Computers & Security 148 (2025) 104115

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Beyond the west: Revealing and bridging the gap between Western and Chinese phishing website detection

Ying Yuan [a,*], Giovanni Apruzzese [b], Mauro Conti [c]

[a] Department of Mathematics, University of Padua, Via Trieste, 63, 35131 Padova, Italy
[b] University of Liechtenstein, Vaduz, Liechtenstein
[c] Department of Mathematics, University of Padua, Padua, Italy

ARTICLE INFO

Dataset link: https://github.com/joanyy/ChiPhish

Keywords:
Phishing
Web pages
Language
Machine learning
Deep learning

ABSTRACT

Phishing attacks are on the rise, and phishing *websites* are everywhere, denoting the brittleness of security mechanisms reliant on blocklists. To cope with this threat, many works proposed to enhance Phishing Website Detectors (PWD) with data-driven techniques powered by Machine Learning (ML). Despite achieving promising results both in research and practice, existing solutions mostly focus "on the West", e.g., they consider websites in English, German, or Italian. In contrast, phishing websites targeting "Eastern" countries, such as China, have been mostly neglected—despite phishing being rampant also in this side of the world.

In this paper, we scrutinize whether current PWD can simultaneously work against Western and Chinese phishing websites. First, after highlighting the difficulties of practically testing PWD on Chinese phishing websites, we create `ChiPhish`—a dataset which enables assessment of PWD on Chinese websites. Then, we evaluate 72 PWD developed by industry practitioners and 10 ML-based PWD proposed in recent research on Western and Chinese websites: our results highlight that existing solutions, despite achieving low false positive rates, exhibit unacceptably low detection rates (sometimes inferior to 1%) on phishing websites of different *regions*. Next, to bridge the gap we brought to light, we elucidate the differences between Western and Chinese websites, and devise an enhanced feature set that accounts for the unique characteristics of Chinese websites. We empirically demonstrate the effectiveness of our proposed feature set by replicating (and testing) state-of-the-art ML-PWD: our results show a small but statistically significant improvement over the baselines. Finally, we review all our previous contributions and combine them to develop practical PWD that simultaneously work on Chinese and Western websites, achieving over 0.98 detection rate while maintaining only 0.01 false positive rate in a cross-regional setting. We openly release all our tools, disclose all our benchmark results, and also perform proof-of-concept experiments revealing that the problem tackled by our paper extends to other "Eastern" countries that have been overlooked by prior research on PWD.

1. Introduction

According to the FBI's report (FBI, 2022), phishing is the top-most form of cybercrime, whose growth has increased by over 1000% since 2018. In the first quarter of 2023, the Anti-Phishing Working Group (APWG) reported over 1.6M phishing attacks—the worst quarter ever observed (APWG, 2024). In this context, phishing *websites* are one of the most common vectors employed by attackers, who aim to reach their goals by tricking victims via apparently legitimate web pages (ProofPoint, 2022). In the first half of 2022, over 200k phishing websites were generated every month (PhishLabs, 2022). These numbers have not improved in 2023 (ProofPoint, 2023) and 2024 (ProofPoint, 2024)—showing that an effective solution to this threat has yet to be found.

What we have just written is the exemplary "introductory paragraph" of papers on phishing website detection (Saha Roy et al., 2023; Althobaiti et al., 2021). Such an opening is typically followed by original analyses (e.g., measurement studies (Peng et al., 2019b)) revealing the brittleness of current anti-phishing schemes; or novel solutions (either human-centred, such as phishing education (Jensen et al., 2017); or machine-centred, such as automated detectors (Sahingoz et al., 2019; Ho et al., 2019)) to mitigate the threat of phishing against Web users. However, we wonder: granted that phishing is a problem worldwide, what is the current status of phishing in the Eastern part of the World—in particular, in China (having 1.4 billion people (worldometers, 2023i))?

* Corresponding author.
E-mail addresses: ying.yuan@unipd.it (Y. Yuan), giovanni.apruzzese@uni.li (G. Apruzzese), mauro.conti@unipd.it (M. Conti).

https://doi.org/10.1016/j.cose.2024.104115
Received 23 June 2024; Received in revised form 14 August 2024; Accepted 9 September 2024
Available online 26 September 2024
0167-4048/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

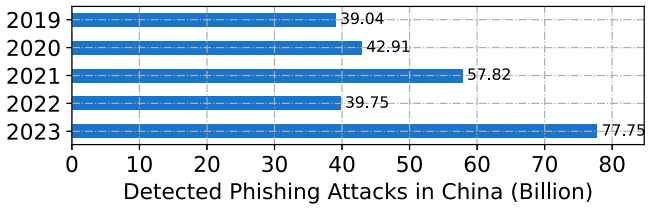


Fig. 1. Snapshot of Chinese phishing landscape.
Phishing intercepted by Qihoo360 (largest Chinese internet security company) in the first half of 2019–2023 in China. These numbers increased by 94% in 2023 (QiHoo360, 2023).

Inspired by the idea of investigating how the World's most populated country is affected by phishing, we searched the Web for evidence on the Chinese phishing landscape. We found that, according to 2016 estimates (Li et al., 2016), China suffers >30B Yuan (≈$4B) in losses every year due to phishing. More recently, we found reports from Qihoo360, the largest Chinese security company (QiHoo360, 2019, 2020, 2021, 2022, 2023), highlighting the yearly trend of phishing attacks intercepted in China—which number is in the *billions* (shown in Fig. 1). Accordingly, over 77 billion phishing attacks have been *blocked* in the first half of 2023, an increase of 94.8% over 2022. Unfortunately, such data only reflects phishing attacks that have been detected by QiHoo360 in China and do not say anything about those that have been missed—potentially outside of China. Hence, we asked ourselves: how can we deal with the spread of phishing websites in China? And what about Chinese phishing websites, which can be visited by Chinese people who reside "in the West" (or even by "Westerners" themselves)?

Intrigued by these dilemmas, we turned our attention to research papers. We found that a large body of literature proposed/investigated methods (reliant either on blocklists (Bell and Komisarczuk, 2020; Oest et al., 2020), or on data-driven heuristics (Tian et al., 2018; Apruzzese et al., 2022a; Liu et al., 2022a)) to detect phishing websites. The conclusions are that—Phishing Website Detectors (PWD) *can work* in some scenarios. However, the majority of such papers assumed that websites (benign or phishing) used to test any given PWD were in *phonological* languages (e.g., English). Such an assumption, despite being the de-facto standard "in the West", does not allow to determine if (and how much) the considered PWD also works "in the East", i.e., for countries having *hieroglyphic* languages—such as China. Indeed, we found that only a handful of papers (see Table 1) on PWD tried to see this problem from the Chinese perspective. However, even these works considered a restricted subset of the Chinese phishing landscape, and did not emphasize whether the corresponding PWD work also on Western sites.

Simply put, there is a large side of the World (i.e., China) that, unfortunately, has been overlooked by prior research on PWD; at the same time, the few works which do consider such side of the World "overlook" the effectiveness of Chinese-specific PWD in the rest of the World. **Such a lack of attention led us to question whether PWD previously shown to be effective for "Western" websites also works for "Eastern" websites—and, specifically, Chinese ones.** Besides obvious differences in languages, Chinese websites present unique characteristics (due to, e.g., some regulations (Cyberspace Administration of China, 2022c)), suggesting that PWD may not work equally well when analysing websites tailored for people from different regions. If found to be true, such a hypothesis would reveal a problem for the real World. Indeed, today *(i)* an increasing number of Western people reside in China (Migration Policy Institute, 2022a) and *(ii)* an increasing number of Chinese people migrated to the West (OECD Ilibrary, 2021a). For instance, an English person can be protected from phishing if they live in the UK and only visit English websites—but what if such a person goes to China and visits Western websites? And, vice-versa, Chinese PWD may be effective as long as they monitor Chinese residents—but what if a Chinese person goes abroad? Moreover, in the era of economic globalization, it is undeniable that people would visit websites from different regions and in various languages. Therefore, in this paper, we scrutinize whether PWD can transfer between different geographic regions—and, if they cannot, devise ways to bridge this "gap" between Western and Eastern PWD.[1]

Contributions and Organization. Our paper's contributions lie at the intersection of *three research domains*: socio-technical aspects of the Web, measurements of the Web, and security of the Web. As such, we write this paper so that it is understandable by any reader interested in one of such domains. At a **high-level**, we provide a threefold contribution to the state of the art:

- ℂ1: We carry out *a measurement study* wherein we assess the effectiveness of existing PWD for websites from diverse regions;
- ℂ2: We propose *ways to enhance PWD*, so that they work both on Western and Chinese websites;
- ℂ3: We provide *factual evidence that this problem has been neglected, and release all our tools and data* to facilitate development of real-world solutions.

Let us explain how these contributions are distributed in this paper.

- In Section 2 we introduce the *essential concepts of modern PWD*; then, through an extensive *literature review*, we showcase the **limited scope of prior research** on Chinese-PWD.
- In Section 3 we discuss our *data-collection* procedures. By analysing the landscape of publicly available resources, we show the *shortage of Chinese-specific data*. To fix such a lack, we **create `ChiPhish`**, the largest open-source dataset to evaluate (or develop) Chinese PWD.
- In Section 4 we assess *72 PWD developed by industry practitioners*. By considering *both blocklist-based and data-driven* PWD, we show that **operational anti-phishing services can only detect phishing websites from their respective "regions."**
- In Section 5 we consider *state-of-the-art PWD proposed in research* and reliant on machine learning (ML). We assess 10 ML-PWD (previously tested on Western websites), and **show their immaturity when analysing Chinese websites**. Then, we *critically analyse the spectrum of visual-based PWD*, highlighting their pros-and-cons for cross-regional PWD.
- In Section 6 we focus on *enhancing feature-based ML-PWD*, to prepare them for cross-regional PWD. After dissecting the anatomy of Chinese websites, we **propose a new feature set** which allows to capture the characteristics of Chinese and Western websites. We *then implement 81 ML-PWD and test them on our data*, showing improvement over the baselines.
- In Section 7 we *find ways to bridge the gap* we brought to light. We *distill all the knowledge and tools* produced during our research, and **develop ML-PWD that work on Chinese and Western websites**, achieving above 0.98 tpr with only 0.01 fpr in cross-regional settings.
- In Section 8 we reflectively discuss our contributions, *pointing out room for improvement*. Then, as an inspiration for future work, we provide *further evidence that China was overlooked by prior study*, and conduct a small **assessment on Japanese and Korean websites**.

We conclude our paper in Section 9. Moreover, we provide the complete results of our assessments, additional experiments, and analyses in our appendix. Finally, for complete reproducibility and transparency, we fully release our resources in an open-source repository (https://github.com/joanyy/ChiPhish).

[1] We focus on phishing *websites*: other forms of phishing (e.g., email (Hasegawa et al., 2021; Roepke et al., 2022; Gao et al., 2021)) are outside our scope.

2. Background and motivation

We summarize existing techniques for phishing website detection (§2.1). Then, we survey the phishing landscape in China (§2.2), highlighting the limited vision of related work. Finally, we provide real-world evidence that inspired our research problem (§2.3).

2.1. Phishing Website Detection (overview, benefits and drawbacks of existing methods)

Phishing is a historical security problem, which has been tackled by abundant research. Reliant on social engineering (Braun et al., 2014), used to lure victims onto malicious webpages, phishing website attacks require the victim to *(i)* be shown the webpage; and *(ii)* be *caught* by providing sensitive data, or clicking on a harmful link (Shusterman et al., 2020). Clearly, the attack fails if the potential victim recognizes the webpage as malicious. Therefore, anti-phishing training programs can reduce the risk of phishing attacks (Jampen et al., 2020), and some research has been carried out (e.g., (Sarno et al., 2022; Lain et al., 2022)). However, according to ProofPoint's 2024 report (ProofPoint, 2024), more than 30% of organizations dedicate *less than 1 hour per year* to educate their employees (some do not have any training program at all). Hence, there is still a need for "machine-based" schemes that provide a first line of defence against phishing for uneducated (or distracted) users (Draganovic et al., 2023). These automated phishing website detectors (PWD) are based on the combination of signatures (i.e., *blocklists*) or data-driven heuristics—among which, many rely on *machine learning* methods. Let us briefly review their pros-and-cons.

2.1.1. Signature-based PWD

Signature-based PWD still represents the preferred countermeasure against phishing, and leverage blocklists of suspicious websites (taken from, e.g., PhishTank (PhishTank, 2022g), or Google Safe Browsing (Google, 2023c)). Before rendering any given website, the browser (or an organization-wide detector) checks if the URL (or a subdomain) is included in the blocklist, thereby alerting the user upon a correct match (Zuraiq and Alkasassbeh, 2019). To avoid triggering annoying false alarms, the websites in these lists must be verified: as a result, signature-based PWD have high precision—which is appreciated in the context of PWD, given that web-users visit hundreds of pages every day (Adebowale et al., 2019). Unfortunately, signature-based detectors are useless (Apruzzese et al., 2023b) against "novel" attacks. Despite huge efforts put in by the maintainers of blocklists to keep them as up-to-date as possible, some websites are bound to evade blocklist-based PWD (Tian et al., 2018). Such a shortcoming (which we empirically confirm in §4) led to the proliferation of complementary PWD that can cope with the ever-changing landscape of phishing websites—which can be accomplished via machine learning (ML).

2.1.2. ML-based PWD

The underlying principle of machine learning is to have "machines that autonomously learn from data." This process is done by *training* an ML *model* over some *training data* by means of a given learning *algorithm*. The successes of ML in various fields (most notably, computer vision and natural language processing (LeCun et al., 2015; Jordan and Mitchell, 2015)) showed the remarkable performance of ML for classification tasks, inspiring researchers to investigate their effectiveness also for cyberthreat detection (Apruzzese et al., 2023b)—which also encompasses PWD (Chiew et al., 2019; Gandotra and Gupta, 2021; Jain and Gupta, 2018a; Makkar et al., 2021; Aydin and Baykal, 2015; Singh et al., 2015). The systematic literature review conducted in Safi and Singh (2023), which analysed 80 papers, revealed that ML techniques enabled to yield PWD with up to 99.98% accuracy. Existing ML-PWD can fall into three categories, depending on the information used as basis to perform the (binary) classification of a given website (Apruzzese et al., 2022a). Specifically, an ML-PWD can use either the URL of a website, its representation (e.g., the image or the HTML), or their combination. Each of these can be elaborated in diverse ways: for instance, some ML-PWD necessitate some preprocessing aimed at extracting some features from a given piece of data (e.g., computing the length of the URL (Mohammad et al., 2014a)); others (typically those relying on deep learning (Abdelnabi et al., 2020)) may analyse a given input in its raw form. We stress that – despite appreciable results shown in research – recent findings revealed that commercial PWD using deep learning can be easily evaded (by real attackers!) via decade-old tricks (Apruzzese et al., 2023a). [2] (We will provide additional low-level details on some exemplary ML-PWD in §5).

Narrow Scope. Despite many papers proposing data-driven countermeasures to phishing, prior efforts (e.g., Jain and Gupta (2018b), Le et al. (2018), Ozcan et al. (2021), Mohammad et al. (2014b), Aljofey et al. (2022), Ariyadasa et al. (2022), Tian et al. (2018), Apruzzese and Subrahmanian (2022)) only focused on Western websites—overlooking that phishing is a long-standing problem also in other areas of the world, such as China.

2.2. The Chinese Phishing Landscape (and shortcomings of prior work)

Reports estimated over one *billion* Chinese netizens as of June 2022 (China Internet Network Information Center, 2022): accordingly, 24% had been scammed by phishing websites in the previous 6 months. According to Qihoo 360 (QiHoo360, 2023), 77.75 *billion* phishing attacks have been intercepted in the first half of 2023—a rate of 430 million per day. Among these, 99.7% target PC devices, with only a minority entailing mobile users.

Intriguingly, however, most of these facts are only accessible "to Chinese": the QiHoo360 reports are not available in English (QiHoo360, 2019, 2020, 2021, 2022, 2023). Moreover, some sources are not accessible from outside of China: for instance, some authors of this paper reside in Europe and could not load the (non-archived) webpage providing a valuable report (360 secure brain, 2021). This suggests that – despite phishing websites being clearly rampant also in China (Liu et al., 2021b) – *such a threat may be difficult to investigate from the perspective of a researcher* "in the West" who may not know Chinese.

2.2.1. High-level analysis

Let us review the landscape of Chinese-focused phishing research over the last 10 years. In 2014, Zhang et al. (2014) proposed 5 domain-specific features to detect phishing websites targeting Chinese eCommerce: despite achieving 96% accuracy, which only focus on eCommerce websites, neglecting the plethora of other websites (e.g., forums, hospitals and government) which can very well be targeted by phishers. Indeed, Chinese websites tend to have different characteristics, as shown in Fig. 2: eCommerce websites must report their business licence (red box in Fig. 2(a)), which is not necessary for Chinese government websites—which, in turn, have a government identification code (blue box in Fig. 2(b)). More recently, Li et al. (2020) proposed five space transformations that disentangle the linear and non-linear interactions between features in malicious URL data. The dataset used in the experiments of Li et al. (2020) includes URL from generic Chinese websites, thereby allowing one to assess the effectiveness of ML-PWD analysing the URL of a website. Even though such an approach may work when analysing Chinese sites, its effectiveness is questionable when Western websites are taken into account. This is because the *URL is a combination of alphanumeric characters*: hence, the URL for Western and Chinese

[2] We stress that themes within the realm of "adversarial machine learning" (e.g., Tian et al. (2024)) are outside our scope.

Table 1
Papers on Chinese PWD. None of these release the source-code (the online tool in Li et al. (2020) is not functional anymore), thereby preventing complete reproducibility of their results.

Paper (1st Author)	Year	Dataset Available	Other Regions	Website Focus	Collection Date	Analysed Features (F)	Chinese Specific	Stat. Sign.
Chu et al. (2013)	2013	✗	✗	eCommerce	2012	F_u	✗	✗
Zhang et al. (2014)	2014	✗	✗	eCommerce	2014	F_c	✓	✓
Zhang et al. (2016)	2016	✗	✓	generic	2011	F_u	✗	✗
Li et al. (2016)	2016	✗	✗	generic	2015	F_u, F_h	✗	✗
Zhang et al. (2017a)	2017	✗	✓	generic	2017	F_c	✗	✗
Zhang et al. (2017b)	2017	✗	✗	generic	2017	F_c	✓	✗
Xiangdong et al. (2017)	2017	✗	✗	finance	2017	F_u, F_h	✗	✗
Feng et al. (2017)	2017	✗	✗	finance	2016	F_c	✓	✓
Li et al. (2020)	2020	✗	✓	generic	2020	F_u	✗	✗
Liu et al. (2021a)	2021	✓	✓	generic	2018	F_c	✗	✗
Jiang and Wu (2022)	2022	✓	✗	generic	2021	F_c	✓	✗

(a) Footer of a Chinese eCommerce website (https://global.jd.com/).

(b) Footer of a Chinese govt. website (https://banshi.beijing.gov.cn/).

Fig. 2. Chinese eCommerce and government websites.
Websites in China have different identifiers, useable for PWD. Red boxes denote the "business license", whereas the blue box denotes the "government identification code". Chinese websites can also display trusted website certifications (issued, e.g., by the police), denoted in yellow boxes.

sites may appear similar despite their HTML being different; plus, PWD analysing the URL can be easily bypassed (e.g., Apruzzese and Subrahmanian (2022)). We argue that HTML is a more valuable information source for PWD; however, as we will show (in §6.1.1), proper usage of the HTML for Chinese websites requires some tweaks that are neglected by existing ML-PWD (typically tested on Western websites).

2.2.2. *Chinese PWD: state of research*

We carry out a literature analysis, aimed at scrutinizing works that consider the problem of PWD in China. To this purpose, we query well-known scientific repositories (e.g., Google Scholar, IEEE Xplore, ACM DigitalLibrary), looking for *peer-reviewed* papers that propose (or evaluate) PWD tailored for Chinese phishing websites (e.g., we categorically exclude any paper that does not mention "China" in the text). Importantly, we only consider works on ML-PWD: as we discussed (§2.1.1), using "signatures" is known to be effective—but only if the blocklist includes the corresponding URL. Hence, our goal is to identify if previously proposed solutions can work to counter "novel" (Chinese) phishing websites—which necessitates data-driven methods. Whenever we find a relevant paper, we use the snowball method (Wohlin, 2014) to look at downstream research. Despite reviewing dozens of papers, we could only find 10 peer-reviewed publications that specifically focus on countering phishing websites in China. We visualize the results of our analysis in Table 1. For each paper, we report: whether the experimental dataset is publicly *available* (✗ or ✓) and whether it also included websites from different *regions* than China (✗ or ✓); the *focus* of the ML-PWD (either generic, or for specific types of websites); the *date* (i.e., year) on which the data was collected; the types of *features* used for the analysis (F_u=URL only, F_h=HTML only, F_c=URL+HTML) and whether these features entail *Chinese-specific* characteristics; and if the conclusions are drawn after making *statistically significant* comparisons (✗ or ✓). To the best of our knowledge, Table 1 represents the state of the art of Chinese-PWD research.

As we can see from Table 1, some work (Chu et al., 2013; Li et al., 2020; Zhang et al., 2016) only consider ML-PWD analysing the URL,

thereby failing to capture the additional information provided by the HTML (which, as we show in our experiments, plays a crucial role).

The authors of Zhang et al. (2017b) devise a ML-PWD analysing HTML features that consider Chinese-specific word embeddings—which are clearly language dependent and are hence not appropriate in a cross-language setting (most Chinese websites have "Western" variants, e.g., Alibaba—cf. Fig. 7), which is crucial for our research (as we will explain in §2.3). The ML-PWD proposed by Zhang et al. (2017a) are assessed on both Chinese and English websites, achieving over 95% accuracy, but the corresponding evaluation is *(i)* not reproducible—since neither the code nor the data are publicly available; and *(ii)* lacks statistical validation—since the experiments are run only once, instead of being repeated many times (preventing a bias-free assessment (Arp et al., 2022)). Worryingly; both of these shortcomings affect most papers (i.e., Chu et al. (2013), Zhang et al. (2016), Li et al. (2016), Zhang et al. (2017b), Xiangdong et al. (2017), Li et al. (2020)) in Table 1. As a matter of fact, our experiments will reveal substantially different results than those reported by Zhang et al. (2017a). Notably, Zhang et al. (2014) proposed five Chinese-specific features and constructed ML-PWD analysing both the URL and HTML of a webpage, but they only focused on Chinese eCommerce websites (similarly to Chu et al. (2013)). Such a narrow focus also affects the research in Feng et al. (2017) and in Xiangdong et al. (2017), whose proposed ML-PWD are assessed only on financial websites. The recent work by Liu et al. (2021a) proposes a complex PWD, but their dataset includes *only 51 Chinese webpages*, which cannot represent the landscape of phishing in China (and, unfortunately, the source-code of Liu et al. (2021a) is not provided). Finally, and very recently, Jiang and Wu (2022) rely on a combination of feature-based and natural language processing techniques to detect "malicious" websites (not necessarily phishing): experiments on a dataset of 954 benign and 521 malicious (of which 221 are phishing) Chinese-only webpages revealed an accuracy of 85% (and 79% F1-score), which is an underwhelming result (especially since false positives are not mentioned in the paper) and although part of the data is publicly available, the source code is not openly released.

Our Goal. Prior research on Chinese PWD is scarce and presents limitations (e.g., lack of statistical validation, reproducibility, or generality). We aim to overcome all such shortcomings and provide reliable results *to assess, and then bridge,* the state of Chinese w.r.t. Western PWD.

2.3. Motivation, problem statement, and research questions

Prior research has shown that: signature-based PWD work well as long as the blocklist is kept up-to-date (§2.1.1); existing ML-PWD work well on Western websites (§2.1.2); and some papers also showed that ML-PWD can be tailored for Chinese websites with some success (§2.2.2). However Chinese websites are different from Western websites (§2.2.1, and we will expand this in §6.1). Such a difference led us to ask ourselves: *How do PWD that are effective on Western websites perform on Chinese websites (and vice versa)?* To the best of our knowledge, previous research cannot provide an answer to our question, since *(i)* Chinese and Western websites have been mostly treated independently; and/or *(ii)* the few works that consider both "regions" simultaneously have limitations.[3] This is because **no prior work scrutinized the "cross-regional" effectiveness of anti-phishing schemes.** Let us explain why this is crucial.

[3] Actually, the results of Zhang et al. (2017a) and Liu et al. (2021a) suggest that these two regions may be compatible from a PWD viewpoint!

2.3.1. A real-world problem

Plenty of Chinese people live in the West, and many Westerners live in China (OECD Ilibrary, 2021a; Migration Policy Institute, 2022a). People living abroad need to browse their home country's website. If Chinese PWD work poorly on western websites, then Westerners who live in China will be more likely to fall victim to Western phishing websites, and vice-versa. Even though it is well known that the Great Firewall prevents (Wikipedia, 2023n; Hoang et al., 2021) Chinese residents from accessing popular Western websites (e.g., Facebook), hence implicitly providing some form of protection to Chinese users against Western phishing websites, some can still be reached (e.g., GitHub). Moreover, usage of VPN services can bypass the Great Firewall, thereby allowing Chinese residents to access any.[4] website—but this will expose them to Chinese phishing websites. Indeed, Google Chrome is the most popular browser even in China (with a share of 56% (Statcounter, 2022i)); however, the anti-phishing schemes of Chrome are not tailored for Chinese phishing websites (as we show in §4.1.2). In turn, also users who live outside China (either Westerners or Chinese expats) and use Chrome (or similar browsers) can fall victim to phishing Chinese websites. Simply put, if PWD exhibit poor compatibility between Chinese and Western websites (as our analysis in §6.1 suggests), then many people can fall victim to phishing attacks that (perhaps inadvertently) exploit such a vulnerability.

2.3.2. Research methodology

Inspired to investigate this real-world problem, we want to verify if the gap between Chinese and Western phishing website detection truly exists—both in *research* and *practice*; and, if so, potentially find ways to close this gap. We tackle this problem by asking ourselves, and then answering, the following three research questions (RQ):

- RQ1: Do closed-source PWD (either blocklist- or ML-based) developed by practitioners (and deployed in the real world) work "equally well" on Chinese and Western websites?
- RQ2: Do open-source ML-PWD proposed in research for Western websites work "equally well" on Chinese websites? (We recall there are no open-source Chinese ML-PWD.)
- RQ3: Is it possible to "adapt" open-source ML-PWD originally tested only on Western websites so that they work equally well on both Chinese and Western websites?

Since Western websites are a broad term, we enrich our RQs by differentiating (a) 'English-only' websites from (b) 'generic' Western websites (e.g., Italian, German). Before focusing on our RQs, however, we must deal with a crucial obstacle: *finding the right data* to even test each of our hypotheses—and, specifically, the lack of publicly available datasets useable for Chinese PWD.

3. Data collection

Answering any of our RQ requires experiments which entail *(i)* assessing the effectiveness of PWD on *(ii)* Chinese and Western websites. Unfortunately, we were not able to find *any* existing resource that provided a representative dataset of the Chinese phishing website landscape. The datasets used by *prior research are not publicly available*, and the (few) *existing dataset only contains limited information*,[5] which point to websites that are outdated or no-longer active and hence do not allow to assess the effectiveness of PWD that analyse information

[4] These users can even fall victim to phishing in the dark web (Yoon et al., 2019)

[5] E.g., there are only 221 phishing URLs (without screenshot) in the dataset included in Yanting Jiang, Di Wu (2022b) (related to Jiang and Wu (2022)); whereas the majority of the entries in CN-Malicious-website-list Contributor (2017) are outdated; finally, Liu et al. (2021a) released their data, but it only has 51 phishing entries—preventing a sound analysis.

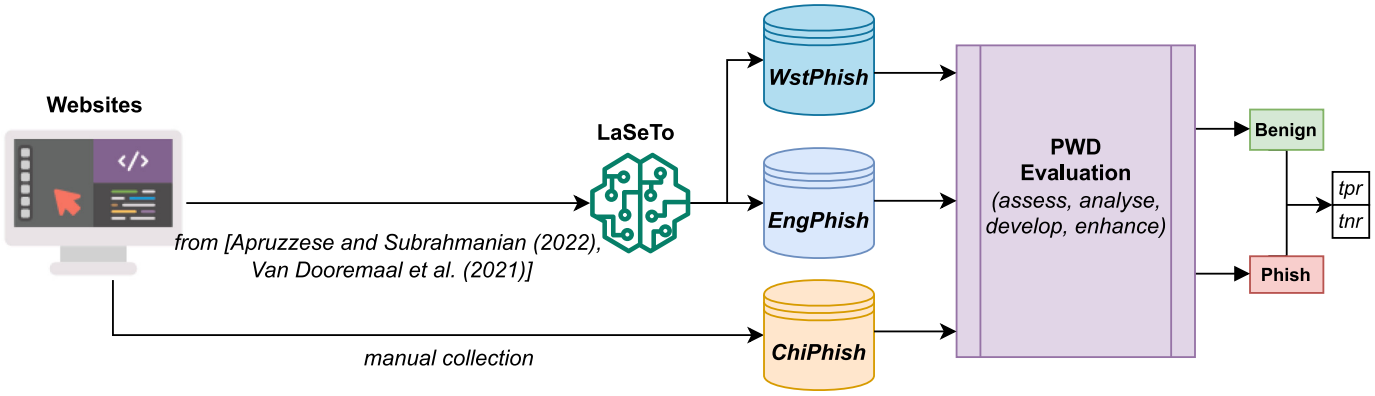


Fig. 3. Overview of our workflow.
We collect three datasets (one containing only Chinese websites, a second one only English websites, and a third one containing a mix of websites in popular Western languages) which we use to evaluate existing phishing website detectors, measuring their *tpr* and *tnr* (defined in §4.2).

not derived from the URL. In contrast, publicly available datasets having websites appropriate for Western PWD are typically provided with more information on each sample (e.g., URL, HTML, and occasionally also the screenshot).

Hence, as a first step towards answering our RQ (and our first contribution), we manually collect a dataset, `ChiPhish`, that enables the assessment of PWD on Chinese websites (§3.1). Then, we collect *two* datasets, `EngPhish` and `WstPhish`, that allow one to gauge how well existing PWD perform on websites "from the West" (§3.2); notably, for our data-collection procedure we also develop an original tool, LaSeTo, which we publicly release and will play a crucial role in bridging the gap revealed in our research (§7). We summarize our workflow in Fig. 3.

3.1. `ChiPhish`: the first comprehensive dataset for Chinese phishing website detection

Among the contributions of this paper is the first public dataset for Chinese-focused PWD, `ChiPhish`. To understand *why* such a contribution is significant, let us elucidate the difficulties we have to face (§3.1.1). We will then describe how we envision (§3.1.2) and create (§3.1.3) our `ChiPhish`.

3.1.1. Challenges

Although finding legitimate Chinese websites is trivial, finding *active Chinese phishing websites* is difficult. Even popular phishing tracking services (PhishTank) hardly report websites from China—likely because their trackers do not visit Chinese websites. Furthermore, a 2021 report by Interisle (Interisle Consulting Group, 2021b) also mentioned that their data under-represent Chinese phishing, since they did not collect any attacks against the four largest Chinese banks and major Chinese eCommerce companies. Such an "oversight" is confirmed by our difficulties in interpreting (and accessing) verified accounts of phishing attacks in China (§2.2), suggesting that this side of the World may (intentionally) be "closed" to foreigners.

As a matter of fact, the APWG (APWG, 2016) indicated that more than half of malicious gTLD registrations worldwide stem from China, and that *six of the top ten* registrars of malicious phishing domains were located in China and had primarily Chinese customers; this data was contributed by APAC which works with phishing targets in China. However, because of the Chinese cybersecurity law in 2022 (Cyberspace Administration of China, 2022c), APAC or other public cybersecurity platforms in China no longer broadcast Chinese phishing data. These difficulties may partly explain why the landscape of Chinese PWD has been mostly unexplored.

3.1.2. Design goals

Despite having to face many challenges, we seek to build a dataset for Chinese PWD that fulfils a twofold objective: *(i)* enable a meaningful analysis *for this paper*; and *(ii)* provide a solid foundation *for future work*. Hence, to create `ChiPhish`, we set three design goals:

- *Generality:* It must include websites (benign and phishing) of various types. This is to overcome the limitations of prior work which considered websites of only one type (refer to Table 1).
- *Representativity:* It must have a sufficient (≥1000) amount of samples (i.e., websites). This serves to allow comparisons with PWD on (more popular) "Western" websites.
- *Completeness:* Samples must be provided with three formats of raw data: URL, HTML, and screenshot. This serves to enable analyses of common phishing detection approaches (cf. §2.1).

Of course, we *cannot ensure* (nor we claim) that our `ChiPhish` dataset is (or will ever be) representative of the entire Chinese phishing landscape. However, by publicly releasing `ChiPhish`, downstream research can contribute to further expanding this dataset with additional samples.

3.1.3. Creation and overview

Let us explain how we collect `ChiPhish`. For the *benign* data, we relied on Chinaz (Chinaz, 2023b), a popular (Li et al., 2021) trusted source which provides a ranking of popular Chinese websites (similarly to Amazon's Alexa rankings). Specifically, our benign samples are taken by using the top-60 websites reported by Chinaz (in Oct–Dec 2022) and scraping the links contained in these websites (which we manually verified pointed to trusted websites). As for the *phishing* data, we had to draw from various sources. We searched across the Threat Intelligence Centers of Chinese IT companies (e.g., VenusEye (Venustech, 2023l), QiHoo 360 (360 secure brain, 2021)), competition platforms and security forums (e.g., kafan (KaFan, 2023d)) to retrieve hundreds of Chinese phishing websites—which we manually checked to ensure that they were still online. Importantly: all phishing samples have been verified by the publishers of the respective source, as well as by ourselves (during our manual checks). Whenever we visit a website, we first check whether it is online; if so, we then store its URL, and save the entire HTML of the landing web page (including potential javascripts) as well as its whole-page screenshot (in high resolution). To provide useful and up-to-date resources for future research, we collect our phishing entries in two different points in time: in Oct–Dec. 2022 (during which we collect 372 entries), and in Jul–Aug. 2024 (during which we collect 193 entries). Overall, we collect 1 055 benign and 565 phishing Chinese websites. A summary of `ChiPhish` is in Table 2. To the best of our

Table 2
Summary of websites in ChiPhish. We only provide examples of *benign* websites (to protect readers).

Category	#Benign	#Phishing	Example
eCommerce	173	124	1688.com
finance	59	84	boc.cn
education	121	15	eol.cn
government	2	4	cwl.gov.cn
health	23	27	99.com.cn
email	10	22	163.com
information	267	84	labs.zol.com.cn
news	96	11	thepaper.cn
search engine	22	10	360.cn
social	35	40	weixin.qq.com
entertainment	247	96	kuwo.cn
other	0	48	n/a

knowledge, ChiPhish is the only publicly available dataset for Chinese PWD with the characteristics described in §3.1.2.

Finally, we emphasize that we carried out most of the experiments described in the remainder of this paper at the beginning of 2023, i.e., when only 372 of the phishing entries in ChiPhish had been collected. To provide more comprehensive results, we expanded our set of 372 phishing entries by also considering 193 phishing entries from the dataset[6] by Jiang and Wu (2022). This means that the set of phishing webpages used in our experiments amounts to 565 entries—all of which are verified phishing websites that were uploaded to the Web up to December 2022. Such a setup provides a temporally consistent timeframe for phishing and benign entries, enabling a fair evaluation. For simplicity, we will still refer to this dataset as "ChiPhish", but we do not claim the 193 phishing samples from Jiang and Wu (2022) to be part of our contributions.

3.2. Datasets for "Western" Phishing Website Detection (in phonological languages)

Our focus is comparing Western and Chinese PWD. However, "Western" is a broad term: recent surveys reveal that English is the global and most spoken language (Statista, 2022d), being also the most commonly used website content language (with a share of over 57% (W3Techs, 2023k)). To account for the predominant usage of English on Web, we must consider it a stand-alone 'population' with reference to (w.r.t.) other Western languages. Besides allowing for a bias-free evaluation, providing such a twofold perspective also allows us to ascertain whether PWD perform similarly across websites using *different phonetic western languages*, thereby serving as a (potential) validation mechanism. Therefore, we seek to collect two datasets: EngPhish, containing only English websites; and WstPhish, containing websites of the most popular Western languages.

3.2.1. Preliminary investigation

There are many datasets for "western" PWD, i.e., having websites in phonological languages. However, *most of such datasets do not enable* an evaluation that can provide an unbiased answer to our RQs. For example, the dataset used in Mowar and Jain (2021) only includes the URL of its websites, preventing retrieval of any data on the corresponding HTML (phishing webpages are taken down quickly) at a later time. The same problem affects the dataset proposed by Abdelnabi et al. (2020), which reports the screenshot but neither the URL nor the HTML of its samples. Finally, the well-known datasets proposed by Hannousse and Yahiouche (2021) and Mohammad et al. (2014b) are only provided as pre-computed features, thereby preventing the retrieval of the original data on the corresponding website. After surveying the few existing datasets that provide complete information on each sample contained therein, we found two recent ones which have been *validated by the research community*: the one in Van Dooremaal et al. (2021) and the one in Apruzzese and Subrahmanian (2022). However, we observed that both of these datasets contain websites of diverse languages—which, for the sake of our assessment, demanded further analyses.

3.2.2. Language Selector Tool (LaSeTo)

To create WstPhish and EngPhish, we must put some order in the "mixture" of websites contained in the datasets in Van Dooremaal et al. (2021) and Apruzzese and Subrahmanian (2022). To this purpose, we develop an original Language Selector Tool—or LaSeTo, for short. LaSeTo fosters two elements of the HTML alongside Google's Compact Language Detector v3 (CLD3), i.e., an open-source system that leverages state-of-the-art techniques for language identification, supporting over 100 languages (Google, 2020). Specifically, LaSeTo considers: *(i)* the 'lang' HTML attribute—used to declare the language of a webpage; and *(ii)* the language used in the HTML 'title' tag—which can also suggest the primary language of the userbase of a given website. Practically, LaSeTo receives the raw HTML of a webpage as input: if it can detect the 'lang' attribute, it will output the corresponding language; otherwise, it will query CLD3 with the 'title' tag, and provide the corresponding language as output. This design choice (i.e., looking for the 'lang' attribute first) is to save computational runtime: we experimented with LaSeTo, and the 'lang' attribute (if present) requires half the time ($0.022s$ vs $0.043s$) to produce an output, which was almost always the same as CLD3. We release the source code of LaSeTo in our repository.

3.2.3. Creating EngPhish and WstPhish

For consistency, we will use the same source to build each of our "Western" datasets. Let us explain our procedure in more detail, motivating our choices.

- **EngPhish**: Since we want an English-only data corpus, we chose the *latest* suitable dataset as a starting point. Specifically, we consider the dataset provided in the 2022 paper from Apruzzese and Subrahmanian (2022), which contains nearly 24k samples (16k benign and 8k phishing). However, not all of these are English websites: we hence submit all 24k samples to LaSeTo, finding that 15 111 are in English (specifically, 4 092 phishing and 11 019 benign). Overall, these samples represent EngPhish.
- **WstPhish**: Our last dataset should include webpages representing a broad coverage of "western" languages. Hence, we use LaSeTo to extract a subset from the websites provided in the 2021 paper by Van Dooremaal et al. (2021). This data corpus entails almost 4M websites, of which 100k are phishing (taken from various repositories); some samples are 'blank' webpages, which we ignore. To create WstPhish, we begin by considering 17 phonologic languages from the list of most common (Wikipedia, 2023e) European languages (besides English), i.e.: German, Italian, French,

[6] We reached out to the creators of Yanting Jiang, Di Wu (2022b) and were given permission to use these samples for our research. The breakdown of these 193 phishing webpages is as follows: 39=eCommerce, 22=finance, 16=education, 14=government, 0=health, 7=email, 24=information, 0=news, 1=search engine, 6=social, 43=entertainment, 21=other.

Table 3
Summary of the datasets used in our evaluation. For ChiPhish, we use 372 phishing samples we collected in Oct–Dec 2022, and 193 phishing samples from Jiang and Wu (2022) (see Section 3.1.3); the parentheses report the total number of phishing samples in ChiPhish as of 2024 (all of which have been collected by us).

Dataset	#Benign	#Phish	Used in
WstPhish	4 269	6 935	Van Dooremaal et al. (2021), Apruzzese et al. (2022a)
EngPhish	11 019	4 092	Apruzzese and Subrahmanian (2022)
ChiPhish	1 055	565 (372)*	Jiang and Wu (2022)

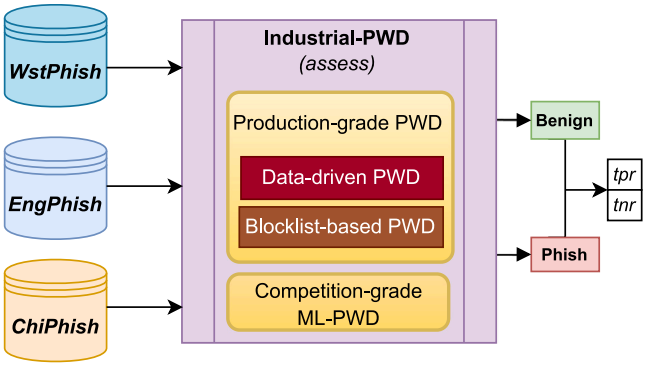


Fig. 4. Assessment of industry-developed PWD.
We test 72 PWD (leveraging various detection methods) developed by industry practitioners on our three datasets, and measure the respective *tpr* and *tnr* (defined in §4.2).

Swedish, Polish, Spanish, Norwegian, Hungarian, Czech, Danish, Dutch, Greek, Turkish, Slovenian, Croatian, Romanian and Luxembourgish. Then, we run LaSeTo on the websites, saving all webpages that match any of these 17 languages. We thus obtain 4 269 benign and 6 935 phishing webpages, which represent our WstPhish.

To make our experiments reproducible, we provide these datasets in our repository. We summarize the statistics of our three considered datasets in Table 3. In what follows, we will use these datasets to assess existing anti-phishing schemes, and develop new ones (see Fig. 3).

4. Assessment of Phishing Website Detectors by practitioners (RQ1)

We begin our assessment by focusing on our first RQ, for which we must test "how well" PWD developed by practitioners can detect phishing/benign webpages "from the West" and "from China" (§2.3). To this purpose, we consider a total of 72 "closed-source" detectors, encompassing both operational products/services and competition-grade systems. We first explain our choices (§4.1), and the present the results of this assessment (§4.2). An overview is in Fig. 4

4.1. Selection of "closed-source" detectors, and experimental workflow

Several industry-developed tools exist that can be used to determine whether a website is malicious. These tools accept as input either the URL or the HTML of a website, and then analyse such input in a black-box manner, and can rely either on up-to-date blocklists, but they may also query third-party services that perform a deeper analysis (e.g., Virustotal (Virustotal, 2023m), Netcraft (Netcraft, 2023f), or PhishDetector (PhishDetector Contributor, 2023h)). To provide a meaningful answer to RQ1, we select 72 PWD developed by industry practitioners. Despite such PWD being "closed-source" (we cannot inspect their low-level implementation), there exist some publicly available information that allows us to infer their overarching functionality. We consider: 62 data-driven PWD (§4.1.2), 2 blocklist-based PWD (§4.1.1), and 8 ML-driven PWD (§4.1.3). Let us elaborate on which, why, and how we use each of these.

4.1.1. Blocklist-based (production-grade) PWD

To provide a complementary perspective to data-driven (production-grade) PWD, we also consider PWD whose output is (to the best of our knowledge) based on blocklist, and hence accept as input the URL of a given website. For this analysis, we rely on two anti-phishing solutions: **VirusTotal** (which we use for its URL variant,[7] and not for the HTML used in §4.1.2), which is popular in the West and widely used by prior work (Peng et al., 2019b; Cheng et al., 2018; Dambra et al., 2023; Choo et al., 2023; Acharya and Vadrevu, 2021; Kondracki et al., 2021); and **VenusEye** which is a PWD *developed by a Chinese vendor* (Venustech, 2023l). Importantly: for our experiments, we searched for (freely available) anti-phishing services provided by Chinese companies and that we could use for our research; unfortunately, we *could not find any such service*. Existing Chinese solutions can only be accessed via paywalls, and we could not find any that could leverage data-driven heuristics. The only one we could find is the (blocklist-based) VenusEye—but even this has a limitation: each query must be manually submitted. Indeed, while other services popular in the West facilitate analyses by providing APIs that can be used to send queries in bulk, for VenusEye there is no working API,[8] and we must submit each URL individually, wait for its response, and then repeat this process again. For this reason, the assessment of blocklist-based (production-grade) PWD will entail a subset of our data—and, specifically, we will only consider a (randomly chosen) set of 200 samples (100 from WstPhish and 100 from EngPhish). This is, however, not a problem: it is well-known that blocklist-based PWD work *only if they include the corresponding URL in their list*. Hence, this analysis serves merely to ascertain if these blocklists are well-maintained by the respective vendors. Please note: we do not consider *ChiPhish* here because many samples in ChiPhish have been validated thanks to VenusEye (making the comparison unfair).

4.1.2. Data-driven (production-grade) PWD

For an exhaustive assessment, we rely on two popular "production-grade" anti-phishing services: **GoogleSafeBrowsing**, which accepts the URL of a web page as input, whose landing webpage is allegedly analysed by various data-driven methods (Tarun Bansal, 2023a; Miao et al., 2023); and **VirusTotal**, which accepts HTML as input, since its output accounts for the responses of dozens of scanners (in contrast, Netcraft and PhishDetector only consider the response of a single tool); it is conjectured (Liras et al., 2021) that some scanners in VT adopt data-driven heuristics to perform their analyses on malicious samples. We use these tools as done by prior work (e.g., Peng et al. (2019b)): we submit the corresponding input (URL for GoogleSafeBrowsing, and HTML for VirusTotal) to each tool and observe their output. For VirusTotal, every query.[9] corresponds to having 61 PWD (each leveraging

[7] The URL version of VirusTotal inquires ≈96 blocklists from various vendors. To assess the maximum effectiveness of VirusTotal, we consider an output to be "malicious" if at least one vendor says so. Such a straightforward detection mechanism is why we consider the URL version of VirusTotal as a single PWD in this paper.

[8] We tried using the provided link https://venuseye.com.cn/api/ but it is not functional when we submitted our samples. We even contacted the developers, explaining the issue, but we received no response.

[9] We perform our analysis in Jan. 2023, but this number can change (Wang et al., 2023) At that point in time, the detectors queried by VirusTotal are 78, but 17 of these returned an error, so we will not consider them.

proprietary detection methods) to analyse the corresponding sample—allowing us to provide a broad perspective on the detection capabilities of real systems.

4.1.3. ML-driven (competition-grade) PWD

We find it instructive to provide a yet another perspective, and hence consider 8 "competition-grade" PWD—for which we have *certainty* that they leverage ML techniques (although we do not know which specific methods, despite being obviously data-driven). Specifically, we consider the anti-phishing detectors provided for the **Machine Learning Security Evasion Competition** (MLSEC) organized by CujoAI (CujoAI, 2022). These ML-PWD (8 in total) analyse the raw HTML of a webpage as input, and provide a 'phishing' confidence (within [0–1] range, with 0 denoting a benign sample and 1 denoting a phishing sample) as output. The organizers of MLSEC allowed the research community to use their ML-PWD for three months after the challenge ended in September 2022. We took this opportunity to test these detectors on the raw HTML of every webpage in our three datasets—thereby ensuring a consistent setup as the one for the HTML of VirusTotal (§4.1.2).

4.2. Results (do industry-developed PWD work well on Chinese and Western websites?)

Let us report the results of our assessment, starting from the production-grade PWD (data-driven in §4.2.1 and blocklist-based in §4.2.2), and finishing with the competition-grade ML-PWD (§4.2.3). The performance metrics of choice are the *true positive rate* (tpr), which is the percentage of phishing websites classified as malicious; and the *true negative rate* (tnr), which is the percentage of benign websites classified as benign—and which can be used to derive the false positive rate (fpr) by doing fpr=1-tnr. A PWD has good quality if both tnr and tpr are close to 1.

4.2.1. Performance of Data-driven (production-grade) PWD

We assess the capabilities of **GoogleSafeBrowsing** (GSB). We submit all the samples in our three datasets (`ChiPhish`, `EngPhish` and `WstPhish`) to the GSB API (Google, 2023c) (which accepts URL as inputs) and we record how many webpages trigger a "suspicious" response. We consider suspicious samples as positive instances while benign samples as negative ones. Hence, we calculate the tpr and tnr of each detector by comparing its prediction with the respective ground truth. The results are as follows:

- `EngPhish`: tpr=0.043 (176 phishing samples are detected), with no false positives (tnr=1.0).
- `WstPhish`: tpr=0.004 (26 phishing samples are detected), with one false negative (tnr=0.999).
- `ChiPhish`: tpr=0.002 (only 1 phishing sample is detected) with no false positives (tnr=1.0).

At the same time, we submit the raw HTML of every sample in each of our three datasets to **VirusTotal** (VT), which automatically forwards it to 61 cyber detectors (provided by security companies) and then reports the label ('malicious' or 'benign') of each sample. The results (see Table B.11) align with GSB's, showing that all detectors achieve a perfect tnr, but perform *terribly* in identifying the phishing samples in `ChiPhish`, with an average tpr of 0.004.[10] And, these detectors perform not-very-well also on `WstPhish` and `EngPhish`: the average tpr is 0.04 and 0.11, respectively (which are, however, 10 and 30x better than the average tpr on `ChiPhish`)[11]. These (potentially underwhelming) results echo those in prior work, showing that such anti-phishing schemes have many blind-spots (Peng et al., 2019b).

It is apparent that even important anti-phishing tool, deployed in popular web-browsers, is unable to identify Chinese phishing websites,

Takeaway. Out of 62 production-grade (data-driven) PWD, none can reliably detect Chinese phishing websites. In contrast, some exhibit a subpar detection rate for Western phishing websites. The rate of false alarms is always low, showing that these PWD might be tuned to minimize false positives.

4.2.2. Performance of Blocklist-based (production-grade) PWD

We turn the attention to blocklist-based PWD. As we explained (in §4.1.1), we only consider a select subset of our datasets for this proof-of-concept experiment. We randomly sample 100 phishing samples from both `WstPhish` and `EngPhish`, and submit the corresponding URL to VenusEye (in March 2023). Accordingly, 74 and 82 of the samples from `WstPhish` and `EngPhish` are flagged as phishing. Then, we submit the exact same samples to VirusTotal (by using its URL variant) which achieved[12] a 97% tpr.

Takeaway. Chinese production-grade PWD (using blocklists) cannot detect 20% of our submitted Western phishing samples—which are perfectly detected by "Western" counterparts. This suggests that the blocklists of our considered Chinese-PWD are not kept up-to-date with Western phishing websites.

4.2.3. Performance of ML-driven (competition-grade) PWD

Lastly, we submit the raw HTML of all samples in our three datasets to each of the 8 "black-box" ML-PWD of MLSEC. The output of each of these detectors is a confidence score (from 0.0 to 1.0): for this experiment, we consider a score that is higher than 0.5 to denote a "phishing" prediction (and "benign" otherwise). We then check these predictions against the corresponding ground-truth, and derive the tpr and tnr. The results are in Fig. 5, showing the distribution of the tpr and tnr across the 8 MLSEC detectors for each of our three datasets (the detailed performance of each detector is in Table B.15). From Fig. 5 we see that these detectors perform much better on `WstPhish` (avg tpr=0.60) and `EngPhish` (avg tpr=0.64) compared to `ChiPhish` (avg tpr=0.27). Even though these detectors are for competitions,[13] these results further show that *Chinese websites are rarely accounted for* when designing ML-PWD.

Answer to RQ1: PWD developed by security practitioners exhibit underwhelming performance in cross-regional contexts. Anti-phishing products "from the West" work poorly on Chinese phishing websites; whereas blocklist-based PWD *in China* are not as well-maintained as their Western counterparts for Western phishing websites. Minimizing the fpr takes priority.

5. Assessment of Phishing Website Detectors *by researchers* (RQ2)

Having demonstrated that the current landscape of industry-developed PWD is not equipped to simultaneously cover both Western and Chinese phishing websites, we now turn the attention to our second RQ. We select 10 ML-PWD, analysing various types of information (cf. §2.1.2), proposed by related research, and whose code is open-source (§5.1). Such a characteristic allows us, after assessing their performance (§5.2), to perform a critical analysis (§5.3) of their detection mechanism, thereby highlighting the pros-and-cons of these state-of-the-art PWD for the sake of our goal. An overview of our experimental workflow is in Fig. 6.

[10] The best detector, *AVG*, has a tpr of 0.03 in `ChiPhish`.

[11] The best detector achieves a tpr of 0.49 on `WstPhish` and 0.52 on `EngPhish`.

[12] N.B.: these results should *not be compared* to those in Table B.11: we are submitting the URL here, and the HTML for Table B.11.

[13] The organizers of MLSEC admittedly tweaked their ML-PWD to make evasion harder (explaining the underwhelming tnr).

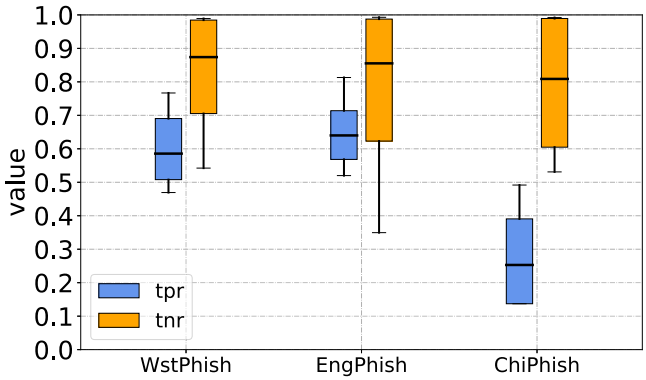


Fig. 5. Performance of the ML-PWD of MLSEC.
Boxplots show the distribution of the *tpr* and *tnr* among the 8 ML-PWD in the MLSEC anti-phishing evasion challenge on our three datasets: `WstPhish`, `EngPhish` and `ChiPhish`.

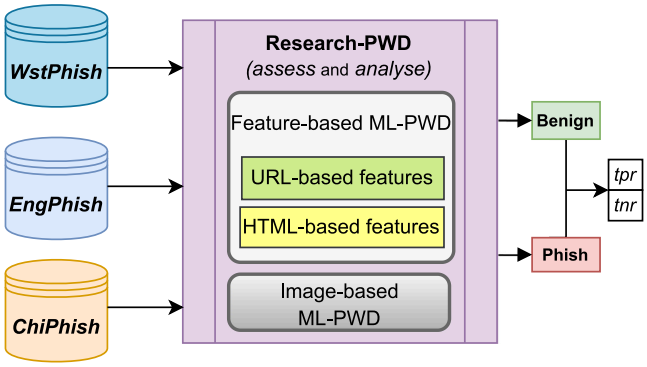


Fig. 6. Assessment of ML-PWD proposed in research.
We test 10 ML-PWD (leveraging various detection methods) proposed in prior research on our three datasets, and measure the respective *tpr* and *tnr*.

5.1. Considered "open-source" ML-based detectors, and experimental workflow

To provide a comprehensive answer to RQ2, we consider 10 representative ML-PWD that have been recently proposed in research papers accepted at top-venues. Specifically, 9 are from SpacePhish (§5.1.1), whose artifact received a *Reusable Badge* at ACSAC'22 (Apruzzese et al., 2022a); and 1 is from PhishIntention (§5.1.2), from USENIX SEC'22 (Liu et al., 2022a). Importantly: we only consider ML-based PWD because there is no point in assessing those based on blocklists (an inspection of the corresponding blocklist would immediately reveal whether they would work or not); and we do not consider Chinese-specific ML-PWD because no previous paper has its source-code openly accessible today (cf. §2.2).

5.1.1. SpacePhish (feature-based)

This research paper (Apruzzese et al., 2022a) was published in ACSAC'22, its source-code is fully available on GitHub (Apruzzese et al., 2022h), and its results have been reproduced by downstream research (e.g., Montaruli et al. (2023), Yuan et al. (2023)). The paper entailed a replication of well-known "feature-based" ML-PWD, i.e., whose detection mechanism relies on the analysis of 57 (mostly binary) features proposed by acclaimed prior work (e.g., Mohammad et al. (2014a), Xiang et al. (2011), Marchal et al. (2016), Jain and Gupta (2018a)). At a high-level, such features can be divided into two groups: URL-based, which are computed by using the URL-string as input (e.g., length of the URL, or number of subdomains); and HTML-based, which are computed by analysing the HTML of the web page (e.g., internal objects, DOM elements, or javascript). Using the source-code of SpacePhish (Apruzzese et al., 2022a) enables one extract 57 features describing a given web page. After this extraction process, the feature representation of the web page (i.e., a sample) can then be provided as input to an ML model, tasked to analyse the sample and determine whether it is benign or malicious. SpacePhish implemented a total of 9 ML-PWD, each considering a specific group of features (F_u=URL-only, F_h=HTML-only, or F_c=both URL and HTML) and a specific ML algorithm (RF=random forest, LR=logistic regression, or CNN=convolutional neural network). To train these ML-PWD, the authors of SpacePhish use a subset of 4000 websites (2k benign, 2k phishing) taken from the dataset proposed in Van Dooremaal et al. (2021), i.e., the same source that we used to create `WstPhish` (albeit our `WstPhish` is 2.5x larger, and the choice of websites follows a specific criteria instead of being randomly drawn—see §3.2.3). For our assessment, we consider all these 9 ML-PWD, thereby allowing for a comprehensive (and fair) evaluation of such feature-based ML-PWD. The procedure is simple: first, we use the open-source code of SpacePhish (Apruzzese et al., 2022h) and re-implement their 9 ML-PWD (ensuring that they obtain the same performance as in Apruzzese et al. (2022a)). Then, we use the feature extractor of SpacePhish to produce the feature representation of every website in our three datasets (`ChiPhish`, `WstPhish`, `EngPhish`). Finally, we submit every sample (i.e., a feature vector of 57 numerical values) to each of the 9 ML-PWD of SpacePhish, and measure the corresponding *tpr* and *tnr* for each of our datasets.

5.1.2. PhishIntention (image-based)

In their USENIX SEC'22 paper, Liu et al. (2022a) propose PhishIntention—an updated variant of a (relatively) new class of anti-phishing schemes, which rely on the capabilities of deep learning to analyse images. Following the trend initiated by Abdelnabi et al. (2020) and Lin et al. (2021), PhishIntention seeks to identify phishing webpages that are visually similar to popular (benign) webpages that are frequently visited by a large pool of users. The intuition is that

phishers try to lure their victims to websites they are familiar with (e.g., PayPal) and, if such users land on a webpage which is very similar to what they expect (e.g., the real PayPal), then the users could fall for the phishing trap and input their credentials. PhishIntenion aims to detect such (malicious) webpages, which is done by analysing (via deep learning) the visual representation of any given webpage, and checking whether it is similar to the one of a popular website (e.g., PayPal, Google). If the result of such an analysis exceeds a certain similarity threshold (θ), then the page is considered similar, and this will trigger a domain verification mechanism (i.e., if the page is similar to PayPal, and it is hosted under the same domain as PayPal, then it is a benign page by PayPal!): if the domain matches, then no action is taken (i.e., the web page is likely benign); otherwise, an alarm is raised (i.e., the web page is likely phishing). Therefore, to test PhishIntention, we proceed as follows. First, we take the exact models provided in the corresponding GitHub repository (Liu et al., 2022f), which are pre-trained. Then, we submit the screenshot of every web page in ChiPhish [14] (we will explain in §5.3 why we consider only ChiPhish) to the models of PhishIntention. Finally, we analyse the results: if the output exceeds the similarity threshold (for which we use the exact same value determined in PhishIntention, i.e., θ=0.87) then the page is phishing (we already know that the domain does not match); and benign otherwise.

5.2. Results (do ML-PWD "from research" work well on Chinese and western websites?)

Let us report the results of our empirical assessment of these "open-source" ML-PWD.

5.2.1. Do feature-based ML-PWD work?

First, we consider the 9 detectors from SpacePhish (Apruzzese et al., 2022a). To provide a statistically-significant testbed, we repeat our assessment 10 times—each by training any given ML-PWD (using a specified algorithm and feature-set) anew on a different (randomly chosen) portion of the training dataset used in SpacePhish (which, we recall, is different from any of our three dataset). We report the results in Table 4, wherein cells report the average (and std) tpr and tnr across our 10 trials on each of our datasets. By observing these results, we see that the performance (in terms of tpr and tnr) tends to be highest for the detectors analysing all features (i.e., F_c), which makes sense because they use a superior amount of information to make their decisions. Furthermore, we also see that the RF-based models tend to be slightly better than LR and CNN, as evidenced by an overall lower tnr (which translates to a lower false positive rate—which is preferred for practical PWD). Nonetheless, we also see that the best performance is obtained on WstPhish: this is an important observation, because WstPhish is drawn from the same distribution as the samples used to train these ML-PWD (Apruzzese et al., 2022a), hence such a result confirms that our implementation is correct. Then, we see that the performance on EngPhish is appreciable, with the F_c variant of all these detectors being able to achieve perfect tnr and above 0.8 tpr. Finally, and worryingly, the performance on ChiPhish is underwhelming: despite achieving a (relatively) good tnr (for F_c: RF=0.95, LR=0.98, CNN=0.86), the tpr is unacceptably low (for F_c: RF=0.39, LR=0.42, CNN=0.37).

Takeaway. The ML-PWDs of Apruzzese et al. (2022a) cannot detect >58% of the phishing webpages in ChiPhish.

[14] For the phishing webpages from Jiang and Wu (2022) we manually extract their screenshot by displaying them in our browser.

5.2.2. Do image-based ML-PWD work?

Next, we focus on PhishIntention (Liu et al., 2022a), which we test on ChiPhish. Out of the 565 phishing samples in ChiPhish, 561 (99,3%) trigger a "no target" response by PhishIntention: they are too different from any web page "seen" by the models of PhishIntention, and are hence flagged as benign (i.e., they evade detection). The remaining 4 trigger some similarity, and we further inspect these results: 3 are (phishing) webpages that mimic the Chinese version of Apple, whereas 1 is mimicking Netease (all of which are brands whose web pages are included in the training data of in PhishIntention). However, the similarity of these is: 0.69, 0.84, 0.84, 0.72: all such values are *below the threshold* (θ=0.87) that would induce PhishIntention to proceed with the domain checking (and which would lead to a phishing output). Hence, these webpages are also classified as benign. Notably, in ChiPhish there are 20 more phishing samples that mimic the Chinese Apple, but they also yielded "no target" (i.e., they also evaded PhishIntention).

Takeaway. None of the 565 phishing samples in ChiPhish are detected by PhishIntention (Liu et al., 2022a).

Answer to RQ2: Open-source ML-PWD proposed by prior research (which did not specifically account for Chinese websites) are not well-equipped to detect Chinese-phishing websites.

5.3. Critical Analysis (why are the results the way they are?)

The extensive documentation and open-source code of these ML-PWD from research allows us to explain our results. Hence, we take a step back, and review how our considered state-of-the-art ML-PWD work to perform their detection. Our underlying objective is identifying whether there is room for improvement (and also justifying why we only considered ChiPhish in §5.2.2).

5.3.1. Review: using ML to detect phishing webpages

According to Corona et al. (2017), ML-PWD approaches can be divided in two categories: target *dependent*, and target *independent*. The former aim to detect phishing samples that focus on a specific target, whereas the latter seek to detect phishing without making any assumption whatsoever. For instance, the ML-PWD of SpacePhish (Apruzzese et al., 2022a) (considered in our evaluation) are all target independent: after training on a broad set of benign and phishing samples, they aim to infer whether any 'test' sample is benign or phishing. In contrast, state-of-the-art image-based PWD mostly follow a target dependent approach (even in practice, e.g., Draganovic et al. (2023), Apruzzese et al. (2023a)). Albeit there exist target-dependent ML-PWD that do not use images (e.g., Tan et al. (2016)), *we are not aware of target-independent ML-PWD that do use images*: interestingly, Marchal et al. (2016) propose ML-PWD that relies on various features (most of which overlap with those in SpacePhish), and despite stating that screenshots are an "information source", the proposed features *do not use the screenshot*. To further justify this claim, we perform an original experiment which yielded *negative results* (discussed in Appendix E).

5.3.2. Target-dependent ML-PWD using images

These approaches focus on catching phishing websites that "target a specific brand". The intuition is that most phishing attacks try to lure victims to (malicious) websites that resemble a reputable brand. Specifically, instead of inferring whether a website is benign or phishing, these approaches seek to identify whether a website (or a part of it) is "similar" to another website (or a part of it) that is known to be benign. If this is true, then this finding is used to verify whether other elements of the website (e.g., its domain) match with those of the known brand.

The reason of this two-step approach is due to *efficiency*. Indeed, querying third-party websites for domain is expensive, so it is only

Table 4
Performance of the ML-PWD of SpacePhish. We replicate the exact ML-PWD of Apruzzese et al. (2022a) and test them on our datasets, reporting the avg (and std.) *tpr* and *tnr* across 10 independent trials.

Alg.	Feature Set	WstPhish		EngPhish		ChiPhish	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	F_c	0.92 ± 0.013	0.98 ± 0.005	0.84 ± 0.010	1.00 ± 0.000	0.39 ± 0.044	0.95 ± 0.021
	F_u	0.92 ± 0.010	0.97 ± 0.007	0.86 ± 0.010	1.00 ± 0.000	0.36 ± 0.037	0.96 ± 0.014
	F_h	0.64 ± 0.026	0.96 ± 0.008	0.59 ± 0.030	0.98 ± 0.004	0.19 ± 0.041	0.92 ± 0.016
LR	F_c	0.90 ± 0.007	0.98 ± 0.004	0.80 ± 0.020	1.00 ± 0.000	0.42 ± 0.051	0.98 ± 0.013
	F_u	0.92 ± 0.005	0.97 ± 0.007	0.81 ± 0.020	1.00 ± 0.000	0.53 ± 0.052	0.92 ± 0.027
	F_h	0.62 ± 0.009	0.86 ± 0.010	0.61 ± 0.010	0.91 ± 0.010	0.35 ± 0.048	0.87 ± 0.012
CNN	F_c	0.91 ± 0.019	0.99 ± 0.005	0.84 ± 0.020	1.00 ± 0.001	0.37 ± 0.069	0.86 ± 0.038
	F_u	0.92 ± 0.007	0.97 ± 0.007	0.85 ± 0.010	1.00 ± 0.002	0.38 ± 0.028	0.78 ± 0.048
	F_h	0.53 ± 0.024	0.85 ± 0.025	0.55 ± 0.030	0.94 ± 0.010	0.34 ± 0.043	0.81 ± 0.030

done if there is risk that the page is actively trying to mimic a well-known website (Lee et al., 2023). Abundant works have proposed target dependent approaches reliant on visual similarity. Notable examples include the seminal work by Fu et al. (2006), and the one by Geng et al. (2013) focusing on *favicons*. More recently, we mention Corona et al. (2017), Abdelnabi et al. (2020) and, our considered PhishIntention (Liu et al., 2022a). Unfortunately, the main limitation of these approaches is that they only work if the phishing webpage tries to resemble one of the targeted brands—which is typically referred to as *protected set* (PS). Such a peculiarity makes them "similar" to blocklist-based approaches: they will work only as long as such list (i.e., the PS) covers the brands that will be seen at test-time. Hence, although we acknowledge that phishers tend to target well-known brands, *target-dependent ML-PWD will fail by design to detect* any phishing webpage that targets a brand not included in the PS.[15]

5.3.3. Shortcomings of visual PWD: a case study

Let us link the information provided insofar to the problem tackled by our paper: the gap between Chinese and Western PWD. To provide evidence that *existing* (target dependent) PWD reliant on visual similarity are not-well-equipped to handle Chinese websites, we perform an in-depth look at the brands included in the PS of the most acclaimed works. We do so by asking ourselves the two ancillary questions (AQ):

AQ1: how many of these brands are Chinese?
AQ2: how many of these Chinese brands are in the top30 Chinese websites (Chinaz, 2023b)? (June 2023)

The rationale is that if these methods entail many (top-visited) Chinese websites in their PS, then these methods would be (somewhat) effective to counter Chinese phishing websites. Unfortunately, the results of this case study, shown in Table 5, reveal that this is not the case.

Table 5
Case study on image-based ML-PWD. We scrutinize how many brands included in the "training" datasets of popular image-based (and target-dependent) ML-PWD are from China. (N/A=data not public)

Work	PS size	# Chinese in PS (AQ1)	# top30 Chinese in PS (AQ2)
Fu et al. (2006)	8	1	0
Geng et al. (2013)	81	N/A	N/A
Corona et al. (2017)	1012	2	0
Dalgic et al. (2018)	14	1	0
Van Dooremaal et al. (2021)	8	N/A	N/A
Abdelnabi et al. (2020)	155	3	0
Lin et al. (2021)	181	5	1
Liu et al. (2022a)	277	5	1
Apruzzese et al. (2023a)	40	1	0

We can see that most approaches have a PS with variable size, spanning between less than 10 to few hundreds brands (the exception is DeltaPhish (Corona et al., 2017), which focuses on compromised websites and has a slightly different focus). However, the corresponding PS have *at most five Chinese brands* in them, and none of these are included in the top30 Chinese websites. To provide further evidence, let us focus on VisualPhishNet (Abdelnabi et al., 2020) and PhishPedia (Lin et al., 2021) (and also PhishIntention (Liu et al., 2022a)): the former has only 3 Chinese brands (Alibaba, Aliexpress, made-in-china), whereas the latter has 5 (Alibaba, SFexpress, Netease, made-in-china, global sources HK). This means that, at best, the corresponding PWD models can detect only Chinese phishing websites mimicing those of these six brands. However, we make two interesting observations (which we explain through Figs. 7):

- *Five out of these six brands are not in the top30 Chinese websites.* (The exception is NetEase, which is included in Lin et al. (2021) and Liu et al. (2022a).) This is because Chinese websites that are also visited in the West have a different domain: for instance, "alibaba.com" (included in Lin et al. (2021)) is less popular than "1688.com" (not included in Lin et al. (2021)) in China—despite referring to the exact same brand.
- *The visual content in these datasets has a mismatch between the Chinese and Western versions of a brand.* E.g., Lin et al. (2021) includes the logo for "chinese.alibaba.com" (Fig. 7(a)) but not the one for the Western version of Alibaba (i.e., "alibaba.com", Fig. 7(b)) nor the one for 1688 (Fig. 7(c)).

This means that these approaches are very unlikely to work in a "cross-regional" setting (even if the corresponding PS includes some Chinese brands—since they are tailored for the Western version of such websites). Conversely, the only way to make these approaches "work" is by *(i)* expanding their PS by also including Chinese brands; and *(ii)* by including in the corresponding dataset the Chinese version of the websites within the PS. The drawback, however, is that this will inevitably increase the computational effort to analyse any given webpage at test-time.

[15] **Why do such ML-PWD work in this way?** Image-based PWD are trendy in research, and are now being deployed also in practice (Apruzzese et al., 2023a; Draganovic et al., 2023). However, it is almost paradoxical that their biggest strength is also their main weakness. Indeed, to meet "operational" requirements, PWD must be fast: a user is not willing to wait seconds before their browser renders a given webpage *just because there is a risk of such a webpage being phishing*. Consequently, in a very short time-frame, a given PWD that employs (target dependent) image-based techniques must: *(i)* capture the screenshot of a website; *(ii)* extract the relevant information (e.g., the logo); *(iii)* make a pairwise comparison of such information with each element in the PS—note that for each protected website there may be multiple elements associated to it (e.g., multiple logos are associated to PayPal); *(iv)* if a match is found, check the domain—note that the DNS query is done only after determining which brand is the one most likely associated to the given webpage, i.e., the PS must *always* be checked in its entirety (according to the co-authors of Apruzzese et al. (2023a)); *(v)* after receiving the response, decide whether to block the webpage or not. This long set of operations is computationally expensive. To make such an analysis feasible, the PS typically includes around 200 brands (Lin et al., 2021).

(a) chinese.alibaba.com (b) alibaba.com (c) 1688.com

Fig. 7. Logos of three versions of the same Chinese brand (in 2023).
The same brand has distinct websites, which can be accessed by Chinese or Western people—also depending on where the products are meant to be shipped.

Lesson Learned: Image-based ML-PWD only work against phishing websites that try to mimic a shortlist of well-known brands. Unfortunately, most existing methods do not include Chinese brands in such a shortlist. Gauging (or improving) the effectiveness of these ML-PWD only requires inspecting (or expanding[a]) such a shortlist—or combining them with target-independent methods (e.g., Corona et al. (2017) and Van Dooremaal et al. (2021)) which do not analyse images (proof in Appendix E).

[a] Of course, there are works that have explored such a possibility (e.g., Liu et al. (2023)), but these are outside our scope.

The above explains *why we only tested PhishIntention* (Liu et al., 2022a) on ChiPhish: like any target-dependent ML-PWD, PhishIntention *works well* against phishing websites mimicing (top-ranked) websites—but its PS only covers Western brands Liu et al. (2023). For this reason, in the remainder of this paper, *we will focus on improving target-independent ML-PWD*—since they seek to work on "any" phishing websites.

6. The step forward: Improving target-*in* dependent ML-PWD

We have empirically shown the shortcoming of state-of-the-art ML-PWD proposed in research (§5), and demonstrated the pros-and-cons of target-dependent ML-PWD reliant on visual similarity (§5.3). We now seek to find ways to improve current *target-independent* ML-PWD, since they can provide protection of "generic" phishing websites—including those that are outside a pre-defined reference list. To this end, we first dissect the anatomy of Chinese websites (§6.1). Based on the analysis, we then propose and practically implement an enhanced feature-set covering both Chinese and Western websites (§6.2). Finally, we re-develop the ML-PWD of SpacePhish by using our feature-set and datasets, and assessing their cross-regional effectiveness (§6.3).

6.1. Analysis: Chinese vs Western websites (from a PWD perspective)

As a starting point to improve feature-based ML-PWD, we elucidate the differences that set Chinese websites apart from Western ones. These differences lie in: the language itself and, in particular, its *text* (§6.1.1); and the *structure* of the website (§6.1.2). Both of these influence the representation of the website (i.e., its HTML), thereby suggesting that ML-PWD analysing features extracted from such information are likely to respond differently on websites of different regions.

6.1.1. Chinese & Western texts

English or Western languages (e.g., Italian) are phonetic languages. Their smallest sememe words (Niu et al., 2017) are a combination of 26 alphabet letters. For instance, a generic English word (e.g., "hello") can be easily pronounced with the help of its glyphs. However, Chinese texts (and other Eastern texts, e.g., Japanese) are more complex: there can be little or no correlation between the pronunciation and the glyph of a given word. As an example, the Chinese word '参' has multiple pronunciations: 'cān', 'cēn', and 'shēn'. However, even native speakers cannot determine how to pronounce '参' just by observing its glyph.

Indeed, Chinese is both a kind of hieroglyphics and phonetic language, which has three unique linguistic characteristics: *pinyin*, *glyph* and *tone* (Liu et al., 2022b). Only by knowing all of these it is possible to determine the exact Chinese character. As shown in Fig. 8, words in group (a) have the same *pinyin* and *tone*, but their *glyph* and semantics are different, which means the Chinese texts cannot be confirmed only by the pronunciation. The words in group (b), have the same glyph, '长', but they differ in the *pinyin*, *tone* and semantics. Finally, words in group (c), '离' and '里' have the same *pinyin*, 'li', but different *tones* ˊ, ˇ and different *glyphs*; furthermore, their semantics are different.

In practice, the difference between Chinese and Western texts is likely to affect ML-PWD "trained" on either of these languages. For example, many ML-PWD (e.g., Hannousse and Yahiouche (2021), Li et al. (2019), including those in SpacePhish (Apruzzese et al., 2022a)) analyse a feature that denotes whether the website's title includes the domain of its URL. Let 'H_titBr' denote such a feature: we represent this extraction procedure in Fig. 9. For a Western website, we (step 1) get the domain from the URL and (2) get the title from the HTML, then (3) check if the title includes the domain; these procedures are those followed by the extractor of SpacePhish. However, for Chinese websites, the title is in Chinese hieroglyphs: hence, *it would be misleading to use the same procedure*. Indeed, to correctly extract the 'H_titBr' feature from a Chinese website, we (4) need to 'convert' the title to its corresponding pronunciation, e.g., *pinyin*,[16] a combination of letters; and then (5) compare it with the URL's domain. Unfortunately, to the best of our knowledge, we are not aware of PWD that implement such an extraction procedure.

Importantly: this difference also exists between Chinese and Western versions of the *same* website or brand. As an example, consider Amazon, for which we provide an illustration in Figs. 10. The URL of the Western variant contains the string "amazon", which also appears in the title (as HTML) of the webpage (Fig. 10(a)). Therefore, the extraction of the 'H_titBr' feature (i.e., steps 1, 2, 3 in Fig. 9) is straightforward for "western" PWD (e.g., Apruzzese et al. (2022a)). However, this extraction procedure does not work on the Chinese version of Amazon. As shown in Fig. 10(b), the HTML's title tag is "亚马逊-网上购物商城：要网购，就来z.cn!", which clearly does not include the string "amazon" (which is present in the URL). Therefore, to *correctly* extract H_titBr, it is necessary to convert the title to its pronunciation. Not doing so (and applying the 'straightforward' Western procedure) leads to a mismatch that induces an ML-PWD to believe that the Chinese version of Amazon to be a suspicious website (since it would include *pinyin* and the foreign language of transliterated words).

6.1.2. Chinese & Western websites structure

According to China's network security law and the Administration for Industry and Commerce regulations (Cyberspace Administration of China, 2022c), Chinese websites must be registered with the Ministry of Industry and Information Technology of the Chinese government (i.e., ICP records). Chinese websites engaged in different activities must apply for qualification certificates from the corresponding government departments. E.g., "JD.com" is an eCommerce that mainly sells electronic merchandise and (to a lower extent) medicines, thus it received a telecommunication business licence from the Chinese Ministry of

[16] Among the top30 popular Chinese websites, 15 use pinyin (2023).

Fig. 8. Combination of Chinese texts. Chinese words can be identified only by knowing *pinyin*, *glyph* and *tone*.

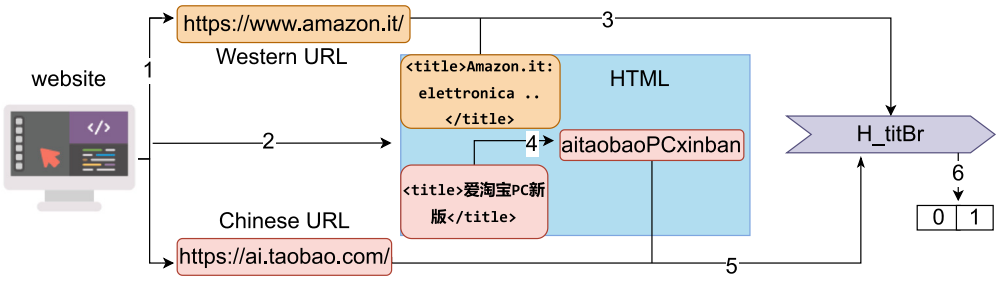


Fig. 9. Extracting 'H_titBr' from Chinese and Western websites.
Current feature-based ML-PWD must be updated to correctly rely on this feature, since it would be incorrectly computed for some Chinese websites.

```
<title>Amazon.com. Spend less. Smile more.</title>
```

(a) Title tag (HTML) of https://amazon.com/.

```
<title>亚马逊-网上购物商城：要网购，就来z.cn!</title>
```

(b) Title tag (HTML) of https://amazon.cn/.

Fig. 10. URL and HTML (title) of the Chinese and Western version of the same website.

Industry and Information Technology, and a qualification certificate for pharmaceutical services approved by Beijing Municipal Medical Products Administration (e.g., see red box in Fig. 2(a)). In addition, Chinese websites may display trusted website certifications to increase their credibility (yellow box in Fig. 2). However, Western websites do not have (nor require) these certificates. Even world-renown websites lack them (see Fig. 11).

In practice, analysing the ICP record *can* be used to identify malicious websites in China (this is done, e.g., in Zhang et al. (2014)). However, the absence of this form of "certification" for Western websites creates an *intrinsic incompatibility* between Chinese and Western ML-PWD that analyse the HTML of a webpage. For instance, ML-PWD for Chinese websites will search for the ICP record on Western websites, but will never be able to find it—thereby inducing the ML-PWD to believe that any Western webpage is "suspicious". In contrast, ML-PWD for Western websites will also be adversely impacted by the presence of the ICP record: it is well-known (Mohammad et al., 2012; Jain and Gupta, 2018b; Yang et al., 2021) that phishing webpages have many objects that point to "external" items—and, of course, the ICP embedded in a Chinese website points to an external resource. As a result, ML-PWD for Western websites will also be more likely to be "suspicious" of any Chinese website.

6.2. Implementation: towards developing ML-PWD that account for Chinese websites

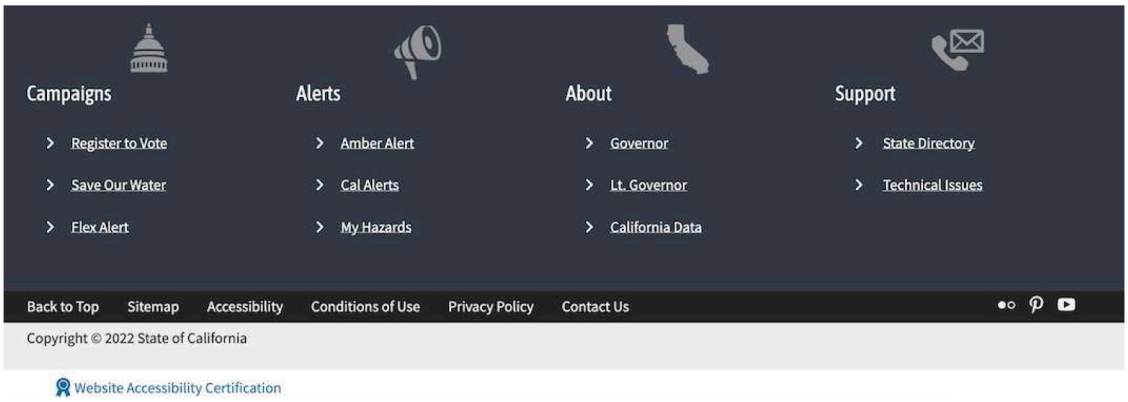
As our first technical contribution, we now propose a new feature set which allows feature-based ML-PWD to simultaneously account for both Chinese and Western websites (§6.2.3). To this end, we *enhance* the feature extractor of SpacePhish (Apruzzese et al., 2022a), which extracts a total of 57 features from the URL and HTML of any given website. However, as we argued (§6.1) such features may not capture the nuances of Chinese websites, and may be incorrectly computed. We hence design 10 new Chinese-specific features (§6.2.1), and change 2 of the existing ones (§6.2.2). Overall, our features follow the same logic as Apruzzese et al. (2022a), i.e., the value of each feature denotes whether the corresponding sample is more likely benign (0) or phishing (1). Let us explain our new features at a high-level. For simplicity, we use 'H_' and 'U_' to denote a feature that is based on the HTML and URL, respectively.

6.2.1. New features

Among the 10 new features, five (i.e., *H_icpReg*, *H_icpDom*, *H_icpApp*, *H_icpCode* and *H_ecert*) are Chinese-specific, and are inspired by the guidelines in Zhang et al. (2014); while the other five (i.e., *U_unicode*, *H_nullItem*, *H_exItem*, *U_SER*, *U_tldNum*) are based on best practices of prior work (Hannousse and Yahiouche, 2021; Rao and Pais, 2019; Huh and Kim, 2011), and should work on "any" website (both Chinese and Western).

- *H_icpReg*. If the domain is in the Ministry of Industry and Information Technology of the Chinese government, then *H_icpReg*=0; and 1 otherwise.
- *H_icpCode*. If the website includes an ICP code and it exists in its ICP recorder (obtainable by checking the domain), then *H_icpCode*=0; and 1 otherwise.
- *H_ecert*. We capture all links in the website. If none of such links point to Trustworthy website certification platforms, then *H_ecert*=1; and 0 otherwise.
- *H_icpApp*. If the domain applicant on the ICP record of the Chinese Ministry of Industry and Information Technology is "enterprise", then we set *H_icpApp*=0; and to 1 otherwise.
- *H_icpDom*. If the domain is consistent with the ICP record, then *H_icpDom*=0, and 1 otherwise.
- *U_unicode*. According to Zhang et al. (2014), phishing website is more likely to use UNICODE in its URL. If true, then *U_unicode*=1, and 0 otherwise.
- *H_nullItem*. This feature extends the "HTML_nullLnkWeb" of Apruzzese et al. (2022a). Specifically, we leverage the guidelines by Hannousse and Yahiouche (2021) and factor in also other "null" elements, such as Login forms with external actions, that are typical indicators of suspiciousness (besides just blank links of "HTML_nullLnkWeb").

(a) Footer of a Western eCommerce website (https://www.amazon.com/).



(b) Footer of the Califonia government website. (https://www.ca.gov/).

Fig. 11. Exemplary eCommerce and Government websites "from the West".
The footers of these websites are substantially different than those of Chinese websites (cf. Fig. 2). For instance, they lack ICP records.

- *H_exItem*. We compute this feature by counting the elements in the HTML that point to external websites. The value is an integer, which will be used to compute the feature *H_obj* (which is the equivalent of the "HTML_objectRatio" of Apruzzese et al. (2022a)).
- *U_SER*. We follow the guidelines of Rao and Pais (2019), Huh and Kim (2011), suggesting that it is possible to use search engine results to detect phishing websites. If the URL matches any of the top-10 websites in a Google search results (by querying Google with the sample's URL), *U_SER*=0; and a 1 otherwise.
- *U_tldNum*. We follow the guidelines of Hannousse and Yahiouche (2021), suggesting that phishing websites may have more than one top-level domain (TLD) located in another position within the URL (e.g., the subdomain). This numerical feature represents the number of TLD in the URL of the sample (this information is not captured by the features of SpacePhish).

6.2.2. Changed features

To account for the nature of Chinese websites, the extraction procedures of two features have been changed w.r.t. those in Apruzzese et al. (2022a).

- *H_titBr*. This feature checks if the website's title includes the domain of its URL. We follow the workflow described in §6.1.1. We extract the title from the corresponding HTML tag, and then we check if it includes any Chinese words via regex and, if so, we convert the Chinese words to their pronunciation (i.e., *pinyin*). Finally, if the domain of the URL is included in the title or in the *pinyin*, *H_titBr*=0 (likely benign); and 1 otherwise (likely phishing).
- *H_DominCopr*. We search for Chinese words in the website's copyright information, convert them to their pronunciation, and finally, we check if the website's copyright information includes the website's domain: *H_DominCopr*=0 if so, and 1 otherwise.

6.2.3. Enhanced feature set

We implement our feature extractor and publicly release it in our repository (Yuan, Ying and Apruzzese, Giovanni and Conti, Mauro , 2023g), so that downstream research can use it for future analysis. Afterwards, we pre-process every website in our three datasets (`ChiPhish`, `WstPhish`, `EngPhish`) with our extractor, generating its feature representation (a vector of 67 features). We find that two features from SpacePhish (i.e., *URL_fakeHTTPS* and *URL_dataURI*) are redundant since the value is the same for all samples in our datasets, so we do not consider these for our evaluation. We summarize the 65 features considered in our evaluation in Table 6. We further discuss some fairness and robustness properties of our features in Appendix A.

6.3. Assessment (is there any improvement w.r.t. the baseline ML-pwds?)

We now practically develop 81 "new" (target-independent) feature-based ML-PWD by using our enhanced feature extractor. We will then test these 81 ML-PWD in a cross-regional setting through our three datasets. Our assessment, besides being useful to verify if there is any improvement (w.r.t. §5.2.1), serves to answer the following ancillary research question:

AQ3: Do ML-PWD specifically developed for, and trained *only* on Chinese websites work well on Western/English websites (and vice-versa)?

AQ3 is different from RQ2 (i.e., there is no publicly available ML-PWD for Chinese websites!). We first describe our experimental workflow (§6.3.1), and then present the results (§6.3.2).

6.3.1. Workflow

For a broad assessment, we consider 81 ML-PWD, given by: 3 (datasets) × 9 (ML algorithms) × 3 (feature sets). Let us explain and motivate our choices.

Table 6
The features considered in our evaluation. Features in boldface are *new*; grey cells denote (new) Chinese-specific features; features in italics have been changed (w.r.t. SpacePhish (Apruzzese et al., 2022a)). Features whose name starts with *U_* are extracted from the URL, and those starting with *H_* are extracted from the HTML. The names of the remaining features reflect those in SpacePhish (extensively described in its publicly available artifact (Apruzzese et al., 2023j)).

#	Feature Name	#	Feature Name	#	Feature Name
1	U_dash	23	U_ip	45	H_rClick
2	U_tldinSub	24	U_at	46	H_brokenLin
3	U_pageRank	25	U_pt	47	H_loginForm
4	U_ssl	**26**	U_unicode	48	H_hidDiv
5	U_abn	27	U_age	49	H_statBarMod
6	U_numerical	28	U_rdr	50	H_css
7	U_tldinPath	29	U_dns	51	H_anchors
8	U_shortestWrdPath	30	U_tldNum	52	H_commRatioFt
9	U_lngHost	31	U_punycode	53	*H_DominCopr*
10	U_regLen	32	U_lngWrdPath	54	H_hidInp
11	U_senwrd	33	U_avgHost	55	H_iframe
12	U_totwrdUrl	34	U_avgWrdPath	56	H_favicon
13	U_shortestWrdUrl	35	U_SER	57	H_exItem
14	U_shortestWrdHost	36	U_GI	58	H_icpCode
15	U_lngWrdUrl	37	H_SFH	59	H_ecert
16	U_avgWrdUrl	38	H_popUp	60	H_freqDom
17	U_statsRep	39	H_nullItem	61	H_obj
18	U_len	40	H_metaScrpLin	62	H_commPage
19	U_shorter	41	H_icpReg	63	H_nulLin
20	U_sub	42	H_icpDom	64	H_nulLinFt
21	U_commItemNum	43	H_icpApp	65	H_hidBtn
22	U_pathExtend	44	*H_titBr*		

- **Feature sets.** We consider 3 feature sets: F_u, corresponding to the 36 URL-based features in Table 6 (starting with 'U_'); F_h, corresponding to the 29 HTML-based features in Table 6 (starting with 'H_'); and F_c, corresponding to all features in Table 6. We consider these three perspectives for both a "research and practical" reason. First, because it allows one to conduct an *ablation study* (we will do this in Appendix D). Second, because some ML-PWD may not analyse the URL, whereas others may not analyse the HTML (this can be done to make the analysis faster, or to enable PWD when some information is missing, or even to create "adversarially robust" ML-PWD, according to Apruzzese et al. (2022a)).
- **Learning algorithm.** We expand the space of ML algorithms considered in SpacePhish (which only included RF, LR, CNN—cf. §5), and consider 9 ML algorithms that support binary classification—all of which have been used in previous ML-PWDs (HR et al., 2020; Tian et al., 2018; Sahingoz et al., 2019; Sharma et al., 2020; Janet et al., 2020; Apruzzese et al., 2022a; Corona et al., 2017; Apruzzese and Subrahmanian, 2022). Specifically, 7 are "shallow" ML algorithms: Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB), AdaBoost (AB), Support Vector Machines (SVM) and K-Nearest Neighbours (KNN); while 2 are deep learning algorithms: Multi-Layer Perception (MLP) and Convolutional Neural Network (CNN). Such a wide array allows one to better appreciate the strengths and weaknesses of these various classification algorithms.
- **Datasets and Setup.** The evaluation of our ML-PWD entails our three 'regional' datasets (§3): `ChiPhish`, `EngPhish`, `WstPhish`. After having generated the feature representation of every website in each dataset, we consider three variants—each corresponding to a specific feature-set (i.e., F_u, F_h, F_c). This procedure yields 9 different sets of samples (i.e., 3 feature sets × 3 language datasets). Each of these sets is then divided into train:test partitions with an 80:20 split (common in ML-PWD (Bac et al., 2021; Al-Qurashi et al., 2021; Apruzzese et al., 2022a)). We then use the training partition of each of these 9 sets of samples to train 9 ML-PWD (each using one among our 9 considered ML algorithms), thereby yielding 81 ML-PWD. The test partition will be used for our assessment (discussed in the remainder of this section). To account for randomness in the split and reduce the chances of biased results, we repeat all our experiments 10 times—thereby allowing one to derive statistically significant conclusions.[17]

These experiments are done on Ubuntu 20.04 system with CPU Intel Xeon W-2223 @ 3.60 GHz; we report the training and testing runtime in Table B.14. We release our (documented) source code for reproducibility. A schematic representation of this workflow is in Fig. 12.

6.3.2. Results

After having trained our 81 ML-PWD, we test their performance in a cross-regional setting. In other words: we test each of the 27 ML-PWD trained on (80% of) `ChiPhish` on (20% of) `ChiPhish`, `EngPhish` and `WstPhish`; then we do the same for the 27 ML-PWD trained on (80% of) `WstPhish` and `EngPhish`. We report the complete results of this evaluation in B.2, showing the tpr and tnr of all our 81 ML-PWD. In what follows, we will provide a high-level analysis and then focus on the performance of the best ML-PWD.

- **High-level.** We provide in Figs. 13 a comprehensive overview of our results. These figures report the F1-score (aggregated across the 10 trials and 9 learning algorithms) achieved by our ML-PWD for different test set. For example, Fig. 13(c) shows the distribution of the $F1$-score for the ML-PWD (which varies either for their training dataset or feature set) when tested on `ChiPhish` (assuming a matching feature set). From Figs. 13, we can see that our ML-PWD, when tested on websites having the same language as their training data, work well—and this is especially the case for ML-PWD using F_c, which always outperform those analysing fewer features. We also appreciate that the ML-PWD trained on either `EngPhish` or `WstPhish` exhibit similar performance when tested on (respectively) `WstPhish` or `EngPhish`. However, the situation changes when our ML-PWD on phonological (resp. hieroglyphical) languages must analyse hieroglyphical (resp. phonological) languages. This is evident when testing on samples from `ChiPhish` (Fig. 13(c)): despite the good performance of the "Chinese" ML-PWD (the $F1$ is almost always>0.85), the "Western" and "English" detectors have a significant drop (almost never above 0.70 $F1$), which is inappropriate to analyse Chinese websites. Similarly, in Fig. 13(a), we can see that the "Chinese" ML-PWD work poorly on English websites; interestingly, however, they still retain around 0.75 $F1$ on the generic Western websites in `WstPhish`. Unfortunately, an in-depth look at our results reveals that such "encouraging" (aggregated) F1-scores are concealing an impractical tnr.
- **Best ML-PWD.** We now focus on the best ML-PWD of our evaluation, i.e., those analysing F_c. We report the detailed tnr and tpr for each of our 9 ML algorithms in Table 7. First, we can see that, as we anticipated, in a cross-regional (`ChiPhish` to `EngPhish`/`WstPhish`, and vice-versa) setting, the average tnr is unacceptably low (i.e., 0.15, 0.47, 0.55); the only decent one is an average 0.87 tnr for the ML-PWD trained on `WstPhish` and tested on `ChiPhish`, but the tpr=0.49. However, let us take a deeper look at the *very-best ML-PWD*, which is the one using the RF algorithm (a result which aligns with prior work, e.g., Apruzzese et al. (2022a) and Tian et al. (2018)). This is because only the best ML-PWD would be (hypothetically) deployed in reality, and hence its results are more appropriate to derive sensible conclusions. We make three observations.

[17] Note that, to follow best-practices (Arp et al., 2022; Apruzzese et al., 2022b) and ensure consistency, we use the *same* test-set to each ML-PWD for each trial. E.g., we use the same 20% of `ChiPhish` to test all the ML-PWD, and then start a new trial by randomly sampling a new training and test partitions—which we use to develop and assess new ML-PWD.

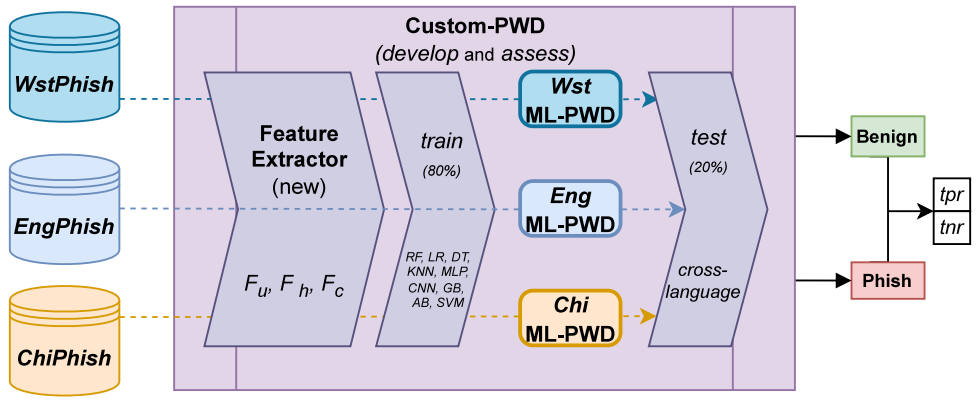
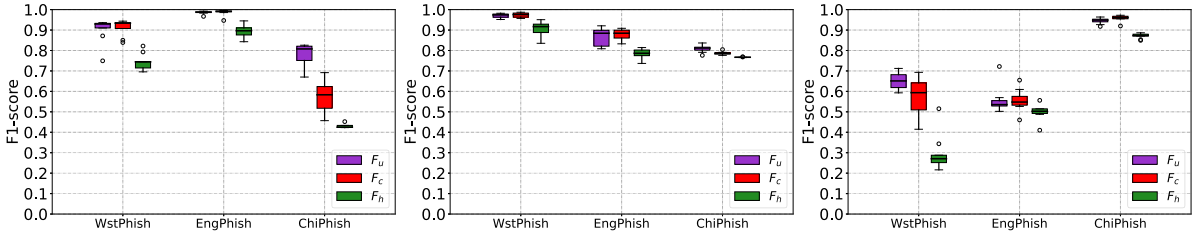


Fig. 12. Overview of our workflow.
We collect three datasets (one containing only Chinese websites, a second one only English websites, and a third one containing a mix of websites in popular Western languages) which we use to evaluate existing phishing website detectors, measuring their *tpr* and *tnr*.



(a) Test partition: `EngPhish`. (b) Test partition: `WstPhish`. (c) Test partition: `ChiPhish`.

Fig. 13. Cross-language performance of our ML-PWD.
We show the distribution of the $F1$-score (y-axis) of our ML-PWD, trained on a specific dataset (x-axis) and analysing a given feature set (legend), on the test partition of each dataset (subfigure). Each boxplot aggregates the results of all our learning algorithms across the 10 trials.

1. In a "same-region" context, the performance is always very high—closing to 1 *tnr* and *tpr*. This result is encouraging from the perspective of Chinese-specific ML-PWD.
2. When cross-tested across `EngPhish` and `WstPhish`, the performance is still good: there is a moderate drop to 0.7 *tnr* for the ML-PWD trained on `EngPhish` and tested on `WstPhish`, and a slight drop to 0.9 *tpr* for the ML-PWD trained on `WstPhish` and tested on `EngPhish`. This result is encouraging for Western ML-PWD, since they show that there is high inter-compatibility among different Western languages.
3. When cross-tested across `ChiPhish` and `EngPhish`/`WstPhish`, the performance is poor. Though the ML-PWD trained on `ChiPhish` retains ≥0.97 *tpr* on `WstPhish` and `EngPhish` (meaning that it can detect *phishing* websites "from the West"), its *tnr* drops to an unusable 0.13 on `WstPhish` and 0.49 on *EngPhish* (meaning that the majority of *benign* websites "from the West" are classified as malicious). The situation is not better when considering ML-PWD trained on `WstPhish` and tested on `ChiPhish` (*tpr*=0.37, *tnr*=0.97, meaning that most Chinese phishing websites are misclassified) and the one trained on `EngPhish` and tested on `ChiPhish` (*tpr*=0.71, *tnr*=0.55, meaning too many false alarms).

In summary: there is an improvement over the results of the vanilla ML-PWD in SpacePhish (cf. Table 4), which are trained on an *inappropriate* feature-set, and on a different dataset (see §5.2); these results are also superior to those achieved by the solution by Jiang and Wu (2022) (which achieved at most 79% F1-score). We carry out an ablation study (discussed in Appendix D) wherein we statistically prove that our enhanced feature set yields superior performance w.r.t. the original one in SpacePhish (assuming the same training/test datasets). However, the performance in a cross-regional setting is still not ideal to bridge the gap between Western and Chinese PWD.

Lesson Learned. Combining our new features (deriving from our analysis of Chinese websites) with our datasets led to *(i)* an improvement of the baselines from prior work, and to *(ii)* an efficient Chinese-specific ML-PWD (0.96 *tpr* and 0.99 *tnr*). However, the cross-regional performance is still underwhelming.

7. Bridging the gap between Chinese and western ML-PWD (RQ3)

We now have all the elements necessary to tackle RQ3. In this section, we will attempt to bridge the gap by piecing together all our previous contributions. We first examine the results of our ML-PWD by focusing on the *feature importance* (§7.1). Then, by using our examination as a scaffold, we pragmatically assess the most straightforward solution to our problem: *combining our three datasets into a single dataset* used to develop an "universal" ML-PWD (§7.2). Finally, we provide a more practical solution entailing the application of our *LaSeTo to devise an ensemble* of ML-PWD (§7.3).

7.1. Examination of the feature importances (explainability analysis of our results)

Intuition. The reason why Chinese (resp. Western) ML-PWD work poorly on Western (resp. Chinese) websites can be traced back to the semantic difference between Chinese and Western websites (cf. §6.1). Such a difference leads to samples having a different feature distribution, thereby preventing a correct analysis by any ML-PWD that is trained and tested on websites from different "regions." To identify potential mitigations to the problem elucidated by our paper, we examine our results by focusing on the features analysed by our (feature-based) ML-PWD.

Method and Analysis. Investigating the most relevant features for classification is a well-known technique for studying the underlying

Table 7
Performance of our "best" ML-PWD. We report the *tpr* and *tnr* (avg and std across 10 trials) of our ML-PWD analysing F_c features (URL+HTML). Overall, RF is the best *Alg* (grey cells). Results for F_u and F_h are in Tables B.12 and B.13.

Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%		Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*			*tpr*	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	WstPhish	0.99 ± 0.003	0.97 ± 0.006	0.90 ± 0.011	1.0 ± 0.001	0.37 ± 0.038	0.97 ± 0.006	*CNN*	WstPhish	0.98 ± 0.003	0.97 ± 0.008	0.88 ± 0.014	0.99 ± 0.003	0.45 ± 0.05	0.84 ± 0.064
	EngPhish	0.99 ± 0.002	0.7 ± 0.014	0.98 ± 0.001	1.0 ± 0.001	0.71 ± 0.02	0.55 ± 0.025		EngPhish	0.99 ± 0.005	0.68 ± 0.053	0.98 ± 0.005	1.0 ± 0.002	0.61 ± 0.174	0.49 ± 0.05
	ChiPhish	0.99 ± 0.005	0.13 ± 0.036	0.97 ± 0.016	0.49 ± 0.152	0.96 ± 0.015	0.99 ± 0.005		ChiPhish	0.97 ± 0.011	0.27 ± 0.054	0.9 ± 0.022	0.65 ± 0.113	0.95 ± 0.021	0.99 ± 0.005
LR	WstPhish	0.96 ± 0.006	0.95 ± 0.007	0.88 ± 0.013	1.0 ± 0.001	0.52 ± 0.056	0.95 ± 0.017	*GB*	WstPhish	0.99 ± 0.003	0.98 ± 0.006	0.9 ± 0.014	0.99 ± 0.004	0.59 ± 0.034	0.94 ± 0.017
	EngPhish	0.99 ± 0.002	0.65 ± 0.021	0.98 ± 0.004	1.0 ± 0.001	0.73 ± 0.04	0.52 ± 0.019		EngPhish	0.99 ± 0.002	0.37 ± 0.032	0.99 ± 0.003	1.0 ± 0.001	0.85 ± 0.03	0.52 ± 0.029
	ChiPhish	0.99 ± 0.003	0.13 ± 0.014	0.93 ± 0.007	0.39 ± 0.049	0.95 ± 0.014	0.99 ± 0.005		ChiPhish	0.99 ± 0.006	0.15 ± 0.064	0.96 ± 0.025	0.57 ± 0.155	0.96 ± 0.014	0.98 ± 0.009
DT	WstPhish	0.98 ± 0.006	0.96 ± 0.006	0.86 ± 0.016	0.9 ± 0.222	0.56 ± 0.077	0.82 ± 0.062	*AB*	WstPhish	0.98 ± 0.004	0.96 ± 0.007	0.89 ± 0.011	0.99 ± 0.002	0.62 ± 0.058	0.84 ± 0.03
	EngPhish	0.98 ± 0.007	0.41 ± 0.075	0.99 ± 0.003	0.99 ± 0.002	0.76 ± 0.067	0.42 ± 0.046		EngPhish	0.99 ± 0.003	0.51 ± 0.024	0.99 ± 0.002	1.0 ± 0.001	0.72 ± 0.059	0.51 ± 0.049
	ChiPhish	0.98 ± 0.01	0.16 ± 0.076	0.95 ± 0.035	0.69 ± 0.165	0.96 ± 0.015	0.97 ± 0.015		ChiPhish	0.99 ± 0.007	0.13 ± 0.053	0.93 ± 0.024	0.41 ± 0.127	0.96 ± 0.018	0.99 ± 0.005
KNN	WstPhish	0.95 ± 0.005	0.94 ± 0.008	0.79 ± 0.017	0.97 ± 0.005	0.31 ± 0.045	0.98 ± 0.011	*SVM*	WstPhish	0.96 ± 0.006	0.95 ± 0.008	0.88 ± 0.011	1.0 ± 0.0	0.55 ± 0.071	0.92 ± 0.022
	EngPhish	0.84 ± 0.009	0.9 ± 0.008	0.92 ± 0.011	0.99 ± 0.002	0.55 ± 0.055	0.94 ± 0.018		EngPhish	0.99 ± 0.003	0.59 ± 0.031	0.99 ± 0.004	1.0 ± 0.001	0.81 ± 0.04	0.49 ± 0.018
	ChiPhish	0.99 ± 0.004	0.09 ± 0.011	0.97 ± 0.008	0.2 ± 0.029	0.9 ± 0.022	0.97 ± 0.015		ChiPhish	0.97 ± 0.014	0.21 ± 0.044	0.89 ± 0.02	0.6 ± 0.126	0.94 ± 0.019	0.98 ± 0.007
MLP	WstPhish	0.98 ± 0.005	0.95 ± 0.01	0.87 ± 0.013	0.98 ± 0.005	0.45 ± 0.057	0.6 ± 0.182		WstPhish	0.97 ± 0.014	0.96 ± 0.011	0.87 ± 0.033	0.98 ± 0.028	0.49 ± 0.099	0.87 ± 0.113
	EngPhish	0.99 ± 0.004	0.67 ± 0.037	0.99 ± 0.003	1.0 ± 0.001	0.78 ± 0.069	0.51 ± 0.023	Avg (std)	EngPhish	0.97 ± 0.045	0.61 ± 0.152	0.98 ± 0.022	1.0 ± 0.002	0.72 ± 0.09	0.55 ± 0.142
	ChiPhish	1.0 ± 0.003	0.08 ± 0.017	0.94 ± 0.009	0.19 ± 0.066	0.96 ± 0.016	0.99 ± 0.009		ChiPhish	0.98 ± 0.007	0.15 ± 0.055	0.94 ± 0.025	0.47 ± 0.172	0.95 ± 0.019	0.98 ± 0.007

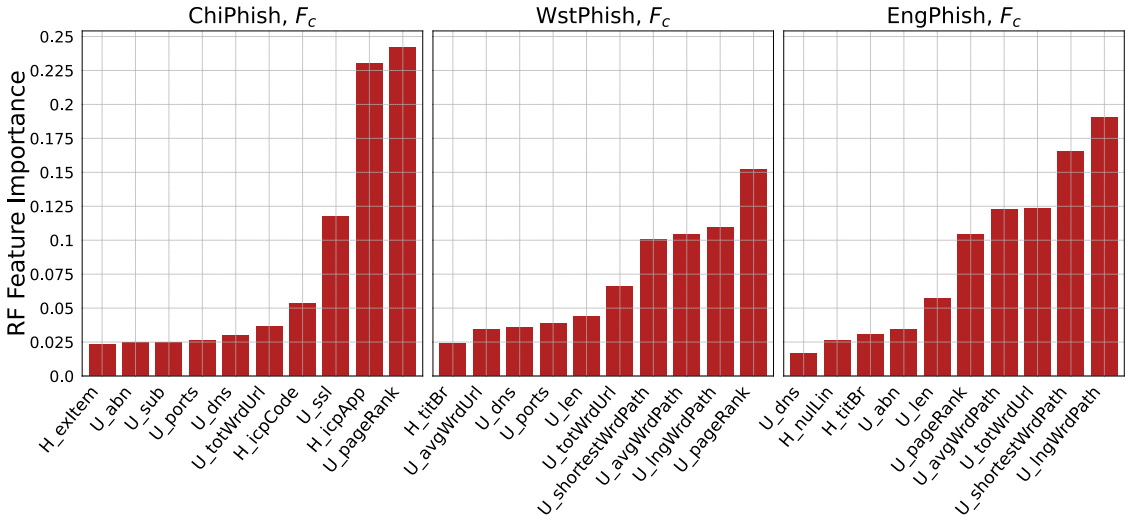


Fig. 14. Top-10 features of RF (F_c) trained on each dataset. We report these plots for F_u and F_h in Figs. C.16, C.17.

logic learned by an ML model (Apruzzese et al., 2022c). To align such an analysis to our previous discussion (§6.3.2), we report in Fig. 14 the ranking (as given by scikit-learn) of the top10 features for the best ML-PWD, i.e., RF analysing F_c. (We provide an analysis of the feature ranking for the RF analysing F_u and F_h in Appendix C.) From Fig. 14, we see that two Chinese-specific features, 'H_icpApp' and 'H_icpCode', appear in the top10 of ChiPhish, and that 'H_icpApp' is the second most important feature; in contrast, neither of these features are relevant for the classifiers trained on WstPhish and EngPhish. This situation can explain why classifiers trained in ChiPhish work poorly on Western websites (i.e., WstPhish and EngPhish). At the same time, by observing the rankings for the classifiers trained on WstPhish and EngPhish, we see that both have 8 features in the top-10: this suggests why classifiers trained on WstPhish and EngPhish perform similarly. Moreover, both 'H_icpApp' and 'H_icpCode' are the features extracted from the HTML, which can verify our intuition that the gap between Western and Chinese PWD is more manifested in HTML.

Considerations. A possible way to reduce the performance gap between our Chinese ML-PWD and the ML-PWD focusing on phonological languages is by considering a feature set that is "less specific" to Chinese websites. This could be done by developing ML-PWD that analyse *only the URL*, i.e., F_u. In fact, by looking at the boxplots in Figs. 13, we can see that the ML-PWD analysing F_u (purple bins) tend to have a higher $F1$ than those using F_h and F_c trained on Chinese (resp. non-Chinese) and tested on non-Chinese (resp. Chinese) websites. For instance, the ML-PWD trained on ChiPhish using F_u obtain nearly 0.8 $F1$ (up from 0.6 of F_c) when analysing EngPhish (see Fig. 13(a)); whereas the ML-PWD trained on WstPhish obtain 0.65 $F1$ (up from 0.60 of F_c) when tested on ChiPhish (see Fig. 13(c)). However, in both cases, such a gain comes at the expense of a reduced $F1$ when analysing websites of the same language: classifiers trained on F_c are *always* statistically superior to those using F_u on the respective language dataset (t-test reveals this hypothesis to be true: $p<0.05$). A much better alternative is developing a ML-PWD that "learns" the characteristics of Chinese and Western websites during its training phase: this requires the availability of a dataset that contains websites from different regions—which we have.

Takeaway. Using ML-PWD analyse only the URL can work, but presents tradeoffs (w.r.t. ML-PWD analysing URL and HTML) on same-language websites. The ideal solution entails a ML-PWD that uses both URL and HTML, and also learns the patterns of Chinese and Western websites during its training.

7.2. Towards an Universal ML-PWD: combining ChiPhish, WstPhish, and EngPhish

Intuition. As suggested by our previous analysis, the most straightforward way to "bridge the gap" between Western and Eastern PWD is to create an universal dataset containing samples in various languages.

Table 8
Universal ML-PWD. We train and test an RF (F_c) on all our datasets (80:20 split), and we measure their performance (avg and std.dev) on each dataset. Overall (on a generic webpage): tpr=0.95±0.043, tnr=0.98±0.022.

	ChiPhish	WstPhish	EngPhish
tpr	0.89 ± 0.023	0.99 ± 0.002	0.97 ± 0.004
tnr	0.99 ± 0.004	0.95 ± 0.006	0.99 ± 0.001
$F1$	0.94 ± 0.001	0.98 ± 0.002	0.98 ± 0.004

Table 9
Ensemble ML-PWD integrating LaSeTo. We report the performance (avg and std.dev., computed over 10 trials) of our most original solution. Overall (on a generic webpage): tpr=0.98±0.0029, tnr=0.99±0.0022.

	ChiPhish	WstPhish	EngPhish
tpr	0.85 ± 0.025	0.99 ± 0.003	0.99 ± 0.003
tnr	0.95 ± 0.012	0.98 ± 0.006	1.00 ± 0.000
$F1$	0.88 ± 0.017	0.99 ± 0.002	0.99 ± 0.001

Indeed, in our previous assessments (§6.3.2) we have considered ML-PWD trained on websites of specific language group. Hence, we now merge all three datasets and scrutinize the effectiveness of an ML-PWD trained (and tested) on the resulting "universal" dataset.

Method and Results. We merge `ChiPhish`, `WstPhish`, `EngPhish`, and then split the resulting dataset into train: test partitions (using the same 80:20 ratio as we did in §6.3.1). We then extract the F_c feature of each sample (which is the one yielding the best results—§6.3.2), and use the corresponding training partition to develop a ML-PWD using RF as classification algorithm (because it outperformed other algorithms, see Table 7). We use the test partition to measure its performance (tpr, tnr, $F1$) and repeat it 10 times to reduce bias. We report the results in Table 8, where rows denote a given metric, and columns denote a specific subset of the test partition.

Considerations. From Table 8, we can see that our universal ML-PWD works well: on a generic webpage, tpr=0.95 and tnr=0.98. In more detail, the $F1$ is always above 0.94, and the worst fpr (i.e., 1-tnr) is of only 0.05 for websites in `WstPhish`. However, by comparing Table 8 with the detailed results in Table 7 (for RF), we see that the there is a significant (verified with a t-test) degradation in two cases: the tpr on `ChiPhish` (which drops from 0.96 to 0.89), and the fpr on `WstPhish` (which increases from 0.025 to 0.05). Nonetheless, such a degradation is expected in the machine learning context: an improved generalizability often comes at the expense of a reduced specificity. For instance, the tpr=0.89 on Chinese websites of our universal ML-PWD requires training on `ChiPhish`, `WstPhish`, `EngPhish`, and also allows to achieve ≥0.95 tnr and ≥0.97 tpr on Western and English websites. In contrast, the Chinese-specific ML-PWD (trained only on `ChiPhish`) had a 0.96 tpr on Chinese websites—but an impractical 0.87 (or 0.51) fpr on English (or Western) websites!

Takeaway. Mixing datasets of Chinese and Western languages improves the cross-regional tpr, but can double the false positive rate on "Western" websites (and requires training datasets having samples from different regions). Such is the price to pay for deploying our universal ML-PWD.

7.3. Exploiting our laseto: a novel "ensemble-based" system for cross-regional ML-PWD

Intuition. As a final solution to bridge the gap, we propose an intuitive solution rooted in our self-developed LaSeTo (§3.2.2); such a solution is orthogonal to the "universal" ML-PWD (discussed in §7.2). Here, we are inspired by the remarkable results achieved by our "language-specific" ML-PWD (see §6.3.2): indeed, our ML-PWD work well *if they analyse websites in a language they "have seen."* We use this observation as a scaffold and develop an *original phishing website detection system* which integrates an "ensemble" of our custom ML-PWD which are put in a pipeline to our self-developed LaSeTo. This system is shown in Fig. 15. We are not aware of existing anti-phishing schemes that entail a "language selector" before the detection model.

Method and Results. We partition each of our datasets in train:test with the usual 80:20 split. We train three "language-specific" ML-PWD (one per dataset) on 80% of each language dataset (we use F_c and RF). Then, we merge the remaining 20% of each dataset into a single Test Dataset. Next, we use LaSeTo to analyse the language of each sample in this Test Dataset: if the language is Chinese, the sample is analysed by the ML-PWD trained on `ChiPhish`; if the language is English, it is analysed by the ML-PWD trained on `EngPhish`; otherwise, it is analysed by the ML-PWD trained on `WstPhish`. We repeat this process 10 times, and report the results in Table 9.

Analysis and Feasibility. At a high-level, our "ensemble" solution represents a superior alternative to the "universal" ML-PWD (§7.2): this is confirmed by carrying out a t-test on the overall performance of these two contenders: the tnr and tpr of the "ensemble" are statistically significantly better than those of the "universal" ML-PWD. Indeed, by using LaSeTo it is possible to develop a system that works much better on "Western" websites, while still achieving a satisfactory performance on Chinese websites—albeit slightly inferior (e.g., 0.88 $F1$ vs 0.94). We also stress that these results are slightly inferior than the "baseline" ones (cf. Table 7) because LaSeTo presents a small margin of error (as we measured in §3.2.2), which may, e.g., lead LaSeTo to forward a Chinese website to the incorrect model. However, by improving LaSeTo, it would be possible to approximate the near-perfect performance of the language-specific ML-PWD on their respective datasets. Moreover, another advantage of such a solution is *flexibility*: while the "universal" ML-PWD must be trained on a single and large dataset, the "ensemble" requires training on multiple but small datasets—thereby enabling one to quickly "add" new models to the ensemble.[18] Finally we note that operational PWD must be fast at processing a webpage (Divakaran and Oest, 2022; Lee et al., 2023). Hence, to demonstrate the feasibility of our solution, we have measured the *runtime* for using LaSeTo: on average, it requires 0.04s to output the language of a given webpage (measured this on commodity hardware and after recording the time required to process all samples in our datasets). Such a low overhead makes our tool appropriate for real-time analyses (we also measured the runtime to train and test all our ML-PWD in Table B.14).

Answer to **RQ3.** Using a custom tool (e.g., LaSeTo) to infer the language of a webpage, and then use the output of such a tool to select the appropriate ML-PWD (organized in an ensemble fashion) allows the development of practical phishing website detection systems that work well in cross-regional contexts—despite still requiring websites from diverse regions (for training).

8. Discussion

Insofar, our paper revealed that the gap between Chinese and Western PWD *exists*, and it is significant since it affects both PWD proposed in *research* (§5) and those in *industry* (§4); and we have also devised, implemented, and pragmatically assessed some ways to bridge this gap (§7). Here, we reflectively analyse our major findings (§8.1), discussing limitations and room for improvement in future work (§8.2). Then, we provide additional evidence suggesting that the gap we brought to light is wider than it seems (§8.3), also serving as inspiration for future work.

[18] E.g., it is possible to quickly add a new Japanese-specific ML-PWD to the ensemble, whereas for the "universal" ML-PWD it is necessary to add the Japanese websites to the universal dataset and train the entire ML-PWD anew—potentially decreasing the performance on websites of other regions (as we discussed in §7.2, generalizability comes at the expense of specificity).

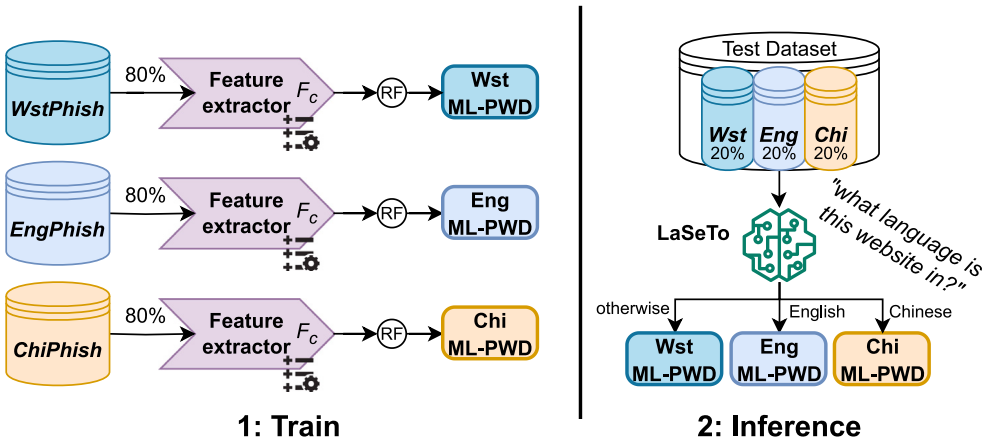


Fig. 15. Proposed "ensemble-based" phishing website detection system.
We train language-specific ML-PWD (left), and then use our self-developed LaSeTo to determine the language of any given webpage, which is forwarded to the most suitable ML-PWD (right). We use RF and F_c given that they yield ML-PWD with best results.

8.1. Findings

Our research casts light on a real-world problem (§Section 2.3). We do this by piecing together original (theoretical and empirical) analyses revolving around various domains: the conclusions of such analyses may be "expected" by experts in these domains, but our novelty stems from underscoring some aspects that would otherwise remain hidden. Let us elaborate.

- Experts in Web-based measurement studies should know that some anti-phishing services may adopt a conservative approach leading to an underwhelming tpr... but, to the best of our knowledge, no prior work pointed out that such tpr *is 10x to 30x times lower for Chinese websites*.
- Experts in visual-based ML-PWD should know that these systems would work poorly on websites outside the reference list... but, to the best of our knowledge, we are the first to provide factual evidence that the reference lists of in prior work *do not include Chinese websites*.
- Experts in feature-based ML-PWD should know that the poor performance was due to lack of Chinese websites in the training set... but, to the best of our knowledge, we are the first to point out that *current feature extractors may be inappropriate* to deal with Chinese websites, and that there is a *shortage of publicly available datasets* for Chinese ML-PWD.
- Experts in socio-technical aspects of the Web should know that some geographical regions may be "underinvestigated" from a research perspective... but, to the best of our knowledge, we are the first to *point out such a discrepancy in the context of PWD and for Chinese websites*.
- Experts in machine learning should know that ensemble architectures are a pragmatic solution to classification problems... but, to the best of our knowledge, we are the first to *design an ensemble-based "multi-language" ML-PWD* built on top of an original language-selection tool.

Finally, our resources (data, and code) are novel, and they can be leveraged by future work to further investigate the problem we brought to light.

8.2. Limitations and future work

To avoid creating misunderstandings, we provide some disclaimers that underscore three potential limitations of our research. Then, we provide some avenues for future research.

- **Study Scope.** Our primary goal in this study is to provide *factual evidence* (theoretical and empirical) on the gap between Chinese and Western PWD. To this purpose, we scrutinize *existing* state-of-the-art techniques (open source and proprietary) for PWD and highlight their pros-and-cons. However, we *cannot realistically assess the entire spectrum of PWD* that has been proposed since the dawn of phishing website detection (especially given that most are not publicly accessible, or may rely on tools not active today). Hence, some PWD "may" work well in a cross-regional context. Also, in our research *we consider web pages designed for (and rendered by) desktop browsers* due to mobile phishing being less prominent in China (QiHoo360, 2023). Hence, our analysis does not cover the extent to which this problem (i.e., the gap between Chinese and Western PWD) affects detectors devoted to mobile users.
- **Deployment in Practice.** To bridge the gap between Chinese and Western Phishing website detection, we propose mitigations and release our code, tools, and data (including our proposed `ChiPhish` dataset, the first of its kind). However, *we cannot claim that our solutions are ready for operational deployment*: despite the good performance and low overhead, our results are drawn from the data we collected, and real-world assessments are necessary to demonstrate the efficacy of our methods in practice. As researchers, we cannot do such an assessment (also due to ethical reasons,[19]). Nonetheless, we hope that our resources will spearhead the development of novel techniques that can bridge the gap we brought to light.
- **Datasets Size.** For our evaluation, the size of our three datasets varies considerably (Table 3), with 15 111 samples for `EngPhish`; 11 204 for `WstPhish`; and 1 620 for `ChiPhish`. We acknowledge that such differences may hinder the generalizability of our findings *for Chinese websites*; indeed, this is the major reason why we do not claim that our proposed solutions will work if deployed "today" (and if trained on the exact same data). However, as we discussed (§3.1.1), there is a lack of publicly available repositories that can be used to create Chinese-specific datasets.

[19] **Ethical Statement:** for our research, we collected websites (benign and phishing) from well-known sources, and which were *publicly accessible* so we did not perform any sort of copyright infringement and also did not violate any existing regulation. We release all our data publicly, but some web pages may be inactive (or the URL may point at different resources), which is why complete replication of our results may not be possible. To protect internet users, *we do not deploy any phishing web page*, nor carry out on-field experimental campaigns focused on monitoring the browsing activities of unaware users.

Nowadays, the widespread adoption of mobile devices, especially smartphones, is driven by their wide range of functionalities. Cross-regional mobile Phishing defence can be an interesting avenue for future research. To further mitigate the threat of Chinese phishing attacks and bridge the gap between Chinese and Western PWD, we endorse future work to build upon our resources and expand our findings by, e.g., using `ChiPhish` to create a larger (and up-to-date) dataset of Chinese websites. Moreover, future endeavours can use `ChiPhish` as a basis for an "universal" ML-PWD (which improves our attempt in §7.2); or enhance LaSeTo so that the performance of the corresponding "ensemble" (§7.3) better approximates the one of the individual ML-PWD (§6.3.2).

8.3. Additional evidence and experiments (how "wide" is this gap, exactly?)

We conclude this section by carrying out two original analyses to incentivize future work to focus on these real-world problems. The first is a systematic literature review focused on revealing the extent to which prior research papers (published at top venues) on phishing (in any format) accounted for China (§8.3.1). The second is a proof-of-concept experiment showcasing how well anti-phishing services work on websites from other Eastern countries (§8.3.2).

8.3.1. How much awareness is there of "China" in phishing research?

We carry out a systematic literature review in February, 2024. We consider the proceedings from 2014 to 2023 of 11 top-venues related to Security, Human Factors and the Web: WWW, S&P (and EuroS&P), CCS, USENIX SEC, NDSS, AsiaCCS, ACSAC, IMC, WSDM, CHI. We searched for papers having "phish" in the title (We only considered full-papers, and not short or workshop papers) and found 56 papers. Then, we inspected the text of these 56 papers, scrutinizing if there were any occurrences of the terms "China", "Chinese" or "Eastern". Only 9 papers mentioned one of these terms in the text. Let us report the *context in which these terms are used*, and see whether they relate to "phishing website detection."

- (from ACM CCS): Aonzo et al. (2018) mentions "China" to refer that in this country there are third-party markets for Android apps; while Thomas et al. (2017) mention "China" once to specify that account hijakers are predominantly located in China.
- (from USENIX SEC): Hu et al. (2021) mention "Chinese" twice to refer that Chinese characters can be mixed with Latin characters in some browsers.
- (from ACM AsiaCCS): Peng et al. (2019a) mention "China" once, to specify that, during their investigation, they had an unusual login from "Beijing, China."
- (from ACSAC): Koide et al. (2023) mention "China" in a Table, and highlights that many websites with "squatting" domain names are in the East; however, it is not about phishing website detection. Liu et al. (2021b) mention "Chinese" 11 times, "China" 32 times and "Eastern" once: *this is indeed a paper about Chinese phishing*; however, it is about SMS spearphishing attacks—which is an orthogonal problem to phishing website detection.
- (from IMC): Saha Roy et al. (2023) mention "Chinese" once to denote that a coder "could not identify the intention of websites in Chinese language".
- (from ACM CHI): Althobaiti et al. (2021) mentioned "Chinese" once, stating that "in Chinese culture red is considered a happy color".
- (from IEEE EuroS&P): Ruggia et al. (2023) mention "China" once, stating that "in China, the Google Play market share is less than 4% while MyApp has 25% of the share".

Three venues (IEEE S&P, NDSS, WWW) had no papers with either of our search terms in the text; whereas WSDM did not have any paper with "phish" in the title. Nevertheless, it is surprising that, among the 56 papers that had "phish" in the title, there were more occurrences of the term "China" or "Chinese" ...but most of these referred to the countries of the authors/research grants.

Takeaway. In reputable venues that consider phishing-related themes, "China" is not a common term.

8.3.2. What about other "Eastern" countries?

The gap highlighted in this paper pertains to Chinese w.r.t. Western languages—and respective regions. However, phishing websites are also a problem in other areas besides China, each with its own language and regulations. Recent reports show an increasing trend of phishing websites in countries such as India (TrendMicro, 2022), Japan (Manichi, 2022), and the Middle East (Trellix, 2022). Unfortunately, these regions are vastly underrepresented in the PWD context (e.g., Verma (2013), Ahmad and Erdodi (2021)). Moreover, from a generic phishing perspective, few researches (e.g., Smeal et al. (2022), Tembe et al. (2014)) attempt to analyse the differences between such "minorities" and "Western" countries. Given the huge migration waves that interest Western countries (e.g., from the Middle East (Tausch, 2016), China, India (Lo et al., 2019) or Africa (Simko et al., 2018)), we hope our findings can inspire future efforts to scrutinize whether such issues also affect other geographical areas. **[Motivational Experiment.]** We scrutinize the effectiveness of existing PWD on *Japanese* (JP) and *Korean* (KR) websites. We collected a small sample of 200 webpages—of which 109 are benign (50 for JP and 59 for KR, taken from SimilarWeb (2023)) and 91 are phishing (50 for JP and 41 for KR, gathered from OpenPhish (2022e), PhishTank (2022g), Liu et al. (2022a)). We train our best ML-PWD (RF using F_c on `WstPhish`) and test it on them; we repeat this ten times. For JP: $F1{=}0.82{\pm}0.024$ ($tpr{=}0.74$; $tnr{=}0.94$); for KR: $F1{=}0.93{\pm}0.01$ ($tpr{=}0.93$; $tnr{=}0.95$). We also submit these to GSB (for which: $tnr{=}1.0$, and tpr is 0.04 for JP, and 0.0 for KR) and to VirusTotal (for which the average tpr is 0.12 for both JP and KR, but the tnr is 0.67 for JP, and 0.84 for KR). Our repository also includes these webpages.

Takeaway. Operational PWD work poorly also on websites of other Eastern countries (Japan & Korea).

9. Conclusion

This paper aims to reveal, assess, and mitigate the performance gap between Chinese and Western Phishing Website Detection. After collecting the *first* dataset for Chinese-focused PWD, we *practically* demonstrate the existence of such a gap in modern PWD. We assess the performance of state-of-the-art PWD, spanning across: 62 operational security services and 8 competition-grade ML-PWD developed by industry practitioners; and 10 open-source ML-PWD proposed in recent research. Our large evaluation reveals that real systems (tuned to minimize false positives) can detect at best 3% of phishing Chinese websites—whereas they can detect around 50% for Western languages. Such an imbalance also affects ML-PWD from research papers, which are hardly tested on Chinese websites (which appear to be very hard to collect). To bridge this gap, we propose, implement and pragmatically assess some possible solutions, whose overall performance in cross-regional settings achieves 0.98 tpr with 0.01 fpr on our datasets.

Takeaway. Existing PWD "in the West" are poorly equipped to detect phishing websites "in China" (and vice-versa). This is not acceptable given the constant migratory waves from/to these two sides of the World. We release all resources to facilitate future endeavours (https://github.com/joanyy/ChiPhish).

Our paper casts light on a hidden problem that likely also exists for other languages beyond Chinese. We encourage future efforts to **build**

upon our work, potentially by considering phishing websites targeting other "underrepresented" geographical areas in the PWD context.

CRediT authorship contribution statement

Ying Yuan: Writing – original draft, Visualization, Resources, Methodology, Investigation, Data curation. **Giovanni Apruzzese:** Writing – review & editing, Supervision, Project administration, Methodology. **Mauro Conti:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The resources (data and tools) used and created for this paper are provided in our Github: https://github.com/joanyy/ChiPhish.

Acknowledgments

This work was supported by the European Commission under the Horizon Europe Programme, as part of the project LAZARUS (Grant Agreement no. 101070303), project SERICS (PE00000014) under the NRRP MUR program funded by the EU-NGEU, and the project Privacy Aware Anti Malware (PAAM) funded by PRIN 2022 PNRR. This research was also partially supported by Hilti.

Appendix A. Validation of our new features

Let us justify why our implemented features (discussed in §6.2) are appropriate for a fair assessment of ML-PWD, and also robust from an "adversarial" perspective.

Diversity. Given the breadth of our enhanced feature set, it is important to determine if the values of our "new" features among the samples is sparse enough to prevent the occurrence of "evaluation artifacts" that bias the results (Arp et al., 2022). To this purpose, we extract features of all samples in our `ChiPhish` dataset (shown in Table A.10) and perform a quantitative analysis. We find that the resulting distribution makes it "challenging" for an ML model to classify a sample on the basis of a single feature. For instance, although all benign webpages have *H_icpDom*=0, the same holds for 97% of phishing samples. Furthermore, while 98% of phishing samples have *H_ecert*=1, the same holds for 83% of benign samples. These results suggest that our testbed represents a reliable way to test the proficiency of state-of-the-art ML-PWD proposed in research.

Robustness. We recall that many of our features rely on the ICP record: one may think that phishers may try to "spoof" such ICP record in order to trick an ML-PWD. We argue this is not simple: the ICP codes are released and verified by the issuer, and they frequently change. To give an example, consider the link provided in Fig. 2(a) (https://global.jd.com/). At the bottom of the page, there is a "business licence" as a code *which is linked to an image that shows the approved licence* (in this case, it can be viewed here and here). These licences change everytime they are renewed. (Note that we do not claim our ML-PWD to be robust against "adaptive" attackers, which is outside our scope). Moreover, we have reached out to the *maintainers of a Chinese ICP records' search tool* (https://www.tianapi.com/), who confirmed that the records change frequently, making it hard for phishers to keep up.

Table A.10
Distribution of our Chinese-specific features values (0s and 1s) in the samples of `ChiPhish` used in our evaluation (with 372 phishing samples collected by us, and 193 samples from Jiang and Wu (2022)).

Feature	Benign		Phishing	
	# 0s	# 1s	# 0s	# 1s
H_icpApp	1024	31	110	455
H_icpDom	1055	0	547	18
H_icpReg	1054	1	474	91
H_icpCode	533	522	518	47
H_ecert	183	872	7	558

Appendix B. Complete evaluation results

We report the complete results of our massive assessments, which we deferred to the Appendix, but which are still important for reproducibility, transparency, and benchmarking.

B.1. Performance of VirusTotal

We submit the raw HTML of each website of our datasets and report the performance of VirusTotal in B.11.

B.2. State-of-the-art, open-source and target-dependent ML-PWD

First, we report the detailed tpr and tnr of all our self-developed ML-PWD in Tables B.12, B.13 (for F_h, F_u, respectfully), which integrate Table 7 (for F_c, in the main paper). Specifically, the rows of these tables denote a specific ML-PWD, identified by its learning algorithm (Alg.) and training dataset (Train 80%). The values report the average performance (and std. dev, averaged over 10 trials) of the corresponding ML-PWD on the test partition of each 'language' dataset.

Then, we report in Table B.14 the *runtime* (in seconds) for training and testing our ML-PWD (on F_c), discussed in §6.3.2. For the CNN, we do not use GPU acceleration for a fair comparison.

B.3. Mlsec's ML-PWD (competition-grade)

We report in Table B.15 the exact tpr and tnr of each competition-grade ML-PWD of MLSEC.

Appendix C. Feature importance for ML-PWD using URL- or HTML-only features

Let us extend our analysis in §7.1 by studying the feature ranking for the best classifiers using F_u and F_h in our three datasets: `ChiPhish`, `WstPhish`, `EngPhish`.

Analysis of $\mathbf{F_u}$. We report in Fig. C.16 the top-10 features of the ML-PWD using RF (i.e., the best classifiers also for F_u, see Table B.12). From Fig. C.16, we see that there are five common features between `WstPhish` and `ChiPhish`, and six common features between `EngPhish` and `ChiPhish`. This is consistent with the results in Table B.12, i.e., the classifier trained on `ChiPhish` has a higher performance when tested on `EngPhish` than on `WstPhish`. Interestingly, 'U_pageRank' is the most important feature learned by the RF trained on `ChiPhish`, being *three times* more important than the second ranked feature (i.e., 'U_ssl'). In contrast, for `WstPhish`, 'U_pageRank' is also the first-ranked feature, but it is not as dominating as on `ChiPhish`, since it has a similar importance than three other features (i.e., 'U_shortestWrdPath, U_lngWrdPath, U_avgWrdPath'): this can explain why the RF trained on `ChiPhish` has high fpr when tested on `WstPhish`. Finally, the rankings between the RF trained on `WstPhish` and `EngPhish` are strikingly similar, suggesting why they also exhibit a good performance when tested on different datasets in phonological languages (refer to Table B.12).

Analysis of $\mathbf{F_h}$. We report in Fig. C.17 the top-10 features of the ML-PWD using RF (which, as shown in Table B.13, are the best classifiers also for F_h). By focusing on the ranking for `ChiPhish`, we observe that 'H_icpApp' and 'H_icpCode' are the most important

Table B.11
Performance of VirusTotal's anti-phishing services. We report the *tpr* and *tnr* of the 61 detectors queried by VirusTotal on each of our three datasets. Boldface show the best detector. Grey cells denote the average.

Anti-phishing service	WstPhish		EngPhish		ChiPhish		Anti-phishing service	WstPhish		EngPhish		ChiPhish	
	tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*		*tpr*	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
Bkav	0.00	1.00	0.00	1.00	0.00	1.00	*Lionic*	0.01	1.00	0.08	1.00	0.00	1.00
MicroWorld	0.05	1.00	0.20	1.00	0.01	1.00	*Panda*	0.00	1.00	0.00	1.00	0.00	1.00
ClamAV	0.00	1.00	0.00	1.00	0.00	1.00	*CMC*	0.00	1.00	0.00	1.00	0.00	1.00
QuickHeal	0.09	1.00	0.12	1.00	0.01	1.00	*McAfee*	0.02	1.00	0.04	1.00	0.00	1.00
Malwarebytes	0.00	1.00	0.00	1.00	0.00	1.00	*Zillya*	0.01	1.00	0.03	1.00	0.00	1.00
K7AntiVirus	0.00	1.00	0.00	1.00	0.00	1.00	*AVG*	**0.49**	**1.00**	**0.52**	**1.00**	**0.03**	**1.00**
K7GW	0.00	1.00	0.00	1.00	0.00	1.00	*NOD32*	0.14	1.00	0.20	1.00	0.00	1.00
Baidu	0.01	1.00	0.04	1.00	0.00	1.00	*VirIT*	0.00	1.00	0.01	1.00	0.01	0.99
Cyren	0.05	1.00	0.17	1.00	0.00	1.00	*Fortinet*	0.11	1.00	0.22	1.00	0.00	1.00
Symantec	0.01	1.00	0.11	1.00	0.00	1.00	*AhnLab*	0.01	1.00	0.04	1.00	0.00	1.00
HouseCall	0.00	1.00	0.04	1.00	0.00	1.00	*Avast*	0.49	1.00	0.52	1.00	0.03	1.00
Cynet	0.08	1.00	0.18	1.00	0.01	1.00	*Kaspersky*	0.01	1.00	0.04	1.00	0.00	1.00
BitDefender	0.05	1.00	0.19	1.00	0.01	1.00	*NANO*	0.02	1.00	0.20	1.00	0.01	1.00
SuperAntiSpyw.	0.00	1.00	0.00	1.00	0.00	1.00	*Tencent*	0.01	1.00	0.06	1.00	0.00	1.00
Ad-Aware	0.05	1.00	0.17	1.00	0.01	1.00	*Sophos*	0.00	1.00	0.06	1.00	0.00	1.00
Comodo	0.01	1.00	0.16	1.00	0.00	1.00	*F-Secure*	0.01	1.00	0.02	1.00	0.00	1.00
DrWeb	0.00	1.00	0.05	1.00	0.00	1.00	*VIPRE*	0.03	1.00	0.15	1.00	0.01	1.00
TrendMicro	0.00	1.00	0.04	1.00	0.00	1.00	*McAfee-GW*	0.02	1.00	0.07	1.00	0.01	1.00
FireEye	0.05	1.00	0.20	1.00	0.01	1.00	*Emsisoft*	0.05	1.00	0.17	1.00	0.01	1.00
Ikarus	0.09	1.00	0.23	1.00	0.03	1.00	*GData*	0.07	1.00	0.22	1.00	0.01	1.00
Jiangmin	0.00	1.00	0.00	0.99	0.00	1.00	*ZoneAlarm*	0.01	1.00	0.03	1.00	0.00	1.00
Avira	0.08	1.00	0.18	1.00	0.01	1.00	*Antiy-AVL*	0.02	1.00	0.12	1.00	0.00	1.00
Kingsoft	0.00	1.00	0.00	1.00	0.00	1.00	*Gridinsoft*	0.00	1.00	0.00	1.00	0.00	1.00
Arcabit	0.04	1.00	0.15	1.00	0.01	1.00	*ViRobot*	0.00	1.00	0.00	1.00	0.00	1.00
Microsoft	0.02	1.00	0.17	0.99	0.00	1.00	*Google*	0.11	1.00	0.28	1.00	0.02	1.00
Acronis	0.00	1.00	0.00	1.00	0.00	1.00	*VBA32*	0.00	1.00	0.00	1.00	0.00	1.00
ALYac	0.04	1.00	0.15	1.00	0.01	1.00	*MAX*	0.05	1.00	0.19	1.00	0.01	1.00
Zoner	0.00	1.00	0.01	1.00	0.01	1.00	*BitDefenderΘ*	0.00	1.00	0.00	1.00	0.00	1.00
Rising	0.01	1.00	0.05	1.00	0.00	1.00	*Yandex*	0.00	1.00	0.00	1.00	0.00	1.00
Tachyon	0.00	1.00	0.00	1.00	0.00	1.00	*MaxSecure*	0.01	1.00	0.04	1.00	0.00	1.00
Sangfor	0.04	1.00	0.17	1.00	0.00	1.00	AVERAGE	0.04	1.00	0.11	1.00	0.004	1.00

Table B.12
Performance of our ML-PWD analysing the F_u feature set (URL only). RF is the best *Alg*.

Alg.	Train 80%	*WstPhish* 20%		*EngPhish* 20%		*ChiPhish* 20%		Alg.	Train 80%	*WstPhish* 20%		*EngPhish* 20%		*ChiPhish* 20%	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*			*tpr*	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	WstPhish	0.99 ± 0.003	0.96 ± 0.006	0.89 ± 0.011	0.99 ± 0.002	0.53 ± 0.039	0.95 ± 0.012	*CNN*	WstPhish	0.98 ± 0.005	0.96 ± 0.006	0.89 ± 0.014	0.98 ± 0.014	0.68 ± 0.06	0.71 ± 0.077
	EngPhish	0.99 ± 0.003	0.6 ± 0.036	0.98 ± 0.003	1.0 ± 0.001	0.73 ± 0.038	0.55 ± 0.022		EngPhish	0.99 ± 0.007	0.68 ± 0.031	0.97 ± 0.006	1.0 ± 0.001	0.7 ± 0.128	0.45 ± 0.021
	ChiPhish	0.95 ± 0.006	0.38 ± 0.034	0.87 ± 0.014	0.9 ± 0.047	0.96 ± 0.019	0.98 ± 0.009		ChiPhish	0.92 ± 0.011	0.39 ± 0.041	0.84 ± 0.013	0.86 ± 0.037	0.94 ± 0.013	0.98 ± 0.008
LR	WstPhish	0.96 ± 0.005	0.95 ± 0.004	0.88 ± 0.012	1.0 ± 0.001	0.72 ± 0.032	0.81 ± 0.027	*GB*	WstPhish	0.98 ± 0.002	0.97 ± 0.005	0.89 ± 0.012	0.97 ± 0.019	0.68 ± 0.054	0.84 ± 0.026
	EngPhish	0.99 ± 0.001	0.65 ± 0.024	0.98 ± 0.004	1.0 ± 0.0	0.7 ± 0.041	0.51 ± 0.018		EngPhish	0.99 ± 0.002	0.26 ± 0.025	0.99 ± 0.004	1.0 ± 0.001	0.84 ± 0.022	0.43 ± 0.026
	ChiPhish	0.96 ± 0.005	0.29 ± 0.021	0.88 ± 0.013	0.9 ± 0.047	0.93 ± 0.024	0.98 ± 0.008		ChiPhish	0.95 ± 0.007	0.36 ± 0.025	0.88 ± 0.012	0.91 ± 0.033	0.95 ± 0.013	0.98 ± 0.007
DT	WstPhish	0.98 ± 0.004	0.95 ± 0.008	0.88 ± 0.015	0.74 ± 0.348	0.67 ± 0.067	0.77 ± 0.054	*AB*	WstPhish	0.97 ± 0.005	0.95 ± 0.005	0.88 ± 0.012	0.99 ± 0.002	0.75 ± 0.039	0.72 ± 0.033
	EngPhish	0.98 ± 0.006	0.35 ± 0.065	0.99 ± 0.004	0.99 ± 0.001	0.76 ± 0.079	0.41 ± 0.038		EngPhish	0.99 ± 0.003	0.29 ± 0.044	0.99 ± 0.003	1.0 ± 0.001	0.79 ± 0.045	0.38 ± 0.046
	ChiPhish	0.89 ± 0.044	0.33 ± 0.057	0.84 ± 0.064	0.74 ± 0.133	0.94 ± 0.025	0.97 ± 0.011		ChiPhish	0.94 ± 0.007	0.39 ± 0.032	0.86 ± 0.015	0.89 ± 0.111	0.95 ± 0.016	0.97 ± 0.015
KNN	WstPhish	0.95 ± 0.005	0.93 ± 0.008	0.83 ± 0.019	0.97 ± 0.006	0.6 ± 0.032	0.83 ± 0.034	*SVM*	WstPhish	0.96 ± 0.005	0.95 ± 0.005	0.88 ± 0.012	1.0 ± 0.001	0.71 ± 0.03	0.86 ± 0.025
	EngPhish	0.95 ± 0.006	0.81 ± 0.019	0.95 ± 0.006	0.99 ± 0.002	0.69 ± 0.049	0.89 ± 0.018		EngPhish	0.99 ± 0.002	0.48 ± 0.048	0.98 ± 0.004	1.0 ± 0.001	0.77 ± 0.05	0.44 ± 0.035
	ChiPhish	0.93 ± 0.007	0.44 ± 0.057	0.87 ± 0.016	0.84 ± 0.027	0.91 ± 0.017	0.97 ± 0.013		ChiPhish	0.92 ± 0.013	0.54 ± 0.042	0.85 ± 0.018	0.92 ± 0.029	0.91 ± 0.02	0.96 ± 0.015
MLP	WstPhish	0.98 ± 0.004	0.96 ± 0.009	0.88 ± 0.016	1.0 ± 0.002	0.55 ± 0.039	0.84 ± 0.04	Avg (std)	WstPhish	0.97 ± 0.013	0.95 ± 0.011	0.88 ± 0.019	0.96 ± 0.079	0.65 ± 0.072	0.81 ± 0.069
	EngPhish	1.0 ± 0.002	0.63 ± 0.037	0.98 ± 0.005	1.0 ± 0.001	0.74 ± 0.053	0.48 ± 0.027		EngPhish	0.99 ± 0.012	0.53 ± 0.181	0.98 ± 0.011	1.0 ± 0.002	0.75 ± 0.047	0.51 ± 0.144
	ChiPhish	0.95 ± 0.013	0.27 ± 0.025	0.88 ± 0.014	0.8 ± 0.037	0.94 ± 0.018	0.98 ± 0.01		ChiPhish	0.94 ± 0.019	0.38 ± 0.076	0.86 ± 0.017	0.86 ± 0.056	0.94 ± 0.015	0.97 ± 0.009

features—both of which are Chinese-specific features. Surprisingly, the first feature (i.e., 'H_icpApp') is *ten times* more important than the second (i.e., 'H_icpCode'): this can explain why – despite all three classifiers sharing some features in the respective top10 (six are common between `ChiPhish` and `EngPhish`, whereas five for `ChiPhish` and `WstPhish`) – they exhibit different performance when tested on websites of a different language group. Finally (and similarly to the RF analysing F_u), there are nine common features among the RF trained on `WstPhish` and those trained on `EngPhish`: this can explain why these classifiers perform similarly even on samples of a different dataset.

Appendix D. Comparison with SpacePhish (ablation study)

We find it instructive to assess the performance of the vanilla version of the ML-PWD developed in SpacePhish (Apruzzese et al., 2022a) when trained (and tested) on our datasets, but by using the original feature extractor of SpacePhish (Apruzzese et al., 2022h). Recall that the ML-PWD we considered in our evaluation analyse *(i)* 55 features from Apruzzese et al. (2022a) (2 of which changed by us) and *(ii)* 10 additional features created by us (cf. §6.2). Hence, we question whether our enhancements provide any substantial advantage w.r.t. the original 55 features of Apruzzese et al. (2022a). Given that our ML-PWD use additional features (including Chinese-specific ones by Zhang

Table B.13
Performance of our ML-PWD analysing the F_h feature set (HTML only). RF is the best *Alg.*

Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%		Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*			*tpr*	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	WstPhish	0.95 ± 0.006	0.93 ± 0.009	0.75 ± 0.029	0.97 ± 0.002	0.17 ± 0.038	0.96 ± 0.011	*CNN*	WstPhish	0.93 ± 0.004	0.89 ± 0.016	0.69 ± 0.021	0.93 ± 0.006	0.2 ± 0.053	0.85 ± 0.137
	EngPhish	0.7 ± 0.018	0.84 ± 0.014	0.93 ± 0.006	0.99 ± 0.003	0.45 ± 0.039	0.81 ± 0.016		EngPhish	0.69 ± 0.018	0.83 ± 0.027	0.89 ± 0.014	0.98 ± 0.005	0.38 ± 0.244	0.86 ± 0.111
	ChiPhish	1.0 ± 0.0	0.0 ± 0.001	1.0 ± 0.0	0.0 ± 0.0	0.84 ± 0.028	0.97 ± 0.007		ChiPhish	1.0 ± 0.001	0.01 ± 0.008	1.0 ± 0.001	0.01 ± 0.011	0.82 ± 0.034	0.97 ± 0.009
LR	WstPhish	0.86 ± 0.011	0.67 ± 0.022	0.83 ± 0.012	0.82 ± 0.007	0.44 ± 0.052	0.86 ± 0.024	*GB*	WstPhish	0.93 ± 0.008	0.88 ± 0.016	0.78 ± 0.018	0.93 ± 0.007	0.18 ± 0.046	0.96 ± 0.013
	EngPhish	0.78 ± 0.008	0.77 ± 0.01	0.82 ± 0.009	0.95 ± 0.005	0.57 ± 0.031	0.76 ± 0.019		EngPhish	0.76 ± 0.01	0.8 ± 0.012	0.9 ± 0.007	0.98 ± 0.003	0.46 ± 0.05	0.81 ± 0.013
	ChiPhish	1.0 ± 0.001	0.0 ± 0.002	1.0 ± 0.0	0.0 ± 0.001	0.83 ± 0.03	0.96 ± 0.012		ChiPhish	1.0 ± 0.003	0.01 ± 0.01	1.0 ± 0.001	0.03 ± 0.024	0.84 ± 0.035	0.97 ± 0.008
DT	WstPhish	0.95 ± 0.006	0.85 ± 0.011	0.7 ± 0.031	0.88 ± 0.011	0.27 ± 0.047	0.85 ± 0.039	*AB*	WstPhish	0.91 ± 0.01	0.78 ± 0.015	0.79 ± 0.014	0.87 ± 0.009	0.15 ± 0.024	0.97 ± 0.019
	EngPhish	0.66 ± 0.024	0.78 ± 0.015	0.93 ± 0.005	0.95 ± 0.004	0.48 ± 0.034	0.75 ± 0.021		EngPhish	0.77 ± 0.035	0.75 ± 0.011	0.87 ± 0.009	0.96 ± 0.004	0.51 ± 0.039	0.75 ± 0.019
	ChiPhish	0.97 ± 0.016	0.07 ± 0.042	0.99 ± 0.011	0.12 ± 0.091	0.84 ± 0.036	0.93 ± 0.02		ChiPhish	1.0 ± 0.001	0.02 ± 0.011	1.0 ± 0.002	0.05 ± 0.039	0.83 ± 0.035	0.97 ± 0.011
KNN	WstPhish	0.92 ± 0.01	0.84 ± 0.015	0.72 ± 0.014	0.88 ± 0.01	0.13 ± 0.019	0.97 ± 0.015	*SVM*	WstPhish	0.84 ± 0.01	0.77 ± 0.02	0.79 ± 0.014	0.87 ± 0.009	0.15 ± 0.024	0.97 ± 0.019
	EngPhish	0.72 ± 0.012	0.82 ± 0.014	0.87 ± 0.01	0.96 ± 0.006	0.48 ± 0.048	0.79 ± 0.019		EngPhish	0.77 ± 0.035	0.75 ± 0.011	0.82 ± 0.009	0.96 ± 0.005	0.51 ± 0.039	0.75 ± 0.019
	ChiPhish	1.0 ± 0.001	0.01 ± 0.003	1.0 ± 0.006	0.01 ± 0.003	0.81 ± 0.042	0.95 ± 0.011		ChiPhish	1.0 ± 0.001	0.02 ± 0.011	1.0 ± 0.002	0.05 ± 0.039	0.81 ± 0.033	0.97 ± 0.007
MLP	WstPhish	0.91 ± 0.013	0.88 ± 0.023	0.73 ± 0.042	0.91 ± 0.014	0.2 ± 0.034	0.86 ± 0.064	Avg (std)	WstPhish	0.91 ± 0.033	0.83 ± 0.075	0.76 ± 0.046	0.9 ± 0.043	0.21 ± 0.089	0.92 ± 0.054
	EngPhish	0.7 ± 0.025	0.83 ± 0.015	0.86 ± 0.013	0.98 ± 0.005	0.46 ± 0.036	0.81 ± 0.015		EngPhish	0.73 ± 0.042	0.8 ± 0.032	0.88 ± 0.037	0.97 ± 0.012	0.48 ± 0.049	0.79 ± 0.034
	ChiPhish	1.0 ± 0.001	0.0 ± 0.003	1.0 ± 0.0	0.0 ± 0.002	0.82 ± 0.032	0.96 ± 0.006		ChiPhish	1.0 ± 0.008	0.02 ± 0.02	1.0 ± 0.003	0.03 ± 0.037	0.83 ± 0.01	0.96 ± 0.014

Table B.14
Runtime. We report the avg (and std) seconds to train/test our ML-PWD (F_c) on our datasets (across 10 trials).

Alg.	WstPhish		EngPhish		ChiPhish	
	train (80%)	*test* (20%)	*train* (80%)	*test* (20%)	*train* (80%)	*test* (20%)
RF	1.2 ± 0.093	0.06 ± 0.003	1.37 ± 0.146	0.06 ± 0.004	0.19 ± 0.026	0.01 ± 0.003
LR	0.39 ± 0.151	0.0006 ± 0.0	0.768 ± 0.2672	0.00084 ± 0.00035	0.122 ± 0.0619	0.0021 ± 0.0014
DT	0.07 ± 0.005	0.0006 ± 0.0001	0.0656 ± 0.01156	0.00069 ± 0.00015	0.008 ± 0.0011	0.0013 ± 0.0001
KNN	0.0012 ± 0.00013	0.2796 ± 0.0055	0.0015 ± 0.00019	0.47866 ± 0.01578	0.001 ± 0.0001	0.0174 ± 0.0032
MLP	7.17 ± 1.433	0.0018 ± 0.0007	16.0134 ± 11.42109	0.00287 ± 0.00167	5.9 ± 1.9532	0.0026 ± 0.0015
CNN	695.23 ± 3.74	0.36 ± 0.0171	241.87 ± 15.13	0.47 ± 0.042	28.78 ± 3.242	0.23 ± 0.13
GB	4.58 ± 0.28	0.01 ± 0.0005	5.96 ± 0.57	0.011 ± 0.001	0.279 ± 0.0384	0.002 ± 0.0002
AB	1.73 ± 0.13	0.08 ± 0.0043	3.44 ± 0.41	0.16 ± 0.01	0.27 ± 0.034	0.018 ± 0.0023
SVM	2.57 ± 0.196	0.056 ± 0.0028	2.45 ± 0.44	0.014 ± 0.001	0.07 ± 0.024	0.002 ± 0.0002

Table B.15
Performance of each ML model (M) of the competition-grade ML-PWD considered in MLSEC.

M	WstPhish		EngPhish		ChiPhish	
	tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
*m*0	0.50	0.99	0.58	0.99	0.14	0.99
*m*1	0.69	0.72	0.76	0.66	0.35	0.62
*m*2	0.47	0.99	0.52	0.99	0.14	0.99
*m*3	0.63	0.77	0.66	0.73	0.38	0.63
*m*4	0.54	0.98	0.62	0.99	0.14	0.99
*m*5	0.77	0.54	0.81	0.35	0.49	0.53
*m*6	0.51	0.98	0.54	0.99	0.15	0.99
*m*7	0.70	0.65	0.70	0.52	0.43	0.57
Avg (std)	0.60 ± 0.104	0.83 ± 0.169	0.65 ± 0.097	0.78 ± 0.236	0.28 ± 0.140	0.79 ± 0.203

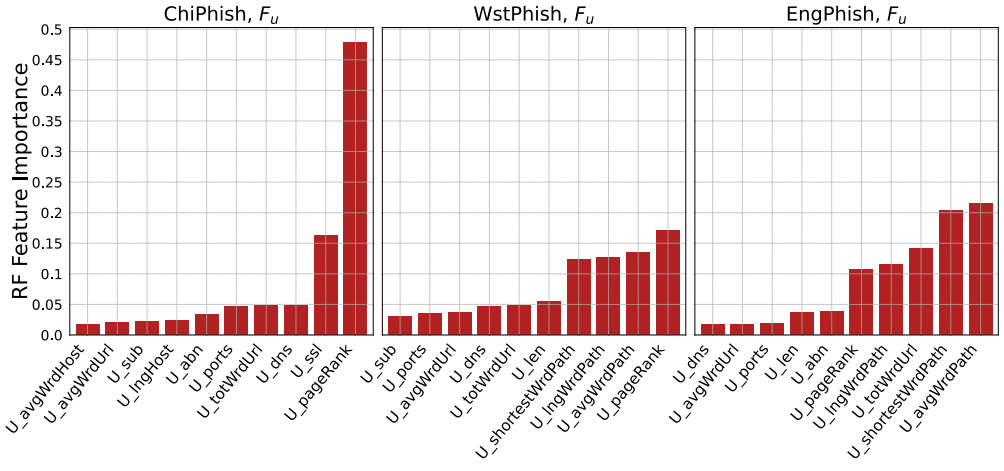


Fig. C.16. Feature rankings (top10) of RF (the best) using F_u.

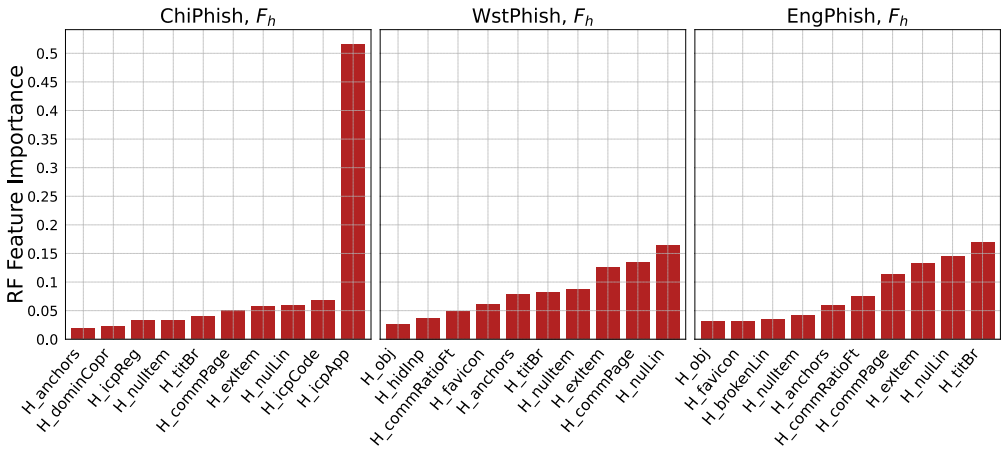


Fig. C.17. Feature rankings (top10) of RF (the best) using F_h.

Table D.16
Vanilla PWD of SpacePhish (Apruzzese et al., 2022a), analysing its F_c (trained/tested on our datasets).

Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	WstPhish	0.99 ± 0.004	0.97 ± 0.007	0.90 ± 0.007	1.00 ± 0.001	0.37 ± 0.052	0.97 ± 0.010
	EngPhish	0.98 ± 0.004	0.70 ± 0.020	0.98 ± 0.004	1.00 ± 0.000	0.71 ± 0.034	0.53 ± 0.034
	ChiPhish	0.94 ± 0.007	0.43 ± 0.016	0.86 ± 0.011	0.93 ± 0.02	0.95 ± 0.019	0.99 ± 0.006
LR	WstPhish	0.96 ± 0.005	0.95 ± 0.004	0.88 ± 0.017	1.00 ± 0.001	0.66 ± 0.046	0.86 ± 0.047
	EngPhish	0.99 ± 0.002	0.68 ± 0.021	0.98 ± 0.003	1.00 ± 0.001	0.72 ± 0.042	0.51 ± 0.029
	ChiPhish	0.94 ± 0.005	0.50 ± 0.020	0.84 ± 0.011	0.86 ± 0.019	0.94 ± 0.016	0.98 ± 0.010
CNN	WstPhish	1.00 ± 0.002	0.99 ± 0.002	0.88 ± 0.013	0.98 ± 0.007	0.50 ± 0.045	0.90 ± 0.022
	EngPhish	0.98 ± 0.006	0.54 ± 0.072	1.00 ± 0.002	1.00 ± 0.000	0.79 ± 0.048	0.49 ± 0.042
	ChiPhish	0.93 ± 0.007	0.45 ± 0.026	0.85 ± 0.013	0.90 ± 0.031	0.93 ± 0.020	0.98 ± 0.006

Table D.17
Vanilla PWD of SpacePhish (Apruzzese et al., 2022a), analysing its F_u (trained/tested on our datasets).

Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	WstPhish	0.98 ± 0.004	0.96 ± 0.006	0.90 ± 0.007	0.99 ± 0.001	0.50 ± 0.035	0.95 ± 0.011
	EngPhish	0.99 ± 0.003	0.63 ± 0.027	0.98 ± 0.004	1.00 ± 0.001	0.73 ± 0.046	0.54 ± 0.034
	ChiPhish	0.95 ± 0.006	0.38 ± 0.025	0.88 ± 0.009	0.88 ± 0.044	0.96 ± 0.018	0.99 ± 0.005
LR	WstPhish	0.96 ± 0.006	0.94 ± 0.008	0.88 ± 0.021	1.00 ± 0.001	0.72 ± 0.019	0.80 ± 0.073
	EngPhish	0.99 ± 0.002	0.68 ± 0.019	0.98 ± 0.004	1.00 ± 0.001	0.75 ± 0.048	0.49 ± 0.031
	ChiPhish	0.95 ± 0.007	0.43 ± 0.020	0.88 ± 0.008	0.93 ± 0.013	0.93 ± 0.022	0.98 ± 0.008
CNN	WstPhish	0.99 ± 0.002	0.97 ± 0.007	0.89 ± 0.012	0.98 ± 0.014	0.65 ± 0.073	0.73 ± 0.046
	EngPhish	0.99 ± 0.007	0.47 ± 0.107	0.99 ± 0.004	1.00 ± 0.001	0.78 ± 0.071	0.42 ± 0.046
	ChiPhish	0.94 ± 0.008	0.38 ± 0.048	0.86 ± 0.011	0.85 ± 0.046	0.94 ± 0.014	0.98 ± 0.008

et al. (2014)), we expect some form of improvement when a ML-PWD is tested on websites from the same dataset; however, the additional information may lead to overfitting.

Setup. We take the *exact* feature extractor of SpacePhish (from their repository (Apruzzese et al., 2022h)), and we use it to generate the feature representation of every sample in our three language datasets, i.e., `ChiPhish`, `EngPhish`, `WstPhish`. Then, we consider the *exact* same learning algorithms (i.e., RF, CNN, LR.) used in SpacePhish. Finally, we adopt the *exact* procedure described in our workflow (§6.3.1). We report the results of this "cross-regional" assessment in Tables D.16, D.17, D.18 (for F_c, F_u, F_h).

Results. Many insightful observations can be drawn by comparing these results with those of our main evaluation (shown in Tables 7, B.13, B.12). Let us focus on the most significant ones, i.e., those entailing F_c. First, as expected, each classifier of "our" ML-PWD tends to have a slightly superior performance (w.r.t. the vanilla ones in SpacePhish)[20] when tested on samples coming from the same language dataset; the best improvement is on the ML-PWD using LR (which is the learning algorithm allegedly used by Google (Liang et al., 2016)). However, we also note an intriguing phenomenon: the classifiers in SpacePhish, when trained on `ChiPhish` have a remarkably better performance when tested on `EngPhish` (w.r.t. the "enhanced" variant we used in our main evaluation—see Table 7). As an example, the vanilla RF of SpacePhish has a 0.93 *tnr*, whereas ours has 0.49 (even though ours has a 0.97 *tpr* against the 0.86 of SpacePhish). Finally, our RF and LR classifiers using F_c tend to be better than those in SpacePhish when tested on `ChiPhish`.

[20] Such improvement is statistically significant, e.g., a Welch t-test entailing both the *tpr* and *tnr* achieved by RF, LR, CNN trained and tested on `ChiPhish` and analysing F_c reveals that $p < 0.001$, therefore our variants are different (i.e., better) than SpacePhish's.

Table D.18
Vanilla PWD of SpacePhish (Apruzzese et al., 2022a), analysing its F_β (trained/tested on our datasets).

Alg.	Train 80%	WstPhish 20%		EngPhish 20%		ChiPhish 20%	
		tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
RF	WstPhish	0.94 ± 0.006	0.90 ± 0.008	0.69 ± 0.023	0.93 ± 0.004	0.32 ± 0.058	0.87 ± 0.019
	EngPhish	0.59 ± 0.023	0.84 ± 0.020	0.90 ± 0.011	0.98 ± 0.004	0.26 ± 0.055	0.87 ± 0.017
	ChiPhish	0.42 ± 0.042	0.75 ± 0.029	0.38 ± 0.034	0.62 ± 0.038	0.54 ± 0.045	0.87 ± 0.022
LR	WstPhish	0.86 ± 0.007	0.63 ± 0.015	0.82 ± 0.014	0.74 ± 0.007	0.61 ± 0.036	0.61 ± 0.030
	EngPhish	0.61 ± 0.010	0.79 ± 0.009	0.68 ± 0.013	0.94 ± 0.003	0.44 ± 0.044	0.79 ± 0.017
	ChiPhish	0.44 ± 0.025	0.79 ± 0.017	0.40 ± 0.019	0.59 ± 0.031	0.45 ± 0.028	0.88 ± 0.020
CNN	WstPhish	0.94 ± 0.005	0.91 ± 0.008	0.65 ± 0.017	0.88 ± 0.012	0.35 ± 0.066	0.85 ± 0.025
	EngPhish	0.62 ± 0.046	0.86 ± 0.027	0.90 ± 0.011	0.98 ± 0.003	0.28 ± 0.051	0.87 ± 0.025
	ChiPhish	0.50 ± 0.071	0.70 ± 0.051	0.44 ± 0.061	0.61 ± 0.043	0.48 ± 0.042	0.87 ± 0.027

Table E.19
Performance of VGG and CNN for analysing the screenshot of a webpage in our datasets.

M	WstPhish		EngPhish		ChiPhish	
	tpr	*tnr*	*tpr*	*tnr*	*tpr*	*tnr*
CNN	0.60 ± 0.021	0.40 ± 0.037	0.28 ± 0.020	0.74 ± 0.028	0.33 ± 0.054	0.68 ± 0.022
VGG	1.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	1.00 ± 0.000	0.00 ± 0.000	1.00 ± 0.000

Table E.20
Runtime (s) to train/test VGG and CNN. We train each model for 20 epochs (on a Tesla V100).

M	WstPhish		EngPhish		ChiPhish	
	train	test	train	test	train	test
CNN	4213.8	55.1	23048.5	298.2	2507.4	105.5
VGG	7228.9	849.9	22583.9	286.8	2185.3	26.2

We can hence make the following **considerations**:

- Our "improved" feature sets *(i)* employ strategies proposed by reputable prior work, and *(ii)* lead to a superior baseline performance...
- ...however, in some cases, such a higher performance in a same-language setting comes at the expense of reduced performance in a cross-language setting.

In summary, this experiment confirm the "no free lunch". By sacrificing some performance, it may be possible to improve the generalizability of the PWD. Our decision to develop an *ensemble* ML models (jointly with our LaSeTo) for PWD (§7.3) is inspired also by this result.

Appendix E. Negative result: a target *IN* dependent image-based ML-PWD

Motivation. In our paper, we have covered image-based PWD reliant on target dependent approaches. A question arises: "what about target INdependent PWD that use visual similarity?". To the best of our knowledge, *there is no paper that managed to do so effectively*. The reason is that, even by leveraging the capabilities of deep learning, it is difficult to design a PWD that can capture the nuances of benign/phishing websites just by, e.g., looking at its screenshot—given the immense variability that modern websites tend to have. Nonetheless, to provide additional proof that image-based ML-PWD are still immature for "target independent" PWD – and hence inappropriate to investigate our RQ2 (§5.2) – we perform an original proof-of-concept experiment.

Setup. We seek to develop an image-based PWD that leverages deep learning (DL) to discriminate benign from malicious webpages—i.e., a binary classification problem. For this purpose, we rely on our three datasets (§3) and, specifically, on the screenshots of each webpage included therein (as we did in §5.2.2, for the Chinese webpages, we use the 372 samples we collected in 2022 for ChiPhish, and the 193 samples from Jiang and Wu (2022); for the latter, we extract the screenshot manually by rendering the webpages' HTML). We chose two well-known DL algorithms as decision components: VGG16 (Simonyan and Zisserman, 2014) (we add dropout layers to improve generalization) and a CNN; we provide the exact implementation in our repository. We partition our datasets in train:test with the usual 80:20 split, and train and test each model on the respective dataset. We measure the performance with the *tpr* and *tnr*. We repeat this assessment 5 times. We report the detection results in Table E.19, and the runtime in Table E.20.

Results. From these (negative) results, we can see that these DL models are terrible at discriminating benign from malicious webpages—even in "same-region" context. Indeed, the performance is always skewed, showing either a perfect *tpr* but null *tnr* (and vice-versa) for VGG16; or just an unacceptably low *tpr* or *tnr* for the CNN. Furthermore, both the train and test runtime is much higher compared than our "feature-based" models (cf. Table B.14 with Table E.20): for instance, on WstPhish, the CNN analysing the screenshot requires 70 m to train (on GPU), whereas the CNN analysing F_c requires 11 m (on CPU). Simply, image-based PWD that are not target-dependent are not yet ready for practical deployment—which is why we did not include these in §5.

Lesson Learnt: Image-based PWD that perform binary classification (via the screenshot) are still *immature*: their performance is impractical, and the runtime is excessively high. This can be an *avenue for future research*, given the never-ending advances of deep learning.

References

360 secure brain, 2021. China mobile security status report for the first quarter of 2021. https://web.archive.org/web/20210802132226/https://www.freebuf.com/articles/paper/273527.html. Accessed in Dec 2022.

Abdelnabi, S., Krombholz, K., Fritz, M., 2020. VisualPhishNet: Zero-day phishing website detection by visual similarity. In: Proc. of CCS.

Acharya, B., Vadrevu, P., 2021. PhishPrint: evading phishing detection crawlers by prior profiling. In: USENIX Security Symposium.

Adebowale, M.A., Lwin, K.T., Sanchez, E., Hossain, M.A., 2019. Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text. Expert Syst. Appl..

Ahmad, H., Erdodi, L., 2021. Overview of phishing landscape and homographs in arabic domain names. Security and Privacy 4 (4), e159.

Al-Qurashi, R., AlEroud, A., Saifan, A.A., Alsmadi, M., Alsmadi, I., 2021. Generating optimal attack paths in generative adversarial phishing. In: Proc. of ISI.
Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., Wang, Y., 2022. An effective detection approach for phishing websites using URL and HTML features. Sci. Rep..
Althobaiti, K., Meng, N., Vaniea, K., 2021. I don't need an expert! making url phishing features human comprehensible. In: ACM CHI.
Aonzo, S., Merlo, A., Tavella, G., Fratantonio, Y., 2018. Phishing attacks on modern android. In: ACM CCS.
Apruzzese, G., Anderson, H., Dambra, S., Freeman, D., Pierazzi, F., Roundy, K., 2023a. Position:"real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice. In: Proc. of SaTML.
Apruzzese, G., Conti, M., Yuan, Y., 2022a. SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning. In: Proc. of ACSAC.
Apruzzese, G., Conti, M., Yuan, Y., 2022h. SpacePhish artifact. URL https://github.com/hihey54/acsac22_spacephish.
Apruzzese, G., Conti, M., Yuan, Y., 2023j. SpacePhish website. https://spacephish.github.io/.
Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Burdalo Rapa, L., Vasileios Grammatopoulos, A., Di Franco, F., 2023b. The role of machine learning in cybersecurity. ACM DTRAP 4, 1–38.
Apruzzese, G., Laskov, P., Tastemirova, A., 2022b. SoK: The impact of unlabelled data in cyberthreat detection. In: IEEE EuroS&P.
Apruzzese, G., Pajola, L., Conti, M., 2022c. The cross-evaluation of machine learning-based network intrusion detection systems. IEEE Trans. Netw. Serv. Manag. 19, 5152–5169.
Apruzzese, G., Subrahmanian, V., 2022. Mitigating adversarial gray-box attacks against phishing detectors. IEEE TDSC.
APWG, 2016. Phishing activity trends report. Tech. rep., https://docs.apwg.org//reports/apwg_trends_report_q4_2016.pdf.
APWG, 2024. Phishing activity trends report. Tech. rep., https://docs.apwg.org/reports/apwg_trends_report_q1_2024.pdf.
Ariyadasa, S., Fernando, S., Fernando, S., 2022. Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML. IEEE Access.
Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., Rieck, K., 2022. Dos and don'ts of machine learning in computer security. In: USENIX Sec..
Aydin, M., Baykal, N., 2015. Feature extraction and classification phishing websites based on URL. In: Proc. of IEEE CNS.
Bac, T.N., Duy, P.T., Pham, V.-H., 2021. PWDGAN: Generating adversarial malicious URL examples for deceiving black-box phishing website detector using GANs. In: Proc. of ICMLANT.
Bell, S., Komisarczuk, P., 2020. An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank. In: Proceedings of the Australasian Computer Science Week Multiconference.
Braun, B., Johns, M., Koestler, J., Posegga, J., 2014. PhishSafe: leveraging modern JavaScript API's for transparent and robust protection. In: ACM CODASPY.
Cheng, B., Ming, J., Fu, J., Peng, G., Chen, T., Zhang, X., Marion, J.-Y., 2018. Towards paving the way for large-scale windows malware analysis: Generic binary unpacking with orders-of-magnitude performance boost. In: Proc. of CCS.
Chiew, K.L., Tan, C.L., Wong, K., Yong, K.S., Tiong, W.K., 2019. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Inf. Sci..
China Internet Network Information Center, 2022. The 50th statistical report on China's internet development. Tech. rep., https://www.cnnic.net.cn/NMediaFile/2022/0926/MAIN1664183425619U2MS433V3V.pdf.
Chinaz, 2023b. Website ranking. https://top.chinaz.com.
Choo, E., Nabeel, M., Kim, D., De Silva, R., Yu, T., Khalil, I., 2023. A large scale study and classification of virustotal reports on phishing and malware urls. In: ACM POMACS.
Chu, W., Zhu, B.B., Xue, F., Guan, X., Cai, Z., 2013. Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs. In: IEEE ICC.
CN-Malicious-website-list Contributor, 2017. CN malicious websites. https://github.com/zzhihao2017/CN-Malicious-website-list.
Corona, I., Biggio, B., Contini, M., Piras, L., Corda, R., Mereu, M., Mureddu, G., Ariu, D., Roli, F., 2017. Deltaphish: Detecting phishing webpages in compromised websites. In: Proc. of ESORICS.
CujoAI, 2022. Machine learning security evasion competition (MLSEC). https://mlsec.io/.
Cyberspace Administration of China, 2022c. Cyber security law of the People's Republic of China. http://www.cac.gov.cn/2022-01/04/c_1642894602182845.htm.
Dalgic, F.C., Bozkir, A.S., Aydos, M., 2018. Phish-iris: A new approach for vision based brand prediction of phishing web pages via compact visual descriptors. In: Proc. of ISMSIT.
Dambra, S., Han, Y., Aonzo, S., Kotzias, P., Vitale, A., Caballero, J., Balzarotti, D., Bilge, L., 2023. Decoding the secrets of machine learning in malware classification: A deep dive into datasets, feature extraction, and model performance. In: ACM CCS.
Divakaran, D.M., Oest, A., 2022. Phishing detection leveraging machine learning and deep learning: A review. IEEE Secur. Priv..
Draganovic, A., Dambra, S., Iuit, J.A., Roundy, K., Apruzzese, G., 2023. "Do users fall for real adversarial phishing?" investigating the human response to evasive webpages. In: ECrime.
FBI, 2022. Internet crime report. Tech. rep., https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf.
Feng, Z., Xiangdong, H., Jiafu, L., Zhihui, G., Jun, F., Ke, L., 2017. Method of detecting the financial phishing webpage based on SVM. J. Chongqing Univ. Posts Telecommun..
Fu, A.Y., Wenyin, L., Deng, X., 2006. Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). IEEE Trans. Dependable Secure Comput. 3, 301–311.
Gandotra, E., Gupta, D., 2021. An efficient approach for phishing detection using machine learning. In: Multimedia Security: Algorithm Development, Analysis and Applications. pp. 239–253.
Gao, P., Shao, F., Liu, X., Xiao, X., Qin, Z., Xu, F., Mittal, P., Kulkarni, S.R., Song, D., 2021. Enabling efficient cyber threat hunting with cyber threat intelligence. In: IEEE ICDE.
Geng, G.-G., Lee, X.-D., Wang, W., Tseng, S.-S., 2013. Favicon-a clue to phishing sites detection. In: Proc. of ECrime.
Google, 2020. Compact language detector v3 (CLD3). https://github.com/google/cld3.
Google, 2023c. Google safe browsing. https://developers.google.com/safe-browsing/.
Hannousse, A., Yahiouche, S., 2021. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Eng. Appl. Artif. Intell. 104, 104347.
Hasegawa, A.A., Yamashita, N., Akiyama, M., Mori, T., 2021. Why they ignore english emails: The challenges of {non-native} speakers in identifying phishing emails. In: Proc. of SOUPS.
Ho, G., Cidon, A., Gavish, L., Schweighauser, M., Paxson, V., Savage, S., Voelker, G.M., Wagner, D., 2019. Detecting and characterizing lateral phishing at scale. In: USENIX Security.
Hoang, N.P., Niaki, A.A., Dalek, J., Knockel, J., Lin, P., Marczak, B., Crete-Nishihata, M., Gill, P., Polychronakis, M., 2021. How great is the great firewall? Measuring China's {dNS} censorship. In: USENIX Security.
HR, M.G., MV, A., et al., 2020. Development of anti-phishing browser based on random forest and rule of extraction framework. Cybersecurity 3, 1–14.
Hu, H., Jan, S.T., Wang, Y., Wang, G., 2021. Assessing browser-level defense against {IDN-based} phishing. In: USENIX SEC.
Huh, J.H., Kim, H., 2011. Phishing detection with popular search engines: Simple and effective. FPS 11, 194–207.
Interisle Consulting Group, 2021b. Phishing landscape 2021: An annual study of the scope and distribution of phishing. https://www.interisle.net/PhishingLandscape2021.html.
Jain, A.K., Gupta, B., 2018a. PHISH-SAFE: URL features-based phishing detection system using machine learning. In: Proc. of CSI.
Jain, A.K., Gupta, B.B., 2018b. Towards detection of phishing websites on client-side using machine learning based approach. Telecommun. Syst. 68, 687–700.
Jampen, D., Gür, G., Sutter, T., Tellenbach, B., 2020. Don't click: towards an effective anti-phishing training. a comparative literature review. Human-centric Comp. Inf. Sci..
Janet, B., Reddy, S., et al., 2020. Anti-phishing system using LSTM and CNN. In: Proc. of INOCON.
Jensen, M.L., Dinger, M., Wright, R.T., Thatcher, J.B., 2017. Training to mitigate phishing attacks using mindfulness techniques. J. Manage. Inf. Syst..
Jiang, Y., Wu, D., 2022. An integrated Chinese malicious webpages detection method based on pre-trained language models and feature fusion. In: International Conference on Web Information Systems and Applications. Springer, pp. 155–167.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science.
KaFan, 2023d. Kafan forum. https://bbs.kafan.cn/.
Koide, T., Fukushi, N., Nakano, H., Chiba, D., 2023. PhishReplicant: A language model-based approach to detect generated squatting domain names. In: ACSAC.
Kondracki, B., Azad, B.A., Starov, O., Nikiforakis, N., 2021. Catching transparent phish: Analyzing and detecting mitm phishing toolkits. In: ACM CCS.
Lain, D., Kostiainen, K., Čapkun, S., 2022. Phishing in organizations: Findings from a large-scale and long-term study. In: IEEE S&P.
Le, H., Pham, Q., Sahoo, D., Hoi, S.C., 2018. URLNet: Learning a URL representation with deep learning for malicious URL detection. arXiv preprint arXiv:1802.03162.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature.
Lee, J., Xin, Z., Ng, M.P.S., Sabharwal, K., Apruzzese, G., Divakaran, D.M., 2023. Attacking logo-based phishing website detectors with adversarial perturbations. In: ESORICS.
Li, X., Geng, G., Yan, Z., Chen, Y., Lee, X., 2016. Phishing detection based on newly registered domains. In: IEEE Big Data.
Li, T., Kou, G., Peng, Y., 2020. Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. Inf. Syst. 91, 101494.
Li, T., Tang, J., Xiao, L., Cai, M., 2021. Evaluation of smart library portal website based on link analysis. Procedia Comput. Sci. 188, 114–120.
Li, Y., Yang, Z., Chen, X., Yuan, H., Liu, W., 2019. A stacking model using URL and HTML features for phishing webpage detection. Future Gener. Comput. Syst. 94, 27–39.

Liang, B., Su, M., You, W., Shi, W., Yang, G., 2016. Cracking classifiers for evasion: a case study on the google's phishing pages filter. In: WWW.
Lin, Y., Liu, R., Divakaran, D.M., et al., 2021. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In: Proc. of USENIX Security.
Liras, L.F.M., de Soto, A.R., Prada, M.A., 2021. Feature analysis for data-driven APT-related malware discrimination. Comput. Secur..
Liu, D.-J., Geng, G.-G., Jin, X.-B., Wang, W., 2021a. An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment. Comput. Secur..
Liu, R., Lin, Y., Yang, X., Ng, S.H., Divakaran, D.M., Dong, J.S., 2022a. Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In: Proc. of USENIX Security.
Liu, R., Lin, Y., Yang, X., Ng, S.H., Divakaran, D.M., Dong, J.S., 2022f. PhishIntention artifact. https://github.com/lindsey98/PhishIntention.git.
Liu, R., Lin, Y., Zhang, Y., Lee, P.H., Dong, J.S., 2023. Knowledge expansion and counterfactual interaction for reference-based phishing detection. In: USENIX Security.
Liu, M., Zhang, Y., Liu, B., Li, Z., Duan, H., Sun, D., 2021b. Detecting and characterizing SMS spearphishing attacks. In: ACSAC.
Liu, M., Zhang, Z., Zhang, Y., Zhang, C., Li, Z., Li, Q., Duan, H., Sun, D., 2022b. Automatic generation of adversarial readable Chinese texts. IEEE TDSC.
Lo, L., Li, W., Yu, W., 2019. Highly-skilled migration from China and India to Canada and the United States. Int. Migr. 57, 317–333.
Makkar, A., Kumar, N., Sama, L., Mishra, S., Samdani, Y., 2021. An intelligent phishing detection scheme using machine learning. In: Proc. of ICMC.
Manichi, 2022. Phishing scams surge in Japan. https://mainichi.jp/english/articles/20220715/p2a/00m/0na/013000c.
Marchal, S., Saari, K., Singh, N., Asokan, N., 2016. Know your phish: Novel techniques for detecting phishing sites and their targets. In: Proc. of ICDCS.
Miao, C., Feng, J., You, W., Shi, W., Huang, J., Liang, B., 2023. A good fishman knows all the angles: A critical evaluation of google's phishing page classifier. In: ACM CCS.
Migration Policy Institute, 2022a. China's rapid development has transformed its migration trends. https://www.migrationpolicy.org/article/china-development-transformed-migration.
Mohammad, R.M., Thabtah, F., McCluskey, L., 2012. An assessment of features related to phishing websites using an automated technique. In: Proc. of ICITST.
Mohammad, R.M., Thabtah, F., McCluskey, L., 2014a. Intelligent rule-based phishing websites classification. IET Inf. Secur. 8, 153–160.
Mohammad, R.M., Thabtah, F., McCluskey, L., 2014b. Predicting phishing websites based on self-structuring neural network. Neural Comput. Appl. 25, 443–458.
Montaruli, B., Demetrio, L., Pintor, M., Biggio, B., Compagna, L., Balzarotti, D., 2023. Raze to the ground: Query-efficient adversarial HTML attacks on machine-learning phishing webpage detectors. In: Proc. of AISec.
Mowar, P., Jain, M., 2021. Fishing out the phishing websites. In: Proc. of CyberSA.
Netcraft, 2023f. Netcraft's cybercrime detection platform. https://www.netcraft.com/.
Niu, Y., Xie, R., Liu, Z., Sun, M., 2017. Improved word representation learning with sememes. In: Proc. of ACL.
OECD Ilibrary, 2021a. International migration outlook 2021 – China. https://www.oecd-ilibrary.org/sites/a13d0bc2-en/index.html?itemId=/content/component/a13d0bc2-en, Accessed in Dec 2022.
Oest, A., Safaei, Y., Zhang, P., Wardman, B., Tyers, K., Shoshitaishvili, Y., Doupé, A., Ahn, G.-J., 2020. Phishtime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists. In: Proc. of USENIX Security.
OpenPhish, 2022e. OpenPhish. https://openphish.com/.
Ozcan, A., Catal, C., Donmez, E., Senturk, B., 2021. A hybrid DNN–LSTM model for detecting phishing URLs. Neural Comput. Appl. 1–17.
Peng, P., Xu, C., Quinn, L., Hu, H., Viswanath, B., Wang, G., 2019a. What happens after you leak your password: Understanding credential sharing on phishing sites. In: ACM AsiaCCS.
Peng, P., Yang, L., Song, L., Wang, G., 2019b. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In: Proc. of IMC.
PhishDetector Contributor, 2023h. PhishDetector. https://www.moghimi.net/phishdetector.
PhishLabs, 2022. Financials see increase in phishing attacks, compromised sites lead staging methods in Q3. Tech. rep., https://www.phishlabs.com/blog/financials-see-increase-in-phishing-attacks-compromised-sites-lead-staging-methods-in-q3/.
PhishTank, 2022g. PhishTank. https://phishtank.org/.
ProofPoint, 2022. State of the phish 2022. Tech. rep., https://www.proofpoint.com/it/resources/threat-reports/state-of-phish.
ProofPoint, 2023. State of the phish 2023. Tech. rep., https://www.proofpoint.com/us/resources/webinars/2023-state-phish.
ProofPoint, 2024. State of the phish 2024. Tech. rep., https://www.proofpoint.com/uk/blog/security-awareness-training/2024-state-of-phish-report.
QiHoo360, 2019. China's mobile phone security status report in the first half of 2019. Tech. rep., https://web.archive.org/web/20210201131802/http://www.199it.com/archives/924497.html.
QiHoo360, 2020. China mobile security report 2020. Tech. rep., https://finance.sina.com.cn/tech/2020-09-30/doc-iivhuipp7243886.shtml.
QiHoo360, 2021. China mobile security report 2021. Tech. rep., https://pop.shouji.360.cn/safe_report/Mobile-Security-Report-202106.pdf.
QiHoo360, 2022. China mobile security report 2022. Tech. rep., https://pop.shouji.360.cn/safe_report/Mobile-Security-Report-202206.pdf.
QiHoo360, 2023. China mobile security report 2023. Tech. rep., https://pop.shouji.360.cn/safe_report/Mobile-Security-Report-202306.pdf.
Rao, R.S., Pais, A.R., 2019. Detection of phishing websites using an efficient feature-based machine learning framework. Neural Comput. Appl. 31, 3851–3873.
Roepke, R., Drury, V., Peess, P., Johnen, T., Meyer, U., Schroeder, U., 2022. More than meets the eye-an anti-phishing learning game with a focus on phishing emails. In: GALA.
Ruggia, A., Possemato, A., Merlo, A., Nisi, D., Aonzo, S., 2023. Android, notify me when it is time to go phishing. In: EuroSP.
Safi, A., Singh, S., 2023. A systematic literature review on phishing website detection techniques. J. King Saud Univ.-Comput. Inf. Sci. 35 (2), 590–611.
Saha Roy, S., Karanjit, U., Nilizadeh, S., 2023. Phishing in the free waters: A study of phishing attacks created using free website building services. In: IMC.
Sahingoz, O.K., Buber, E., Demir, O., Diri, B., 2019. Machine learning based phishing detection from URLs. Expert Syst. Appl. 117, 345–357.
Sarno, D.M., McPherson, R., Neider, M.B., 2022. Is the key to phishing training persistence?: Developing a novel persistent intervention. J. Exp. Psychol.: Appl..
Sharma, S.R., Parthasarathy, R., Honnavalli, P.B., 2020. A feature selection comparative study for web phishing datasets. In: Proc. of CONECCT.
Shusterman, A., Avraham, Z., Croitoru, E., Haskal, Y., Kang, L., Levi, D., Meltser, Y., Mittal, P., Oren, Y., Yarom, Y., 2020. Website fingerprinting through the cache occupancy channel and its real world practicality. IEEE TDSC.
SimilarWeb, 2023. Top websites ranking. https://www.similarweb.com/top-websites/.
Simko, L., Lerner, A., Ibtasam, S., Roesner, F., Kohno, T., 2018. Computer security and privacy for refugees in the United States. In: IEEE S&P.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Singh, P., Jain, N., Maini, A., 2015. Investigating the effect of feature selection and dimensionality reduction on phishing website classification problem. In: Proc. of IEEE NGCT.
Smeal, W., Kumar, Y., Vishwanath, V., Camp, L.J., Alexeev, A., 2022. Phishing resiliency across socio-cultural spheres: Cyrillic orthographic zone vs. The five eyes. In: Proc. of ACSAC'22 Poster Session.
Statcounter, 2022i. Statcounter: Browser Market Share China. https://gs.statcounter.com/browser-market-share/all/china.
Statista, 2022d. The most spoken languages worldwide 2022. https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide.
Tan, C.L., Chiew, K.L., Wong, K., et al., 2016. PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder. Elsevier Decis. Support Syst. 88, 18–27.
Tarun Bansal, 2023a. Building a more helpful browser with machine learning. https://blog.google/products/chrome/building-a-more-helpful-browser-with-machine-learning/.
Tausch, A., 2016. Muslim immigration continues to divide Europe: A quantitative analysis of European social survey data. Middle East Rev. Int. Affairs 20.
Tembe, R., Zielinska, O., Liu, Y., Hong, K.W., Murphy-Hill, E., Mayhorn, C., Ge, X., 2014. Phishing in international waters: exploring cross-national differences in phishing conceptualizations between Chinese, Indian and American samples. In: Proc. of HotSos.
Thomas, K., Li, F., Zand, A., Barrett, J., Ranieri, J., Invernizzi, L., Markov, Y., Comanescu, O., Eranti, V., Moscicki, A., et al., 2017. Data breaches, phishing, or malware? Understanding the risks of stolen credentials. In: ACM CCS.
Tian, K., Jan, S.T., Hu, H., Yao, D., Wang, G., 2018. Needle in a haystack: Tracking down elite phishing domains in the wild. In: Proc. of IMC.
Tian, J., Shen, C., Wang, B., Xia, X., Zhang, M., Lin, C., Li, Q., 2024. LESSON: Multi-label adversarial false data injection attack for deep learning locational detection. IEEE Trans. Dependable Secure Comput..
Trellix, 2022. Email cyberattacks on arab countries rise in lead to global football tournament. Tech. rep., https://www.trellix.com/en-us/about/newsroom/stories/research/email-cyberattacks-on-arab-countries-rise.html.
TrendMicro, 2022. Massive phishing campaigns target India banks clients. https://www.trendmicro.com/en_us/research/22/k/massive-phishing-campaigns-target-india-banks-clients.html.
Van Dooremaal, B., Burda, P., Allodi, L., Zannone, N., 2021. Combining text and visual features to improve the identification of cloned webpages for early phishing detection. In: ARES.
Venustech, 2023l. VenusEye. https://www.venuseye.com.cn/.
Verma, A., 2013. Effects of phishing on e-commerce with special reference to india. In: Interdisciplinary Perspectives on Business Convergence, Computing, and Legality. pp. 186–197.
Virustotal, 2023m. VirusTotal API. https://www.virustotal.com.
W3Techs, 2023k. Usage statistics of content languages for websites. https://w3techs.com/technologies/overview/content_language.
Wang, J., Wang, L., Dong, F., Wang, H., 2023. Re-measuring the label dynamics of online anti-malware engines from millions of samples. In: IMC.

Wikipedia, 2023e. Languages of Europe. https://en.wikipedia.org/wiki/Languages_of_Europe.
Wikipedia, 2023n. Websites blocked in China. https://en.wikipedia.org/wiki/List_of_websites_blocked_in_mainland_China.
Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: EASE.
worldometers, 2023i. Population comparison: China, EU, USA, and Japan. https://www.worldometers.info/population/china-eu-usa-japan-comparison/.
Xiang, G., Hong, J., Rose, C.P., Cranor, L., 2011. Cantina+: A feature-rich machine learning framework for detecting phishing web sites. ACM Trans. Inf. Syst. Secur. 14, 21.
Xiangdong, H., Ke, L., Feng, Z., Jiafu, L., Jun, F., Zhihui, G., 2017. Financial phishing detection method based on sensitive characteristics of webpage. Chinese J. Netw. Inf. Secur..
Yang, L., Zhang, J., Wang, X., Li, Z., Li, Z., He, Y., 2021. An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. Expert Syst. Appl. 165, 113863.
Yanting Jiang, Di Wu, 2022b. Chinese malicious web pages dataset and detection. https://github.com/JiangYanting/Chinese_Malicious_Web_Pages_Dataset_And_Detection.
Yoon, C., Kim, K., Kim, Y., Shin, S., Son, S., 2019. Doppelgängers on the dark web: A large-scale assessment on phishing hidden web services. In: Proc. of WWW.
Yuan, Y., Apruzzese, G., Conti, M., 2023. Multi-SpacePhish: Extending the evasion-space of adversarial attacks against phishing website detectors using machine learning. Digital Threats: Research and Practice.
Yuan, Ying and Apruzzese, Giovanni and Conti, Mauro , 2023g. Our repo. https://github.com/joanyy/ChiPhish.
Zhang, W., Jiang, Q., Chen, L., Li, C., 2017a. Two-stage ELM for phishing web pages detection using hybrid features. World Wide Web 20, 797–813.
Zhang, J., Pan, Y., Wang, Z., Liu, B., 2016. URL based gateway side phishing detection method. In: IEEE Trustcom/BigDataSE/ISPA. pp. 268–275.
Zhang, D., Yan, Z., Jiang, H., Kim, T., 2014. A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. Inf. Manag. 51, 845–853.
Zhang, X., Zeng, Y., Jin, X.-B., Yan, Z.-W., Geng, G.-G., 2017b. Boosting the phishing detection performance by semantic analysis. In: Proc. of IEEE Big Data. pp. 1063–1070.
Zuraiq, A.A., Alkasassbeh, M., 2019. Phishing detection approaches. In: Proc. of ICTCS.

Ying Yuan is a Postdoctoral researcher at the University of Padua. She completed her Ph.D. in the Department of Brain, Mind & Computer Science at the University of Padua in 2024. Her research interests include Phishing detection and the Security of Machine Learning.

Giovanni Apruzzese is an Assistant Professor at the University of Liechtenstein. His primary interests lie at the intersection of cybersecurity and data analytics, with a specific focus on networking, phishing, machine learning, and adversarial machine learning.

Mauro Conti is a Full Professor at the University of Padua. His research interests are mainly in the area of security and privacy.