**INTRODUCTION**

In the last years, web services usage has grown drastically due to the current digital transformation. Companies motivate the change by providing their services online, like e-banking, e-commerce or SaaS (Software as a Service) [1]. Nowadays, due to the COVID-19 pandemic, restrictions have spread out the work-from-home model, which implies extra millions of workers, students, and teachers developing their activities remotely [2], leading to a substantial additional workload for services such as email, student platforms, VPNs or company portals. Therefore, there are even more potential targets exposed to phishing attacks, where phishers try to mimic legitimate websites to steal users' credentials or payment information [3], [4]. Recent studies [5], [6] concluded that phishing is one of the most significant attacks based on social engineering during the COVID-19 pandemic, together with spam emails and websites to execute these attacks.

Identifying phishing sites through their HTTP protocol is no longer a valid rule. In the 3*rd* quarter of 2017 [7], the APWG reported that less than 25% of phishing websites were hosted under HTTPS protocol, whilst this amount has increased up to 83% in 1*st* quarter of 2021 [8]. These websites provide secure end-to-end communication, which transmits a false safe impression to the user while making an online transaction [9]. Furthermore, the Anti- Phishing Working Group (APWG) [10] has reported a significant increase in phishing attacks, i.e. from 165; 772 to 611; 877 websites, just between the first quarter of 2020 and 2021 respectively. A reason behind this increase might be that people have resorted (and still are) to online services during the COVID-19 pandemic.

One of the most popular solutions for phishing detection is the list-based approach, which analyzes the requested URL against a phishing database [11]. Some examples of this solution are Google SafeBrowsing,1 PhishTank,2 OpenPhish3 or SmartScreen.4 If a requested URL matches any record, the request is blocked, and a warning is displayed to the user before visiting the website. However, despite the capabilities of the list-based approach, it would fail if the phishing URL was not reported previously [12]\_[14], and it will require a continuous effort to update the database with newer phishing data. Bell and Komisarczuk [11] observed that many phishing URLs were removed after day five from Phishtank while Open Phish removed all URLs after seven days from its report. This issue allows attackers to reuse the same URL when it is removed from different lists.

Due to the mentioned drawbacks with the blacklist-based methods, automatic detection of phishing URLs based on machine learning, have attracted attention in research [15], [16]. These approaches can be grouped into four classes according to the type of data used for the detection: the text of the URL, the page content, the visual features and networking information [17]. Methods based on the page content and visual features require visiting the website to collect the source code and render it, which is a time-consuming task. Other availability limitations can be found in studies that rely on networking and 3*rd* party information such as WHOIS or search engine rankings. To overcome these limitations, we focus on phishing detection through URLs since it implies advantages such as fast computation -because no websites are loaded- and 3*rd* party and language independent, since features are extracted only from the URLs.

Existing URL datasets use the homepage URL from well-known websites as the legitimate [18], [19]. However, we think that the challenge is to determine if a *login form* of a website is legitimate or phishing. From our perspective, and to the best of our knowledge, publicly available datasets are not rejecting conditions that represent some real problems for phishing URL detection. Fig. 1 displays the differences between a homepage, a login page and a phishing website. Furthermore, it is observed that recent machine learning proposals obtained high accuracy using outdated datasets, i.e., typically containing URLs collected from 2009 to 2017. We demonstrate that models trained with old URLs decrease their performance when they are tested with URLs coming from recent phishing pages.

This paper presents a phishing URL dataset using legitimate login websites to obtain the URLs from such pages. Then, we evaluate machine and deep learning techniques for recommending the method with higher accuracy. Next, we show how models trained with legitimate homepages struggle to classify legitimate login URLs, demonstrating our hypothesis about phishing detection and legitimate login URLs. Additionally, we show how the accuracy decrease with the time on models trained with datasets from 2016 and evaluated on data collected in 2020. Finally, we provide an overview of current phishing encounters, explaining attacker tricks and approaches.

The main contributions of the paper can be summarized as

follows V

\_ We extended our previous dataset PILU-60K (Phishing Index Login URL) [20], from 60K to 90K URLs equally distributed among three classes: phishing, the

legitimate homepage, and legitimate login. We make this extended dataset, PILU-90K, publicly available for research purposes5

\_ Using PILU-90K, we implemented and evaluated three pipelines for URL phishing detection: (i) we use the 38 handcrafted feature descriptors proposed by Sahingoz *et al.* [21] for training eight supervised machine learning classifiers and also (ii) automatic feature extraction using Term Frequency Inverse Document Frequency (TF-IDF) at character N-gram level combined with Logistic Regression (LR) algorithm, and (iii) a Convolutional Neural Network (CNN) at character level too.

\_ We demonstrated empirically how an URL phishing detection model struggles in classifying login URLs when it was trained on the URLs of the homepage of

phishing and legitimate URLs.

\_ We evaluated the robustness of the proposed phishing detection over time. We trained the model on a dataset collected between March 2016 and April 2016, and we evaluated the model on other datasets collected between 2017 and 2020.

\_ Phishing websites were analyzed using domain frequency. We found six different phishing domains depending on the service hired by the attacker.

The organization of the paper is as follows: Section II reviews the literature on phishing detection. Next, Section III describes the proposed dataset and its content. Then, we explain the used features and the proposed classifiers in Section IV. The carried out experiments are covered in Section V. Section VI presents and discusses the obtained results. Finally, the main conclusions are drawn in Section VII, where we also point to our future work.