phishing and legitimate URLs were acquired and merged from different reliable and valid sources. To gather the data, we followed similar strategies followed by similar previous studies. Initially, we gathered raw unstructured URLs, encompassing both phishing and legitimate ones, from reputable publicly available sources. Among the 274,446 URLs (before undergoing preprocessing), 48,009 legitimate URLs and 48,009 phishing URLs were obtained from Aalto University's research data (Marchal, 2014), while 86,491 phishing URLs were collected from OpenPhish (OpenPhish, n.d.) and 91,937 legitimate URLs collected from DomCop (Top 10 million Websites Based on Open Data from Common Crawl and Common Search, n.d.). These URLs were in their original form (e.g., https://www.facebook.com/), lacking any specific structure or organization where analysis can be performed. All these data were collected between 01/03/2022 and 31/05/2023.

TABLE 1 Data specification table.

Specifications	Description
Subject	Computer Science
Specific subject area	Artificial Intelligence (Cybersecurity)
Type of data	Table
Date of Collection	01/03/2022–31/05/2023
Data format	Raw (CSV file)
Dataset size	247,950 instances (Phishing URLs =, Legitimate URLs=)
Number of features	42
Number of novel features (proposed by this study)	10 (having_repeated_digits_in_url, number_of_digits_in_domain, having_repeated_digits_in_domain, number_of_subdomains, average_subdomain_length, average_number_of_dots_in_subdomain, average_number_of_hyphens_in_subdomain, having_repeated_digits_in_subdomain, entropy_of_url, entropy_of_domain)
Target feature	1 (Type)
Data accessibility	**Repository name:** Mendeley Data **Data identification number:** 10.17632/6tm2d6sz7p.1 **Direct URL to data:** https://doi.org/10.17632/6tm2d6sz7p.1
Data source	The raw data (unstructured phishing and legitimate URLs) were acquired and merged from different reliable and valid publicly available sources.

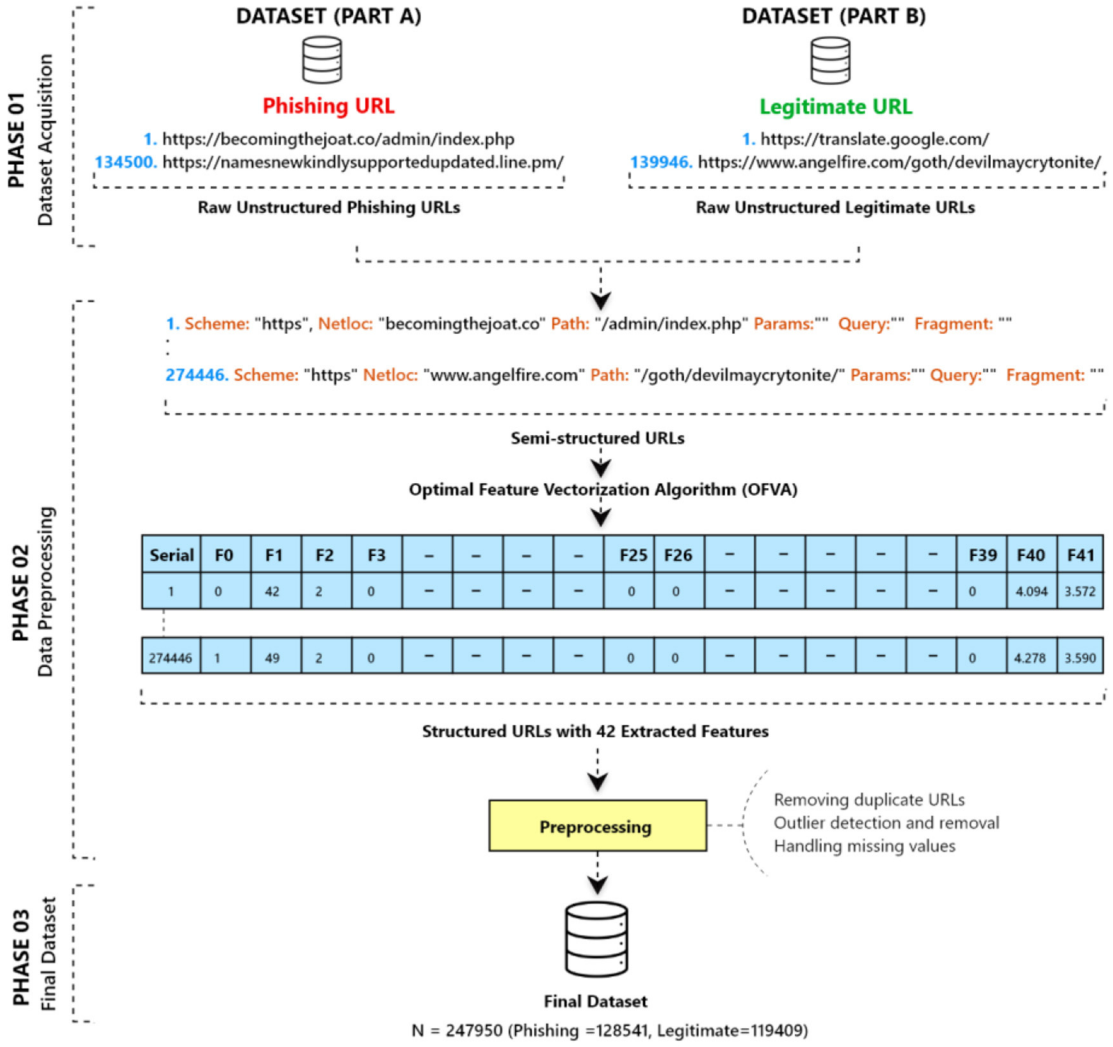

PHASE 01
Dataset Acquisition
DATASET (PART A)
Phishing URL
1. https://becomingthejoat.co/admin/index.php
134500. https://namesnewkindlysupportedupdated.line.pm/
Raw Unstructured Phishing URLs
DATASET (PART B)
Legitimate URL
1. https://translate.google.com/
139946. https://www.angelfire.com/goth/devilmaycrytonite/
Raw Unstructured Legitimate URLs
PHASE 02
Data Preprocessing
1. Scheme: "https", Netloc: "becomingthejoat.co" Path: "/admin/index.php" Params:"" Query:"" Fragment: ""
274446. Scheme: "https" Netloc: "www.angelfire.com" Path: "/goth/devilmaycrytonite/" Params:"" Query:"" Fragment: ""
Semi-structured URLs
Optimal Feature Vectorization Algorithm (OFVA)
Serial | F0 | F1 | F2 | F3 | – | – | – | – | F25 | F26 | – | – | – | – | – | F39 | F40 | F41
1 | 0 | 42 | 2 | 0 | – | – | – | – | 0 | 0 | – | – | – | – | – | 0 | 4.094 | 3.572
274446 | 1 | 49 | 2 | 0 | – | – | – | – | 0 | 0 | – | – | – | – | – | 0 | 4.278 | 3.590
Structured URLs with 42 Extracted Features
Preprocessing
Removing duplicate URLs
Outlier detection and removal
Handling missing values
PHASE 03
Final Dataset
Final Dataset
N = 247950 (Phishing =128541, Legitimate=119409)

FIGURE 1
Step by step process of data preparation.

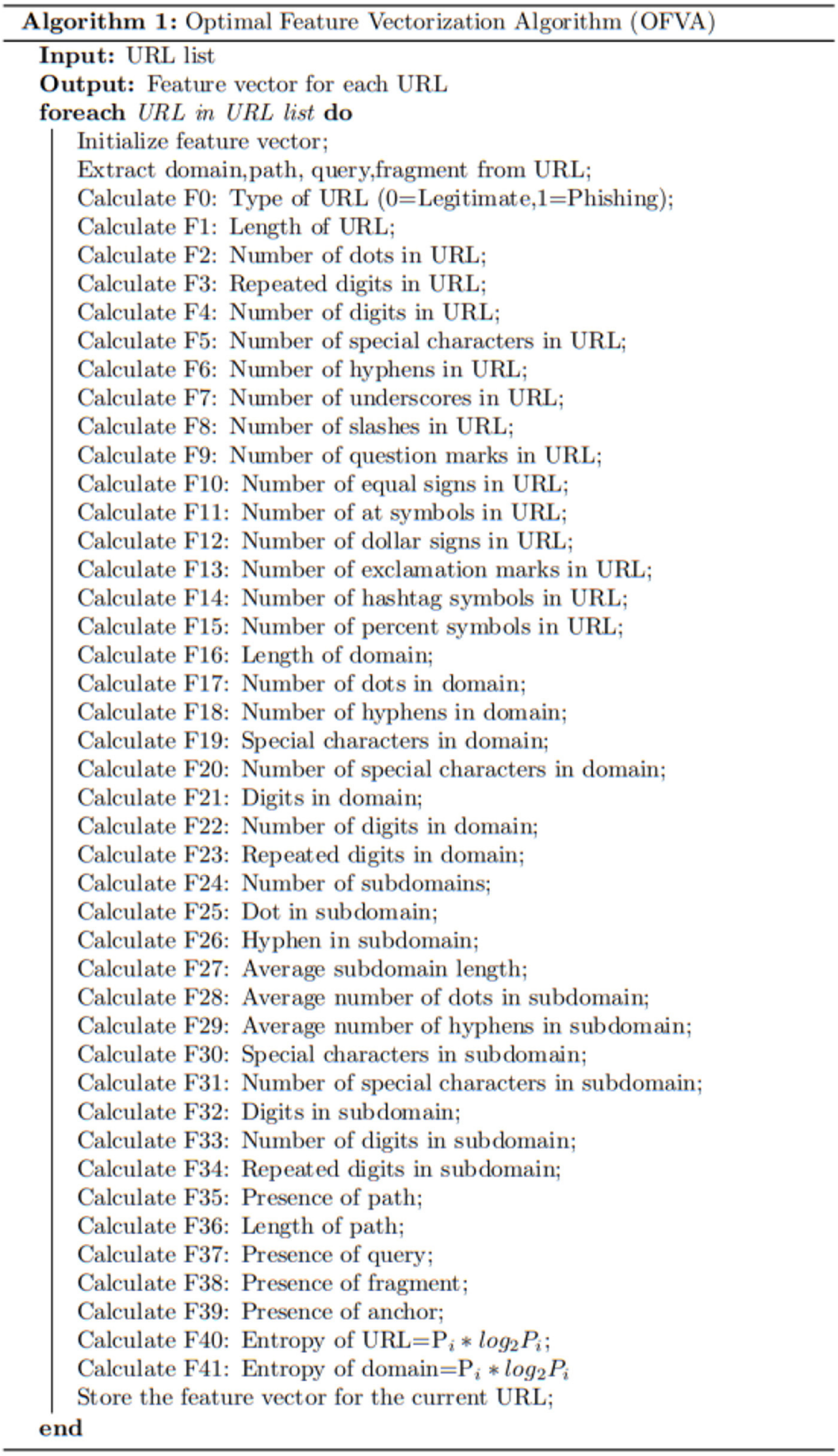
3.2 Dataset preprocessing

3.2.1 Feature generation

In the second phase, unstructured raw URLs (strings) were initially transformed into semi-structured components (scheme, network location, path, etc.) using the “urllib.parse” python module (urllib.parse - Parse URLs into components, n.d.). Subsequently, a list of 41 features was extracted to generate a particular feature vector ($x = F_1, F_2, F_3, \ldots\ldots\ldots\ldots F_{41}$) for each of the URLs to create a labeled dataset using a self-developed Optimal Feature Vectorization Algorithm (OFVA) (see Figure 2). The key purpose of the OFVA was to extract the optimal intra-URL features from a given semi-unstructured URL list (see Phase 2 of Figure 1). Table 2 depicts the extracted feature list with a detailed explanation. Among the 41 features, 31 features ($F_1 - F_2$, $F_4{-}F_{21}$, $F_{25}{-}F_{26}$, $F_{30} - F_{33}$, $F_{35}{-}F_{39}$) were extracted based on findings of the prior studies (Jeeva and Rajsingh, 2016; Singh, 2020; Vrbančič et al., 2020; Mourtaji et al., 2021). These features capture known red flags related to URL, host, domain, sub-domain, path, query, network location components, etc. However, while adopting these features, we performed few features removals, modifications and adjustments to optimize their relevance and improve the overall performance of our feature set. These modifications were informed by an analysis of current phishing trends and emerging threat vectors. Additionally, recognizing the evolving nature of phishing tactics, we introduced 10 novel features (F_3, $F_{22}{-}F_{24}$, $F_{27} - F_{29}$, F_{34}, $F_{40} - F_{41}$) (for details, see Table 2). These features encapsulate nuanced aspects that are not traditionally considered in feature sets, providing a unique contribution to the anti-phishing tool landscape.

3.2.2 Data cleansing and curation

After feature generation, data cleansing and curation were performed. As data was obtained from multiple sources, there was a possibility of having duplicate URLs. Hence, in order to achieve optimal data quality, the data cleansing phase involved the removal of a total of 9,725 duplicate URLs. Moreover, to maintain



```
Algorithm 1: Optimal Feature Vectorization Algorithm (OFVA)
Input: URL list
Output: Feature vector for each URL
foreach URL in URL list do
    Initialize feature vector;
    Extract domain,path, query,fragment from URL;
    Calculate F0: Type of URL (0=Legitimate,1=Phishing);
    Calculate F1: Length of URL;
    Calculate F2: Number of dots in URL;
    Calculate F3: Repeated digits in URL;
    Calculate F4: Number of digits in URL;
    Calculate F5: Number of special characters in URL;
    Calculate F6: Number of hyphens in URL;
    Calculate F7: Number of underscores in URL;
    Calculate F8: Number of slashes in URL;
    Calculate F9: Number of question marks in URL;
    Calculate F10: Number of equal signs in URL;
    Calculate F11: Number of at symbols in URL;
    Calculate F12: Number of dollar signs in URL;
    Calculate F13: Number of exclamation marks in URL;
    Calculate F14: Number of hashtag symbols in URL;
    Calculate F15: Number of percent symbols in URL;
    Calculate F16: Length of domain;
    Calculate F17: Number of dots in domain;
    Calculate F18: Number of hyphens in domain;
    Calculate F19: Special characters in domain;
    Calculate F20: Number of special characters in domain;
    Calculate F21: Digits in domain;
    Calculate F22: Number of digits in domain;
    Calculate F23: Repeated digits in domain;
    Calculate F24: Number of subdomains;
    Calculate F25: Dot in subdomain;
    Calculate F26: Hyphen in subdomain;
    Calculate F27: Average subdomain length;
    Calculate F28: Average number of dots in subdomain;
    Calculate F29: Average number of hyphens in subdomain;
    Calculate F30: Special characters in subdomain;
    Calculate F31: Number of special characters in subdomain;
    Calculate F32: Digits in subdomain;
    Calculate F33: Number of digits in subdomain;
    Calculate F34: Repeated digits in subdomain;
    Calculate F35: Presence of path;
    Calculate F36: Length of path;
    Calculate F37: Presence of query;
    Calculate F38: Presence of fragment;
    Calculate F39: Presence of anchor;
    Calculate F40: Entropy of URL=$P_i * log_2 P_i$;
    Calculate F41: Entropy of domain=$P_i * log_2 P_i$
    Store the feature vector for the current URL;
end
```

FIGURE 2
Optimal feature vectorization algorithm (OFVA).

TABLE 2 Feature description.

SN	Feature	Description	Type
F0	Type	Indicating the type of the URL. It is a Boolean feature with 0 representing a legitimate URL and 1 representing a phishing URL.	Boolean
F1	url_length	Representing the number of characters in a URL, including the domain name, path, and any query parameters.	Numeric
F2	number_of_dots_in_url	Indicating the number of dots (“.”) in the UR	Numeric
F3	having_repeated_digits_in_url	A Boolean feature that denotes whether the URL has repeated digits (e.g., 2232)	Boolean
F4	number_of_digits_in_url	Representing the number of digits (0-9) in the URL.	Numeric
F5	number_of_special_char_in_url	Indicating the number of special characters (e.g., ”, #, $, %, &, ∼) in the URL.	Numeric
F6	number_of_hyphens_in_url	Representing the number of hyphens (“-”) in the URL.	Numeric
F7	number_of_underline_in_url	Indicating the number of underscores (“_”) in the URL.	Numeric
F8	number_of_slash_in_url	Representing the number of forward slashes (“/”) or backward slashes (”\“) in the URL.	Numeric
F9	number_of_questionmark_in_url	Indicating the number of question marks (“?”) in the URL.	Numeric
F10	number_of_equal_in_url	Representing the number of equal signs (“=”) in the URL. It is a numeric feature	Numeric
F11	number_of_at_in_url	Indicating the number of at symbols (“@”) in the URL.	Numeric
F12	number_of_dollar_sign_in_url	Representing the number of dollar signs (“$”) in the URL.	Numeric
F13	number_of_exclamation_in_url	Indicating the number of exclamation marks (“!”) in the URL.	Numeric
F14	number_of_hashtag_in_url	Representing the number of hashtags (“#”) in the URL.	Numeric
F15	number_of_percent_in_url	Indicating the number of percent signs (%) in the URL.	Numeric
F16	domain_length	Representing the length of the domain name in the URL.	Numeric
F17	number_of_dots_in_domain	Representing the number of hyphens (“-”) in the domain name.	Numeric
F18	number_of_hyphens_in_domain	It is a Boolean feature that denotes whether the domain name contains special characters (e.g., !, ”, #, $, %, &, ∼).	Numeric
F19	having_special_characters_in_domain	Having special characters (e.g., !, ”, #, $, %, &, ∼ etc.) in domain.	Boolean
F20	number_of_special_characters_in_domain	Indicating the number of special characters in the domain name.	Numeric
F21	having_digits_in_domain	It’s a Boolean feature that denotes whether the domain name contains digits (e.g., 0-9).	Boolean
F22	number_of_digits_in_domain	Representing the number of digits in the domain name.	Numeric
F23	having_repeated_digits_in_domain	A Boolean feature that denotes whether the domain name has repeated digits (e.g., 223321).	Boolean
F24	number_of_subdomains	Representing the number of subdomains in the URL.	Numeric
F25	having_dot_in_subdomain	Denoting whether the subdomain contains a dot (“.”).	Boolean
F26	having_hyphen_in_subdomain	It’s a Boolean feature that denotes whether the subdomain contains a hyphen (“-”).	Boolean
F27	average_subdomain_length	Representing the average length of the subdomains in the URL.	Continuous
F28	average_number_of_dots_in_subdomain	Indicating the average number of dots (“.”) in the subdomains.	Continuous
F29	average_number_of_hyphens_in_subdomain	Representing the average number of hyphens (“-”) in the subdomains.	Continuous
F30	having_special_characters_in_subdomain	Having special characters (e.g., ”, #, $, %, &, ∼ etc.) in subdomain	Boolean
F31	number_of_special_characters_in_subdomain	Number of special characters (e.g., !, ”, #, $, %, &, ∼ etc.) in subdomain	Numeric
F32	having_digits_in_subdomain	It’s a Boolean feature that denotes whether the subdomain contains special characters (e.g., ”, #, $, %, &, ∼).	Boolean
F33	number_of_digits_in_subdomain	Representing the number of digits in the subdomain.	Numeric

(Continued)

TABLE 2 (Continued)

SN	Feature	Description	Type
F34	having_repeated_digits_in_subdomain	It's a Boolean feature that denotes whether the subdomain has repeated digits (e.g., 223342).	Boolean
F35	having_path	Denoting whether the URL has a path.	Boolean
F36	path_length	Representing the length of the path in the URL	Numeric
F37	having_query	It's a Boolean feature that denotes whether the URL has a query.	Boolean
F38	having_fragment	It's a Boolean feature that denotes whether the URL has a fragment.	Boolean
F39	having_anchor	It's a Boolean feature that denotes whether the URL has an anchor.	Boolean
F40	entropy_of_url	Representing the Shannon entropy of the URL. It is a continuous feature calculated based on the probabilities of each character in the URL. entropy_of_url, $E = \sum P_i{*}log_2P_i$. Here, P_i = probability of each character in the URL, and $\log_2$ is the binary logarithm.	Continuous
F41	entropy_of_domain	Representing the Shannon entropy of the domain. It is a continuous feature calculated based on the probabilities of each character in the domain name. entropy_of_domain, $E = \sum P_i{*}log_2P_i$. Here, P_i = probability of each character in the domain, and $\log_2$ is the binary logarithm.	Continuous

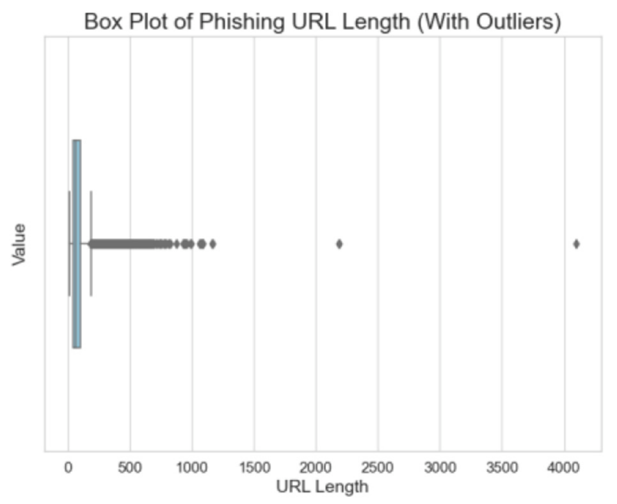


FIGURE 3
Distribution of phishing URLs (with outliers).

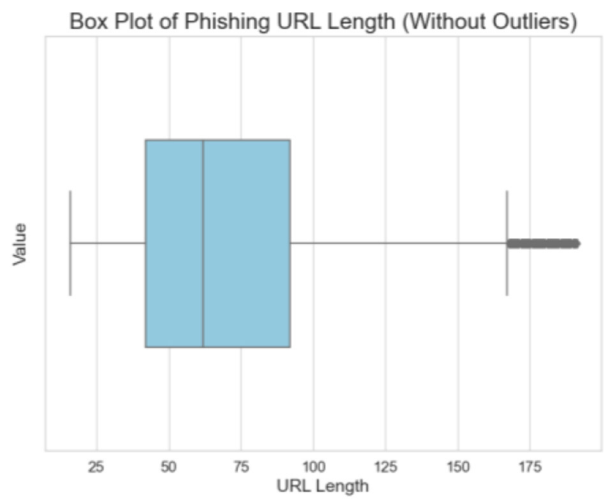


FIGURE 4
Distribution of phishing URLs (without outliers).

the robustness of the dataset, rigorous outlier detection techniques were employed, focusing particularly on the interquartile range (IQR) (Mohr et al., 2022) and box plot analysis (McGill et al., 1978). The rationale behind this approach was to identify and address outliers, with specific attention given to URL length as a key variable. Through the application of the IQR method, data points that fell outside the acceptable range were flagged as outliers. A total of 16,771 such outliers were identified and subsequently removed from the dataset. This process is illustrated in detail in Figures 3–6.

3.3 Final dataset

After data cleansing and outlier removal the final dataset was uploaded in Mendeley Data [66] and made publicly accessible. The final data is comprised of 2,47,950 records (phishing URLs = 119,409, legitimate URLs = 128,541).

4 Data description

The dataset available in the repository consists of a single CSV file with a total of 247,950 instances. Among these instances, 128,541 are classified as phishing URLs, while 119,409 are classified as legitimate URLs. Table 2 provides a comprehensive overview of the dataset, including 42 features associated with both phishing and legitimate URLs. Here, the target feature in the dataset is the "Type" column, which indicates whether a URL is classified as phishing (1) or legitimate (0). This binary classification nature of the target feature makes the dataset suitable for binary classification

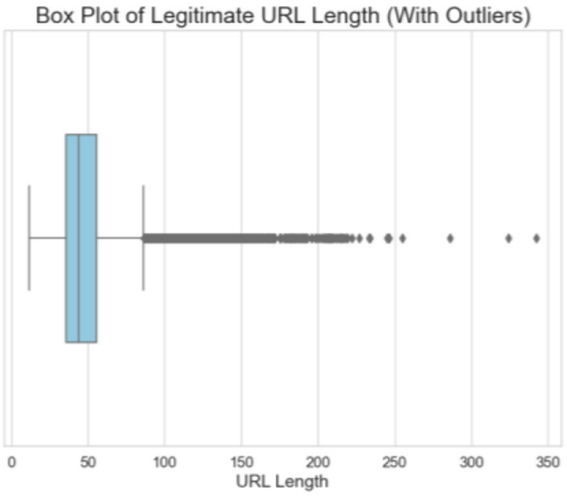


FIGURE 5
Distribution of legitimate URLs (with outliers).

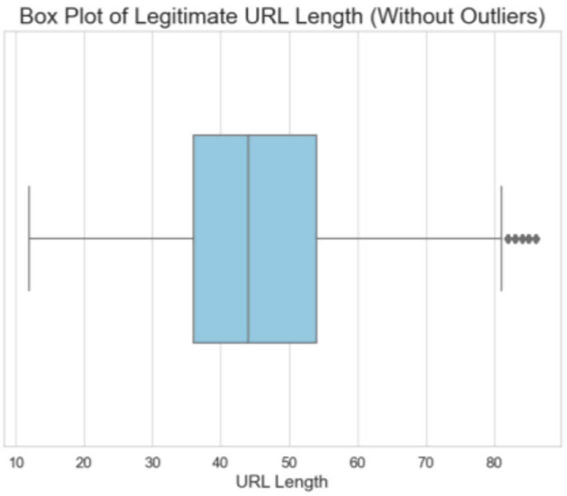


FIGURE 6
Distribution of legitimate URLs (without outliers).

tasks. The remaining features are organized based on their distinct characteristics. For instance, the URL-related features, represented by columns F1–F16, offer valuable insights into the URLs. These features provide information such as the length of the URL, the presence of specific characters or symbols (e.g., dots, hyphens, slashes), and the count of digits or special characters within the URL. Most of these features consist of numeric values representing counts or lengths.

On the other hand, the domain-related features span from F16–F24 and focus on attributes associated with the domain within the URL. These attributes include the length of the domain, the presence of dots or hyphens, the occurrence of special characters or digits, and the number of subdomains. These domain-related features incorporate both Boolean values and numeric counts, providing a comprehensive perspective on the characteristics of the domain.

The subdomain-related features (F24–F34) specifically examine the subdomain section of the URL. These features provide information about the presence of dots, hyphens, special characters, and digits within the subdomain. Additionally, these features calculate averages and counts of these elements. The subdomain-related features contribute to a more detailed analysis of the URLs.

Furthermore, the dataset includes a few other features (F35–F39) that determine the presence of a path, query, fragment, and anchor in the URL. These features employ Boolean values to indicate the existence or absence of these components. Lastly, the table incorporates two continuous features (F40 and F41) that calculate the Shannon entropy of the URL and domain, respectively. These features quantify the randomness or complexity of characters within the URL or domain. Higher values of these features indicate a higher degree of entropy.

5 Comparison with exiting datasets

Table 3 provides a comprehensive comparison between the proposed dataset and existing datasets, highlighting the distinctive features of the former in the realm of phishing detection. In contrast to the limited datasets presented by Orunsolu et al. (2022) and Aljofey et al. (2022), which consist of 5,041 and 60,252 samples, respectively, the proposed dataset sets itself apart by offering a substantially larger volume of data, comprising 247,950 samples. Comparatively, Zouina and Outtaj (2017) and Chiew et al. (2019) present more modest datasets, containing 2,000 and 10,000 samples, respectively. Notably, Vrbančič et al. (2020) boasts a larger dataset with 88,647 samples, however, it lacks information on novel features and preprocessing applied, making it difficult to directly compare its effectiveness. Furthermore, the proposed dataset excels in its feature richness, providing a diverse set of 42 features. This includes the incorporation of 10 novel features that are absent in other datasets. This comprehensive feature set spans numeric, Boolean, and continuous data types, thereby creating the potential for the development of more sophisticated and effective phishing detection models.

6 Limits and suggestions for future works

While the proposed dataset boasts several strengths, it is crucial to recognize and address its inherent limitations. Firstly, despite the dataset's innovation with 10 novel features, there is a lack of novelty in the approach to dataset preparation. Our methodology aligns with common practices used in the preparation of similar existing datasets. Future efforts should explore alternative approaches to dataset creation to enhance originality. Secondly, in the pursuit of a streamlined, efficient model that prioritizes simplicity, speed, and responsiveness, certain content-related features, such as web images, logos, the Document Object Model (DOM), as well as HTML and CSS structural elements, were deliberately excluded. Although this design decision was made to optimize speed and responsiveness, it is essential to acknowledge that the inclusion of

TABLE 3 Comparison with exiting datasets.

Dataset/ References	Dataset type	Experimental data volume	Features	Novel features	Feature types	Preprocessing applied
Proposed dataset	URL-based	Total = 247,950, Phishing URLs= 119,409, Legitimate URLs= 128,541	42	10	Numeric, Boolean, Continuous	Yes
Orunsolu et al. (2022)	Mixed (URL, web document and web behavior attributes)	Total = 5,041, Phishing URLs= 2,541, Legitimate URLs= 2,500	15	Not mentioned	Numeric, Boolean	Yes
Aljofey et al. (2022)	Mixed (URL, and HTML features)	Total = 60,252, Phishing URLs= 27,280, Legitimate URLs= 32,972	15	8	Numeric	Yes
Zouina and Outtaj (2017)	URL-based	Total = 2,000, Phishing URLs= 1,000, Legitimate URLs= 1,000	6	0	Numeric, Boolean, Continuous	Yes
Chiew et al. (2019)	Mixed (URL and HTML features)	Total = 10,000, Phishing URLs= 5,000, Legitimate URLs= 5,000	48	0	Numeric, Boolean, Categorical, Continuous	Yes
Vrbančič et al. (2020)	URL-based	Total = 88,647, Phishing URLs= 30,647, Legitimate URLs= 58,000	111	Not mentioned	Numeric, Boolean	No

these features could potentially contribute to improved accuracy. To this end, future research endeavors should investigate the impact of incorporating these omitted features, exploring whether their inclusion enhances the overall performance of the model.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://doi.org/10.17632/6tm2d6sz7p.1.

Author contributions

MT: Conceptualization, Writing – original draft. MI: Supervision, Writing – review & editing. TB: Writing – review & editing. AS: Data curation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to extend our deepest gratitude to the ICT Division, People's Republic of Bangladesh, for their invaluable support in the successful completion of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alahmari, S., Renaud, K., and Omoronyia, I. (2022). Moving beyond cyber security awareness and training to engendering security knowledge sharing. *Inform. Syst. E-Busi. Manage.* 21, 123–158. doi: 10.1007/s10257-022-00575-2

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., and Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. *Sci. Reports* 12, 1. doi: 10.1038/s41598-022-10841-5

APWG | Phishing Activity Trends Reports (2022). *Phishing Activity Trends Report, 3rd Quarter 2022.* Available online at: https://docs.apwg.org/reports/apwg_trends_report_q3_2022.pdf (accessed May 9, 2022).

Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., and Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. *Knowl. Inform. Syst*. 64, 1457–1500. doi: 10.1007/s10115-022-01672-x

Chiew, K. L., Chang, E. H., Sze, S. N., and Tiong, W. K. (2015, October). Utilisation of website logo for phishing detection. *Comp. Secur*. 54, 16–26. doi: 10.1016/j.cose.2015.07.006

Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S., and Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Inform. Sci.* 484, 153–166. doi: 10.1016/j.ins.2019.01.064

Daengsi, T., Pornpongtechavanich, P., and Wuttidittachotti, P. (2021). Cybersecurity awareness enhancement: a study of the effects of age and gender of thai employees associated with phishing attacks. *Educ. Inform. Technol.* 27, 4729–4752. doi: 10.1007/s10639-021-10806-7

Dhanavanthini, P., and Chakkravarthy, S. S. (2023). Phish-armour: phishing detection using deep recurrent neural networks. *Soft Comp*. doi: 10.1007/s00500-023-07962-y

Ejaz, A., Mian, A. N., and Manzoor, S. (2023). Life-long phishing attack detection using continual learning. *Sci. Reports* 13, 1. doi: 10.1038/s41598-023-37552-9

Gupta, B. B., Tewari, A., Jain, A. K., and Agrawal, D. P. (2016). Fighting against phishing attacks: state of the art and future challenges. *Neural Comp. Appl.* 28, 3629–3654. doi: 10.1007/s00521-016-2275-y

Hoehe, M. R., and Thibaut, F. (2020). Going digital: how technology use may influence human brains and behavior. *Dial. Clini. Neurosci.* 22, 93–97. doi: 10.31887/DCNS.2020.22.2/mhoehe

Internet and Social Media Users in the World 2023 | Statista (2023). *Statista*. Available online at: https://www.statista.com/statistics/617136/digital-population-worldwide/ (accessed October 16, 2023).

Jeeva, S. C., and Rajsingh, E. B. (2016). Intelligent phishing url detection using association rule mining. *Human-Centric Comp. Inform. Sci.* 6. doi: 10.1186/s13673-016-0064-3

Marchal, S. (2014). *PhishStorm - phishing/legitimate URL Dataset*. Espoo: Aalto University.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *Am. Statist.* 32, 12. doi: 10.2307/2683468

Mohr, D. L., Wilson, W. J., and Freund, R. J. (2022). *Data and Statistics*. Gainesville, FL: Elsevier eBooks.

Mourtaji, Y., Bouhorma, M., Alghazzawi, D., Aldabbagh, G., and Alghamdi, A. (2021). Hybrid Rule-Based Solution for Phishing URL Detection Using Convolutional Neural Network. *Wirel. Commun. Mob. Comp.* 2021, 1–24. doi: 10.1155/2021/8241104

Nagaraj, K., Bhattacharjee, B., Sridhar, A., and GS, S. (2018). Detection of phishing websites using a novel twofold ensemble model. *J. Syst. Inform. Technol.* 20, 321–357. doi: 10.1108/JSIT-09-2017-0074

OpenPhish (n.d.). *Phishing Intelligence*. Available online at: https://openphish.com/ (accessed May 31, 2023).

Orunsolu, A., Sodiya, A., and Akinwale, A. (2022). A predictive model for phishing detection. *J. King Saud University – Comp. Inform. Sci.* 34, 232–247. doi: 10.1016/j.jksuci.2019.12.005

Quinkert, F., Degeling, M., and Holz, T. (2021). "Spotlight on phishing: a longitudinal study on phishing awareness trainings," in *Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2021. Lecture Notes in Computer Science, Vol. 12756*, eds L. Bilge, L. Cavallaro, G. Pellegrino, and N. Neves (Cham: Springer). doi: 10.1007/978-3-030-80825-9_17

Rao, R. S., and Pais, A. R. (2017). An enhanced blacklist method to detect phishing websites. *Inform. Syst. Secur*. 2017, 323–333. doi: 10.1007/978-3-319-72598-7_20

Sahingoz, O. K., Buber, E., Demir, O., and Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Syst. Appl.* 117, 345–357. doi: 10.1016/j.eswa.2018.09.029

Salloum, S., Gaber, T., Vadera, S., and Shaalan, K. (2022). A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access* 10, 65703–65727. doi: 10.1109/ACCESS.2022.3183083

Singh, A. (2020). Malicious and benign webpages dataset. *Data in Brief* 32, 106304. doi: 10.1016/j.dib.2020.106304

Top 10 million Websites Based on Open Data from Common Crawl and Common Search (n.d.) *Download list of top 10 million domains based on Open data from Common Crawl and Common Search.* Available online at: https://www.domcop.com/top-10-million-domains

urllib.parse - Parse URLs into components (n.d.). *Python Documentation*. Available online at: https://docs.python.org/3/library/urllib.parse.html (accessed October 6, 2023).

Vrbančič, G., Fister, I., and Podgorelec, V. (2020). Datasets for phishing websites detection. *Data in Brief* 33, 106438. doi: 10.1016/j.dib.2020.106438

Zamir, A., Khan, H. U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., and Hamdani, M. (2020). Phishing web site detection using diverse machine learning algorithms. *Elect. Libr.* 38, 65–80. doi: 10.1108/EL-05-2019-0118

Zieni, R., Massari, L., and Calzarossa, M. C. (2023). Phishing or not phishing? a survey on the detection of phishing websites. *IEEE Access* 11, 18499–18519. doi: 10.1109/ACCESS.2023.3247135

Zouina, M., and Outtaj, B. (2017). A novel lightweight URL phishing detection system using SVM and similarity index. *Human-Centric Comp. Inform. Sci.* 7, 1. doi: 10.1186/s13673-017-0098-1