IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 April 2024, accepted 7 April 2024, date of publication 11 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387437

RESEARCH ARTICLE

A Framework for Preparing a Balanced and Comprehensive Phishing Dataset

IVAN SKULA AND MICHAL KVET, (Member, IEEE)
Department of Informatics, University of Žilina, 010 26 Žilina, Slovakia

Corresponding authors: Ivan Skula (skula@dobraadresa.sk) and Michal Kvet (michal.kvet@uniza.sk)

This work was supported in part by the Erasmus+ Project [Project title: Including EVERyone in GREEN Data Analysis (EVERGREEN)] funded by European Union under Project 2022-1-SK01-KA220-HED-000089149 and in part by the Vedecká grantová agentúra Ministerstva školstva, výskumu, vývoja a mládeže Slovenskej republiky a Slovenskej akadémie vied (VEGA) Project–Developing and Applying Advanced Techniques for Efficient Processing of Large-Scale Data in the Intelligent Transport Systems Environment under Grant 1/0192/24.

ABSTRACT It is not uncommon for people to face phishing attempts on a daily basis, usually via email containing a malicious URL pointing towards a phishing landing page. In recent years, numerous studies have been conducted using machine-learning techniques to detect phishing webpages. These techniques require real-world data from which they extract underlying distinctive patterns that are not easily visible to humans. Capturing and collating such data plays a fundamental role in the overall process. Supervised machine learning algorithms rely on accurate and balanced data for training. Despite the proliferation of research in this field, comparing different studies is a common challenge due to varying data sources, transformations and data cleansing techniques applied when preparing the training dataset. This paper presents a framework for creating a comprehensive and balanced dataset for training machine learning models detecting phishing webpages. The framework covers the process of identifying and gathering the data - phishing and legitimate, data cleansing and highlights important considerations related to the structural composition of the final dataset, like the ratio between phishing and legitimate records or optimal dataset size. Though there is no universal way of preparing a balanced and efficient dataset, the proposed framework provides comprehensive guidelines for constructing one, addressing aspects specific to phishing detection. The practical benefits of applying the framework are accurate, non-skewed, and balanced data, which lead to an accurate model and transparency of data transformation, enabling comparability of the results between different studies.

INDEX TERMS Phishing, framework, dataset design.

I. INTRODUCTION

After almost thirty years of its presence - since the use of the first automated phishing script in 1995 to steal access to America Online (AOL) accounts or collect credit card details that were further used to register new users [1] - phishing became a commonly known term, with six out of ten people familiar with its meaning [2]. During this time, phishing expanded from emails on computer screens to mobile phones, smart TVs, and other media channels via which it's being spread and also fought against through awareness campaigns. The most common distribution channels are depicted in Fig. 1 within green-colored stage ① - short message service (SMS), email, voice call, social networks and quick response (QR) code. As per [3], in 2020, 75% of organizations observed at least one email phishing attempt, 60% at least one SMS phishing attempt, 59% at least one phishing attack via social networks, and 53% recorded at least one vishing attempt.

After this initial stage, which leverages one of the described channels, a phishing attack usually moves to the next stage, which is handled by a phishing landing page (Fig. 1, red-colored stage ②). The three distinctive webpage types represent the three most common phishing webpage objectives. The first is the webpage collecting credentials (e.g., by imitating well-known brands such as Netflix, local banks, or parcel delivery companies). The

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita.

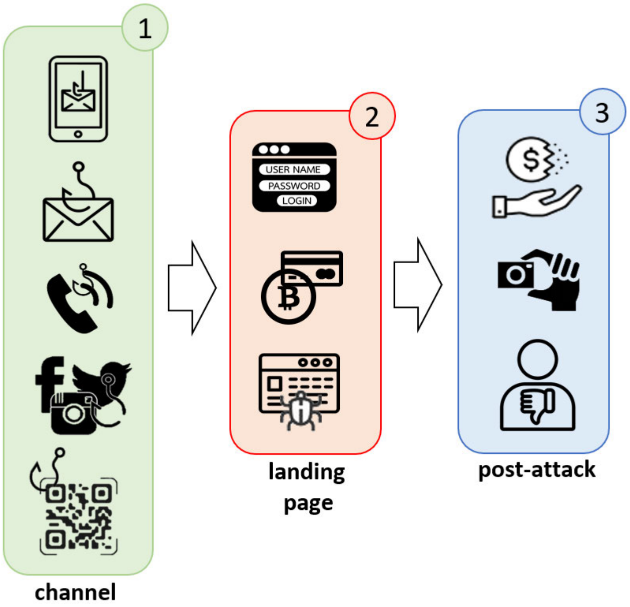


FIGURE 1. Stages of the phishing attacks.

second is a webpage collecting details of payment cards (or Bitcoin wallets). The third is a webpage hosting malware or malicious files (or zero-day exploits) that are supposed to infect the victim's device. The second phase is critical as if the victim doesn't recognize the phishing and continues, the third stage occurs. The third stage is the post-attack phase, where the attacker has collected intended details or infected the victim's device. This stage (Fig. 1 blue-colored stage ③) most commonly results in financial loss for the victim, but depending on the objective of the attacker, it could also be theft of valuable intellectual property, and/or potentially a reputation loss.

As phishing techniques have evolved from simple email scams to sophisticated schemes that target various digital platforms, becoming the most prevalent type of cybercrime [4], the necessity for precise and adaptable phishing detection mechanisms has intensified across various industries. Particularly vulnerable are sectors that have not historically been the focus of phishing, such as healthcare and logistics, including Intelligent Transportation Systems (ITS). ITS integrates communication, control, and information processing across transportation networks, where the integrity and availability of data are paramount. These systems directly influence traffic management and safety measures and are crucial to preventing the paralysis of vital transportation flows. Enhancing phishing detection is thus essential not only for securing ITS against cyber manipulations but also for protecting against data breaches analogous to traditional phishing attacks aimed at personal and financial data theft.

In the early days, when phishing was conducted primarily through emails, deploying an email filter to eliminate most phishing attacks was practical and efficient. With the spread of social networks and instant messaging platforms, which have also become available on smartphones, game consoles, and smart TVs, efficiently monitoring all potential threat vectors has become more challenging. However, email is still the most commonly used channel, and conscious design of phishing email content can positively impact phishing attack efficacy; for example, a phishing email promising a picture of a pretty girl was twice as efficient as a tech support email related to database crash [5]. The susceptibility to phishing worsens when the attacker employs a more focused approach by picking specific targets (spear phishing). Such an attack is conducted by targeting the potential victim with more personalized content (ratio of employees who fell victim to such attacks went up to ≈60% from usual 5% click-ratio of mass phishing attack [6]) 2. On a positive note - a systematic and prolonged approach to awareness training can significantly reduce the susceptibility, e.g., from a baseline(no prior awareness training) ratio of 33.2% (one out of three employees would click on a phishing link) to only 5.4% (approximately one out of nineteen would fall victim) [7].

Though awareness plays an important part in phishing prevention, various techniques are being used to mitigate the risk of phishing on the technical side. This area has gained much attention recently from researchers and commercial companies trying to find the most efficient way to detect and block phishing attacks. Phishing can take many forms based on the objective of the attacker (credentials harvesting, financial theft, extortion, etc.). Some attacks can leverage email with a malicious file attached; others can be spread through an SMS with a shortened Uniform Resource Locator (URL) link pointing to the public cloud with the infected file as a target, and others might be a direct link to a webpage utilizing zero-day exploit and deploying file-less malware. The most common form of phishing is a message containing a URL link to a phishing webpage as depicted in Fig. 1, red-colored stage ②. The diagram shows how a solution that can accurately classify phishing webpages can mitigate the risk of phishing irrespective of the channel used to deliver the message.

The earliest techniques applied to help with phishing detection were blacklists [8], [9]. Blacklists are relatively easy to implement but have limited efficacy as they can capture only re-used domains. Analysis of 10 years of phishing domains data (2013-2022) shows that the share of re-occurring phishing webpages is gradually decreasing - from 21.5% in 2013 to only 6.9% in 2022 [10]. Though blacklists might have been relatively successful in the past, their efficacy is gradually decreasing. Another crucial aspect of blacklists is the need for an additional technique that accurately classifies the visited webpages as confirmed phishing or legitimate webpages, when the domain hasn't been found among the blacklist records. Based on the result of the classification - the webpage is added to the blacklist or ignored in case it's not phishing (alternatively can be added to the whitelist to reduce false-positive alerts in the future).

The most widely researched techniques in detecting phishing webpages are machine learning algorithms - specifically predictive analytics algorithms. These are trained using real-world data with relevant characteristics and a binary identifier distinguishing phishing occurrences from legitimate webpages. The most commonly used are the ones well suited for classification tasks like logistic regression algorithms, decision trees, and support vector machines (SVM). Especially fitting is SVM due to its high accuracy and its ability to work with high-dimensional data [11]. Accurate and commonly used are also algorithms of artificial neural networks. The accuracy of these techniques depends heavily on the accuracy, quality, and comprehensiveness of the data used to train the model.

Hundreds of articles and conference papers have been published (between 2010 and 2017, more than 700 research papers related to phishing detection were available, while the growth trend was clearly recognizable [12]). Despite this high number, it is challenging, if not impossible, to compare the results of one study with another. There are various reasons, but the main ones observed are the insufficient level of detail about the source of data (many times, the details about gathering the legitimate records are insufficient or missing) used by these studies or the lack of details related to data transformation and cleansing before using machine learning techniques [13]. Studies often overlook the importance of describing the data collection process and the adjustments performed, which are crucial to validate or compare the results between various researchers. There are publicly available datasets that can be used:

- Sahingoz et al. [14] - dataset contains 36400 legitimate URLs collected from Yandex Search application programming interface (API) and 37175 phishing URLs collected from PhishTank, though the period during which the data were collected is not provided (probably from 2017)
- Lee et al. [15] - dataset contains 110090 legitimate URLs from the top 300000 Alexa URLs and 32159 phishing pages from PhishTank collected over the period from May till July 2019
- Vrbančič et al. [16] - dataset contains 58000 records of legitimate pages collected from Alexa and 30647 phishing records collected from PhishTank. The period during which the data were collected is not provided, and the dataset doesn't contain the original URLs but only the derived 111 features.
- Marchal et al. [17] - dataset contains 48009 legitimate URLs collected from the Open Directory Project (DMOZ) and 48009 phishing URLs collected from PhishTank during the period from October to November 2012.
- El-Alfy [18], [19] - dataset contains 4898 legitimate URLs collected from Google, Yahoo, and 6157 phishing URLs from PhishTank, MillerSmiles, and other sources. The data collection period is not clear, and the URL is not present in the dataset; only the derived features and a flag indicating whether the record belongs to a phishing or legitimate webpage are present.
- Tan [20] - dataset contains 5000 legitimate records sourced from Alexa and Common Crawl and 5000 phishing records from PhishTank and OpenPhish. Data were collected between May and June 2017, but original URLs are not present in the data; only the derived features are present.
- Yasin et al. [21] - dataset claims to contain 190000 records of phishing URLs collected from PhishTank, but we found only 88084 records. Data are divided into three Excel files: the first file contains 9068 records from May 2013, the second file contains 53668 records from December 2013 to February 2014, and the third file contains 25348 records from March and April 2015. Data doesn't contain any legitimate URLs.

Only a few of the above-mentioned datasets work with the more recent data [15], [16]. And though these can be used for research and academic purposes, they are not sufficient to be used for actual real-world applications, which require even more up-to-date data.

This paper describes a proposed framework for preparing new datasets or validating existing ones for a particular use case. The paper summarizes important stages in designing and creating a dataset for training a predictive analytics model to distinguish phishing from legitimate webpages. In more detail, the framework

- lists all steps relevant to the collection and preparation of data
- discusses various considerations important to creating a comprehensive and balanced dataset
- provides guidelines and best practices to mitigate common mistakes

II. PHISHING DATASET DESIGN FRAMEWORK

Building a single universal phishing dataset that could be used for any machine learning scenario in predictive analytics is impossible in the same way as providing granular yet sufficiently generic steps needed to create a balanced and accurate dataset. Many (if not all) aspects of the dataset will depend on the particular scenario for which the dataset would be used. Particular use cases will impact the expected size of the dataset, the content and the granularity within the dataset, sensitivity to the period from which the data are collected, the length of this period, data cleansing, and data transformation steps required before training the model, etc. Though the weight or importance of a particular area might differ from one use case to another, the proposed framework provides generic steps that should be considered concerning the specifics of the given use case. The first question to answer is the source of the phishing data.

A. PHISHING DATA FEEDS

Though the decision about a relevant data source and the required extent for gathering phishing data depends on the

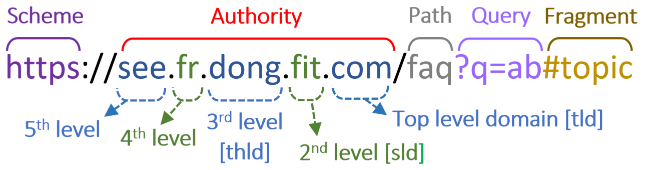


FIGURE 2. Structure and components of the URL.

TABLE 1. Comparison of selected phishing data websites.

	PhishTank	PhishStats	OpenPhish
Real-time interface	Web scrapping	API	Web scrapping
Batch interface	API	API	-
Archive	Accessible	Accessible	Inaccessible
Available features	***	*****	*
Daily volume[1]	≈700	≈2600	≈1000

[1] Daily volumes are calculated using year 2023 data.

intended use case, the minimum data to collect is the list of phishing URLs. For datasets for which the URL (components of the URL in Fig. 2) of the phishing webpage will suffice, phishing data can be gathered from a relevant data source for the desired period without any special considerations. There are already attempts to use artificially generated phishing URLs, but the authors themselves stated a limitation of variation of artificially generated URLs [22]. If the dataset requires characteristics beyond the phishing URL - for example, features derived from the hypertext markup language (HTML) content of the phishing webpage - the collection process will be different as it is vital also to consider the short lifespan of phishing pages.

Availability of the phishing webpage drops quickly; by ≈10% (from 64.9% to 55.8%) within the first five minutes after being reported [23]. After 24 hours ≈34% [24], or in more recent analysis ≈41% [23] of pages are still active and only ≈25% [25] or ≈20% [23] of webpages are still active after 12 days. While the mean lifespan value is measured in days due to the few long-lasting phishing webpages, the median value is measured in hours. A ≈10 hours value was reported in [23] and [26], meaning that only half of the reported phishing webpages were active after this period. Therefore, gathering the data related to the phishing webpage as soon as possible is desirable. This creates a constraint on the relevant source for phishing data, which has to provide reported phishing in real-time or near real-time, and the data collection solution has to be able to capture the required details as soon as they are reported.

Viable and most commonly used data sources are PhishTank (phishtank.org), PhishStats (phishtats.info), and OpenPhish (openphish.com).

PhishTank - The most widely used source of phishing data (in [12] PhishTank was used in 25 out of 45 evaluated research papers, while the second most used data source was used in 6 papers, which shows how often the researchers are leveraging Phishtank) that has been available for many years (since 2005). PhishTank provides data in a format in which the users report them. Registered users can participate in the manual review process of reported suspicious URLs and help classify them as confirmed phishing or legitimate webpage. Each reported phishing has to be evaluated at a minimum by two people. The positive aspect of the manual classification approach is the highest possible classification accuracy. The negative side is a non-negligible volume of reported URLs that remain without the final classification (in Table 1, daily volume of suspicious webpages in PhishTank is ≈700, but these are only records classified as confirmed phishing; the actual overall reported volume is ≈1150 records).

PhishStats - started in 2014, though the archive data go back to 2009. PhishStats receives the highest daily volume of reported phishing pages from all three selected data sources. PhishStats also provides the most comprehensive number of characteristics for each reported URL, though many characteristics are missing, and the actual details of how the characteristics are derived are not explained. In [10], we performed an overlap analysis between PhishTank and PhishStats. While initially, PhishStats contained almost all reported URLs from PhishTank, PhishStats (since 2017) contains a lot of unique records that are not present in PhishTank (approx 40% of PhishStats phishing URLs are unique), which would point towards the preference of using PhishStats to PhishTank.

OpenPhish - started in 2014 and is a free service providing a continuously updated feed of phishing URLs. Free service provides only basic information consisting of three columns - reported URL, targeted brand, and time when the URL was reported. There is an option to upgrade to a paid subscription, which provides more detailed information. In [27], the researcher performed an overlap analysis but focused primarily on which site had the URL captured sooner. No comprehensive analysis of the data overlap has been published.

1) PHISHING FEEDS OVERLAP ANALYSIS

Phishing webpages can be reported via various channels, and the same suspicious URL can be shared or reported to various phishing lists, which causes data overlap between these data sources. We analyzed the overlap between PhishTank, PhishStats, and OpenPhish, which can help decide the preferred data source.

For all three data sources, we analyzed complete 2023 year data and followed the same approach described in [10]. We divided one year of data into monthly parts and compared each month-part while using only the first five levels of the domain part of the URL. Match was found if all five domain levels (Fig. 2) matched in the given month. The results of overlap analysis between the selected data sources show that the highest ratio of unique records has OpenPhish (Fig. 3). And though the PhishStats has the highest daily volume, only 18% of records are unique; the remaining 82% can also be found in PhishTank or OpenPhish.

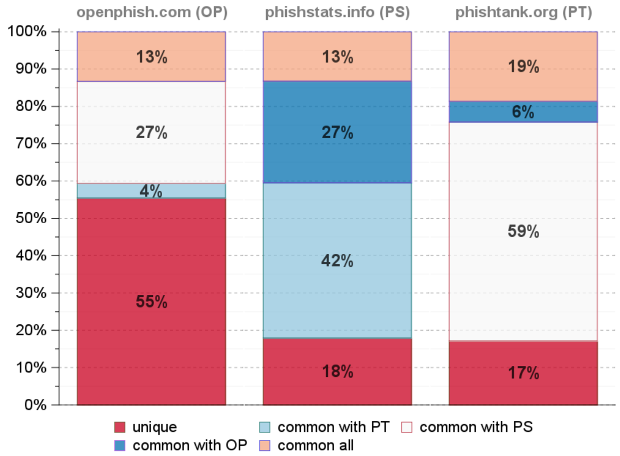


FIGURE 3. **Overlap between data from OpenPhish, PhishStats and PhishTank.**

When deciding which data source provided the suspicious records earlier, we used only the overlapping data across all datasets. The results were that PhishTank presented the earliest data, and after approximately ≈3 hours, the records were available in PhishStats. Then, after ≈20 hours, the records showed up in OpenPhish (hours are derived from median hours difference). The above analysis provides details to decide which phishing data source best fits the intended use case. As there is no clear visibility into how these data feeds(PhishTank, PhishStats, and OpenPhish) source their phishing data, it is advisable to use as many data sources as possible to get the most versatile and comprehensive phishing data.

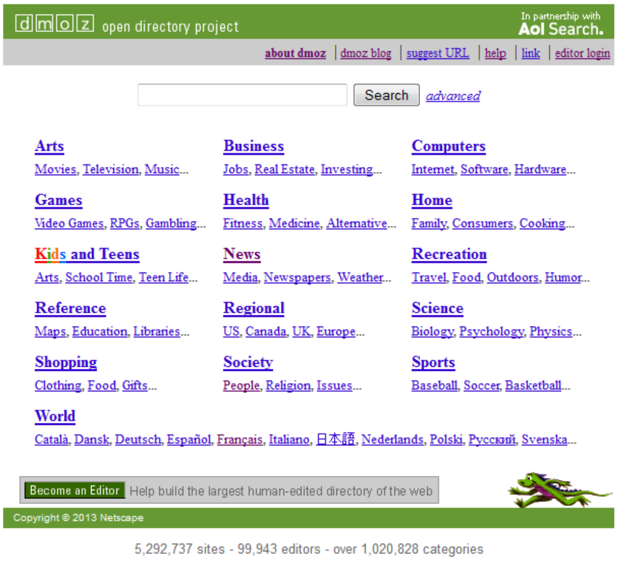
2) PHISHING DATA DE-DUPLICATION

Phishing page URLs can be reported to multiple phishing lists, but they can also be reported to the same phishing feed multiple times by various users. Therefore, one of the initial mandatory steps should be a deduplication process (unless our use case requires duplicate data to be present). This step ensures that the weight of the same reported phishing attack in the final dataset is not multiplied or increased due to the repeated presence of the record in the data [28].

De-duplication can be performed in various ways. We apply deduplication on the records that have all five domain levels of the "Authority" component the same (Fig. 2). The threshold of five levels was derived from empirical analysis of 10 years of data in [10], which represented more than 95% of all records.

B. FEATURES AND CHARACTERISTICS

Phishing raw data, as described in the previous section, are the input to the next step, which encompasses the process of creating and deriving various relevant phishing characteristics, which can be grouped based on the source from which they are derived:


dmoz open directory project

In partnership with Aol Search.

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search advanced

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web

Copyright © 2013 Netscape

5,292,737 sites - 99,943 editors - over 1,020,828 categories

FIGURE 4. **DMOZ Homepage in 2013; (dmoz.org).**

- **URL-based features** are derived from URL components (Fig. 2). Commonly used features within this group are the length of the URL, subdomain levels, scheme, presence of special characters indicating a potential obfuscation, IP address instead of domain name, etc.
- **HTML-based features** are derived from the content of the phishing webpage. Derived features can be linked to the page's visible content - text, images, links, or invisible parts like meta tags, presence of HTML form, scripts, hidden objects, redirect commands, favicon, page title, etc.
- **Externally-linked features** are derived using external data providers (free or paid) that can be linked to the domain, registrar, or hosting IP. Additional data commonly used to enrich the dataset are various domain reputation sites, search index ratings, etc.

Detailed description and calculation logic of the thirty common features representing all three groups is presented in [19]. These characteristics have to provide a distinction between legitimate and phishing webpages. If certain characteristics are the same or very similar for phishing and a legitimate webpage, such characteristics won't be useful in the model training. The primary focus of this step is to identify and create characteristics that reflect the difference between legitimate and phishing web pages.

C. LEGITIMATE PAGES DATA

Phishing pages, though on the rise, constitute only a fraction of the 359 million domains across all top-level domains [29]. There are many ways to gather a sufficient volume of legitimate (non-phishing) webpages, but there are a few considerations to remember. To train a predictive model, it is required to provide actual phishing data and equally relevant non-phishing data. In the research papers, we often see repeated instances of gathering the data from the following sources:

- **DMOZ** (dmoz.org) - also known as Open Directory project owned by AOL and maintained by a community of volunteers (Fig. 4). The web directory site used a hierarchical structure to organize site listings into categories and subcategories. AOL closed the project in 2017; since then, there have been only archived old versions of the database. DMOZ is often used as it contains URLs from across diverse industries and countries, though the language prevalence is skewed with mostly English and European languages [30]. DMOZ was a relevant resource while it was maintained, though the URLs rarely contained the path and query part, which would practically limit the applicable use cases.
- **Alexa 1M** (alexa.com/topsites) - was a list of 1 million domains ranked by the traffic data collected via Alexa toolbar and other traffic data sources. The list was often used as a reputation ranking database or whitelist. The limitation of this list was that it contained only registered domain names (second-level domain SLD and top-level domain TLD components; Fig. 2), which limited its usefulness for deriving features based on the URL characteristics. Alexa 1M list was discontinued in May 2022, but similar alternative lists like "Majestic Million" or "Umbrella 1 Million" from Cisco exist.
- **Yahoo** (yahoo.com) - another common source of URLs with legitimate webpages as it maintained its "Yahoo Directory" - a hierarchically organized database of links grouped into categories similar to DMOZ. Yahoo also provided another function that returned a random URL from its directory. Both of the Yahoo functions were discontinued in December 2014.
- **Common Crawl** (commoncrawl.org)- is a humongous web archive collected by automated crawlers containing billions of URLs spanning across millions of domains. This is still maintained and available.

As web technologies(new frameworks, script libraries, etc.) and web practices mature and change over time, so do legitimate web pages. There are many ways to collect relevant non-phishing data; consider your particular use case and ensure that the non-phishing data you use aligns with the phishing data (e.g., don't combine recent phishing data with historical non-phishing data or vice versa). Also, ensure that the granularity and structure of the data are the same (e.g., URL with all components vs. URL with only registrable domain name).

Last but not least, aspects like phishing webpage language, age of the data, and representation of various industries should be considered. Phishing is a form of social engineering attack and, as such, relies on the impersonation of reputable brands. If your non-phishing data does not contain the records from common industries used by the phishing, such a model will underperform in the real-world setup. The same logic applies to particular parts of the websites. Phishing often uses login pages for various services (banks, entertainment, social networks, etc.). Ensure that your legitimate data contains not only the default landing page of the brand but also the login page. So that the algorithm can extract and capture the difference between the phishing lure webpage and the legitimate login page of a reputable service. These dataset enrichment techniques are described in [31], where the researcher adds sets of specific webpages to ensure the dataset is balanced and represents the common phishing targets. The same approach can be seen in [32], where the dataset was intentionally infused with data of online payment service providers as one of the most common targets of phishing. An analysis of targeted industries and their share within the overall phishing landscape can be found in Anti-Phishing Working Group (APWG) reports [26].

D. SIZE OF THE DATASET

As stated in the [31], machine learning models detect more phishing pages when provided with more patterns (meaning increasing the absolute - number of records and relative - ratio of the phishing pages in the dataset). The important term here is - pattern - not records or observations. Since the predictive model will correlate the observed patterns and their prevalence with a particular class - phishing or not, it makes sense to provide the algorithm with as many patterns as possible and in sufficient numbers to mirror their commonality in the real world. Factors impacting dataset size are:

- **Validation and testing** - training and validation datasets have to be sufficiently sized to be representative (considering the planned ratio of phishing vs. non-phishing data during the model training stage)
- **Machine learning algorithm** - Different predictive algorithms have varying data requirements. Algorithms like neural networks can efficiently ingest and also usually use larger datasets for effective training compared to, e.g., decision trees, which can partition the space and train the model on smaller datasets.
- **Data diversity** - dataset should represent various types of phishing attacks, whereas more diversity usually requires more data to cover variable phishing techniques and tactics.
- **Data dimensionality** - The number of features (columns) can influence the required number of records; more features often require more data samples to accurately model the prevalence of values for all the features and their relationships.
- **Data availability** - this is extremely relevant for particular sub-classes of phishing (like spear phishing or phishing against specific uncommon types of industries, or when we plan to do comparative analysis further back to the past, etc.) where the availability of legitimate and phishing examples also constrains size. Real-world data availability might limit the dataset size.

Among the researchers are those who use a few hundred records for each class [33], those who use a bit more than a thousand records [32], those who use a few thousand [34], and then a few who use tens of thousands of records [14]. Using a few hundred or thousand records might not be sufficient, especially considering the above-mentioned aspects. It is possible to conduct a simple exercise that starts training the model with the smaller size of the data and gradually increases and observes the change in the KPIs (True-Positive Ratio, False-Positive Ratio, Accuracy, Balanced accuracy, F-1 score, etc.). You should observe decreasing gains as the data volume is increased to the point where no further data increase will positively impact the results. The bigger the dataset, the better the detection outcome, as stated in [13], is not necessarily always true. The more representative the dataset, the more comprehensive the features collected and the better the detection performance [35].

1) EXPERIMENTAL EVALUATION OF DATASET SIZE, DIMENSIONALITY, AND ALGORITHM ON MODEL ACCURACY

Analysis was conducted using a dataset with 58000 records of legitimate webpages and 30647 phishing webpages [16]. The dataset contains a column indicating whether the record is phishing or a legitimate webpage and another 111 derived features which can be grouped into the following areas:

- 19 features based on URL - e.g. number of various characters within complete URL like dot, hyphen, at sign, hash sign, percent sign, length of URL, etc.
- 21 features based on domain - e.g. number of various characters within domain part of URL like dot, hyphen, at sign, hash sign, percent sign, length of domain, domain as IP, etc.
- 18 features based on the directory - e.g., number of various characters within the directory part within path component like dot, hyphen, at sign, hash sign, percent sign, directory length, etc.
- 18 features based on the file - e.g. number of various characters within file part within path component like dot, hyphen, at sign, hash sign, percent sign, directory length, etc.
- 20 features based on the parameters - e.g. number of various characters within file part within query and fragment component like dot, hyphen, at sign, hash sign, percent sign, parameters length, tld present flag, etc.
- 15 features from external sources - e.g. ASN IP, days since domain activation, Number of resolved IPs, number of redirects, URL shortener flag, etc.

We separated a validation dataset of 10000 records from the original dataset while keeping the phishing and legitimate records ratio. The remaining data were used to train the model using various training dataset sizes. In the first step, we evaluated model accuracy by using a training dataset of size from 1% to 10% of the size of the dataset. At the same time, we evaluated the model accuracy with respect to the dimensionality of the data. We created 3 variants based on the number of features within the training dataset - the first variant with 10 features, the second with 40 features, and the final with all 111 features (Fig. 5, blue colored area).

Finally, we did this experiment for the following three algorithms:

1) *Logistic Regression*
2) *Decision Tree*
3) *Support Vector Machine*

After evaluating the 1 to 10% range, we also evaluated the models using training data of size from 10% to 100% of the dataset. For this scenario, we used all 111 features available in the dataset (Fig. 5, purple colored area).

Each trained model was validated against the same training dataset, and the resulting KPIs were captured. Since the dataset was slightly skewed (the ratio of phishing vs. legitimate webpages was approximately 1:1.9 we decided to use a balanced accuracy measure (1) as the main qualitative measure.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (1)$$

A confusion matrix summarizes the performance of various decision-making processes or models by showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications. It is a tool used for evaluating the accuracy and effectiveness of a classification algorithm. In our current scenario, these figures represent:

TP - number of correctly classified phishing webpages (phishing classified as phishing)

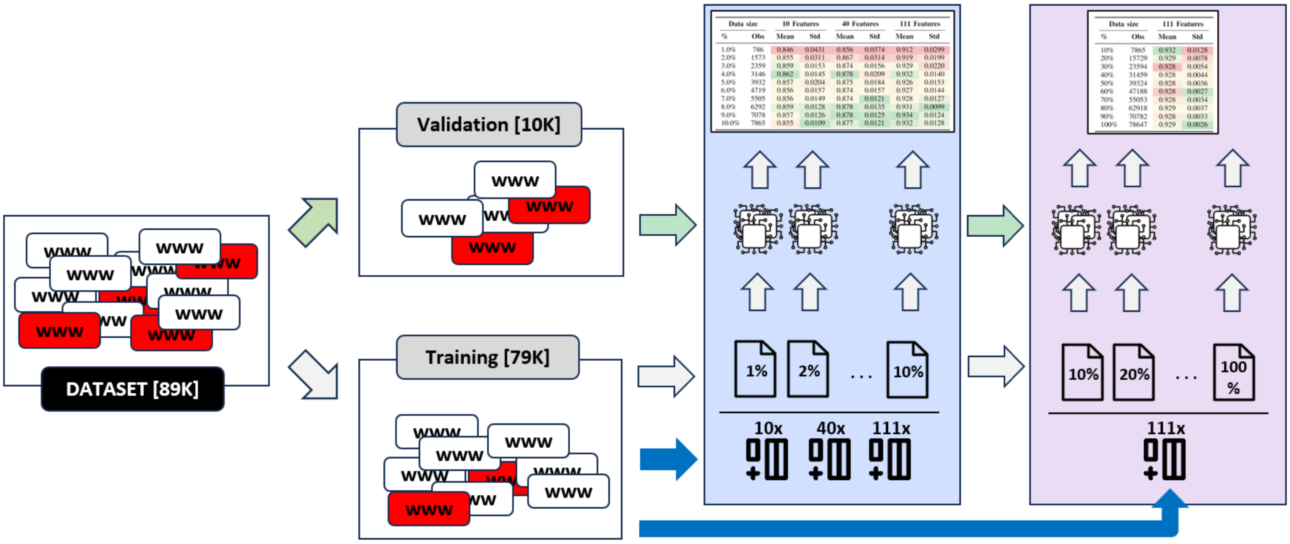
FP - number of legitimate webpages incorrectly classified as phishing

TN - number of correctly classified legitimate webpages (legitimate classified as legitimate)

FN - number of phishing webpages incorrectly classified as legitimate webpages

For every configuration of the trained model, we calculated 10 variants with randomly selected training data from the training dataset while keeping the same size. Therefore, we also measured the standard deviation of balanced accuracy of these 10 model versions.

Analysis findings - the results of the experiments are available in Table 2 for the *Logistic regression* model, in Table 3 for the *Decision Tree* model, and in Table 4 for the *Support Vector Machine* model. In the results, we observed the positive impact of the size, especially within the size between 1% and 4% of dataset size. Gradual improvements across all three models, as well as across all feature variants, can be observed. In the range between 5% and 10%, we observe mixed results, where only the Decision tree algorithm is gradually improving. At the same time, the remaining two models slightly deteriorate, though we observe the improvement of standard deviation figures. Comparing the results for the even bigger training data yields similar findings: only the Decision Tree algorithm improves with additional records within the training dataset. The remaining two models are stagnant, though the standard



Data size		10 Features		40 Features		111 Features	
%	Obs	Mean	Std	Mean	Std	Mean	Std
1.0%	786	0.846	0.0431	0.856	0.0374	0.912	0.0299
2.0%	1573	0.855	0.0311	0.867	0.0314	0.919	0.0199
3.0%	2359	0.859	0.0153	0.874	0.0156	0.929	0.0220
4.0%	3146	0.862	0.0145	0.878	0.0209	0.932	0.0140
5.0%	3932	0.857	0.0204	0.875	0.0184	0.926	0.0153
6.0%	4719	0.856	0.0157	0.874	0.0157	0.927	0.0144
7.0%	5505	0.856	0.0149	0.874	0.0121	0.928	0.0127
8.0%	6292	0.859	0.0128	0.878	0.0135	0.931	0.0099
9.0%	7078	0.857	0.0126	0.878	0.0125	0.934	0.0124
10.0%	7865	0.855	0.0109	0.877	0.0121	0.932	0.0128

Data size		111 Features	
%	Obs	Mean	Std
10%	7865	0.932	0.0128
20%	15729	0.929	0.0078
30%	23594	0.928	0.0054
40%	31459	0.928	0.0044
50%	39324	0.928	0.0036
60%	47188	0.928	0.0027
70%	55053	0.928	0.0034
80%	62918	0.929	0.0037
90%	70782	0.928	0.0033
100%	78647	0.929	0.0026

FIGURE 5. Steps of analysis of dataset size and dimensionality impact on model accuracy.

deviation figures are reduced. This observation confirms that adding more data beyond a certain point might be impractical and sometimes even counterproductive. The analysis of the impact of dimensionality is rather straightforward. We see that an increased number of features brought incremental gain for accuracy and reduction of standard deviation, though the added features must be relevant and bear at least some unique characteristics complementing the other features. It is also important to note that some algorithms are more sensitive to higher dimensionality (e.g., Support Vector Machine compared to the other two algorithms) and might result in increased training time needed, even to the point that would not be practical.

Via the experiments, we also confirmed that more features might require a bigger dataset, which is visible when we compare the best result achieved with the dataset with only 10 features, with the best result achieved for the dataset having all 111 features. While the dataset with the smallest number of features achieved its best result with the dataset of 4% size, the full dataset with 111 features achieved the best results with the 9%-10% sized dataset. This also confirms a logical assumption that a dataset with more features would require more data observations to provide samples for all relevant combinations of these features.

E. STRUCTURE OF THE DATASET

The previous section stated that having more patterns available within the training data allows the trained model to approximate the underlying correlations better and, therefore, be more accurate when classifying new records. The structure of the data also impacts the variability of the patterns. The structure of the data means understanding the share of industries targeted by phishing, as some are more prevalent than others. It also means looking at the language of the phishing targets. If our planned use case revolves around domain structure, aligning with the distribution of top-level domains and representation of domains from various registrars would be relevant. But this is not an exhaustive list of relevant dataset structure considerations - just the most common ones.

1) RATIO BETWEEN THE PHISHING AND LEGITIMATE RECORDS

We would get a hugely imbalanced dataset if we collected all the URLs on the web and could identify all the phishing pages among these URLs. The ratio between legitimate and phishing web pages could easily be 1:1000 or even more. Therefore, what should the ratio between phishing and non-phishing pages in the dataset be? Researchers have asked the same question in [36], and they decided to train the data on a balanced dataset, but evaluation and testing were performed on an imbalanced dataset. In general, it is advised to construct and train the model on a balanced dataset so that the algorithm can have an equal chance to extract the characteristics of phishing pages and those legitimate. The balanced dataset was also used in [14] and [34].

In [31], researchers performed an analysis where they calculated the True-Positive Rate (TPR) and False-Positive Rate (FPR) for various ratios of phishing records in the dataset. The result of this analysis was that the TPR grew gradually from 93% to 98% for 10% to 50% and stayed almost the same for 60% and 70% ratio of phishing records in the dataset, but at the same time, the FPR grew from 0.5% to 1.25% from 10% share to 50% share and continued to grow to 2% for 70%.

Researchers in [33] performed a test with two different ratios of legitimate vs. phishing - 60:40 and 82:18. The outcome was that the PhiDMA algorithm performed with higher accuracy on more skewed data. But since Accuracy as

TABLE 2. Performance metrics across training data sizes and feature counts - logistic regression.

Data size		10 Features		40 Features		111 Features		Data size		111 Features	
%	Obs	Mean	Std	Mean	Std	Mean	Std	%	Obs	Mean	Std
1.0%	786	0.846	0.0431	0.856	0.0374	0.912	0.0299	10%	7865	0.932	0.0128
2.0%	1573	0.855	0.0311	0.867	0.0314	0.919	0.0199	20%	15729	0.929	0.0078
3.0%	2359	0.859	0.0153	0.874	0.0156	0.929	0.0220	30%	23594	0.928	0.0054
4.0%	3146	0.862	0.0145	0.878	0.0209	0.932	0.0140	40%	31459	0.928	0.0044
5.0%	3932	0.857	0.0204	0.875	0.0184	0.926	0.0153	50%	39324	0.928	0.0036
6.0%	4719	0.856	0.0157	0.874	0.0157	0.927	0.0144	60%	47188	0.928	0.0027
7.0%	5505	0.856	0.0149	0.874	0.0121	0.928	0.0127	70%	55053	0.928	0.0034
8.0%	6292	0.859	0.0128	0.878	0.0135	0.931	0.0099	80%	62918	0.929	0.0037
9.0%	7078	0.857	0.0126	0.878	0.0125	0.934	0.0124	90%	70782	0.928	0.0033
10.0%	7865	0.855	0.0109	0.877	0.0121	0.932	0.0128	100%	78647	0.929	0.0026

Notes: Summary of Logistic Regression algorithm performance - mean value and standard deviation - across different training data sizes and feature counts. The color gradient within the results columns indicates value ranges for easier comparison - from red (less favorable) to green (more favorable).

TABLE 3. Performance metrics across training data sizes and feature counts - decision tree.

Data size		10 Features		40 Features		111 Features		Data size		111 Features	
%	Obs	Mean	Std	Mean	Std	Mean	Std	%	Obs	Mean	Std
1.0%	786	0.861	0.0336	0.849	0.0345	0.910	0.0307	10%	7865	0.929	0.0134
2.0%	1573	0.872	0.0284	0.873	0.0193	0.909	0.0198	20%	15729	0.932	0.0051
3.0%	2359	0.881	0.0198	0.877	0.0183	0.922	0.0171	30%	23594	0.935	0.0061
4.0%	3146	0.881	0.0185	0.887	0.0162	0.929	0.0159	40%	31459	0.941	0.0053
5.0%	3932	0.881	0.0254	0.879	0.0225	0.921	0.0190	50%	39324	0.943	0.0034
6.0%	4719	0.876	0.0204	0.887	0.0207	0.922	0.0117	60%	47188	0.944	0.0025
7.0%	5505	0.880	0.0191	0.881	0.0221	0.924	0.0154	70%	55053	0.946	0.0027
8.0%	6292	0.885	0.0149	0.885	0.0121	0.924	0.0101	80%	62918	0.947	0.0030
9.0%	7078	0.890	0.0113	0.887	0.0148	0.929	0.0098	90%	70782	0.947	0.0028
10.0%	7865	0.886	0.0092	0.891	0.0116	0.929	0.0134	100%	78647	0.950	0.0028

Notes: Summary of Decision Tree algorithm performance - mean value and standard deviation - across different training data sizes and feature counts. The color gradient within the results columns indicates value ranges for easier comparison - from red (less favorable) to green (more favorable).

a qualitative measure doesn't perform well with skewed data, we also calculated balanced accuracy, which also performed slightly better for a more skewed ratio of 82:18 (95.63% vs. 92.36%).

Experimental evaluation of ratio between legitimate and phishing records on the accuracy of selected models Analysis was conducted using the same dataset described in the previous section [16]. We separated 10K records from the dataset used as a validation dataset. We created a balanced dataset from the remaining data containing 30K legitimate and 30K phishing records. This dataset of 60K records was used as a pool from which we derived the training dataset used to train the models. All three models were trained on top of the freshly created dataset with 30K records while varying the ratios of legitimate and phishing records - starting with 90% legitimate and 10% phishing and gradually moving towards 10% legitimate and 90% phishing. We used the smallest number of features - first 10 - and gathered the model's mean balanced accuracy figures - similar to the previous analysis.

Analysis findings - the results of the experiments are available in Table 5 for the *Logistic regression* model, in Table 6 for the *Decision Tree* model, and in Table 7 for the *Support Vector Machine* model. In the results, we observed the best results around the balanced ratio only for the Decision Tree model. In the results, we can also observe that the number of phishing records in the dataset results in very similar balanced accuracy figures across various sizes of datasets and ratios of phishing records. For regression and SVM, the results show that a higher ratio of phishing records positively impacts the balanced accuracy of the trained model. While for 10 features, we observed in the first analysis that the model didn't improve further beyond the 3000 records dataset (this dataset had a ratio of phishing vs. legitimate records 1:1.9) and balanced accuracy 0.862, in the second experiment with varying ratios, we achieved even higher balanced accuracy as we moved to the higher ratio of phishing records within the dataset across all dataset sizes. The same results were achieved for SVM. Training models with a balanced dataset helps pay equal attention to all classes but may cause the model to focus too much on random variations (noise) within those classes. On the other hand, using an imbalanced dataset could result in not learning enough about the less common class. Yet, it might lead to a simpler model that works better overall, particularly if the more common class reflects the general trends in the data. This analysis shows that experimenting with the ratios of

TABLE 4. Performance metrics across training data sizes and feature counts - SVM.

Data size		10 Features		40 Features		111 Features		Data size		111 Features	
%	Obs	Mean	Std	Mean	Std	Mean	Std	%	Obs	Mean	Std
1.0%	786	0.836	0.0487	0.861	0.0390	0.901	0.0381	10%	7865	0.932	0.0121
2.0%	1573	0.854	0.0315	0.864	0.0307	0.919	0.0213	20%	15729	0.930	0.0080
3.0%	2359	0.858	0.0149	0.871	0.0140	0.929	0.0216	30%	23594	0.929	0.0052
4.0%	3146	0.864	0.0138	0.877	0.0197	0.929	0.0130	40%	31459	0.929	0.0048
5.0%	3932	0.859	0.0210	0.872	0.0212	0.926	0.0145	50%	39324	0.929	0.0036
6.0%	4719	0.857	0.0160	0.870	0.0179	0.925	0.0134	60%	47188	0.928	0.0025
7.0%	5505	0.856	0.0162	0.871	0.0139	0.930	0.0142	70%	55053	0.929	0.0034
8.0%	6292	0.857	0.0137	0.874	0.0113	0.931	0.0112	80%	62918	0.929	0.0038
9.0%	7078	0.857	0.0138	0.874	0.0128	0.934	0.0113	90%	70782	0.929	0.0031
10.0%	7865	0.854	0.0118	0.872	0.0145	0.932	0.0121	100%	78647	0.929	0.0028

Notes: Summary of Support Vector Machine algorithm performance - mean value and standard deviation - across different training data sizes and feature counts. The color gradient within the results columns indicates value ranges for easier comparison - from red (less favorable) to green (more favorable).

TABLE 5. Performance metrics across training data sizes and ratios - logistic regression.

Training Data Size		Ratio of phishing and legitimate records within training dataset								
%	obs	10:90	20:80	30:70	40:60	50:50	60:40	70:30	80:20	90:10
10%	7865	0.767	0.781	0.836	0.847	0.857	0.888	0.891	0.895	0.896
20%	15729	0.750	0.784	0.842	0.848	0.855	0.876	0.890	0.890	0.881
30%	23594	0.755	0.788	0.842	0.852	0.855	0.879	0.883	0.887	0.893
40%	31459	0.761	0.794	0.841	0.852	0.855	0.877	0.890	0.887	0.885
50%	39324	0.754	0.792	0.841	0.855	0.852	0.875	0.891	0.889	0.892
60%	47188	0.749	0.789	0.840	0.852	0.854	0.881	0.891	0.887	0.895
70%	55053	0.751	0.792	0.844	0.850	0.856	0.878	0.888	0.892	0.888
80%	62918	0.746	0.789	0.840	0.850	0.856	0.877	0.886	0.888	0.893
90%	70782	0.754	0.787	0.844	0.852	0.855	0.878	0.888	0.888	0.889
100%	78647	0.755	0.788	0.841	0.852	0.857	0.878	0.888	0.888	0.891

Notes: Summary of the mean value of balanced accuracy across different training data sizes (rows) and ratios (columns). The color gradient within the results rows indicates value ranges for easier comparison - from red (lower accuracy) to green (higher accuracy). Each model was validated using 10000 records with the ratio of phishing:legitimate 1:1.9 and first 10 features.

TABLE 6. Performance metrics across training data sizes and ratios - decision tree.

Training Data Size		Ratio of phishing and legitimate records within training dataset								
%	obs	10:90	20:80	30:70	40:60	50:50	60:40	70:30	80:20	90:10
10%	7865	0.749	0.833	0.874	0.888	0.893	0.888	0.886	0.892	0.894
20%	15729	0.765	0.824	0.881	0.884	0.885	0.892	0.890	0.882	0.886
30%	23594	0.752	0.837	0.865	0.888	0.897	0.892	0.893	0.890	0.894
40%	31459	0.766	0.838	0.869	0.891	0.894	0.891	0.893	0.886	0.886
50%	39324	0.765	0.836	0.872	0.893	0.895	0.894	0.895	0.888	0.890
60%	47188	0.775	0.834	0.871	0.895	0.894	0.895	0.892	0.896	0.894
70%	55053	0.779	0.835	0.888	0.892	0.895	0.896	0.893	0.893	0.888
80%	62918	0.776	0.837	0.873	0.895	0.896	0.896	0.896	0.889	0.887
90%	70782	0.779	0.837	0.887	0.897	0.897	0.896	0.896	0.893	0.889
100%	78647	0.784	0.833	0.873	0.895	0.896	0.895	0.892	0.893	0.891

Notes: Summary of the mean value of balanced accuracy across different training data sizes (rows) and ratios (columns). The color gradient within the results rows indicates value ranges for easier comparison - from red (lower accuracy) to green (higher accuracy). Each model was validated using 10000 records with the ratio of phishing:legitimate 1:1.9 and first 10 features.

classes might result in higher accuracy and, therefore, should be part of the model training phase.

2) PHISHING BY INDUSTRY

Cybercriminals don't target all industries equally. They tend to focus on some businesses more than others. A summary of the share of phishing by industry can be seen in Table 8. This analysis was conducted on quarterly reports from APWG (similar to [26]) for the last five years. As can be seen, over the years, phishing against certain industries has dropped(Saas/Webmail), while for others, it has increased significantly (social media, logistics, shipping). The most

TABLE 7. Performance metrics across training data sizes and ratios - support vector machine.

Training Data Size		Ratio of phishing and legitimate records within training dataset								
%	obs	10:90	20:80	30:70	40:60	50:50	60:40	70:30	80:20	90:10
10%	7865	0.695	0.790	0.838	0.853	0.860	0.855	0.879	0.891	0.874
20%	15729	0.696	0.789	0.834	0.849	0.858	0.858	0.891	0.891	0.892
30%	23594	0.698	0.794	0.842	0.856	0.854	0.855	0.887	0.889	0.889
40%	31459	0.708	0.795	0.843	0.852	0.854	0.854	0.883	0.887	0.891
50%	39324	0.708	0.789	0.845	0.857	0.851	0.857	0.882	0.892	0.890
60%	47188	0.699	0.795	0.842	0.850	0.851	0.859	0.888	0.891	0.899
70%	55053	0.715	0.789	0.843	0.854	0.854	0.856	0.884	0.891	0.888
80%	62918	0.692	0.793	0.841	0.852	0.852	0.857	0.885	0.890	0.886
90%	70782	0.697	0.792	0.841	0.855	0.852	0.856	0.885	0.889	0.890
100%	78647	0.700	0.793	0.842	0.849	0.854	0.855	0.882	0.889	0.889

Notes: Summary of the mean value of balanced accuracy across different training data sizes (rows) and ratios (columns). The color gradient within the results rows indicates value ranges for easier comparison - from red (lower accuracy) to green (higher accuracy). Each model was validated using 10000 records with the ratio of phishing:legitimate 1:1.9 and first 10 features.

TABLE 8. Average share of phishing per industry per Year.

	2019	2020	2021	2022	2023	Average[*]
SaaS/Webmail	34%	30%	19%	19%	18%	24.7%
Financial inst.	18%	20%	24%	25%	24%	22.0%
Other	14%	11%	10%	19%	11%	13.1%
Payment	22%	13%	9%	5%	6%	11.8%
Social Media	2%	11%	14%	12%	20%	11.2%
Retail/e-comm	4%	7%	12%	7%	5%	7.3%
Logistics/Shipping	1%	4%	5%	6%	7%	4.1%
Telecom	2%	1%	1%	2%	6%	2.3%
Crypto	0%	0%	5%	4%	2%	2.3%
Cloud/File Host	3%	2%	0%	0%	0%	0.2%
Gaming	0%	0%	0%	0%	2%	0.2%
Government	0%	0%	0%	0%	1%	0.1%

[*] The "Average" column is calculated as a mean value across all five years.

consistent and high figures are linked to companies in the finance domain (Financial institutions and Payments).

If the phishing data in the dataset were collected from multiple sources or a single source with sufficient market coverage and during a long enough period, phishing records would have a similar distribution of impacted industries. Ensure that the creation of training and validation datasets contains a sufficient sample of the phishing attack against various industries. With the legitimate data, the distribution of collected records doesn't have to copy the distribution of phishing pages as per Table 8, but since the phishing record will, it is important to represent the legitimate pages from the most targeted industries sufficiently. This will provide pattern variability for the model to distinguish phishing from legitimate pages of a given industry.

3) OTHER CONSIDERATIONS

The above structural considerations are the most common ones, but others might be relevant and depend on your particular use case. One such example might be URL shorteners. Phishing records will most likely contain URL shorteners as they are quite common, with occurrence between 0.2% and 0,7% [37]. So, out of each 1000 phishing records, there will be between 2 and 7 phishing records with URL shorteners. If the dataset contains only legitimate webpages with an actual domain in the URL, whereas there will be phishing records using shorteners, such structural imbalance could impact the model's accuracy as the model will only see phishing records with URL shorteners.

An important consideration impacting the efficacy of the phishing detection model is source data language variability. Given phishing's global reach, a dataset enriched with multilingual content will strengthen the model's ability to discern phishing attempts across various languages, enhancing detection accuracy. Combining webpages in multiple languages eliminates linguistic biases and assures robustness against phishing strategies exploiting language-specific variations.

An example of how important it is to use the recent data for training the model, which should be used in real-world deployment, is the addition of new g-TLD domains (.dad,.phd,.prof,.esq,.foo,.zip,.mov,.nexus) that happened in the first half of 2023. The domain ".zip" captured the highest interest of security researchers as it perfectly mimics the.zip archive extension, which can be easily used for phishing purposes. When we ran a search within the Phishtank and Phsihstats records from 2023, we found already more than 40 unique URLs with the new gTLDs reported as phishing (e.g., url.zip, newdocument.zip, microsoft-office.zip, tax-return-2022.zip, irsrefund.zip, etc.)

As seen above, it is important to consider other aspects of the phishing dataset that could impact the results of our particular use case.

4) LIMITATIONS OF PROPOSED METHODOLOGY

One limitation of the proposed framework is its reliance on available data sources, which might not capture the entirety of phishing activities, especially those targeted at niche or emerging industries. The same applies to phishing attacks, which are extremely perishable and crafted for a narrowly focused target. Due to their rarity and generally low prevalence, these might not show up among the reported phishing web pages. While the framework emphasizes the importance of feature selection, determining the most relevant and effective features for phishing detection is challenging and can significantly impact model performance. Another point being discussed but still posing a challenge is balanced industry representation in the dataset. Achieving a balanced representation accurately reflecting the real-world distribution of phishing attacks across industries is challenging. Furthermore, the dynamic nature of phishing techniques, which continuously evolve to bypass detection, challenges the relevance and effectiveness of the constructed dataset and, by extension, the trained models.

III. CONCLUSION AND FUTURE WORK

The development and evaluation of a balanced and comprehensive dataset for phishing detection underscore the pivotal role of dataset composition in predictive model performance. The research presents a systematic approach to dataset construction, emphasizing the importance of diversity in phishing data feeds, de-duplication, and incorporating a broad spectrum of features and characteristics. Through experimental analysis, it was demonstrated that increasing the overall size of the dataset positively impacts the accuracy only to a certain point beyond which the positive impact diminishes or even reverses (e.g., gradual improvement of balanced accuracy through increasing the size of training dataset from $\approx$800 to $\approx$3000 records utilizing 10 features for training across all three tested models). The same experiment also demonstrated the positive impact of additional features on the balanced accuracy figures (e.g., gradual improvement of the balanced accuracy when increasing the number of features from 10 to 40 and then further to 111 across all three tested models). Training on imbalanced datasets might, in certain use cases, positively impact the model's accuracy - as depicted in the second experimental analysis where the algorithm of logistic regression and SVM improved the balanced accuracy figure when we increased the ratio of phishing records within the training dataset. The study further highlights the varying phishing trends across different industries, underscoring the need for datasets to mirror these variations to train models capable of recognizing the most prevalent industry-specific phishing attempts. The proposed framework contributes to the field by providing insights into dataset preparation that can substantially influence the accuracy and reliability of phishing detection models. This can ultimately aid in developing more effective defenses against phishing attacks and ease the comparability between various types of research.

Future work could take multiple directions, such as examining additional algorithms (e.g., Neural Networks, Naive Bayes, K-Nearest Neighbors) and evaluating these machine learning algorithms' accuracy figures to varying training dataset compositions. Create a fresh phishing dataset as per the framework and compare the experimental analysis results presented in this research with those obtained from the newly created dataset.

ACKNOWLEDGMENT

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of European Union or Slovak Academic Association for International Cooperation (SAAIC). Neither European Union nor SAAIC can be held responsible for them.

Co-funded by the European Union

EverGreen DATA Analytics

REFERENCES

[1] K. Rekouche, "Early phishing," 2011, *arXiv:1106.4692*.

[2] Proofpoint. (2023). *2023 State of the Phish*. [Online]. Available: https://www.proofpoint.com/sites/default/files/threat-reports/pfpt-us-tr-state-of-the-phish-2023.pdf

[3] Proofpoint. (2021). *2021 State of the Phish*. [Online]. Available: https://www.proofpoint.com/sites/default/files/threat-reports/pfpt-us-tr-state-of-the-phish-2021.pdf

[4] *Federal Bureau of Investigation—Internet Crime Report 2023*, document FBI IC3, FBI Internet Crime Complaint Center, 2023. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf

[5] K. Jansson and R. von Solms, "Phishing for phishing awareness," *Behaviour Inf. Technol.*, vol. 32, pp. 584–593, Nov. 2011.

[6] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, "Going spear phishing: Exploring embedded training and awareness," *IEEE Secur. Privacy*, vol. 12, no. 1, pp. 28–38, Jan. 2014.

[7] A. Collard, J. Huisman, J. Jayne, E. Kron, J. Malik, R. Silva, and J. Wieringa, *Phishing by Industry Benchmarking Report—2023 Edition*, document KnowBe4, 2023. [Online]. Available: https://info.knowbe4.com/en-us/phishing-by-industry-benchmarking-report

[8] T. T. P. Thao, T. Makanju, J. Urakawa, A. Yamada, K. Murakami, and A. Kubota, "Large-scale analysis of domain blacklists," in *Proc. 11th Int. Conf. Emerg. Secur. Inf., Syst. Technol.*, Sep. 2017, pp. 161–167.

[9] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, and A. Doupé, "PhishTime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," in *Proc. 29th USENIX Secur. Symp.* USA: USENIX Association, Aug. 2020, pp. 379–396.

[10] I. Skula and M. Kvet, "Domain blacklist efficacy for phishing web-page detection over an extended time period," in *Proc. 33rd Conf. Open Innov. Assoc. (FRUCT)*, May 2023, pp. 257–263.

[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[12] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, "SoK: A comprehensive reexamination of phishing research from the security perspective," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 671–708, 1st Quart., 2020.

[13] 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.

[14] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019.

[15] J. Lee, P. Ye, R. Liu, D. M. Divakaran, and M. C. Chan, "Building robust phishing detection system: An empirical analysis," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, Feb. 2020, pp. 1–12.

[16] G. Vrbancic, I. Fister, and V. Podgorelec, "Datasets for phishing websites detection," *Data Brief*, vol. 33, Dec. 2020, Art. no. 106438. [Online]. Available: https://data.mendeley.com/datasets/72ptz43s9v/1

[17] S. Marchal, J. François, R. State, and T. Engel, "PhishStorm: Detecting phishing with streaming analytics," *IEEE Trans. Netw. Service Manage.*, vol. 11, no. 4, pp. 458–471, Dec. 2014.

[18] E.-S.-M. El-Alfy, "Detection of phishing websites based on probabilistic neural networks and K-medoids clustering," *Comput. J.*, vol. 60, no. 12, pp. 1745–1759, Apr. 2017.

[19] R. M. Mohammad, F. Thabtah, and L. McCluskey. (Jul. 2015). *Phishing Websites Features*. [Online]. Available: https://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf

[20] C. L. Tan. (Mar. 2018). *Phishing Dataset for Machine Learning: Feature Evaluation*. [Online]. Available: https://data.mendeley.com/datasets/h3cgnj8hft/1

[21] A. Yasin, R. Fatima, J. A. Khan, and W. Afzal, "Behind the bait: Delving into PhishTank's hidden data," *Data Brief*, vol. 52, Feb. 2024, Art. no. 109959. [Online]. Available: https://data.mendeley.com/datasets/8py4n46nby/1

[22] M. Sameen, K. Han, and S. O. Hwang, "PhishHaven—An efficient real-time AI phishing URLs detection system," *IEEE Access*, vol. 8, pp. 83425–83443, 2020.

[23] I. Skula and M. Kvet, "Phishing webpage longevity," in *Proc. 12th World Conf. Inf. Syst. Technol. (WorldCIST)*, Mar. 2024.

[24] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. I. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *Proc. Int. Conf. Email Anti-Spam (CEAS)*, Jul. 2009, pp. 50–59.

[25] D. K. McGrath and M. Gupta, "Behind phishing: An examination of phisher modi operandi," in *Proc. USENIX Workshop Large-Scale Exploits Emergent Threats*, Apr. 2008, pp. 1–8.

[26] *Global Phishing Survey: Trends and Domain Name Use in 2H2014*, Phishing Work. Group, Lexington, MA, USA, 2014. [Online]. Available: https://docs.apwg.org/reports/APWG_GlobalPhishingSurvey_2H2014.pdf

[27] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank," in *Proc. Australas. Comput. Sci. Week Multiconference*, Feb. 2020, pp. 1–11.

[28] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2017.

[29] Verisign. (Nov. 2023). *The Domain Name Industry Brief—Volume 20—Issue 4*. [Online]. Available: https://dnib.com/media/downloads/reports/pdfs/2023/domain-name-report-Q32023.pdf

[30] R. Nokhbeh Zaeem and K. S. Barber, "A large publicly available corpus of website privacy policies based on DMOZ," in *Proc. 11th ACM Conf. Data Appl. Secur. Privacy*, Apr. 2021, pp. 143–148.

[31] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 1–28, Sep. 2011.

[32] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 5, pp. 2015–2028, May 2019.

[33] G. Sonowal and K. S. Kuppusamy, "PhiDMA—A phishing detection model with multi-filter approach," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 1, pp. 99–112, Jan. 2020.

[34] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Scherer, and M. Woźniak, "Accurate and fast URL phishing detector: A convolutional neural network approach," *Comput. Netw.*, vol. 178, Sep. 2020, Art. no. 107275.

[35] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, "Systematization of knowledge (SoK): A systematic review of software-based web phishing detection," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2797–2819, 4th Quart., 2017.

[36] K. Rendall, A. Nisioti, and A. Mylonas, "Towards a multi-layered phishing detection," *Sensors*, vol. 20, no. 16, p. 4540, Aug. 2020.

[37] I. Skula and M. Kvet, "URL and domain obfuscation techniques—Prevalence and trends observed on phishing data," in *Proc. IEEE 22nd World Symp. Appl. Mach. Intell. Informat. (SAMI)*, Jan. 2024, pp. 283–290.

IVAN SKULA was born in Zlaté Moravce, Slovakia, in 1981. He received the M.S. degree in software engineering from the University of Žilina, Žilina, Slovakia, in 2004, where he is currently pursuing the Ph.D. degree in software engineering.

Since 2006, he has been a Consultant with SAS Institute, NC, USA, stationed in Bratislava, Slovakia, from 2006 to 2013, and in Dubai, United Arab Emirates, since 2013. He is currently the Principal Consultant of Financial Crimes. His professional research interest includes fraud and financial crimes with an extension to cybersecurity. His current research focuses on the detection of phishing.

Mr. Skula has been an Association of Certified Fraud Examiners Member, since 2019. He holds the Certified Fraud Examiner Certificate.

MICHAL KVET (Member, IEEE) became an Associate Professor of applied informatics from the Faculty of Management Science and Informatics, University of Žilina, Slovakia, in 2020. He is currently a Recognized Researcher, a Conference Speaker, and an Oracle ACE Alumn. He is the author of several textbooks and monographs on temporal database processing. He is the author of more than 70 scientific articles indexed in IEEE Xplore, Scopus, or WOS. He is certified for SQL, PL/SQL, analytics, and cloud databases. He strongly participates with the Oracle Academy and is part of multiple Erasmus+ Projects. Besides, he is the Consortium Leader of the Erasmus+ Project dealing with environmental analytics. He also organizes multiple database workshops annually. His research is devoted to temporal databases, indexing, performance, analytics, and cloud computing.