ISSN:2229-6107

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY

E-mail :
editor.ijpast@gmail.com
editor@ijpast.in

www.ijpast.in

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 13,Issuse 3.july2023

PhishingURL Detection:A Real-Case ScenarioThroughLoginURLs

Mr.P.Rajkumar[1],A.Shruthilaya[2],B.yashaswini[3],CH.Hanusri[4],C.Kavya[5]

ABSTRACT : Phishing is a social engineering cyberattack where criminals deceive users to obtain theircredentials through a login form that submits the data to a malicious server. In this paper, we comparemachine learning and deep learning techniques to present a method capable of detecting phishing websitesthrough URL analysis. In most current state-of-the-art solutions dealing with phishing detection, thelegitimate class is made up of homepages without including login forms. On the contrary, we use URLsfrom the login page in both classes because we consider it is much more representative of a real case scenarioand we demonstrate that existing techniques obtain a high false-positive rate when tested with fromlegitimateloginpagesAdditionally,weusedatasetsfromdifferentyearstoshowhowmodelsdecrease theiraccuracyovertimebytrainingabasemodelwitholddatasetsandtestingitwithrecentURLs.Also, Finally,wepresentaLogisticRegression modelwhich,combined withTerm Frequency - Inverse Document Frequency (TF-IDF) feature extraction, obtains 96.50% accuracy ontheintroducedloginURLdataset.

INDEXTERMSCybercrime,login,machinelearning,phishingdetection,URL.

I.INTRODUCTION

In the last years, web services usage has grown drasticallydue to the current digital transformation. Companies motivatethechangebyprovidingtheirservicesonline,like e-banking,e-commerceorSaaS(SoftwareasaService)[1].Nowadays,duetotheCOVID-19pandemic,restrictionshavespreadoutthe work-from-home model, which implies extra millions ofworkers, students, and teachers developing their activitiesremotely[2],leadingtoasubstantialadditionalworkloadforservicessuchasemail,studentplatforms,VPNs orcompanyportals.Therefore,thereareevenmorepotentialtargetsexposed to phishing attacks, where phishers try to mimiclegitimate websites to steal users' credentials or paymentinformation [3], [4]. Recent studies [5], [6] concluded thatphishingisoneofthemostsignificantattacksbasedonsocial engineering during the COVID-19 pandemic, together withspamemailsandwebsitestoexecutetheseattacks. Identifying phishing sites through their HTTP protocol isnolongeravalidrule.Inthe3rd quarterof2017[7],theAPWGreportedthatlessthan25%ofphishingwebsites

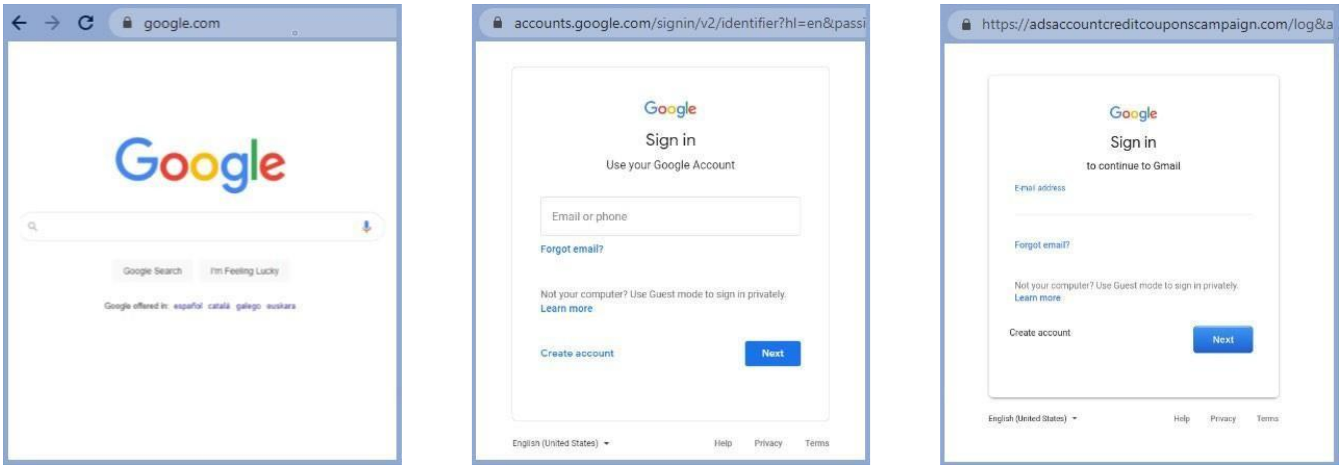
Theassociateeditorcoordinatingthereview of thismanuscriptandapprovingitforpublicationwasSenthilKumar

AssistantProfessor,DeptofComputerScienceand
[2,3,4,5]UGscolarDepartmentofComputerScienceandEngineeringMallaReddyEngineeringCollegeforWomen,Secundarabad
rajkpatil@gmail.com
abbashruthilaya@gmail.com,yashubeldaribgr@gmail.com,hanusrichilagani@gmail.com,
chinnangarikavya@gmail.com

hasincreasedupto83%in1[s]

end-to-end communication,which transmits a false safe impression to the user whilemaking an online transaction [9]. Furthermore, the Anti-PhishingWorkingGroup(APWG)[10]hasreportedasignificant increase in phishing attacks, i.e. from 165, 772 to611,877websites,justbetweenthefirstquarterof2020and2021 respectively. A reason behind this increase might be thatpeople have resorted (and still are) to online services duringtheCOVID-19pandemic.

One of the most popular solutions for phishing detectionisthelist-basedapproach,whichanalyzestherequested



(a) Legitimate homepage (b) Legitimate login page (c) Phishing web page

FIGURE1. Differencebetweenlegitimatehome(a),legitimatelogin(b)andphishing(c)pages.Sampleslike(a)arecommonlyusedinstate-of-the-artapproaches.Weintroduceinour dataset sampleslike(b), whichhasasimilarlook to phishingattackslike(c).

URL againstaphishingdatabase[11].SomeexamplesofthissolutionareGoogleSafeBrowsing,[1]PhishTank,[2]OpenPhish[3]or SmartScreen.[4]If a requested URL matchesanyrecord,therequestisblocked,andawarningisdisplayedto the user before visiting the website. However, despite thecapabilities of the list-based approach, it would fail if thephishing URL was not reported previously [12]–[14], and itwill require a continuous effort to update the database withnewer phishing data. Bell and Komisarczuk [11] observedthat many phishing URLs were removed after day five fromPhishtank while OpenPhish removed all URLs after sevendays from its report. This issue allows attackers to reuse thesameURLwhenitisremovedfrom differentlists.

Due to the mentioned drawbacks with the blacklist-basedmethods, automatic detection of phishing URLs based onmachine learning, have attracted attention in research [15],[16]. These approaches can be grouped into four classesaccording to the type of data used for the detection: the text ofthe URL, the page content, the visual features and networkinginformation [17]. Methods based on the page content andvisual features require visiting the website to collect thesource code and render it, which is a time-consuming task.Other availability limitations can be found in studies that relyonnetworkingand3rdpartyinformationsuchasWHOISor search engine rankings. To overcome these limitations,we focus on phishing detection through URLs since it impliesadvantages such as fast computation -because no websitesare loaded- and 3rdparty and language independent, sincefeaturesare extractedonlyfromthe URLs.

ExistingURLdatasetsusethehomepageURLfromwell-knownwebsitesasthelegitimate[18],[19].However,

[1]https://safebrowsing.google.com/[2]https://www.phishtank.com/[3]https://openphish.com/[4]https://bit.ly/2OJDYBS

wethinkthatthechallengeistodetermineifa*loginform*ofawebsite is legitimate or phishing. From our perspective, andto the best of our knowledge, publicly available datasets arenot reflecting conditions that represent some real problemsfor phishing URL detection. Fig.1 displays the differencesbetween a homepage, a login page and a phishing website.Furthermore, it is observedthat recent machine learningproposals obtained high accuracy using outdated datasets,i.e., typically containing URLs collected from 2009 to 2017.WedemonstratethatmodelstrainedwitholdURLsdecreasetheir performance when they are tested with URLs comingfromrecent phishingpages.

ThispaperpresentsaphishingURLdatasetusinglegit-

Next,we show how models trained with legitimate homepagesstruggle to classify legitimate login URLs, demonstratingourhypothesisaboutphishingdetectionandlegitimateloginURLs. Additionally, we show how the accuracy decreasewiththetimeonmodelstrainedwithdatasetsfrom2016andevaluated on data collected in 2020. Finally, we provide anoverviewofcurrentphishingencounters,explainingattackertricksandapproaches.

:

- We extended our previous dataset PILU-60K (PhishingIndexLoginURL)[20],from60Kto90KURLs equally distributed among three classes: phishing, thelegitimate homepage, and legitimate login. We makethis extended dataset, PILU-90K, publicly available forresearchpurposes[5]
- weimplementedandevaluatedthreepipelinesforURLphishingdetection:(i)we

[5]https://gvis.unileon.es/dataset/pilu-90k/

use the 38 handcraftedfeaturedescriptorsproposedby Sahingoz*et al.* [21] for training eight supervisedmachinelearningclassifiersandalso(ii)automaticfeature extraction using Term Frequency Inverse Doc-ument Frequency (TF-IDF) at character N-gram levelcombinedwithLogisticRegression(LR)algorithm,and (iii) a Convolutional Neural Network (CNN) atcharacterlevel too.
- howanURLphishing detection model struggles in classifying login URLswhen it was trained on the URLs of the homepage ofphishingandlegitimateURLs.
- oftheproposedphishing detection over time. We trained the model on a datasetcollected between March 2016 and April 2016, and Weevaluatedtheirrobustness Wedemonstratedempirically Weevaluatedthemodelonotherdatasetscollectedbetween2017and2020.
- analyzedusingdomainfre-quency.Wefoundsixdifferentphishingdomainsdependingontheservicehiredbytheattacker.

Theorganizationofthepaperisasfollows:SectionII Phishingwebsiteswere reviewstheliteratureonphishingdetection.Next,SectionIIIdescribestheproposeddatasetanditscontent. Then,we explain the usedfeaturesandtheproposedclassi-fiersinSectionIV.ThecarriedoutexperimentsarecoveredinSectionV.SectionVIpresentsanddiscussestheobtainedresults.Finally,the mainconclusionsaredrawn in Section VII, where we also point to our futurework.

II.STATEOFTHEART

Intheliterature,researchershavefocusedonphishingdetection following three main approaches: *List-based* andautomaticdetectionusing*MachineLearning*and*DeepLearning*techniques.

The list-based approach, well-known for detecting phishingURLs [22]–[24], can be based on whitelists or blacklists,dependingiftheystorelegitimateorphishingURLs,respectively.JainandGupta[24]developedawhitelist-basedsystem that blocks all websites which are not on that list.Conversely, the blacklist-based systems, like Google SafeBrowse or PhishNet [23], are more common as they provide azerofalse-positiverate,i.e.nolegitimatewebsiteisclassifiedas phishing. However, they can be compromised if an attackermakes changes on a blacklisted URL. Besides, they dependheavilyontheupdaterateofthesystem’srecords.Therefore,a list-basedapproachis not a robust solution due to thehigh volume of new phishing websites introduced daily andtheir short lifespan, which is estimated to be 21 days onaverage[12].

B.MACHINELEARNINGMETHODS

To overcome blacklist disadvantages, researchers have devel-opedmachinelearningmodelstodetectunreportedphishing encounters. Depending on their input data, these approachescan be classified into two categories: URL-based and content-based.

1)URL-BASED

Buber*etal.*[25]implementedaURLdetectionsystemcom-posedoftwosetsoffeatures.Thefirstwasa209wordvector,obtainedwith‘‘StringToWordVector’’toolfromWeka.[6]Thesecond, 17 NLP (Natural Language Processing) handcraftedfeatures such as the number of sub-domains, random words,digits, special characters and length measurements over theURL words. Combining both feature sets, they obtained ahigh 97.20% accuracy with Weka’s RFC (Random ForestClassifier)ona10%sub-samplesetfromEbbu2017dataset.In the following studies, Sahingoz*et al.* [21] defined threedifferent feature sets: Word vectors, NLP and a hybrid setcombiningboth sets. Theyobtaineda97.98%accuracyonRandomForest (RF) using only 38 NLP features onEbbu2017[25]dataset.Inthiswork,weusedtheNLPfeatures from Sahingoz*et al.* [21], since they reported state-of-the-artperformance inthelaststudies.

JainandGupta[26]builtananti-phishingsystemusing14handcraftedURLdescriptors,includingsomeobtainedusing3rdpartyserviceslikeWHOISregisters orDNSlookups.

BanikandSarma[27]implementedalexicalfeatureselection from URL to optimize thenumberoffeaturesandtheaccuracyoftheirmodel.Theystartedwithasetof 17 descriptors and removed the lesssignificantonesuntil they reached an optimal performance. Using 9 featuresand a Random Forest (RF) classifier they obtained 98.57%accuracyonanextensionofPWD2016[18]dataset.

)CONTENT-BASED

Content-based works use features extracted mainly from thewebsites' source code. However, most of the current workscombine these with URLs and other 3rdparty services suchasWHOIS [28],[29].

Oneofthefirstcontent-basedworkswasCANTINA[30],whichconsistsofaheuristicsystembasedonTF-IDF.CANTINA extracts five words from each website using TF-IDF and introduced them into the Google search engine. If adomain was within the n first results, the page was consideredlegitimate,orphishingotherwise.Theyobtainedanaccuracy of 95% with a threshold of $n = 30$ Google search results.DuetotheuseofexternalserviceslikeWHOIS[7]andthehigh false-positiverate,authorsproposedCANTINA+ [31].Theirnewproposalachieveda99.61%F1-Scoreincludingtwo filters:(i)acomparisonofhashedHTMLtagswithknown

[6]https://www.cs.waikato.ac.nz/ml/weka/[7]https://www.whois.net/

phishing structures and (ii) the discarded websites with noform.

MoghimiandVorjani[32]proposedasystemindependentfromthirdserviceslikeGooglePageRankorWHOIS.Theyused two handcrafted feature sets, extracted from the URLand the Document Object Model (DOM)of the website.The first set has nine legacyfeatures includinga set ofkeywords, while the second has eight novel features whichinform of whether the website's resources are loaded usingSSL protocol or not. They used Levenshtein distance [33] todetect typo-squatting by comparing the website and resourcesURLs. These features were used to train an SVM classifierandobtainedanaccuracyof98.65%ontheirbankingwebsitesdataset.

Adebowale*et al.*[34]createdabrowserextensiontoprotectusersbyextractingfeaturesfromtheURL,thesource code, the images, and features extracted using third-party services like WHOIS. Those features were introducedinto an Adaptive Neuro-Fuzzy Inference System (ANFIS)andcombinedwiththeScale-InvariantFeatureTrans-form (SIFT) algorithm, obtaining an accuracy of 98.30% onRami*et al.*[35]dataset.

Rao and Pais [28] developed a phishing website classifierusing the URL, the hyperlinks on the HTML code andthird-party services including the age of the domain and the pagerankonAlexa.Theyreached99.31%accuracywithaRandomForest classifier.

Yang*etal.*[36]proposedanExtremeLearningMachine (ELM) model and established three different groupsoffeatures:(i)Surfacefeatures,composedof12URLhandcrafted and 4 Domain Name System (DNS) featuresrelated to the registration date and the DNS records for thatdomain; (ii) 28 Topological features that are related to thestructure of the website and (iii) 12 deep features related tothetextandimagesimilarity.CombiningthesesetsoffeaturesandtheELMclassifier,theyobtained97.5%accuracy.

Sadique*et al.* [37] presented a framework for real-timephishing detection using four sets of URL features: (i) Lexicalfeaturesrelatedtothenumberofcharacters,dotsandsymbols found in different parts of the URL, (ii) host-basedfeaturesthatarerelatedtothehost,(iii)WHOISfeaturesarerelated to the registration date and (iv) GeoIP-based featureslikethe AutonomousSystem Number(ASN).A total of142 individual features were evaluated using 98, 000 samplesfrom Phishtank, where legitimate samples are also pickedfrom false positives collected at PhishTank. Theyobtaineda 90.51% accuracy on a Random Forest classifier using theproposeddescriptors.

Li *et al.* [29] presented a stacking model which was thecombination of three models: Gradient Boost Decision Tree(GBDT), eXtreme Gradient Boosting (XGBoost) and LightGradientBoostingModel(LGBM).Thisstackingmodelwasfed with a set of features from different sources: eight fromthe URL, 11 from the HTML and HTML string embeddingsinspired by Word2Vec model [38]. They obtained 97.30%accuracyusinga49,947samplesdataset.

CONTENT-BASED

RegardingthemethodsbasedonDeeplearning,Some-sha*et al.* [39] proposed a model based onLong Short-Term Memory (LSTM) to classify phishing URLs using tenhandcrafted features from Rao and Pais [28]. Those featuresare three URL features based on the number of dots, thelength of the URL, and the presence of HTTPS, six featuresextracted from the HTML, including the internal links andimages, the ratio of broken links and the presence of anchorlinks on the HTML body. Finally, one third-party numericfeature was obtained from Alexa's Page Rank. These featureswereextractedfroma3,526samplesdatasetandintroducedintotheLSTMmodeltoobtain99.57%accuracy.

presented an RCNN model to classifyphishing URLs. They used the URL as input for a tokenizerandthenusedaone-hotencodingtorepresenttheURLasa matrix at a character level. The last step is to set a fixedlength of 200 characters for the model input. If the URL isunderthatthreshold,theremainingcharactersarefilledwithzeros.Otherwise,thecharactersabovethelimitaretrimmed.Finally, they used a 310, 642 URL dataset to feed an RCNNmodel, which obtained 95.02% using the aforementionedcharacterembedding levelfeatures.

Al-AlyanandAl-Ahmadi[41]proposedamodifiedConvolutional Neural Network (CNN). First, they omittedtheURLprotocolandthencroppedURLslargerthan256 characters. Aljofey*etal.*[40] They used a 69 characters alphabet withlower-case letters, numbers and some symbols to obtain a128 embedding vector. Then, a one-dimensional CNN wasapplied to obtain 95.78% accuracy on a 2, 307, 800 URLsdataset.

Zhao *et al.* [42] presented a Gated Recurrent Neural Net-work(GRU)capableoflearningsequencesandpatternswithin the URLs. They compared this approach against a setof 21 handcrafted features combined with an RF classifier.Results showed how automatic feature extraction combinedwith GRUs outperformed RF, reaching 98.5% and 96.4%respectively.

III.DATASET:PHISHINGINDEXLOGINURLs(PILU-90K)

Phishers uselogin formsto retrieveand steal users'data.As far as we are concerned, the legitimate class in mostphishingdatasetsarerepresentedbyURLsfromtheirhomepages [18], [19]. However, most websites have theirlogin form in different locations, making models trainedwith suchpublicdatasetstobebiasedsincetheURLsof homepages tend to be shorterandsimplerthan others.Anexampleofthisisdepictedin Figure2.

Inthispaper,wepresentanextendedversionofthePhishing Index Login URL (PILU-60K) dataset [20] and wename it PILU-90K. PILU-90K contains 90*K* URLs dividedinto three classes (see Figure 2): 30K legitimate URLs ofhomepages, 30K legitimate login URLs and 30K phishingURLs.

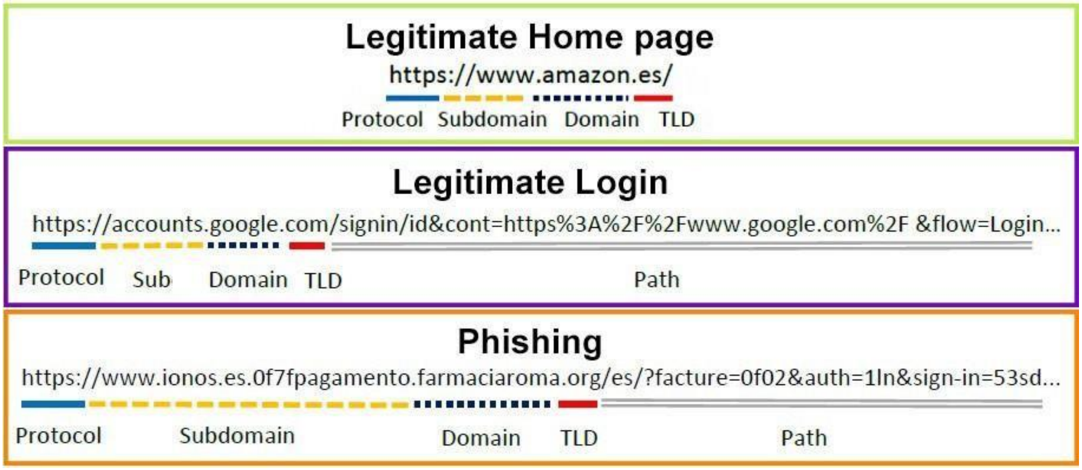


FIGURE2. TypesofURLsinPILU-90Kandtheirparts.AhomepageURL(up),aloginpageURL(middle)andaphishingURL(bottom).The length variationbetweenalegitimateloginpageandaphishingone is minimum.

TABLE 1. Number of samples distributed in the different subsets used in this work.

Subset	Legit Index	Phishing	Legit Login
PIU-60K	30,000	30,000	-
PLU-60K	-	30,000	30,000

Quantcast website,[8] which provides the most visited domains from the United States. The list provided on that website only contains the domain names, so we visited them to extract the complete URL. To reach the login page from a website, we used the Selenium web driver[9] and Python, checking buttons or links that could lead to the login form web page. Once we found the presumptive login, we inspected if the form had a password field in order to confirm whether it was a login form. Otherwise, it was not added to the dataset. We collected reported phishing URLs from Phishtank [21],[36],[39], between November 2019 and February 2020.

In this work, we have built two subsets from the PILU-90K dataset to conduct the proposed experiments. The first one, named PIU-60K (Phishing Index URLs), is built using the URLs of both the homepages of the legitimate samples and the phishing ones, following the configuration of most of the current state-of-the-art approaches. The second one, PLU-60K (Phishing Login URLs), follows our strategy, i.e. it contains URLs of both legitimate login pages and phishing ones. Table 1 shows the distribution of the available URLs into each subset.

To the best of our knowledge, none of the works in the state-of-the-art use legitimate login URLs specifically. By using legitimate login URLs, our work not only reflects the real-world scenario but also shapes an unbiased dataset

[8] https://www.quantcast.com/products/measure-audience-insights/
[9] https://selenium.dev/projects/

in terms of URL length. Table 2 include examples of URLs of each class in PILU-90K, where differences are noticeable between the legitimate index URLs and the other two classes. Specifically, the length of the different parts of the URLs and the usage of keywords like *login*, *signin* or *secure*, are the most remarkable ones. Figure 3 provides an overview of the distribution of the URLs length in the proposed subsets, where PLU-60K displays a more similar distribution between classes than the PIU-60K subset.

On the one hand, a quarter of the legitimate login forms URLs do not have a path, i.e. login forms were located on the homepages, matching its URL structure with the homepage samples. On the other hand, one out of seven samples from the phishing class does not have a path, so they will also match with the legitimate homepage samples, increasing the classification challenge, even for skilled humans.

IV. METHODOLOGY

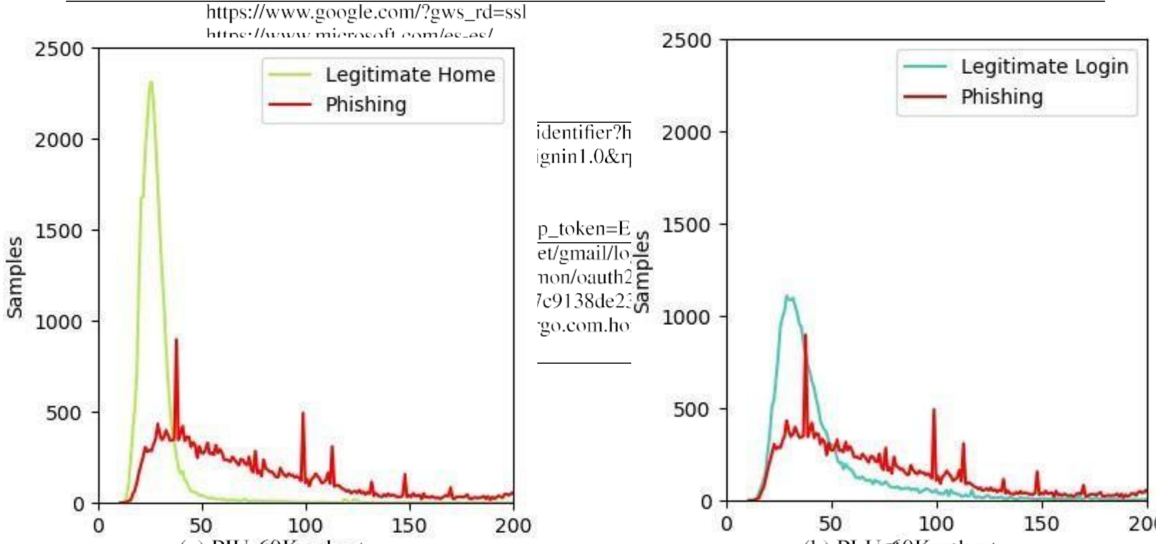
In this paper, we compare the performance of machine learning and deep learning methods for URL phishing classification. Regarding ML techniques, we used for feature extraction the handcrafted features proposed by Sahingoz *et al.* [21] and (ii) statistical features using Term Frequency-Inverse Document Frequency (TF-IDF) combined with character N-gram. Concerning the DL techniques, we adopted the CNN models of Zhang *et al.* [43] and Kim [44].

A. MACHINE LEARNING TECHNIQUES

Text classification based on supervised machine learning consists of three main stages: text preprocessing, text representation to convert the input text into a vector of features and a classifier. In this section, we explain the two techniques we used to extract features along with the evaluated classifiers.

TABLE 2. Examples of URLs within the different sets collected on PILU-90K dataset.

Class	URLs
	https://www.google.com/?gws_rd=ssl
	https://www.microsoft.com/es-es/
	identifier?h
	ignin1.0&r
	p_token=E
	et/gmail/lo
	non/oauth2
	7c9138de2
	go.com.ho



(a) PIU-60K subset (b) PLU-60K subset

FIGURE3. DistributionofURLlengthacrosssubsets.PIU-60Ksubsetwithasignificantdifferencebetweenclasses.Incontrast,PLU-60Ksubsethasacloserlengthdistributionoverbothclasses.

For the handcrafted features, URLs were parsed using *tldextract*[10] library. Then, raw words are extracted from the different parts of the URL by splitting the string using a set of symbols (specifically, ‘/’, ‘-’, ‘.’, ‘@’, ‘?’, ‘&’, ‘=’, ’_’). After preprocessing, we extracted 38 features proposed by Sahingoz *et al.* [21] using URL rules and NLP features: the frequency of aforementioned symbols, number of digits in the domain, subdomain and path (see Figure 2) and their lengths. Other features are evaluated, such as the number of subdomains, domain randomness using the Markov Chain Model, whether it has a common TLD (Top Level Domain), whether ‘www’ or ‘com’ are on other places different from the TLD. From the raw words, the following metrics are extracted: maximum, minimum, average and standard deviation of the words length, number of words, compound words, words equals or similar to famous brands or a keyword

[10] https://pypi.org/project/tldextract/

like ‘secure’ or ‘login’, consecutive characters in the URL and the presence of Punycode.

Given these features, we trained and compared eight supervised classifiers commonly used in the related literature [28], [29], [45], [46]:

In NLP, another popular feature extraction technique is the TF-IDF algorithm [49], a statistical approach that gives more or less weight to a term depending on how many documents such term occur on, i.e. the higher the number of URLs a term occurs on, the lower the weight and vice-versa. A term in the TF-IDF algorithm can be either a word or N-gram of characters. Given that the URLs might not have word terms in common, we resorted to the character N-gram. Therefore, TF-IDF operates on the character N-gram level to find patterns of

TABLE 3. Phishing datasets used in this work.

Dataset	Author	Year	Category	Legitimate samples	Phishing samples
PWD2016	Chiew et al. [18]	2016	A	12,550	15,000
1M-PD	Yuan et al. [19]	2017	A	500,000	500,000
Ebbu2017	Buber et al. [25]	2018	B	36,400	37,175
PIU-60K	This work	2020	A	30,000	30,000
PLU-60K	This work	2020	B	30,000	30,000

We used the averaged values of 10-fold cross-validation, reporting the accuracy (Eq. (3)), the F1-Score (Eq. (4)), the precision (Eq. (1)) and the recall (Eq. (2)) [21], [28], [34].

N consecutive characters of a given URL. Following the work of Al-Nabki *et al.* [50], we extracted grams between two to five characters, i.e. $N \in [2,5]$. The text preprocessing was limited to converting the text to a lowercase. The extracted features were used to train an LR classifier given its good performance on similar noisy text tasks, such as File Name Classification [50], [51].

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$

$$F1 = 2\,\frac{Precision \cdot Recall}{Precision+Recall}$$

B. DEEP LEARNING TECHNIQUES

Besides the machine learning approaches, we explored the use of CNN to classify URLs [19], [41]. We selected the architectures of Zhang *et al.* [43] and Kim *et al.* [44], which operate at a character level.

The model of Kim *et al.* was originally built to function as a character-based language model. To use the model for URLs classification, we replaced the subsequent recurrent layers with a dense layer to perform a softmax operation over the classes. In contrast, the model of Zhang *et al.* did not require modifications to its architecture as it was intended for the text classification. It is worth mentioning that for both models, we did not carry out any text preprocessing step.

TPdenotesthetruepositives,i.e.,howmanyphishingwebsiteswerecorrectlyclassified.FPreferstothefalsepositivesandrepresents thenumberoflegitimatesampleswronglyclassifiedasphishing.TN(i.e.,thetruenegatives)denotesthenumberoflegitimatesamplescorrectlyclassified.Finally,FNrepresentsthefalsenegativesthatrepresentthenumberofphishingwebsitesmisclassifiedaslegitimateones.Regardingtheclusteringexperiments,weusedthesameapproachofAl-Nabki*et al.* [50]fortextrepresentation,asexplainedinSectionIV-Aand,fortheclustering,weusedthe Agglomerative Hierarchical Clustering (AHC) [52]. Theclusteringprocessisrepeatedfourtimes,andeachtimeweinitializedtheAHCwiththenumber$n$ofthedesiredclusters, i.e.$n \in \{4,5,6,7\}$.

VI.RESULTSANDDISCUSSION

A.MACHINELEARNINGANDDEEPLEARNING APPROACHES

V.EXPERIMENTSANDRESULTS

A.DATASETS

Totestthemodelrobustnessagainst URLscollectedindifferent periods, we used the five phishing datasets showninTable3.

These datasets are grouped into two different categoriesdependingontheirrecollectionstrategy:(i)categoryA:PWD2016, 1M-PD and PIU-60K collected legitimate sam-plesbyinspectingthetop-visiteddomainsand(ii)category

Inthefollowing,wereporttheresultofthedesignedmachinelearningclassifiersusingbothhandcraftedandautomaticfeatureextractiontechniques.Then,deeplearningapproaches are presented and compared with the previousones.Finally,weprovedtheimpactofusinglegitimateloginURLsagainstthecurrent state-of-the-artapproach.

1)HANDCRAFTEDFEATUREEXTRACTION

B:Ebbu2017andPLU-60Kvisitedthosewebsitesandperformedfurtheractions:inthecaseofEbbu2017,itsauthors retrieved the inner URLs and, in the case of PLU-60K, we looked for the login form page. Therefore most ofthe URLs include a path. Table 4 shows the distribution ofsamplestructurewithinthedatasets.

B.EXPERIMENTALSETTINGS

Experiments are executed on an Intel Core i3 9100F at 3.6Ghzand 16 GB of DDR4 RAM. We used scikit-learn[11]and Python3fortheimplementationofthedifferentexperiments.

Forthemachinelearningexperiments,weempiricallyassign the parameters that returned the best accuracy on thethree different phishing datasets. These parameters are showninTable5.

[11]https://scikit-learn.org/stable/

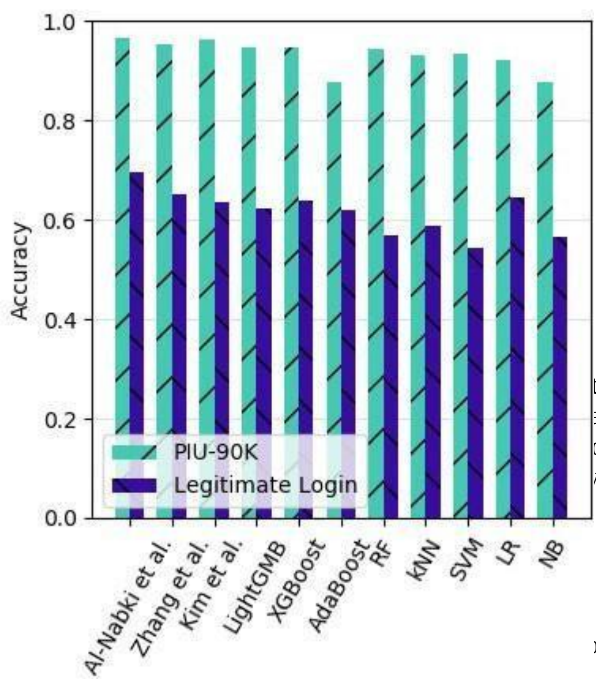
In this configuration, we extracted handcrafted features andbenchmarkedseveralclassifiers,asexplainedinSectionIV-A. Each model was trained and tested on each subset of thePILU-90K dataset. Table 6 reports the performance of eachclassier. It can be seen that XGBoost, LightGBM and RFoutperformtherestoftheclassifiersonbothsubsets,obtain-ing 93.22%, 93.12% and 92.91% accuracy on PLU-60K,respectively.WhileforthePIU-60Ksamplesubset,94.63%,94.67% and 94.42% accuracy were obtained, respectively.Results for the eight machine learning algorithms showed thatSahingoz*et al.* [21] descriptors achieve better performanceon PIU-60K. Length-based features, the number of wordsandthepresenceofkeywordsenhancetheperformancewhen the difference between legitimate and phishing URLsis significant. Using the PLU-60K subset, such descriptorsdecrease their performance since their values are similarbetweenclasses.

TABLE4. PhishingURLsdatasetsdistribution.

		PWD2016	1M-PD	PIU-60K	Ebbu2017	PLU-60K
Legitimate URLs	w/o a path	**10,548 (84.00%)**	**471,728 (94.35%)**	**26,446 (88.15%)**	1,684 (4.62%)	6,693 22.31%
	w/ a path	2,002 (16.00%)	28,272 (5.65%)	3,554 (11.85%)	**34,716 (95.38%)**	**23,307 (77.69%)**
Phishing URLs		15,000	500,000	30,000	37,175	30,000

TABLE5. Parameterconfigurationforthedifferentmodelsanddatasets.

Algorithm	Parameter	PWD2016	Ebbu2017	PILU-90K
LightGBM	n_leaves	100	500	100
	objective	binary	binary	binary
XGBoost	n_estimators	100	100	100
AdaBoost	n_estimators	50	50	50
RF	n_estimators	250	350	250
kNN	k	1	1	3
SVM	kernel	rbf	rbf	rbf
	gamma	0.1	0.1	0.1
NB	algorithm	bernoulli	bernoulli	bernoulli
LR	solver	lbfgs	lbfgs	lbfgs



2)AUTOMATICFEATUREEXTRACTION

In this experiment, we evaluate the cl terN-gramforfeatureextractionandLRforclassification,asexplainedinSecti ained a classificationmodelandreporteditsperformance.Automaticfeatur other methodsintheF1-score,includingthosebasedonDeepLearning.Forthe of 96.93%,while for the PLU-60K, accuracy was 96.50%. Hence that dependonhandcraftedfeatures (seeTable6).

3)EVALUATIONOFDEEPLEARNING-BASEDPHISHING DETECTIONMODELS

Similarly,wetrainedandevaluatedtheproposedCNNcharacter-based models of both subsets of the PILU-90Kdataset.We foundthat the modelof Zhang *et al.* [43] hasan accuracy of 95.22% on the PIU-60K subset and 94.10%on the PLU-60K one. The model of Kim [44] has a slightlybetterresultwithanaverageaccuracyof96.43%onthePIU-60K and 96.00% on the PLU-60K (see Table 6). Comparedto machine learning algorithms, both CNN models obtainedbetter results than handcrafted features but TF-IDF combinedwith N-gram [50] remains as the best classifier for the twoproposedsubsets.

4)IMPACTOFTHEREPRESENTATIONOFTHELEGITIMATE CLASSONTHECLASSIFICATION

We assessed the impact on URL phishing classifiers whenthey are trained with samples where the legitimate class isrepresentedwithhomepageURLs,e.g.PIU-60K.Wetrained 11classifiersandreportedtheiraccuracy,asshowninFigure 4. Then, these models classified 30, 000 legitimateloginURLs and their accuracywas reportedagain. It canbe seen that all the models have suffered from a significantdecreaseintheiraccuracy.Al-Nabki*etal.*[50]model's

FIGURE 4. Accuracy of classification models trained on PIU-60K subsetandthereportedaccuracywhenclassifying30,000legitimateloginURLs.

accuracy decreased 27% and was the most resilient with69.50%accuracy.SVMdecreaseditsaccuracyupto39.12%and obtained the worst result, 54.46% accuracy. CNN modelsofZhang *etal.* [43]and Kim[44] obtained anaccuracyof 65.13% and 63.50%, respectively. Furthermore, modelsbasedonhandcraftedfeatures,obtainedthelowestaccuracy,probably,duetothelength-basedfeatures.

We observed that all models, including those trained withautomaticfeatures,misclassifiedmorethan30%ofthelegitimate login URLs. These results can interfere with theapplication of the model in real-world applications since itpresentsahighfalse-positiverate.WearguethatourTF-IDFand N-gram approach trained with PLU-60K can solve thisissuesince itcan classifylegitimate loginsampleswithhigh accuracy as seen in Table 6. It should be noticed thatthis capability reduces overall accuracy in the advantage ofreducingthefalsepositiveswhenusersvisitloginpages.

B.ANALYSISOFTHEPERFORMANCEOFPHISHING MODELSOVERTIME

Recentmachinelearningproposalshavereportedgoodperformancetrainedwith PWD2016andEbbu2017datasets.Sincephishingattacksand, asaconsequence,phishing

TABLE6. Performanceoftheassessedalgorithmsonthesubsetsof PILU-90Kdatasets.Theeightfirstrowscorrespondtohandcraftedfeatureextractionmethods,whereasthe9thonecorrespondstoautomaticfeatureextractionmethods.Thelasttwocolumnsdepicttheresultsfortheassessed deeplearningmodels.Alltheresultsaregivenin%.

	PIU-60K				PLU-60K			
Algorithm	**Precision**	**Recall**	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**Accuracy**	**F1-Score**
LightGBM	95.38	93.89	**94.67**	**94.63**	93.15	91.60	93.12	92.36
XGBoost	95.21	93.99	94.63	94.59	94.02	92.32	**93.22**	**93.16**
AdaBoost	94.18	91.72	93.03	92.93	89.24	85.82	87.74	87.50
RF	91.57	94.25	94.42	94.40	92.78	93.06	92.91	92.92
kNN	94.06	92.18	93.18	93.11	91.52	89.05	90.40	90.27
SVM	94.15	92.95	93.59	93.55	91.80	89.83	90.91	90.81
LR	93.57	90.91	92.33	92.22	86.64	81.87	84.62	84.19
NB	93.84	80.73	87.72	86.79	78.79	68.99	75.21	73.56
TF-IDF + N-gram	96.57	96.58	**96.93**	**96.93**	96.54	96.48	**96.50**	**96.51**
Zhang et al. [43]	95.93	94.57	95.22	95.24	92.12	95.90	94.10	93.97
Kim et al. [44]	95.22	97.57	96.43	96.38	95.96	96.02	96.00	95.99

websites' URLs get more and more sophisticated over time,we hypothesize that models trained with outdated datasetsmaydecreasetheirperformancewhenanalyzingrecentURLs.

To prove if this hypothesis is correct, we used PWD2016and Ebbu2017 and the features from Sahingoz*et al.* [21] totrain eight machine learning models (see Table 7) and testthem using URLs from recent years. These datasets are 1M-PDfrom2017,PIU-60Kfrom2020andPLU-60Kalsofrom2020.Amongtheproposeddatasetswefoundtwocategories(see Section III). Datasets in category A were built usinglegitimate homepage URLs with no path, whereas in categoryB they include the path. For each category, we created apipeline to avoid biased results. The first pipeline was focusedonclassifyingURLswithnopath,andweusedcategoryAdatasets:PWD2016,1M-PDandPIU-60KcontainingURLscollectedin2016,2017and2020,respectively.Inthispipeline, PWD2016 was used to train the eight machinelearning algorithms and then it was evaluated using 1M-PDand PIU-60K. The second pipeline focused on classifyingURLs with a path and, in this case, we used the datasetsfrom category B: Ebbu2017 and PLU-60K, which containURLs collected in 2017 and 2020, respectively. In this case,Ebbu2017 was used to train the proposed algorithms and thenPLU-60Kwasutilizedtotestitsperformance.

FromtheexperimentalresultsshowninTable7,allmodels struggled to endure over time and their performancedecreased when tested on the followingyears'datasets.The model LightGBM obtained the best accuracy on bothpipelines, but its results were the most affected over time,losing 10.42% and 30.69% accuracy on the first and secondpipelines,respectively.Ontheotherhand,SVMobtainedthebestresultsonrecentdatasetsforthefirstpipeline,achieving89.04% on the PIU-60K test, a 6.24% less than with thePWD2016dataset usedfortraining.

Overall results for the first pipeline, showed how a modeltrained with four years old datasets could not reach 90%accuracy,even whenthey obtainedhigh performance onthe base dataset. Moreover, the second pipeline, involvingURLs classification with paths, also struggled to maintainperformanceonrecent URLs.

TABLE7. Phishingdetectionaccuracyevolutionovertime(in%).

Training set	PWD2016			Ebbu2017	
Test set	**PWD2016**	**1M-PD**	**PIU-60K**	**Ebbu2017**	**PLU-60K**
LightGBM	97.60	91.42	87.18	95.94	65.25
XGBoost	97.47	91.65	87.59	95.27	65.75
AdaBoost	95.27	91.71	87.95	89.77	61.30
RF	97.32	91.72	88.15	95.69	64.02
kNN	95.49	90.02	86.42	92.55	58.92
SVM	95.28	91.87	89.04	93.05	63.43
NB	87.89	86.39	85.18	80.70	60.91
LR	93.37	89.07	86.95	87.90	58.40

C.CLUSTERINGPHISHINGURLs

In this experiment, we attempt to cluster the phishing URLssearching for patterns. By analyzing the obtained clusters,wedidnotidentifysignificantrelations amongsamples,despitethenumbersoftheclusterswetried.Nevertheless, when$n$∓,wenoticedassociationsbetweenURLsbutafurthermanualinspectionoftheclustersleadtouncertain conclusions. URLs were clustered due to similarities betweendifferent parts of the URL, i.e. similar domain or subdomainnames were in the same cluster, but no further conclusionscouldbeextracted.

Trying tolookforphishingcategories,weperformeda term frequency analysis over the domain names of theURLs. First, we parsed the URL and obtained the domainusing tldextract Python library.[12]Then, we sorted the resultsaccording to the domain frequency. We observed that thephishingclass holds 12, 980 unique domain names,where3, 543 of themwererepeatedusing other subdomainor path.In order to identify the different categories, we performed amanual analysis of the 35 most common domains. We visitedthose domains and we evaluated the services provided on eachdomain,resultinginthesixcategoriesreflectedinTable8.

The first group is related to free subdomains, i.e. servicesthat allow phishers to host their fake websites and makethemaccessibletothepublic.Typically,theseservicesallowattackers to create a custom subdomain name to locate theirwebsite. Hence, this feature helps attackers in deceiving usersby adding popular company names or using typosquattingandcombosquattingtechniques[53].Themainadvantageofthesehostingservicesistheirprice,astheyhavefreeplans

[12]https://pypi.org/project/tldextract/

TABLE 8. Most common phishing domains on PILU-90K dataset groupedintothespottedcategories.

Type	Domain	Total domains
Free subdomain	000webhostapp	1415
	weebly	422
	umbler	398
	16mb	304
	godaddysites	197
	webcindario	134
	ddns	98
	joomla	76
	webnode	75
Cloud services	googleapis	125
	appspot	100
	sharepoint	97
	windows	90
	web	62
	xsph	59
	secureserver	55
	kl	49
Fake form	docs.google	713
	typeform	103
	forms.office	69
Standalone domains	update-information	71
	ticari	65
Social media	reddit	71
	steamcommunity	61
	twitter	61
Malware blog posts	imdb	550
	celestini	278
	fundraise	213
	ibm	195
	stackoverflow	110
	medium	107
	bandarrow	80
	toornament	65
	leetchi	55
	hatena	49

where phishers only introduce their email with no identityconfirmation. Another advantage is the free SSL certificatethey offer. However, the only disadvantage could be thelimited free resource offered by these services, in terms ofthebandwidth,storageandcomputationassets.

The second group comprehends cloud services. In thisapproach, phishers hire resources on different cloud plat-forms, such as Google or Azure, to host their phishing websitewithanSSLcertificate.Someoftheseservicesprovidefixedorrandomsubdomains, andonly thepath canbeedited.Themain disadvantages of this strategy are the price and the factthatphishershavetoprovidepaymentinformationtohiretheservice.

Fakeformsarecommonphishingmethods.Intheseattacks,phishersuseformplatformsfromGoogle,MicrosoftorTypeformtolooklegitimate,usinglogosandmessagestoencouragetheusertointroducetheircredentials.Companieshave detected these issues and advise users not to introducetheirpersonalinformation orcredentials.

Social media andmalware blogposts are reported onPhishTank to advise users from entering those sites. Thesedomainsusuallyofferfreerecentfilmsforuserstodownloadand watch. These files are detected as malware by manycommercialantivirussystems,such asAvast.

Finally,mostofthedatasetsamplesarerelatedtostandalone domains bought or compromised by phishers tohost their websites. Within this category, some domains areusedtohostdifferentcampaignsofphishingovertime.They get online on active campaigns and offline when suchcampaignshavefinishedorwhentheyhavebeenreportedtoblacklists.

VII.CONCLUSION

Phishingdetectionmechanismaimstoimprovecurrentblacklistmethods,protectingusersfrommaliciousloginforms.Ourworkprovidesanupdateddatasetl PILU-90Kforresearcherstotrainandtesttheirapproaches.Thisdataset includes legitimate login URLs which are the mostrepresentativescenarioforreal-worldphishingdetection.

We explored several URL-based detection models usingdeep learning and machine learning solutions trained withphishingandlegitimatehomeURLs.The mainadvantageofour approach is the low false-positive rate when classifyingthis type of URL. Among the different evaluated models, TF-IDF combined with N-gram and LR algorithm obtained thebest results with a 96.50% accuracy. In comparison with thecurrent state-of-the-art, reviewed in Section II, our approachpresentthreemainadvantages:

Nodependenceonexternalservices.Alimitationofthe description methods that use features such as WHOISdomain age, page ranking on Google or Alexa or onlineblacklists, is their dependence on those services. Networkslowdownsandserviceshortagescannegativelyimpactanalysis time, making real-time execution infeasible. Sincephishing websites have a short lifespan [12], low detectiontimes are required to warn users before accessing phishingwebsites.

Login website detection. Unlike other methods, whichare trained with homepage URLs as representatives of thelegitimateclass,ourmodelwastrainedwith legitimateloginwebsites. This ensures the correct classificationof thosewebsites.Therefore,ourapproachcanbeappliedtothereal-case scenario where users have to predict whether a loginformpageislegitimateorphishing.

Updated and real-world dataset. PLU-60K is focusedon using updated legitimate login URLs. As demonstrated,modelstrainedwitholddatasetswerenotabletoenduretheirperformance over time. We provide an updated phishing URLdataset for models to learn from nowadays phishing URLsandtrends,whicharecrucialforreal-worldperformance.

We demonstrated that phishing URL detection systemstrainedwithlegitimatelandpageURLsfailtoclassifylegiti-mate login URLs correctly. The best-tested models could onlyclassify 69.50% of these URLs correctly, which implies ahigh false-positive rate. For this reason, we recommend that aphishingdetector,whichintendstobeusedinarealsituation,should be trained using *legitimate login* websites (such asPLU-60K) instead of homepages. The main drawback ofusing login websites for training is that, due to the similaritybetweenphishingandlegitimatesamples,overallaccuracy is slightly reduced. The tradeoff against the state-of-the-artmethodsisstillfairduetotheirhighfalse-positiverate.

Differentcategoriesforcurrentphishingattackswereidentifiedbyusingadomainfrequencyanalysis.Whilestan-dalone and compromised domains were the most commonapproaches, free hosting services, cloud web servers andmalware blog posts represent many current phishing attacksduetotheircostandeffectivenessforphishingcampaigns.

Finally, we demonstrated that machine learning modelsusinghandcraftedURLfeaturesdecreasedtheirperformanceovertime,upto10.42%accuracyinthecaseoftheLightGBM algorithm from the year 2016 to 2020. For thisreason, machine learning methods should be trained withrecent URLs to prevent substantial ageing from the date ofitsrelease.Inthefuture,wewilladdmoreinformationaboutthesamplesintotheanalysis,suchasthesourcecodeofthewebsiteandascreenshot ofits content,whichcouldbe usefultoincreasethe phishingdetection performance.Inaddition,wewillenlargeourdataset,includingsuchinfor-mation.Finally,observingthatdeeplearningtechniquesandautomatic feature extraction obtained promising results overtraditional feature extraction, we intend to explore differentURLcodificationstoimprovedetectionperformance.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NvidiaCorporation for their kind donation of GPUs (GeForce GTXTitanXand K-40)thatwereused inthiswork.

DECLARATIONOFCOMPETINGINTEREST

The authors declare that they have no known competingfinancial interests or personal relationships that could haveappearedtoinfluencetheworkreportedinthispaper.

REFERENCES

[1] Statista.(2020).*AdoptionRateofEmergingTechnologiesinOrganizationsWorldwideasof2020*.Accessed:Sep.12,2021.[Online].Available:https://www.statista.com/statistics/661164/worldwide-cio-survey-operati%onal-priorities/
[2] R. De', N. Pandey, and A. Pal,''Impact of digital surge during COVID-19 pandemic: A viewpoint on research and practice,'' *Int. J. Inf. Manage.*,vol.55,Dec.2020,Art.no.102171.
[3] P.Patel,D.M.Sarno,J.E.Lewis,M.Shoss,M.B.Neider,andC.J.Bohil,''Perceptual representation of spam and phishing emails,'' *Appl. Cognit.Psychol.*,vol.33,no.6,pp.1296–1304,Nov.2019.
[4] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, ''Phishing attacksanddefenses,''*Int.J.Secur.Appl.*,vol.10,no.1,pp.247–256,2016.
[5] M. Hijji and G. Alam, ''A multivocalliteraturereviewongrowingsocialengineeringbasedcyber-attacks/threatsduringtheCOVID-19pandemic: Challenges and prospective solutions,''*IEEE Access*, vol. 9,pp.7152–7169,2021.
[6] A.Alzahrani,''Coronavirussocialengineeringattacks:Issuesandrecommendations,'' *Int.J.Adv.Comput.Sci.Appl.*,vol.11,no.5,pp.154–161,2020.
[7] *PhishingActivityTrendsReport3Q*,Anti-PhishingWorkingGroup,International,2017.Accessed:Sep.12,2021.
[8] *PhishingActivityTrendsReport1Q*,Anti-PhishingWorkingGroup,International,2021.Accessed:Sep.14,2021.
[9] R. Chen, J. Gaia, and H. R. Rao, ''An examination of the effect of recentphishing encounters on phishing susceptibility,'' *Decis.SupportSyst.*,vol.133,Jun.2020, Art.no.113287.
[10] *PhishingActivityTrendsReport4Q*,Anti-PhishingWorkingGroup,International,2020.Accessed:Sep.12,2021.

[11] S. Bell and P. Komisarczuk, ‘‘An analysis of phishing blacklists: Googlesafe browsing, OpenPhish, and PhishTank,’’ in *Proc. Australas. Comput.Sci. WeekMulticonf.*,Feb.2020,pp.1–11.
[12] A.Oest,Y.Safaei,P.Zhang,B.Wardman,K.Tyers,Y.Shoshitaishvili, A. Doupé,andG.-J.Ahn,‘‘Phishtime:Continuouslongitudinalmea-surement of the effectiveness of anti-phishing blacklists,’’ in *Proc. 29thUSENIXSecur.Symp.*,2020,pp.379–396.
[13] L. Li, E. Berki, M. Helenius, and S. Ovaska, ‘‘Towards a contingencyapproach with whitelist- and blacklist-based anti-phishing applications:Whatdousabilitytestsindicate?’’*BehaviourInf.Technol.*,vol.33,no.11,pp.1136–1147,Nov.2014.
[14] N. Samarasinghe and M. Mannan, ‘‘Oncloackingbehaviours of maliciouswebsites,’’*Comput.Secur.*,vol.101,pp.102–114,Feb.2021.
[15] L.Halgas,I.Agrafiotis,andJ.R.C.Nurse,‘‘Catchingthephish:Detectingphishing attacks using recurrent neural networks (RNNs),’’ in *InformationSecurity Applications* (Lecture Notes in Computer Science), vol. 11897.Cham,Switzerland:Springer,2020, pp.219–233.
[16] R. S. Rao and A. R. Pais,‘‘Jail-phish:Animprovedsearchenginebased phishing detection system,’’ *Comput. Secur.*, vol. 83, pp. 246–267,Jun.2019.
[17] Z.Dou,I.Khalil,A.Khreishah,A.Al-Fuqaha,andM.Guizani,‘‘Systematization of knowledge (SoK): A systematic review of software-based webphishing detection,’’*IEEE Commun.Surveys Tuts.*, vol. 19,no.4,pp.2797–2819,4thQuart., 2017.
[18] K. L. Chiew, E. H. Chang, C. Lin Tan, J. Abdullah, and K. S. C. Yong,‘‘Building standard offline anti-phishing dataset for benchmarking,’’ *Int.J.Eng.Technol.*,vol.7,no.4,pp.7–14,2018.
[19] H.Yuan,Z.Yang,X.Chen,Y.Li,andW.Liu, ‘‘URL2 Vec:URLmodelingwithcharacterembeddingsforfastandaccuratephishing website detection,’’ in *Proc.IEEEInt.Conf.ParallelDis-trib.Process.WithAppl.,UbiquitousComput.Commun.,BigDataCloudComput.,SocialComput.Netw.,Sustain.Comput.Commun.(ISPA/IUCC/BDCloud/SocialCom/SustainCom)*,Dec.2018,pp. 265–272.
[20] M. Sánchez-Paniagua, E. Fidalgo, V. González-Castro, and E. Alegre,‘‘Impact of current phishing strategies in machine learning models forphishingdetection,’’in*ComputationalIntelligenceinSecurityforInforma-tion Systems Conference*, vol. 12676. Cham, Switzerland: Springer, 2021,pp.87–96.
[21] O.K.Sahingoz,E.Buber,O.Demir,andB.Diri,‘‘MachinelearningbasedphishingdetectionfromURLs,’’*ExpertSyst.Appl.*,vol.117,pp.345–357,Mar.2019.
[22] Y.Cao,W.Han,andY.Le,‘‘Anti-phishingbasedonautomatedindividualwhite-list,’’ in *Proc. 4th ACM Workshop Digit. Identity Manage. (DIM)*,2008,pp.51–59.
[23] P.Prakash,M.Kumar,R.R.Kompella,andM.Gupta,‘‘PhishNet:Predictiveblacklistingtodetectphishingattacks,’’in*Proc.IEEEINFOCOM*, Mar.2010,pp.1–5.
[24] A.K.JainandB.B.Gupta,‘‘Anovel approachtoprotect againstphishingattacks at client side using auto-updated white-list,’’ *EURASIP J. Inf.Secur.*,vol.2016,no.1,pp.1–11,Dec.2016.
[25] E.Buber,B. Diri,and O. K.Sahingoz,‘‘NLP basedphishingattackdetectionfrom URLs,’’ in*Proc.Int. Conf.Intell. Syst. Design Appl.*,vol.736,2018,pp.608–618.
[26] A. K. Jain and B. B. Gupta, ‘‘PHISH-SAFE: URL features-based phishingdetectionsystemusingmachinelearning,’’in*AdvancesinIntelligentSystemsandComputing*,vol.729.Singapore:Springer,2018,pp.467–474.
[27] B. Banik and A. Sarma, ‘‘Lexical feature based feature selection andphishing URL classification using machine learning techniques,’’ in *Proc.Int.Conf.Mach.Learn.,ImageProcess.,Netw.Secur.DataSci.*,vol.1241.Singapore:Springer,2020,pp.93–105.
[28] R.S.Rao and A. R. Pais, ‘‘Detection of phishing websites using anefficient feature-based machine learning framework,’’ *Neural Comput.Appl.*,vol.31,no.8,pp. 3851–3873,Aug.2019.
[29] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, ‘‘A stacking model usingURLandHTMLfeaturesforphishingwebpagedetection,’’*FutureGener.Comput.Syst.*,vol.94,pp. 27–39,May2019.
[30] Y.Zhang,J.I.Hong,andL.F.Cranor,‘‘Cantina:Acontent-basedapproachtodetectingphishingwebsites,’’in*Proc.16thInt.Conf.WorldWideWeb(WWW)*,2007,pp.639–648.
[31] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, ‘‘CANTINA+: A feature-rich machine learning framework for detecting phishing web sites,’’ *ACMTrans.Inf.Syst.Secur.*,vol.14,no.2,pp.1–28,Sep.2011.
[32] M.Moghimiand A. Y. Varjani,‘‘Newrule-based phishingdetectionmethod,’’*ExpertSyst.Appl.*,vol.53,pp.231–242,Jul.2016.

[33] V.I.Levenshtein,''Binarycodescapableofcorrectingdeletions,insertions, and reversals,'' *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710,1966.
[34] M.A.Adebowale,K.T.Lwin,E.Sánchez,andM.A.Hossain,''Intelligentweb-phishing detectionand protectionscheme using integratedfeaturesof images, frames and text,'' *Expert Syst. Appl.*, vol. 115, pp. 300–313,Jan.2019.
[35] M. Rami, M. Lee, and T. Fadi, ''UCI machine learning repository,'' Univ.Huddersfield,Huddersfield, U.K.,Tech. Rep.,2015.
[36] L.Yang,J.Zhang,X.Wang,Z.Li,Z.Li,andY.He,''AnimprovedELM-based and data preprocessing integrated approach for phishing detectionconsideringcomprehensivefeatures,''*ExpertSyst.Appl.*,vol.165,Mar.2021,Art.no.113863.
[37] F.Sadique,R.Kaul,S.Badsha,andS.Sengupta,''Anautomatedframework for real-time phishing URL detection,'' in *Proc. 10th Annu.Comput.Commun. WorkshopConf.(CCWC)*, Jan.2020, pp. 335–341.
[38] T.Mikolov,I.Sutskever,K.Chen,G.S.Corrado,andJ.Dean,''Distributedrepresentationsofwordsandphrasesandtheircompositionality,''in*Proc.Adv. Neural Inf.Process.Syst.*,vol.26,no.4,Dec. 2013,pp.3111–3119.
[39] M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, ''Efficient deeplearningtechniquesforthedetectionofphishingwebsites,"*Sādhanā*, vol.45,no.1,pp.1–18,Dec.2020.
[40] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. Niyigena, ''An effectivephishing detection model based on character level convolutional neuralnetworkfromURL,''*Electronics*,vol.9, no.9,pp.1–24,2020.
[41] A. Al-Alyan andS. Al-Ahmadi,''Robusturl phishingdetection basedondeeplearning,''*KSIITrans.InternetInf.Syst.*,vol.14,no.7,pp.2752–2768,2020.
[42] J. Zhao, N. Wang, Q. Ma, and Z. Cheng, ''Classifying malicious URLsusing gated recurrent neural networks,'' in *Proc. Int. Conf. Innov. MobileInternetServicesUbiquitousComput.*,vol. 773, 2019,pp.385–394.
[43] X.Zhang,J.Zhao,andY.Lecun,''Character-levelconvolutionalnetworksfortextclassification,''in*Proc.Adv.Neural Inf.Process.Syst.*,Jan.2015,pp.649–657.
[44] Y. Kim, ''Convolutional neural networks for sentence classification,'' in*Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014,pp.1746–1751.
[45] K.L.Chiew,C.L.Tan,K.Wong,K.S.C.Yong,andW.K.Tiong,''Anewhybrid ensemble feature selection framework for machine learning-basedphishingdetectionsystem,'' *Inf.Sci.*,vol.484, pp.153–166,May2019.
[46] R.S.Rao,T.Vaishnavi,andA.R.Pais,''CatchPhish:Detectionofphishingwebsites by inspecting URLs,'' *J. Ambient Intell. Humanized Comput.*,vol.11,no.2,pp.813–825,Feb.2020.
[47] G.Ke,Q.Meng,T.Finley,T.Wang,W.Chen,W.Ma,Q.Ye,andT.-Y.Liu, ''LightGBM: A highly efficient gradient boosting decision tree,'' in *Proc.Adv.NeuralInf.Process. Syst.*,Dec.2017, pp.3147–3155.
[48] T. Chenand C. Guestrin,''XGBoost: Ascalable tree boostingsystem,''in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vols. 13–17,2016,pp.785–794.
[49] A.Aizawa,''Aninformation-theoreticperspectiveoftf–idfmeasures,''*Inf.Process.Manage.*,vol.39,no.1,pp. 45–65,2003.
[50] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and R. Aláiz-Rodríguez, ''Filename classification approach to identify child sexual abuse,'' in *Proc. 9thInt.Conf.PatternRecognit.Appl.Methods*,2020, pp.228–234.
[51] C. Peersman, C. Schulze, A. Rashid, M. Brennan, and C. Fischer, ''ICOP:LiveforensicstorevealpreviouslyunknowncriminalmediaonP2Pnetworks,''*Digit.Invest.*,vol.18,pp.50–64,Sep.2016.
[52] W.H.E.DayandH.Edelsbrunner,''Efficientalgorithmsforagglomerativehierarchical clustering methods,'' *J. Classification*, vol. 1, no. 1, pp. 7–24,1984.
[53] J. Spaulding, S. Upadhyaya, and A. Mohaisen, ''The landscape of domainname typosquatting: Techniques and countermeasures,'' in *Proc. 11th Int.Conf.Availability,Rel.Secur. (ARES)*,Aug.2016,pp.284–289.

MANUEL SÁNCHEZ-PANIAGUA received theB.Sc. degree in computer science and the M.Sc.degree in cybersecurity research from the Uni-versity of León, in 2019 and 2021, respectively.He is currently a Researcher at the GVIS Group,University of León. His research interests includeanti-phishing solutions, fraud e-commerce websitedetection, web scrapping, cyber threat intelligence,andmachinelearning.

EDUARDOFIDALGOFERNÁNDEZreceived theM.Sc. degree inindustrial engineeringandthePh.D. ityofLeón,in 2008 and 2015, respectively. He is currentlytheCoordinatoroftheGroupforVisionandIntellig (GVIS), whose objective istheresearchanddevelopmentofsolutionstoproblemsrela NCIBE(https://www.incibe.es/en),byusingartificialin telligence.Hiscurrentresearchinterestsinclude naturallanguageprocessing,computervision,andmachineanddeeplearning.

ENRIQUE ALEGRE received the M.Sc. degreein elec n the University ofCantabria, in 1994, and the Ph.D. degree from theUniversityofLeón,Spain,in2000.Heiscur Research Group for Vision andIntelligent Systems (GVIS) and an Associate Pro the Department of Electrical, SystemsandAutomationEngineering,UniversityofLeón. ludecomputervisionandpatternrecognitionapplication stomedicaland industrial problems and more recently, machine learning and computer visionapplicat dprevention.

WESAM AL-NABKI received the M.Sc. ics and computer vision from the UniversityofBurgundy,in2013,andthePh.D.degreefromt 2019.HeiscurrentlyaResearcherattheUniversidaddeLeón.His researchinterests include natural ing, machine learning, and deep learningandtheirapplicationincomputerforensics.

VÍCTOR GONZÁLEZ-CASTRO received the B.S. in computer science from theUniversityof León,Spain,in2006and2011,respectively. He is curr Profes-sor with the Department of Electrical, SystemsandAutomationEngineering,UniversityofLeón. research interests include computervision,machinelearning,anddeeplearning,applie uality assessment, andcyber-security.