IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 November 2022, accepted 14 November 2022, date of publication 17 November 2022,
date of current version 1 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3223111

RESEARCH ARTICLE

Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites With Machine Learning Methods

ILKER KARA[1], (Member, IEEE), MURATHAN OK[2], (Member, IEEE), AND AHMET OZADAY[2], (Member, IEEE)

[1]Department of Medical Services and Techniques, Eldivan Medical Services Vocational School, Çankırı Karatekin University, 18100 Cankiri, Turkey
[2]Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey

Corresponding author: Ilker Kara (karaikab@gmail.com)

ABSTRACT Along with the means of communication, it has also prompted the birth of more harmful, and challenging websites in the device of information systems, and electronics. According to current estimates, you can deal with a huge budget to arrange detailed information on attackers. Furthermore, only those that are handled similarly to HTML, DOM, and URL based features in the literature are easily manipulated by attackers. To respond to these attacks, we propose a new method that detects phishing websites by categorizing the Internet URL, and domain names of websites with six different classifier algorithms according to eleven predetermined features. For this method, we created a previously unused list. The list was obtained by analyzing an index created with information obtained from internationally reputable intelligence services, and entire organizations. The proposed method simplifies the process of feature extraction, and reduces processing overhead while going beyond analyzing on HTML, DOM, and URL based features by considering URLs, and domain names. To illustrate the highest accuracy rate among six different classification results, we preferred to use the Random Forest algorithm. In this study, we use a dataset with 32,928 data in which 12,134 data without phishing websites, and 20,614 data with phishing websites to be labeled according to eleven predetermined features. Our experimental results show that phishing websites can be detected with as much as 98.90% accuracy with our proposed method. As a result, it has been demonstrated that RF descriptors with SVM representation can be utilized to accurately mark phishing web pages. In addition, characteristic updates can be followed with a continuously updated source.

INDEX TERMS Cyber-security, website features, phishing, feature extraction, machine learning.

I. INTRODUCTION

With the advancements of e-commerce, online services, and social media fraudsters have embraced a new generation attack type known as "Phishing? to have unlawful gain. According to recent studies, this technique is used for capturing sensitive data such as credit card information, personal information, e-mail accounts, and social network account information [1]. Google had registered 2,145,013 phishing sites as of January 17, 2021 [2]. Compared on 2020, this rate climbed by 27%. Furthermore, IBM said that phishing is the second most expensive source of data breaches. With a breach triggered by the attack costing firms an average of $4.65 million [3].

Moreover, despite significant recent developments in phishing website detection, and prevention technologies this problem continues to generate massive losses each year [4], [9]. In general, strategies for countering phishing websites can be divided into two categories. The first is a list of suspect websites with URL blacklists, hosting providers, antivirus software providers, or other authorized bodies [10], [15]. The Uniform Resource Locator (URL)

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

blacklists method asks blacklists every time the browser loads the page to see if the currently visited URL is on this list. If your URL appears on any blacklists, you will be notified, and necessary action will be taken. Otherwise, this page is considered to be trustworthy. Blacklists can be stored locally on the client or a central server. On the other hand, whitelists list safe websites. The primary concept behind this strategy is to develop a list of safe websites, and block them when they try to access a website that is not on this list.

Second category is, rather than looking at a list, employing machine learning algorithms on websites [16], [17]. These components include URLs, Hypertext Markup Language (HTML) code, and page content to identify phishing websites.

Several features for categorization, and tagging have been presented in the literature, although some of these features are not distinct enough. Because attackers modify URL, and HTML, and it generates inconsistency between real, and phishing websites.

The limited number of features in categorizing, and tagging URLs when classifying them is one of the primary issues of machine learning-based solutions. Furthermore, there are extremely few publicly available training datasets including phishing URLs.

Taking all of this into consideration, we present a vision for detecting phishing websites based on machine learning techniques. We will examine six different machine learning approaches about phishing detection. Furthermore, we will analyze, and classify website internet URLs, and domain names based on eleven different features. We have assumed, and tested our hypothesis, and found out the structure, and organization of phishing websites are more distinguishing from non-phishing websites.

This study mainly presents the two contributions listed below:

- This study will create a phishing domain name dataset using an up-to-date intelligence database. This data set can also be used for future research. The difference from the existing data sets is that the essential list it created included data from national channels, and was developed by security organization experts.
- On the provided data set, we have tested the URLs, and domain names of the websites by classifying them according to eleven determined features, and using six different machine learning algorithms. We tried to determine which machine learning algorithm would get more accurate results with the available data. The algorithm was compared with the data content.

This paper is organized as follows: In Chapter 2, we looked at several relevant research. Chapter three discusses the proposed dataset features, and briefly explains Dataset Preparation. The 4th chapter includes the experimental results as well as a comparison of the classification work using six different machine learning methods. Discussions are offered in the 5th chapter, while Research Challenges are addressed in the 6th. Finally, we wrapped up the research, and discussed potential future directions.

II. RELATED WORK

Machine learning can detect phishing websites by classifying, and labeling the URLs, and domain names of websites based on the identified features. It is possible to extract two features; host-based, and lexical features. Host-based features indicate the location of the website, who manages it, and where the site was loaded from. The text properties of the URL are described by lexical features. URLs can assess the validity of a website based on its file structure as well as components such as protocol, and hostname.

To identify phishing URLs, several machine learning approaches have been published in the literature. Some of the researches classified according to the determined features of URLs, and domain names are highlighted in this section.

Ludl et al. used the J48 decision tree algorithm on 18 features to classify phishing websites based only on HTML, and URL information [18]. The study's dataset contains 4,149 safe pages, and 680 phishing pages. According to the results of the test, it has an accuracy of 83.09%. Approaches relying entirely on HTML DOM, and URL-based features. On the other hand, they've had limited success. Because attackers can manipulate the HTML DOM, and URL.

Kulkarni et al. suggested a machine learning method for detecting phishing attacks. The suggested method makes use of a dataset that contains 1,353 safe website URLs that may be classified as phishing sites [19]. Similarly, the decision tree, Nave Bayesian classifier, support vector machine (SVM), and neural network were used as classifiers in our study. According to the findings of the study, the classifiers categorized real-world websites with 90% accuracy.

Fette et al. created a technique named PILFER to categorize URLs for identifying phishing attacks in their study [20]. They released ten features that were created specifically to expose misleading methods used to commit fraud. The study's dataset contains around 860 phishing emails, and 6,950 non-phishing emails. The classifier in the application was a Support Vector Machine (SVM). They used 10-fold cross-validation to train, and evaluate the classifier, achieving 92% accuracy. Because of the success rate of the suggested PILFER approach, they claimed that it was superior to the SpamAssassin filter, a commonly used spam filter. It is contentious due to the study's restricted data set, and low success rate.

Chiew et al. used data from the Machine Learning Repository (UCI) to detect phishing websites using several machine learning methods [21]. 5,000 URLs from the PhishTank, OpenPhish, Alexa, and Common Crawl archives were used in the study, and The Random Forest algorithm has shown 94.6% accuracy.

Similarly, Parekh et al. used the Random Forest method to detect phishing attacks based on URL identification [22]. The Random Forest method is divided into three stages: parsing, heuristic data classification, and performance analysis. In the

study, eight characteristics were used to parse, and the Random forest algorithm used in the study that has provided 95% accuracy.

Zhang et al. presented a methodology for identifying Chinese phishing e-business website URLs, and website content in another study [23]. The data set of 3,000 website samples was used in the model. In the study, four distinct approaches were used: Sequential Minimal Optimization, Logistic Regression, the Naive Bayes classifier, and Random Forests. The Sequential minimum optimization (SMO) algorithm technique has proven to be the most accurate, with a 95.83% accuracy rate. It is unclear how this method will work if it is restricted to only Chinese or non-Chinese phishing e-business websites. Xiang et al. employed a pre-trained model, CANTINA, in a machine learning framework to classify, and apply an ID to detect phishing websites using URL, HTML DOM, and other data [24]. As a dataset, it used 8118 phishing, and 4,883 legal websites. The CANTINA approach produced over 87% TP (True Positive Rate) on unique sites, with over 95% TP accuracy.

Although this approach is interesting, attackers have underlined that they can simply overcome it by generating phishing web pages completely of images, and analyze them using the CANTINA algorithm suggested in the paper.

Garera et al. classified phishing URLs into four types based [25]. They used a dataset of 2,508 URLs for the study. The study has shown 95% accuracy. The success percentage is debatable since attackers may easily modify the URL.

III. MATERIALS AND METHODS

In this part, we will first introduce the dataset that we have gathered. Then the proposed method is defined. After that the experimental results of the classification algorithms, and preference classification algorithms used are presented.

A. DATASET

It is widely known that a well-organized, and accurate dataset is critical for data-driven studies. When the studies of an approach to phishing attacks that uses machine learning method are reviewed in the literature, it can be seen that there is limited number of data sets; Fette et al., 860 phishing emails, and 6950 non-phishing 7810 [20], Zhang et al., 3,000 Phishing websites [13], Xiang et al., 8,118 phishing, and 4,883 legitimate total 13,001 web pages [24].

However, we discovered two major problems in these datasets. The first is that it lacks sufficient information for feature classification. The second issue is that these databases need more real-world examples. The lists reached in the research are outdated, and handled more hypothetically. Also, the lists in previous surveys were not collected by organizations with access to different sources. In particular, it is aimed to use daily, and various institution/organization data derived from other target kits, and determined by experts as a result of targeted attacks. In this sense, it is obvious that the datasets mentioned above might be useful for specialized inference tasks. To be more specific, we intended to develop a model that could evaluate numerous classification features on real-world samples to address the issue that phishing website characteristics are vulnerable to manipulation by attackers. As a result, the current datasets were insufficient for our study.

To create a suitable dataset, we used the open-source data from the TR-CERT official website which was formed to develop, and share ways to prevent or eliminate the effects of potential cyber-attacks, and events in Turkey. The organization in question has a specialized team, and method for collecting a large number of phishing website samples, which they shared with us. Between September 23, 2020, and October 15, 2021, 50,491 malicious links were detected. Based on this data, it was determined that 34,134 records belonged to the phishing category, and the records were separated. During the feature collecting phase for the data set, all attributes of 20,614 phishing data out of 34,134 records were obtained. In addition, the data collection now includes information from 12,314 non-phishing web pages.

If a feature is evaluated in the phishing category, it is marked with "1? otherwise it is marked with "−1?. Data suspected of phishing is indicated by a "0?. Before starting the machine learning studies, there are 32,928 entries in the dataset.

The dataset has been carefully selected to ensure that it is representative of both phishing, and legal sites. It was attempted to be produced using a well-balanced data set. Table 1 shows the distribution of values with or without phishing domain in the data analysis set.

TABLE 1. Data Distribution.

Total Data	Phishing Data	Non-Phishing Data	Data Rate
32,928	20,614	12,314	0.626

The data is decomposed for training, and testing, and algorithm learning.

TABLE 2. Distribution of training and test data.

Total Data	Phishing Data	Non-Phishing Data	Data Rate
Training-23049	14,429	8,620	0.626
Test-9879	6,185	3,694	0.626

The absence of data in the data that correlate with each other will significantly affect the machine learning results. If the correlation results of two values with each other are too high, this will cause deep learning, and increase the effect on the results. In figure 1, a matching analysis was performed for these correlation results.

To determine the data close to a value that would affect the learning algorithm the definitions have been turned into graphs based on their values. There were no unacceptable

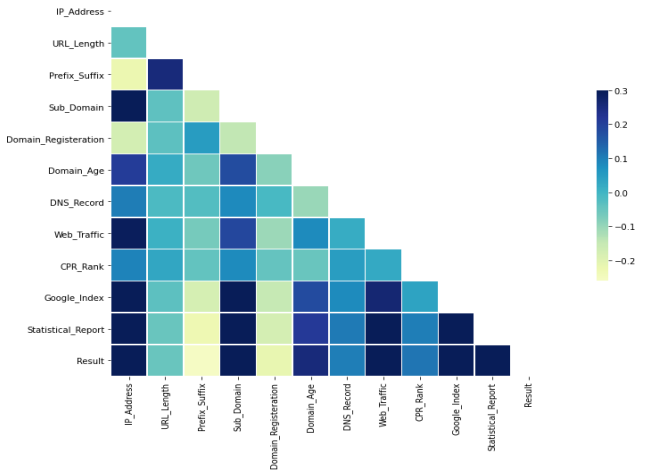

FIGURE 1. Example correlation table.

value variances in the transformed data. The data did not require any manipulation. One of the data distribution examples shows in the second figure.

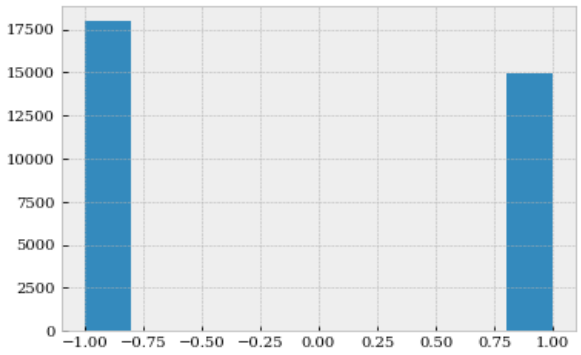

FIGURE 2. Example description value distribution (google indexing).

The following conclusions may be drawn from the Data Distribution results in Tables 1, 2, and the classification algorithms used in Table 3:

TABLE 3. Distribution of training data.

Algorithm	Cross Validation Score	Standard Deviation
LR	0.9594778862376477	0.00340907995
LDA	0.9446397663593636	0.0048696648
KNN	0.9574820739937333	0.0045236736
DT	0.9589571282236683	0.0026385757
SVM	0.9607339645095203	0.0038771969
RF	0.9610849602313811	0.0030192699

Algorithms used Logistic Regression (LR), Linear Discriminant Analysis (LDA), Nearest Neighbor (KNN), Black Tree (DT), Support Vector Machines (SVM), and Random Forest (RF). The RF algorithm produced the maximum accuracy. However, the SVM classification's performance may also be described as competitive.

B. PREPARATION OF DATASET

During a phishing attack to steal corporate account information, we focus on detecting the source page used by the malicious cybercriminal using incoming email, notification, SMS, or a different communication channel.

When current research on detecting phishing sites are evaluated, it is clear that URL, and query-based data are regularly used. Furthermore, in most cases, the combination of URL, and Query-based data is immune to obfuscation, and manipulation techniques. By definition, URL analysis data contains IP address, Sub-Domain, Prefix-Suffix, and URL length. Google Search, Web Traffic, Check Page Rank, Whois Query, and Statistics Report are examples of query-based data (See Figure 3).

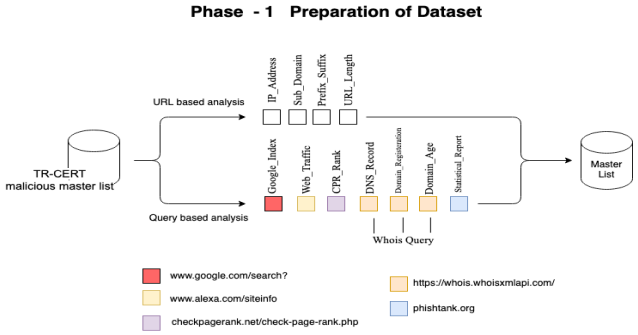

FIGURE 3. Example description value distribution (google indexing).

Figure 3 shows the process of preparing a dataset. First, four indicators are extracted from URL-based static analyses of the TR-CERT harmful list. Is there a known IP address of URL? Is the number of Subdomains in the second step larger than two? Is there a "-" symbol in the third step? Finally, it was determined whether the URL length exceeded 30 characters. Secondly, the study of creating query-based dynamic analyses was carried out in four steps. Is the suspicious website indexed

first in Google at this point? Second, the website's web traffic is retrieved using an Alexa query. Third, the website's CPR score is obtained, and lastly, the domain's Whois query is performed. With this query, we may find out when the DNS record was created, as well as its current status, and record information. In addition, the suspicious domain's similarity to the domain names in the lists of the PhishTank dataset was evaluated. If we look more closely at the steps of analysis;

C. USING LONG URLS TO HIDE SUSPICIOUS PART

Using an IP address as an alternative to the domain name in the URL, such as "http://88.18.221.19/phish.html,? indicates that it is attempting to steal personal information. The IP address may also be translated to hexadecimal at times, as indicated in the link below. "http://0x58.0x

CC.0xCA.0x62.2/phish.html?. If the IP address appears in the URL text, the page is regarded as phishing.

D. USING LONG URLS TO HIDE SUSPICIOUS PART

Phishers can use long URLs to hide the suspicious part in the address bar. For example, http://bimcelltr.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/? If the length of the URL is greater than or equal to 30 characters, the URL is classified as a phishing page.

E. PRESENCE OF PREFIX OR SUFFIX WITH "-? SYMBOL IN THE DOMAIN NAME

The dash symbol is rarely used in legitimate URLs. Phishers tend to add (-) separated prefixes or suffixes to a domain name to make users feel like they're dealing with a legitimate web page. For example, http://www.account-corporate-name.com/. Such uses are common in phishing attacks.

F. MULTIPLE SUBDOMAINS

Domain names made up of three parts: a top-level domain (sometimes called an extension or domain suffix), a domain name (or IP address), and an optional subdomain. Phishers try to deceive users by including the company name, and familiar login information in the subdomains. Domains will contain several subdomains if the number of dots is more significant than two; it is classified as a phishing website. For example, http://login.companyname.xxx.com/.

G. DOMAIN NAME REGISTRATION PERIOD

Even though a phishing website has only just begun to appear online, it has been discovered that trusted domain names were often created several years in advance. It was considered phishing if the domain name was registered in less than a year.

H. DOMAIN NAME CREATION DATE

The dates on which domain names were registered are stored in Whois databases. Domain names registered for phishing purposes may only be active for a brief period. In this dataset, if the domain name registration is less than six months old, the page is classified as phishing.

I. DNS RECORD

If the relevant web page is not recognized by the Whois database or if no records have been created for the domain name, i.e. the DNS record is empty or cannot be found, the website is classified as "Phishing?, otherwise as "Legitimate?.

J. WEBSITE TRAFFIC

The popularity of a website may be determined by counting the number of visitors, and the pages they view.

Websites designed for phishing may not be identified by databases such as Alexa since they are only available for a short period. It is considered phishing if the domain of the website has no history of traffic or is not recognized by the Alexa database.

K. CPR POINT

It can be measured in what order, and how important a web page is on the internet. It is a measure of how well the technical aspects of a web page will result in higher results as compared to search engine optimizations, as well as how well it will reach organic visitors. It has been discovered that websites designed for phishing purposes might have a CPR of "0.2?.

L. GOOGLE INDEX

When Google indexes a website, it appears in search results. Because phishing web pages are only available for a limited period, many phishing websites are not in the Google index. Pages that are not indexed by Google are considered phishing.

M. FEATURE-BASED ON STATISTICAL REPORTS

PhishTank, and StopBadware companies working specifically on phishing, and e-mail security generate a large number of statistical reports about phishing websites at a certain time. In these studies, phishing is defined as a web website that uses the names of organizations working in this sector (post/cargo companies, chain marketplaces, internet service providers etc.) in their domain names. In the study, we use similar names obtained from the TR-CERT list.

IV. USE OF MACHINE LEARNING ALGORITHMS

The methodology, and system of machine learning algorithms used in the study are detailed in this part. The total project is divided into two sections. The first step is to fill the master list with non-phishing URLs. With a candlestick analysis, accurate result analyzes, and near-far deviation inferences were made. In the fourth figure Logistic Regression (LR), Linear Discriminant Analysis (LDA), Nearest Neighbor (KNN), Black Tree (DT), Support Vector Machines (SVM), and Random Forest (RF) methods were used to compare the obtained learning dataset. Because phishing, and legal sites are separated in the dataset, the term "accuracy? is used while categorizing. Cross-validation was performed on the

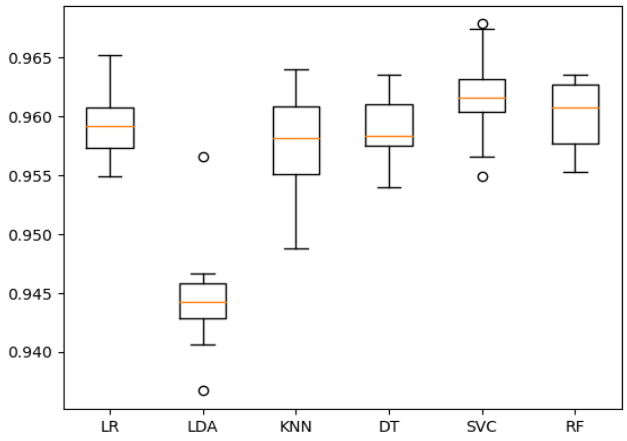

FIGURE 4. Comparison of algorithms.

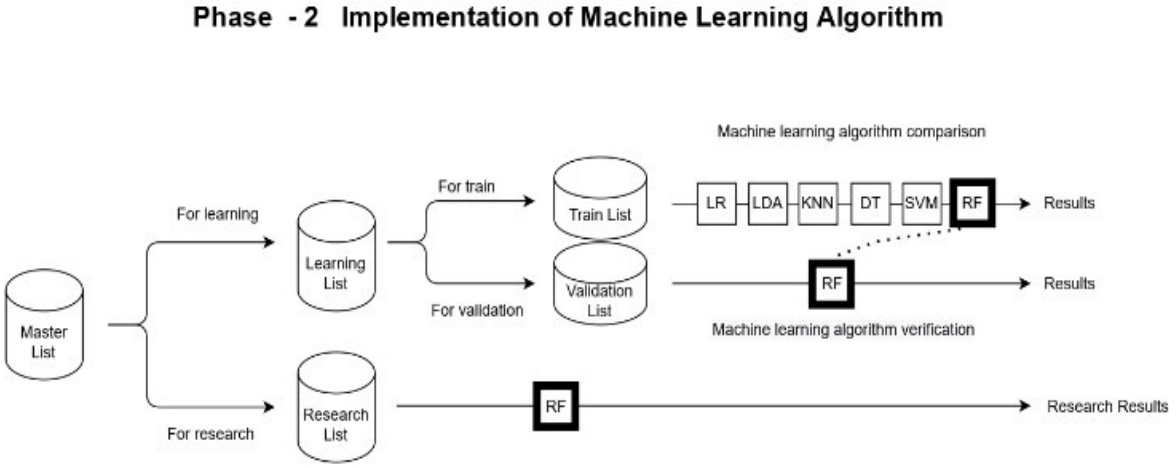

FIGURE 5. Implementation of the machine learning algorithm.

algorithms, and the method with the best success rate was chosen.

Following this stage, the master list is separated into learning, and research sections. The produced learning list is then split into two sections for training, and control. Machine learning algorithms are used to test the generated training list. By comparing the test results with the checklist results, the algorithm with the highest score is determined by comparing the outcomes obtained with the highest score. Figure 5 shows the graph of this distribution. Learning, and research lists were evaluated in two stages. The breakdowns are shown as follows, following the flow.

It was ensured that the learning phase of the algorithm was completed by using the learning, and testing data produced from data set. The disaggregated learning dataset was used at this level. Then, to verify that the learning occurred, results were obtained using test data that had not previously been entered into the algorithm. The key goal is to predict how the algorithm will perform given data that it has never seen before.

A validation phase was carried out with the learning data as the first step. Learning is provided after separating a validation set from the learning dataset. There were 18,439 records used for learning, and 6,610 records used for verification.

With this model, it was concluded that 0.989 successful predictions were made. In the validation set, 2,853 phishing, and 1,688 legitimate websites were successfully detected.

When decision trees, and support vector machine methods were compared, it was discovered that while high performance was reached at some intervals, the random forest approach provided more steady results in average performance, and the average performance was greater.

As a result of the findings, it was decided to continue the machine learning research using the random forest method.

V. RANDOM FOREST ALGORITHM

The random forest technique is capable of both classification, and regression. The random forest algorithm combines two or more algorithms to arrive at the prediction result. It generates a decision tree from a randomly chosen point. This procedure is repeated for ''N? trees. The number of trees in the research was set at 100 [27]. The depth of the tree, and the greatest leaf node value is unconstrained [28].

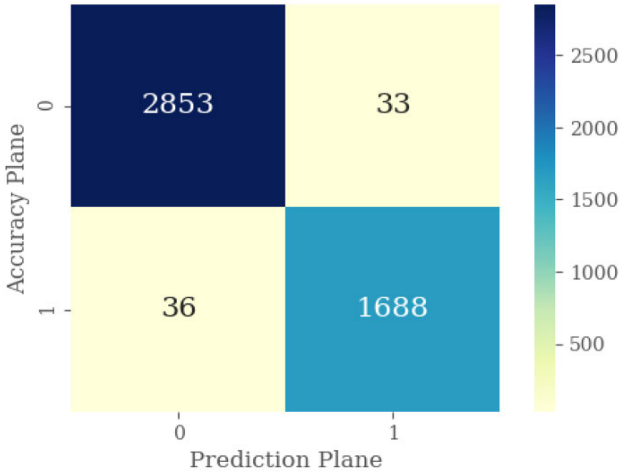

FIGURE 6. Error matrix (value).

In the validation set, 98% success was achieved in phishing pages, and 97% in the detection of legitimate internet pages.

The role definitions play in predicting the feature class indicates the importance of that definition. The feature importance in the dataset, and random forest model is shared in the code output below. The comparison features, including the priority outputs in terms of results, are given in Table 4. Finally, the effects of these properties on the detection values are listed.

We have also compared our best results with four other studies, such as (Kulkarni et al., Fette et al., Parekh et al., Zhang et al.) (Table 5). For example, the study of (Kulkarni et al., Fette et al.) employed the use of SVM machine learning methods, whereas (Parekh et al.) used the

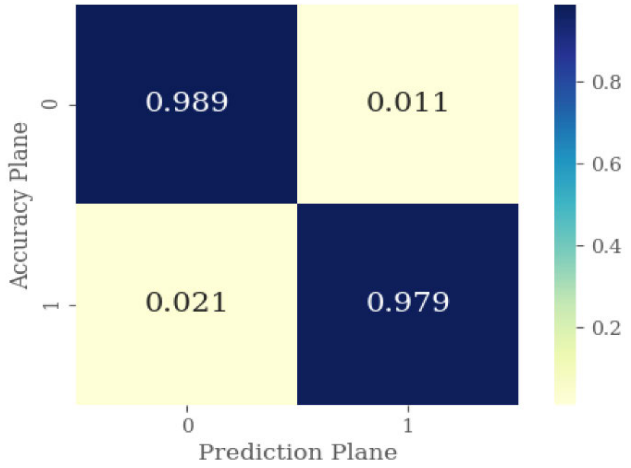

FIGURE 7. Error matrix (ratio).

TABLE 4. Feature Significance.

Label	Significance
Statistical_Report	0.422398
IP_Address	0.236978
Google_Index	0.156484
Sub_Domain	0.100489
Web_Traffic	0.040633
Domain_Age	0.010133
Prefix_Suffix	0.010081
Domain_Registeration	0.007630
URL_Length	0.005531
DNS_Record	0.005241
CPR_Rank	0.004401

Random Forest method and used the (Zhang et al.) SMO machine learning methods.

TABLE 5. Comparison of our best classifier with other references works.

Study	Accuracy
Kulkarni et al., 2019	90%
Fette et al., 2007	92%
Parekh et al., 2018	95%
Zhang et al., 2014	95.83%
Our best	98.90%

VI. DISCUSSION

The threat of phishing attacks is rapidly growing, and causes great harm by targeting unconscious users [29], [30], [31], [32], [33], [34]. Thus, in this study, we propose a new technique to detect, and determine phishing websites by using URL, and domain names specified with eleven features.

The suggested model's applicability was evaluated by examining its capabilities in determining legal, and phishing website detection using a random forest model. Table 4 displays the performance outcomes. Based on Table 4, it is concluded that the suggested model outperforms other traditional models. This great performance demonstrates the suggested model's applicability in making judgments regarding phishing, and legal websites in a short amount of time, and with high capabilities. This approach can be used as a novel ransomware detection approach to protecting systems against new digital threats.

There are some limits to the success of our work. To begin with, the quality, and quantity of the data set used to have a direct impact on its success. As a result, training the model with a smaller dataset may not produce effective results for feature classification with fewer features. Another issue is that the attacker is continually improving his approaches to avoid detection. This circumstance has a direct impact on the success of the developed models, and makes detecting phishing attacks challenging.

VII. RESEARCH CHALLENGES

Based on a review of the literature on phishing websites, we highlight key research challenges in this section. The research challenges identified include user awareness, a lack of Open

Access phishing website samples, and insufficient detection algorithms.

Awareness among users: One of the most difficult issues in researching the impact of phishing websites is raising user awareness. The majority of security flaws are the result of user mistakes [35]. Despite all of the safety precautions, no solution guarantees complete protection from phishing websites [36]. While current antivirus, and spam filters are beneficial, there is always the risk that attackers will compromise the target system via a variety of approaches (such as fake e-mail, and phishing messages) [37], [42]. This threat can be avoided if users are aware of it. Although there is a great amount of research (workshops, programs, or educational internet pages) to support this, it is vital to expand these measures.

Lack of Open Access phishing website examples: For phishing website identification, analysis, and blocking research, up-to-date data sets are required. The researchers' studies with current data sets will lead to a better understanding of the attackers' techniques, and the development of a solution to this problem. We give an up-to-date data collection that we employed in our study for this aim. http://ilkerkara.karatekin.edu.tr/RequestDataset.html is the access link. These datasets, however, must be created. As a result, worldwide collaboration in the fight against phishing websites is required.

Inadequate detection techniques: Instead of using preset blacklists or looking at a list, machine learning algorithms are used to detect phishing websites based on the attributes of the websites. The attacker uses techniques to escape detection in both approaches, and these tactics are continually changing. This situation limits the effectiveness of the methods used. As a result, further research is required to build phishing website detection methods, and expand their application.

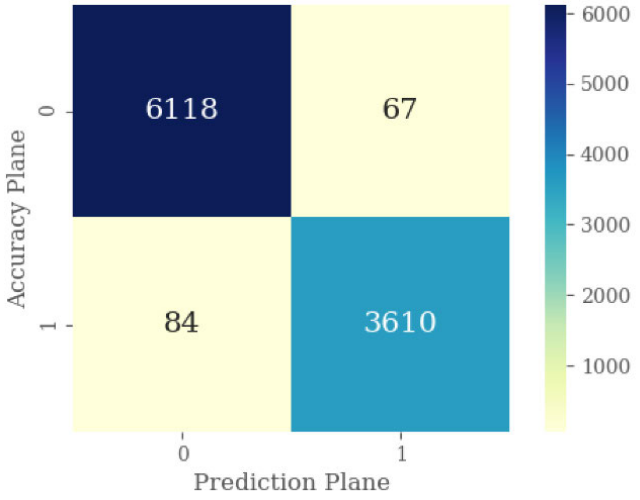

FIGURE 8. Error Matrix (Value) - Test Data.

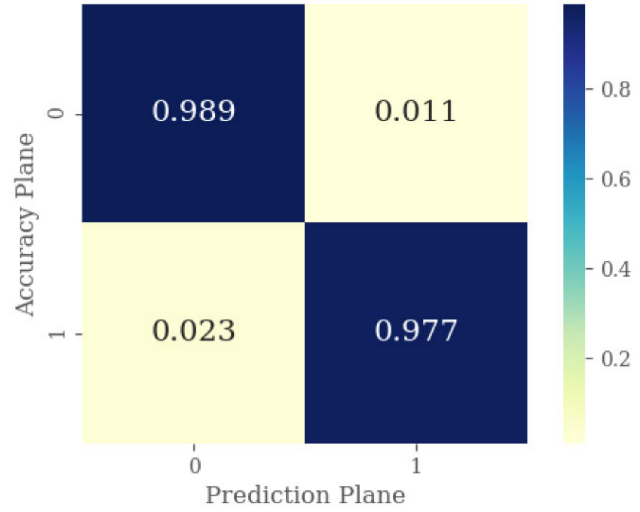

FIGURE 9. Matrix (Ratio) - Test Data.

VIII. CONCLUSION

The data from internet pages were used in the study to try to guess whether the page was prepared for phishing. In the study, the random forest model was chosen. The model was applied to a successful data set, and the prediction rates produced acceptable results. On untaught data, the model likewise has a high prediction rate.

The model's success may be defined as a minimum loss in data conversion, selecting the appropriate machine learning technique, and consistency of definitions in the data set. The final estimation values obtained from the disaggregated test data are shown below.

In the test dataset, 6,118 phishing, and 3,610 legitimate websites were successfully detected out of 9,879 records. 98% success was achieved in phishing pages, and 97% success in detecting legitimate internet pages. The total correct prediction rate was found to be 98%.

In the information security approach, local, and global attack factors, and characteristics may differ. This situation is especially exploited by attackers who are aware of their regional usage habits. In particular, targeted attacks on a certain area may have different characteristics from the current attack characteristics. Here, inferences were made from a national list by evaluating the attributes in global attack vectors. The usability of the methods was evaluated in the inferences made.

In addition, a national list has been provided that local researchers can use in their future studies.

IX. FUTURE WORK

The model's success rate has shown that it may be used in a mechanism to block links in e-mails sent to users after identifying phishing or to notify about the website [14]. With the machine learning system acting as a detection mechanism, it will be able to play a significant part in the infrastructure required for users to access secure connections. With the help of developed code, the link in the e-mail that reaches the user will be converted to the data set needs used in this model, and the model will be able to determine based on the incoming data, and manage the user's communication with the relevant link.

It can be used as a successful barrier mechanism in phishing attacks, which is one of the most important elements for user security.

The national list used for research adds about 2000 new phishing web page registrations daily. These data, which can be dynamically incorporated into learning, will also enable tracking of whether existing qualifications are still available. A decrease in success will indicate that the attackers' approach to using attributes has changed.

DECLARATION OF COMPETING INTEREST

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

REFERENCES

[1] I. Vayansky and S. Kumar, "Phishing: Challenges, and solutions," *Comput. Fraud, Secur.*, vol. 2018, pp. 15–20, Jan. 2018.

[2] (2022). *Tessian*. [Online]. Available: https://www.tessian.com/blog/phishing-statistics-2020/

[3] (2021). *IBM*. [Online]. Available: https://www.ibm.com/security/data-breach

[4] D.-J. Liu, G.-G. Geng, X.-B. Jin, and W. Wang, "An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment," *Comput. Secur.*, vol. 110, Nov. 2021, Art. no. 102421.

[5] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Comput., Appl.*, vol. 31, pp. 3851–3873, Aug. 2019.

[6] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in *Proc. 4th ACM Workshop Digit. Identity Manage. (DIM)*, 2008, pp. 51–60.

[7] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Utilisation of website logo for phishing detection," *Comput. Secur.*, vol. 54, pp. 16–26, Oct. 2015.

[8] W. Zhang, H. Lu, B. Xu, and H. Yang, "Web phishing detection based on page spatial layout similarity," *Informatica*, vol. 37, no. 3, pp. 231–244, 2013.
[9] L. J. P. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[10] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, p. 2019, 2019.
[11] A. Valjarevic and H. S. Venter, "Introduction of concurrent processes into the digital forensic investigation process," *Austral. J. Forensic Sci.*, vol. 48, no. 3, pp. 339–357, May 2016.
[12] A. Georgiadou, S. Mouzakitis, and D. Askounis, "Detecting insider threat via a cyber-security culture framework," *J. Comput. Inf. Syst.*, vol. 62, no. 4, pp. 706–716, Jul. 2022.
[13] M. C. A. M. R. Damodaram and L. Valarmathi, "Phishing website detection, and optimization using particle swarm optimization technique," *Int. J. Comput. Sci., Secur.*, vol. 5, no. 5, p. 477, 2011.
[14] A. V. Bhagyashree and A. K. Koundinya, "Detection of phishing websites using machine learning techniques," *Int. J. Comput. Sci., Inf. Secur.*, vol. 18, no. 7, 2020.
[15] J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cybersecurity," *J. Comput. Syst. Sci.*, vol. 80, no. 5, pp. 973–993, Aug. 2014.
[16] G. K. Birajdar and V. H. Mankar, "Digital image forgery detection using passive techniques: A survey," *Digit. Invest.*, vol. 10, no. 3, pp. 226–245, Oct. 2013.
[17] M. Rami, T. L. McCluskey, and F. Thabtah, "An assessment of features related to phishing websites using an automated technique," in *Proc. Int. Conf. Internet Technol., Secured Trans.*, London, U.K., Dec. 2012, pp. 492–497.
[18] C. Ludl, S. McAllister, E. Kirda, and C. Kruegel, "On the effectiveness of techniques to detect phishing sites," in *Proc. Int. Conf. Detection Intrusions, Malware, Vulnerability Assessment.*, 2007, pp. 20–39.
[19] P. R. Kankrale, "Phishing website detection using machine learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. VI, pp. 3216–3220, Jun. 2021.
[20] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 649–656.
[21] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, May 2019.
[22] S. Parekh, D. Parikh, S. Kotak, and S. Sankhe, "A new method for detection of phishing websites: URL detection," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 949–952.
[23] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-business websites," *Inf. Manage.*, vol. 51, no. 7, pp. 845–853, Nov. 2014.
[24] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 1–28, Sep. 2011.
[25] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection, and measurement of phishing attacks," in *Proc. ACM Workshop Recurring Malcode (WORM)*, 2007, pp. 1–8.
[26] (2022). *Dataset*. [Online]. Available: http://ilkerkara.karatekin.edu.tr/RequestDataset.htmldataset
[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[28] L. Zhang and P. N. Suganthan, "Oblique decision tree ensemble via multisurface proximal support vector machine," *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2165–2176, Oct. 2015.
[29] A. V. Bhagyashree and A. K. Koundinya, "Detection of phishing websites using machine learning techniques," *Int. J. Comput. Sci., Inf. Secur.*, vol. 18, no. 7, 2020.
[30] A. A. Akinyelu, "Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based, and nature-inspired-based techniques," *J. Comput. Secur.*, vol. 29, no. 5, pp. 473–529, 2021.
[31] S. Anupam and A. K. Kar, "Phishing website detection using support vector machines and nature-inspired optimization algorithms," *Telecommun. Syst.*, vol. 76, no. 1, pp. 17–32, Jan. 2021.
[32] A. Almomani, M. Alauthman, M. T. Shatnawi, M. Alweshah, A. Alrosan, W. Alomoush, B. B. Gupta, B. B. Gupta, and B. B. Gupta, "Phishing website detection with semantic features based on machine learning classifiers: A comparative study," *Int. J. Semantic Web Inf. Syst.*, vol. 18, no. 1, pp. 1–24, Jan. 2022.
[33] F. Song, Y. Lei, S. Chen, L. Fan, and Y. Liu, "Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers," *Int. J. Intell. Syst.*, vol. 36, no. 9, pp. 5210–5240, Sep. 2021.
[34] A. Awasthi and N. Goel, "Phishing website prediction: A machine learning approach," in *Progress in Advanced Computing and Intelligent Engineering*. Singapore: Springer, 2021, pp. 143–152.
[35] R. Chiramdasu, G. Srivastava, S. Bhattacharya, P. K. Reddy, and T. R. Gadekallu, "Malicious URL detection using logistic regression," in *Proc. IEEE Int. Conf. Omni-Layer Intell. Syst. (COINS)*, Aug. 2021, pp. 1–6.
[36] C. Rupa, G. Srivastava, S. Bhattacharya, P. Reddy, and T. R. Gadekallu, "A machine learning driven threat intelligence system for malicious URL detection," in *Proc. 16th Int. Conf. Availability, Rel. Secur.*, Aug. 2021, pp. 1–7.
[37] G. H. Lokesh and G. BoreGowda, "Phishing website detection based on effective machine learning approach," *J. Cyber Secur. Technol.*, vol. 5, no. 1, pp. 1–14, Jan. 2021.
[38] P. Pujara and M. B. Chaudhari, "Phishing website detection using machine learning: A review," *Int. J. Sci. Res. Comput. Sci., Eng., Inf. Technol.*, vol. 3, no. 7, pp. 395–399, 2018.
[39] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, p. 2019, 2019.
[40] Y. A. Alsariera, A. V. Elijah, and A. O. Balogun, "Phishing website detection: Forest by penalizing attributes algorithm and its enhanced variations," *Arabian J. Sci. Eng.*, vol. 45, no. 12, pp. 10459–10470, Dec. 2020.
[41] C. Singh and Meenu, "Phishing website detection based on machine learning: A survey," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 398–404.
[42] Y. Chen, F. M. Zahedi, A. Abbasi, and D. Dobolyi, "Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools," *Inf. Manage.*, vol. 58, no. 1, Jan. 2021, Art. no. 103394.

ILKER KARA (Member, IEEE) received the Ph.D. degree from Gazi University, in 2015. He has been an Assistant Professor with the Department of Medical Services and Techniques, Eldivan Medical Services Vocational, School in Çankırı, Karatekin University, since 2019. He has also been a Part-Time Lecturer with the Computer Engineering Department, Hacettepe University, since 2017. He has actively collaborated with researchers from several other disciplines such as computer science and forensics security in particular. He has authored more than 50 technical publications focusing on the applications of cyber security, malware analysis, and data security mechanisms. His research interests include cover digital investigation, malware analysis, and internet security.

MURATHAN OK (Member, IEEE) received the master's degree from the Informatics Institute, Hacettepe University, in 2020. He is working as an Information Systems Deputy Manager with ANADOLU AJANS A.Ş. His research interests include digital image processing, cybersecurity, machine learning, computer vision, and embedded programming.

AHMET OZADAY (Member, IEEE) received the B.Sc. degree in profession of industrial engineering from Eskişehir University, the master's degree in quality and conformity assessment engineering from Hacettepe University, and the associate degree in management information systems from Anadolu University. He is currently pursuing the master's degree in information security with Hacettepe University. His research interests include cyber security, information security management systems, machine learning, and image processing.