A Hybrid Phishing Detection System Using Deep Learning-based URL and Content Analysis

Mehmet Korkmaz[1], Emre Kocyigit[1], Ozgur Koray Sahingoz[2, *], Banu Diri[1]
[1]*Department of Computer Engineering, Yildiz Technical University, 34220 Istanbul, Turkey*
[2]*Department of Computer Engineering, Biruni University, 34010 Istanbul, Turkey*
osahingoz@biruni.edu.tr

***Abstract*—Phishing attacks are one of the most preferred types of attacks for cybercriminals, who can easily contact a large number of victims through the use of social networks, particularly through email messages. To protect end users, most of the security mechanisms control Uniform Resource Locator (URL) addresses because of their simplicity of implementation and execution speed. However, due to sophisticated attackers, this mechanism can miss some phishing attacks and has a relatively high false positive rate. In this research, a hybrid technique is proposed that uses not only URL features, but also content-based features as the second level of detection mechanism, thus improving the accuracy of the detection system while also minimizing the number of false positives. Additionally, most phishing detection algorithms use datasets that contain easily differentiated data pieces, either *phishing* or *legitimate*. However, in order to implement a more secure protection mechanism, we aimed to collect a larger and high-risk dataset. The proposed approaches were tested on this *High-Risk URL and Content-Based Phishing Detection Dataset* that only contains suspicious websites from PhishTank. According to experimental studies, an accuracy rate of 98.37 percent was achieved on a more realistic dataset for phishing detection.**

***Index Terms*—Phishing detection; Deep learning; URL-based; Content-based; Two-stage hybrid system; High-risk dataset.**

I. INTRODUCTION

A cyberattack can be defined as a deliberate attempt to harm computers, steal data, or make a compromised computer system launch some new attacks on other computers. Cyberattacks are composed of four parts: Attacker, Victim, Communication Tool, and Attack Mechanism. First, two of these components are similar in all cyberattacks, while the last two show the characteristics of attacks.

To reach a great number of victims, attackers prefer the use of multicast communication channels such as short message service, voice over Internet protocol, instant message over applications, email, etc. With the anonymous structure of the Internet, these attackers can easily hide themselves by using some third-party services. As one of the popular attack types, phishing is one of the most preferred attack types which imitates a reputable firm or individual to obtain private information of the victims, such as login credentials or financial data. According to the IBM report [1], the highest average cost in the last 17 years is 2021 with $4.24 million in terms of data breach. Remote work owing to COVID-19 is reported to trigger this. According to the report, phishing is the second most effective attack method with 17 % among different methods.

The phishing attack targets the weakest part of the security chain, end-users, and aims to force them to enter some malicious Uniform Resource Locator (URL) addresses, which can collect some sensitive information. The attack is made specifically by creating a fake website that contains an information-stealing kit. Usually, the interface of a trusted firm's website or a website with high traffic is imitated. For this, the visual design and URL address of the fake website should be similar to those from legitimate and trusted ones. Also, the URL address of the fake website should be very similar to the URL of the trusted website. After that, if the user does not notice the attack by investigating the URL and content of the website, their personal information can easily be captured by attackers.

When the reports of Anti-Phishing Working Group (APWG) are examined, it is observed that there is about 400 % increase in the number of unique URLs when the first six months of 2020 and 2021 are compared [2], [3]. In the second half of 2021, as in the first six months, if this number increases, it can be predicted that there will be an increase of 126 %. Looking at these rates, it can be predicted that the increase in the mechanisms used in phishing attacks will continue. Therefore, it is important to emphasize that precautions should be taken against phishing.

Researchers are working on the detection and prevention of phishing attacks in both the academic field and the software industry. They have offered many different solutions, which tend to the use of machine learning-based approaches for the detection of unencountered attacks and

Manuscript received 26 April, 2022; accepted 9 August, 2022.

zero-day attacks. Currently, these learning models are transferred to the deep learning approach which can produce efficient results, especially with the use of the Big Data concept. Therefore, in this paper, it is aimed to implement a Phishing Detection System, which can investigate not only the URL but also the content of the websites in a hybrid approach with the use of deep learning models. The main contributions in the paper are listed as follows:

– The dataset, which contains 87,489 URLs and *PhishTank* content data, was created for the first time in the literature. It contains 51,316 legitimate and 36,173 phishing records. It can be said that it is closer to real-world data.

– A "High-Risk URL and Content-based Phishing Detection Dataset" is created. Possible phishing URLs and content that Internet users want to label were collected from a single source, PhishTank.com (PhishTank), for both classes. In the implementation of a phishing detection system, the dataset is important to measure the efficiency of the system. Generally, datasets are collected from different resources, which can be easily identified as phishing or legitimate. However, in the real world, the "*suspicious"* websites should be considered. Therefore, the used dataset is directly collected from PhishTank, which is a community-based phishing verification system that allows users to submit "*suspected phishing threats"* into the system, and other users can vote to determine whether these threats are authentic. That is why this dataset was named High-Risk Dataset.

– For the first time, a model which is obtained by combining the Generative Adversarial Network (GAN) and Convolutional Neural Network (CNN) models, named as "Generative Convolutional Neural Network" (GCNN), using handcrafted and character embedding features, is proposed for the URL-based approach.

– The Deep Neural Network (DNN) model is proposed with handcrafted features in the content-based approach.

– A new method is proposed, called "Two-Stage Hybrid Phishing Detection System" (TshPhish) with URL and content-based method. In addition, URL- and Content-based Detection Method (UCDeM) has been developed as a new mechanism. Consequently, if the URL-based approach result is phishing, the final decision is given as phishing. If it is legitimate, the content-based detection mechanism is activated.

The rest of the paper is organized as follows. In the next section, there is an overview of phishing attacks and detection systems. In Section III, the academic studies in recent years are summarized. In Section IV, the High-Risk URL and Content Dataset are detailed. The detection models performed by URL and content analysis and the hybrid model, which has not been studied before in the literature, and the experimental results are depicted in Section V and Section VI, respectively. The results of experiments are discussed in Section VII. Finally, the paper concludes by showing conclusions and some future works.

II. BACKGROUND

In the section, we will argue what is a phishing attack and the basic mechanisms used to detect these attacks.

A. Phishing Attacks

PhishTank [4] is an organization that works with many partners to prevent phishing attacks, which are fraud methods to steal user sensitive information, especially disseminated through social media channels such as email. Many organizations work like PhishTank, such as VirusTotal, Google Safe Browsing, InfraGard, Cisco Talos Intelligence, etc. The common goal of these organizations is to prevent phishing attacks that have caused costs in recent years. Although there are many studies on prevention of these attacks, there have been great numbers of efforts in the cybersecurity domain due to its easy creation and large impact.

To organize a phishing attack, attackers need to organize some important steps, as mentioned in the lifecycle of the phishing in Fig. 1.

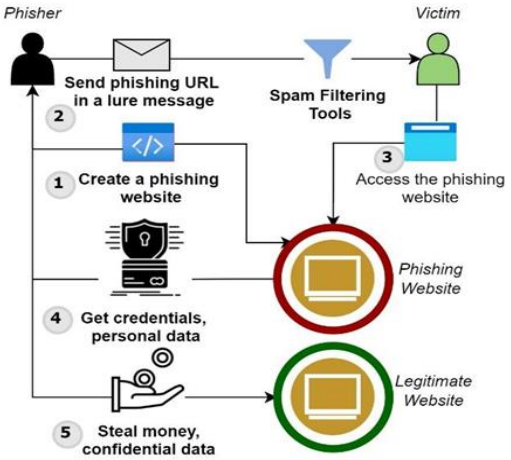

Fig. 1. Phishing lifecycle.

In phishing attacks, the victim is usually deceived by contacting via some trusted social engineering tools, such as email messages. In emails, there are usually texts that contain urgency, offer gifts, or superior-subordinate business relationships. With the URL given at the end of these texts, the victim is forwarded to the website. With the kit prepared on this website, the victim's bank, or personal information is stolen and transmitted to the attacker. The attacker uses this information to gain an unfair advantage (especially financial) from the legitimate website. Unfortunately, every victim who believes this email is vulnerable to possible loss. The main advantage of attackers is to complete transactions as quickly as possible by urging the victim without giving him a chance to think.

B. Phishing Attack Detection System

Basic precaution that can be taken is to increase user awareness. Attackers target the weakest part of the security chain, the end user. Therefore, cybersecurity systems on this domain are focused on minimizing the human factor by investigating these messages, the URLs in the emails, or the

website content hosted in the URL.

Many end users believe that the use of "https" is reliable. However, according to the APWG report [3], the number of websites that use secure protocols reached 82 % in 2021. When examined in terms of Web Design, the similarity rate between the original and the copied website is quite high. Although it cannot be noticed with these features, phishing websites can be distinguished to a certain degree. Examining URLs and content is effective in detecting phishing websites. Research based on machine learning has focused on these two points in recent years. Especially by analyzing the URLs of the websites, effective results were produced [5]. This approach is called "URL-based Phishing Detection". Similarly to this, the approach in which the content of the website is researched is also effective [6]. And this approach is called "Content-based Phishing Detection". Phishing websites can be easily detected with features like Hypertext Markup Language (HTML) language, pop-ups, and scripts running in the background. However, phishing websites are usually zero-day. In other words, their lifespan is about two hours. Therefore, it is quite difficult to find a dataset that contains content from phishing websites. For this reason, the number of studies on content-based phishing detection is also low.

Studies for the detection of phishing websites, whether they are URL-, content-, or image-based, continue at full speed. In addition to these, seminars, conferences, and reports are organized, in which individual measures can be taken against a possible attack are shared. Therefore, studies carried out in both the software industry and the academic field are important to find solutions to the phishing attacks that are encountered in daily life.

III. RELATED WORKS

The literature review shows that there have been studies on phishing attacks for a long time. With the information obtained in these studies, attacks were thought to have been prevented. However, it was understood that as time passed, differences began to occur in phishing attacks. Therefore, studies carried out in recent years to find a solution have focused on artificial intelligence. It is a well-known fact that especially machine learning and deep learning approaches are effective in detecting the attack. In addition, today it can be said that deep learning, with its ability to learn different features, produces an effective solution. It is an endorsed opinion that deep learning is currently the leading model in recent years. Therefore, in this paper, we will share a wide range of publications by limiting them to deep learning. These publications can be classified into three categories as URL-based, content-based, and a hybrid approach of both.

URL-based approaches are widely used because of their short-time operation and high accuracy. In [7], researchers analyzed CNN, Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) models with the characters and words. The accuracy of 97.38 % was achieved by combining CNN and LSTM, 97.5 % accuracy was achieved by combining CNN and BiLSTM, using the dataset that includes 7,484 URLs from PhishTank and Alexa. In [8], the researchers made a feature selection with the stacked auto encoder technique in the dataset they obtained with 17,000 phishing from PhishTank and 20,000 legitimate URLs from the Dmoztools.net website (DMOZ). The tests were performed with Support Vector Machine, Naive Bayes, Regression Tree, K-Nearest Neighbors, and DNN. DNN took the first place in the tests with the accuracy of 94.73 %. In [9], a dataset with a total of 27,700 pieces of data divided into four categories, namely phishing, spamming, malware, and advanced persistent threat was used. Fifty-six features have been studied. Based on these features, the accuracy rate was calculated by making a feature reduction with the Deep Belief Network (DBN). In experiments with different parameters, a 75 % accuracy rate was achieved. In [10], the authors have made keyword-based character embedding in the dataset consisting of 340,000 phishing URLs and 65,000 legitimate URLs, which they have allocated to different attack categories. They used CNN and then Gated Recurrent Units (GRUs) for pooling. They achieved 99.6 % accuracy with their module called "Convolutional GRU". Researchers in [11] used models containing CNN and GRU layers to compare approaches created using lexical features, character embedding, and word embedding. They tested these models in their experiments on 2,585,146 URLs. As a combination of these models, they obtained an accuracy of 94.4 % in their proposed model.

Researchers in [12] created models with Dense layers (86.54 % accuracy) and CNN layers (86.43 % accuracy) using word embedding. According to the combined approach, 86.63 % accuracy was obtained. The tests were carried out with 999,996 legitimate URLs and 523,970 phishing URLs. Some handcrafted features (character embedding, character level term frequency-inverse document frequency, and character level count vector features) were used in the experiments. These experiments performed on four different datasets [13]. The CNN model achieved 95.02 % accuracy in the datasets, which the authors in [13] created. In the other three datasets, they obtained greater accuracy than those of the studies in these datasets. Similarly, the BiLSTM algorithm based on CNN and the independent Recurrent Neural Network (RNN) named bidirectional LSTM algorithm was used by binary character of the domain name, word embedding, twenty host features, twenty-one URL features were used [14]. In the dataset consisting of 13,652 phishing and 10,000 legitimate URLs, the algorithm achieved 98.45 % accuracy. A new approach, CNN-Multi-Head Self-Attention (MHSA), was used in the dataset which researchers created by collecting 45,000 legitimate URLs from best websites and 43,984 phishing URLs from PhishTank [15]. In addition to the CNN model, they created with 84 characters, and they obtained 99.84 % accuracy by using MHSA in weight calculation.

In another paper, CNN, LSTM and models in which these two are used together with a dataset are based on network traffic [16]. In the paper, these models were compared with machine learning algorithms. In the tests performed, 98.67 % accuracy was achieved using Deep learning-based intrusion detection system in the weight calculation. The new model created based on DNN, CNN, and LSTM layers was tested with URLs taken from 2,119 PhishTank and 1,407 Alexa [17]. In the experiments where they used

information gain in feature selection, LSTM produced results that are successful with 99.57 % accuracy, even though they took close values. The researchers developed a new model called "Precise Phishing Detection with Recurrent Convolutional Neural Networks" [18], in which they used BiLSTM and CNN layers. They tested character-based and manually extracted features in experiments with 500,000 URLs obtained from the PhishTank and Alexa database. The results gave a 97 % accuracy for proposed model. Different from the other researchers, the authors in [19] produced a phishing URL to balance the dataset with the GAN. They created a dataset that contained 68,030 legitimate URLs and 12,003 phishing URLs from PhishTank. Then, they obtained 95.6 % accuracy in their experiments with CNN and MHSA.

In another paper, URLs and hyperlinks are passed through certain rules and features are obtained [20]. An accuracy of 99.80 % was obtained as a result of the artificial neural network used in the dataset created with 1,000 legitimate URLs from Alexa database and 1,000 phishing URLs from the PhishTank website. With a new dataset (1 million URLs, half of which was obtained from PhishTank and the rest from the CommonCrawl database, and the dataset contains 10,000 images) which researchers used in [21], CNN and LSTM were tested in Intelligent Phishing Detection System (IPDS) models. The IPDS model produced results with an accuracy of 93.28 %. In [22], detailed URL-based research was carried out using models such as RNN, LSTM, GRU, and CNN. The authors reached a 97.30 % accuracy which they combined all models in their suggestion. They used dataset of 5,000 phishing and 5,000 legitimate data. Researchers in [23] compare CNN and RNN models with some machine learning algorithms. They created a dataset (22,491 phishing and 24,719 legitimate) and got a 99.35 % accuracy rate with CNN + Random Forest model. In [24], different methods of deep learning and machine learning were compared. The datasets were created by collecting data from four different places. Among the results they obtained in their experiments on these different datasets, the RNN + GRU model achieved the highest accuracy with 99.18 % accuracy. In [25], the results were obtained in two publicly shared datasets. In the proposed DNN model, the authors achieved a 96.54 % accuracy in the first dataset and a 96.32 % accuracy in the second.

One of the major approaches that is used in phishing detection problems is the content-based approach, and this approach is significantly exploited in analysis of email and website content. Since phishers view emails as handy communication channels, researchers tend to detect spam and/or phishing emails through a content-based approach [26]. Although email server and browser spam filtering applications perform well, they cannot detect all phishing attacks. At this point, content-based solutions indicate dramatic success in detecting phishing attacks [27]. In particular, machine learning- and deep learning-based models are quite successful compared to the conventional method in these studies [28]. Phishers benefit from social engineering techniques and modify website content to deceive Internet users. Inspecting the website content to analyze whether it is phishing or not is not common and practical in terms of user experience. However, Artificial Intelligence-based algorithms easily retrieve website content and classify it as legitimate and phishing [29]. Content-based phishing website detection solutions use machine learning classifiers and obtain high accuracy scores [30]. Generally, researchers use URL- and content-based features in their models. Contrary to this, only content-based features are used in the content-based model.

Recently, URL- and content-based approaches have been studied together. In [31], researchers created a new dataset taking 1,021,758 phishing URLs from PhishTank and 989,021 legitimate URLs from the DMOZ. A subdataset was created with 22,445 active phishing and 22,390 randomly selected legitimate data from the URLs in this dataset. While there is a structure based on the XGBoost algorithm in the subdataset where both the URL and the content of the website are looked at, CNN and RNN approaches are studied in the main dataset. According to the approach they used dynamic category decision algorithm in the paper, the URL and then the website content, if any, were analyzed, and experiments were carried out. According to this approach, an accuracy of 98.99 % was obtained. In [32], researchers treat the URL, the content of the HTML page, and the structure of the Document Object Model (DOM) as strings of characters. Character and word corpuscles have been created for the URL. The HTML content is based on words and sentences. Only tags were used when creating the dom corpus. The rest have been ignored. The learning model was created using CNN and BiLSTM layers with 99.05 % accuracy rate obtained in the dataset created by taking 24,800 legitimates from the Alexa database and 21,303 phishing URLs and content from the PhishTank website. In [33], the researchers suggested feature extraction in four main titles: URL, abnormal, HTML and JavaScript, Domain features. In the paper, the decision-making mechanism has been developed to choose the features under these headings. Then the neural network was designed and tested in two different datasets. The first one is public, and the second one consists of 14,582 URLs which were taken from the Alexa and PhishTank website. 99.3 % accuracy was obtained in the experiments performed on their own datasets. Researchers in [34] used some machine learning algorithms, the DNN and CNN model with URL and HTML embedding feature vectors. A dataset was created by collecting URLs from PhishTank and Alexa (4,700 phishing and 47,000 legitimate). The experiments were designed using 12 URL-based and 19 content-based features. In addition, by adding the deep learning model they realized with character analysis, they achieved 98.4 % in their dataset with the WebPhish approach they proposed. In [35], two datasets (D1: 4,898 phishing and 6,157 legitimate, D2: 7,044 phishing and 7,049 legitimate) were used. In experiments, a combination of content-based, URL lexical-based, and domain-based features was used. Random forest, support vector machine, logistic regression, artificial neural network, and CNN models were used in the paper. The random forest algorithm achieved 100 % accuracy for

the D1 dataset and 92.83 % accuracy for the D2 dataset.

IV. THE HIGH-RISK URL AND CONTENT DATASET

When the papers in Section III are examined, it is seen that the phishing parts of more than half of the datasets used are taken from the PhishTank website. This organization, which has an important place in analyzing phishing attacks, is a pioneer and reliable, especially in the scope of blacklisting, with its URL list. It shares its database with its partners and supports software companies. The operation on the website of such a large-scale organization is as follows. Users leave URLs to the URL pool to be queried. URLs in this list, which are open to all guests, are checked by users and classified as phishing or legitimate. A URL is tagged according to the number of votes it receives. Thus, three categories of URLs are listed: Phishing, legitimate, and unrated. If the URL is inactive and no user has moderated it, it will be tagged as unrated. These can qualify as neutral elements in the URL list.

URLs that have been inspected and found to be harmful while they are live are labeled "Phishing". The phishing part of the dataset contains these URLs. Those with website content from these URLs listed under Online and Valid Phish on the PhishTank website have been added to the Phishing section of the dataset, along with both the URL and the content. The phishing part of the dataset is similar to studies in the literature. The part where the dataset differs from those in the literature is the legitimate part. In the literature, this part is usually created from the URLs obtained from the Alexa database or website categorization sites, whose web traffic is high and the content is controlled. However, the URLs that were added to query PhishTank are mentioned and tagged as legitimate. These URLs, which are labeled invalid in PhishTank and have content, are the legitimate part of the dataset. Thus, the data that were added to the checklist after being suspicious by the users and then labeled as legitimate constituted a risky legitimate part.

Collecting legitimate and phishing URL addresses and their content is a relatively easy task. However, in the real world, security administrators (and also systems) focus on suspicious websites. Therefore, to collect the dataset, we wrote a script to check for suspicious URLs on PhishTank every 10 minutes. If a URL is found in the list, both the URL and its content are added to our dataset using an ID link. Additionally, this file stores URL ID, URL name, and confirmation time information. By this way, detailed information about the URL was obtained. All datasets are stored in CSV format and legitimate records were tagged as 0, while phishing ones were tagged as 1.

As shown in Fig. 2, a dataset was created with 51,316 legitimate URLs and contents, 36,173 phishing URLs and contents, listed between 2006 and 2021.

Looking at the data obtained, the distribution of data in two categories by years is as in Fig. 3. Approximately 91 % of the phishing data were obtained in 2021. This rate confirms the existence of data used as zero-day attacks in the dataset. However, it can be said that the legitimate data are evenly distributed. Again, it can be deduced with the idea of how accurately the legitimate data are labeled.

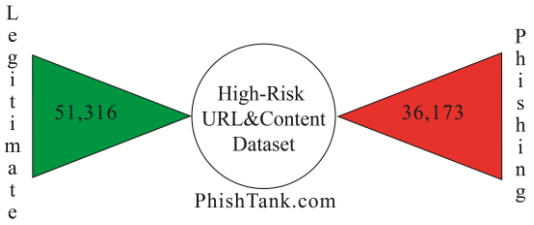

Fig. 2. High-risk URL and content dataset.

The URLs and content were collected with a written script in Python. In data preprocessing, the size of 0 Kb of these URLs is excluded from the dataset. Additionally, all URLs are checked one by one and if we encounter an "Error 403" code in their content, they are removed from the dataset. After creating this dataset, we were able to analyze it to achieve various and valuable results. In this context, the average length of a URL in this dataset is about 76 characters, and the average length of a hostname is about 19. 32,246 of them contain 19 characters or more. 560 of them are less than 5 characters. The dataset has an average of 9.1 digits. The average domain length was found to be 9.4.

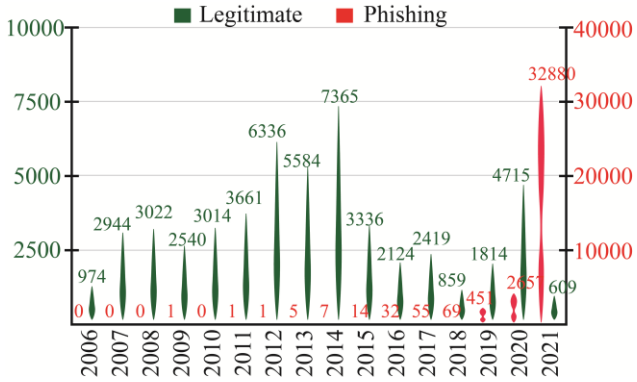

Fig. 3. Distribution of data by years.

From these collections and preprocessing approaches, it can be said that this dataset, called the "High-Risk URL and Content Dataset", is closer to real-world data than other datasets in the literature, and it is shared in the link [36] for other researchers.

V. PROPOSED SYSTEM

To increase the efficiency of the detection system, it is aimed to use both URL- and content-based approaches in a hybrid model.

A. URL-based Phishing Detection Model

To make a fast detection, many phishing detection systems are focused on the use of URLs which are formed as in Fig. 4 by containing protocol, subdomain, main domain, top-level domain, directory, file name, and parameter parts. Some security systems classify domain names; however, some web servers can provide services for their users to create their own pages. Therefore, instead of domain names, making an analysis on the whole URL address gives a detailed result.

End users can easily distinguish whether a URL is legitimate or not. However, attackers aim to distract the end users by urging them to enter the website and share their

sensitive information. Therefore, a URL-based detection system needs to analyze the URL and act as a decision support system for the user by warning about suspicious websites. To gain related knowledge, these systems should be trained with the previous data using some machine learning algorithms.

Fig. 4. Form of URL.

In this paper, two different approaches are used for this training. First, a feature extraction process is examined and 73 features are identified from the URL text as detailed in [5], and then some machine learning/deep learning models are used. Figure 5 shows the deep learning model used.

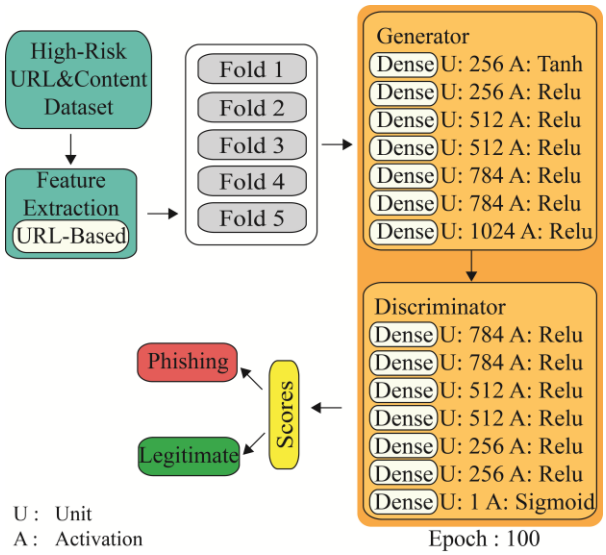

Fig. 5. Workflow of URL handcrafted features based on GAN model.

Second, the URLs are analyzed by their characters. As seen in Fig. 6, character-based embedding is used in the URL-based approach.

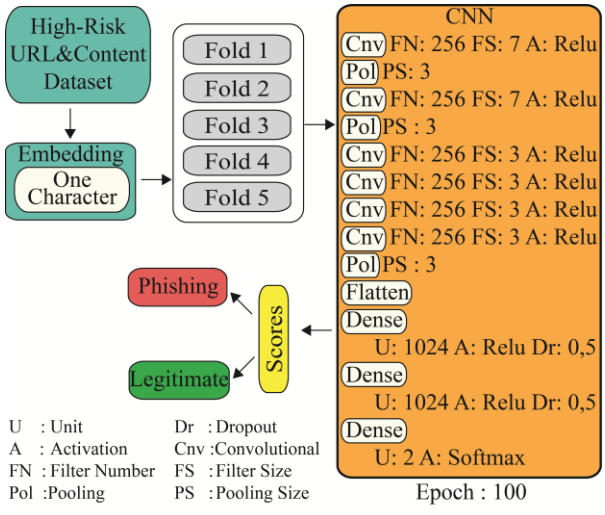

Fig. 6. Workflow of the character-based CNN model on URL.

All letters are converted to lower case in the preprocessing step. The number of words used for embedding is taken as mentioned in [37]. Then, URL-based optimal architecture was created by testing different deep learning models.

Detection systems are implemented for these mentioned models one by one, and then, to increase the efficiency of the system, a hybrid model is tested. Figure 7 shows the proposed architecture for URL-based phishing detection.

In the URL-based approach, the best performance is reached with the GAN model, which uses 73 features, and the CNN model, which uses one character embedding, are merged.

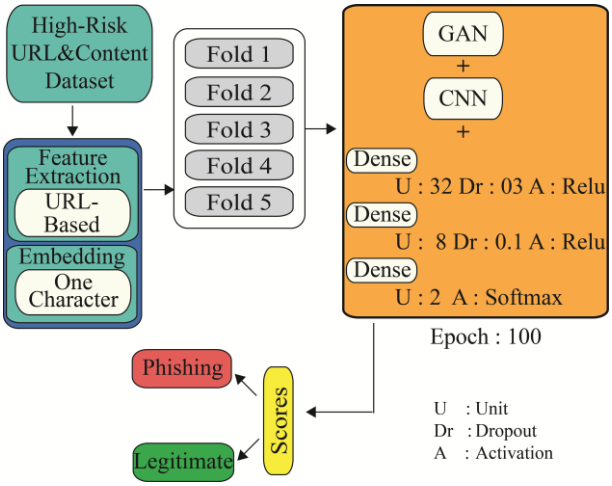

Fig. 7. Workflow of the URL-based model for handcrafted features and one character embedding.

B. *Content-based Phishing Detection Model*

The main components of website content are HTML, CSS, and JavaScript. Particularly, some HTML components and their quantities can be distinctive when classifying phishing and legitimate websites. In [6], 27 content-based features are obtained. After manual analysis of phishing website files, seven new content-based features are generated, such as “comment line, text size”. Therefore, 57 content-based features are listed in Table I.

TABLE I. CONTENT-BASED FEATURES.

Names of content-based features				
Has Submit Input		Has Email Input		Has Hidden Tag
Has Password Input		Has Popup		Has Onmouseover
Has Div Upper Case		Has Favicon		Has Iframe
Image Count		Href Count		Link Count
Submit Input Count		Footer Count		Hidden Tag Count
Email Input Count		Popup Count		Title Count
Iframe Count		Form Count		Title Length
Div Count		H1 Count		H2 Count
Video Count		Canvas Count		Span Count
Stylesheet Count		BR Count		Option Count
Table Count		TH Count		TR Count
Li Count		Ul Count		P Count
Style Count		Button Count		Meta Count
Label Count		Select Count		Base Count
Audio Count		Script Count		Address Count
Nav Count		Figure Count		Section Count
Insert Count		Total Line Count		Total Text Length
Meta Content Count		Div Class Count		Comment Count
Server Form Handler Count		Has Script Upper Case		Script Language Count

Features were classified as two types: One of them is "If a feature name contains "Has", that feature can take only 1 or 0 as a value and it represents if the content contains that feature". And the other one is "If a feature name contains "Count", that feature can take any integer value and represents the total number of that feature in the content".

First, the content of each website is retrieved from the dataset and then preprocessing is done. Second, a function for each feature is defined and executed for each example. Hence, numerical values for each example are obtained and used in the content-based model as depicted in Fig. 8.

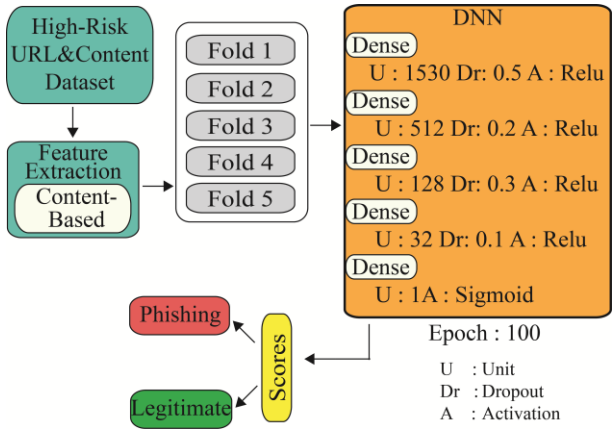

Fig. 8. Workflow of content-based DNN model.

Additionally, the performance of each feature is measured before finalizing the content-based feature set and it was observed that the average runtime of each feature was within a reasonable range. Therefore, all these 57 content-based features are used in the model.

C. Two-Stage URL- and Content-based Hybrid Phishing Detection System

Past research [5], [6] shows that URL and content analysis differ in the time taken for detection. Based on this, it should not be thought that "*URL analysis is faster and can be used in studies, while content analysis works slowly and does not make sense to use in studies*". It is important to minimize the cost of phishing attacks within the scope of cyberattacks. Therefore, the more accurate phishing attacks can be detected, the better. It seems to us that it is acceptable to have a suspicious-looking URL actually be legitimate and predicted as phishing, rather than actually being phishing and predicted as legitimate. When predicted as phishing, the warning mechanism works and warns the victim. If the victim still trusts (no big problem if it is legitimate), they can continue to view this website. However, when it is phishing and guessed legitimate, undesirable situation happens, and the victims can lose all their data by trusting the system (phishing URL). Therefore, it is thought that the two approaches can be optimally used together, as seen in Fig. 9 to obtain a new detection system with UCDeM.

According to the proposed model, the system consists of two control stages. The first is URL-based analysis and the second is content-based analysis. The URL to be checked is first put into the URL-based analysis and the prediction result is obtained. If this result is phishing, it is considered as a system result phishing without switching to content-based analysis. This saves time in analysis.

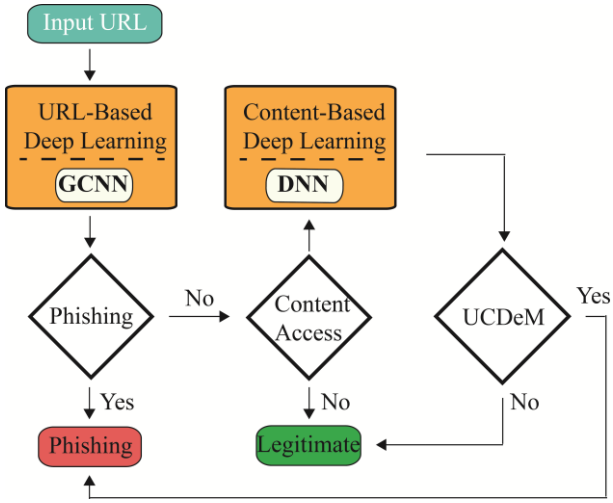

Fig. 9. Architecture of two-stage URL- and content-based hybrid phishing detection system.

If the result is legitimate, it moves to the second stage, the content-based analysis step. If the URL content is not available, the result is considered a URL-based result without review. If there is content, a content-based analysis is also performed. If results are obtained from both stages, UCDeM works as follows. The predicted results for both stages are taken from the models as a percentage. The final legitimate percentage value is obtained by proportioning the percentage of legitimateness from the first and second stages according to the threshold value determined according to the ensemble approach. The same process is calculated for the percentage of phishing. Thus, the new classification values are obtained with the ensemble approach. If the URL is legitimate in the first stage, the results of the ensemble approach in the second stage are valid.

VI. EXPERIMENTAL RESULTS

To measure the efficiency of the proposed models, some experiments are executed, and their results are depicted in the following part.

A. Test Results on URL-based Phishing Detection Model

It was previously explained that the features obtained by URL analysis were considered in two categories: handcrafted features and deep learning features. Different models have been tested by using them. Handcrafted features have been tested in both machine learning and deep learning algorithms with 5-fold cross-validation and the results are listed in Table II.

TABLE II. TEST RESULTS OF MACHINE LEARNING CLASSIFIERS WITH HANDCRAFTED FEATURES.

	Decision Tree	**Random Forest**	**ANN**	**LSTM**	**DNN**	**GAN**
Accuracy (%)	87.78	91.79	90.75	84.31	94.20	94.99

Although the Random Forest algorithm achieved the highest accuracy among the machine learning algorithms used with 91.79 %, the DNN model gave the second best result with 94.20 % accuracy, and the GAN model gave the best result with 94.99 % accuracy among all models.

Characters in a URL address can be used in analysis of the phishing URL's classification. Therefore, it is aimed to use a character embedding for the classification of URL addresses with some deep learning models such as GRU, LSTM, BiLSTM, and CNN. Table III shows the results obtained using a single character with the mentioned deep learning models. According to the test results, the model constructed with single-character embedding and CNN layers achieved the highest accuracy with 97.17 %.

TABLE III. TEST RESULTS OF DEEP LEARNING MODELS WITH ONE CHARACTER EMBEDDING.

Embedding	GRU	LSTM	BiLSTM	CNN
Accuracy (%)	94.89	95.18	95.93	97.17

Table IV presents the results of two models concatenated. Accordingly, it can be said that there is a small increase in all models compared to the single ones, except GRU. The GCNN model was created by merging GAN and CNN, which obtained the highest accuracy.

TABLE IV. TEST RESULTS OF URL-BASED MERGED MODELS.

	GAN + GRU	GAN + LSTM	GAN + BiLSTM	GAN + CNN
Accuracy (%)	94.04	96.14	96.51	97.68

Based on the results obtained up to this point, a model that detects phishing attacks with 97.68 % accuracy is proposed in the first step of the hybrid approach with GCNN model.

B. Test Results on Content-based Phishing Detection Model

In this paper, different DL models are used for a content-based approach, such as LSTM, GAN, and DNN. In a similar paper [6], the GAN model has been proposed. Therefore, the performance of the GAN model is calculated and compared with various DNN models. After several experiments, the DNN model, as shown in Table V, produced an accuracy score of 93.39.

TABLE V. TEST RESULTS OF CONTENT-BASED MODEL.

	LSTM	GAN	DNN
Accuracy (%)	86.01	91.60	93.39

Moreover, it was detected that the success rate of the content-based model was not good from the URL-based model. But content-based model could detect some of the phishing websites that were not detected by URL-based model. In other words, false classified phishing examples of URL-based models can be detected by a content-based model and vice versa.

C. Test Results on a Two-Stage URL- and Content-based Hybrid Phishing Detection System

A phishing detection system is expected to detect phishing attacks quickly and with high accuracy. This is extremely important in real-world data. As mentioned in the sections above, only the URL text is analyzed in the first stage. If it is predicted as phishing, the suggested approach will not move to the second step, thus saving time. Therefore, the accuracy and speed of this stage are important. At this stage, the features to be obtained from third-party applications were not used to run faster. The 97.68 % accuracy obtained in the test results for a high-risk URL and content dataset.

However, compared to the content-based approach based on DNN model with the accuracy of 93.39 %, the results obtained in the URL-based approach are slightly higher. After the investigations, it was observed that these two models complement each other to a certain extent. Table VI presents the results obtained by combining these two approaches with the ensemble model.

TABLE VI. ACCURACY SCORES OF ENSEMBLE MODELS FOR TshPhish.

		URL-based Models			
		GAN + GRU	GAN + LSTM	GAN + BiLSTM	GAN + CNN
Content-based Models	DNN	96.73	97.57	97.74	98.37
	GAN	96.61	97.46	97.67	98.31

Based on Table VI, the TshPhish model is proposed, which can provide maximum performance with the UCDeM approach which was created. As can be seen in Table VI, the new model obtained by combining DNN in the content-based approach and GCNN in the URL-based approach achieved 98.37 % accuracy. Also, it can be seen in the results that the accuracy is increased from the URL-based model by approximately 0.69 % and from the content-based model by approximately 4.98 %, as depicted in Table VII.

TABLE VII. TEST RESULTS OF STAGE ONE, STAGE TWO, AND TshPhish.

Metrics	Value		
	URL	Content	TshPhish
Sensitivity	0.9672	0.8971	0.9764
Specificity	0.9836	0.9598	0.9888
Precision	0.9765	0.9403	0.9840
F1 Score	0.9718	0.9182	0.9802
Negative Predictive Value	0.9770	0.9297	0.9834
False Positive Rate (FPR)	0.0164	0.0402	0.0112
False Discovery Rate	0.0235	0.0597	0.0160
False Negative Rate	0.0328	0.1029	0.0236
Accuracy	0.9768	0.9339	0.9837
Error Rate	0.0231	0.0661	0.0112

In systems designed to detect phishing attacks, the riskiest situation is when the URL, which is identified as legitimate, is phishing. This means the value is False Positive. Therefore, the FPR is important in the confusion matrix. In the TshPhish model, this rate has been reduced from 0.0164 to 0.0112, which shows that the system improves the FPR by about 30 %. Additionally, there is a considerable decrease in the error rate of the proposed model (about 49 %).

Also, there are phishing attack detection studies in the literature that use URL- and content-based approaches in addition to our paper. These studies are detailed in our article. Table VIII also compares the results obtained in these studies with the approach we propose. If the features such as dataset size, number of phishing data, obtaining data from a single source, and using deep learning approaches

are taken into account, the distinctive properties of our study can be clear. Another important point is that the success lies in the high-risk category in our dataset, whose legitimate URLs are reported as suspicious to the PhishTank. Therefore, this paper produced a real-world dataset and reached a detection accuracy of 98.37 %.

TABLE VIII. COMPARISON WITH URL- AND CONTENT-BASED PHISHING DETECTION SYSTEMS.

Ref.	Dataset		Method	Accuracy (%)
	Phishing	Legitimate		
[31]	22,445 PhishTank	22,390 DMOZ	XGBOOST(Content) CNN + LSTM (URL)	98.99
[32]	21,303 PhishTank	24,800 Alexa	CNN + BiLSTM	99.05
[33]	1,476 UCI	13,106 UCI	ANN	99.3
[34]	4,700 PhishTank	47,000 Alexa	CNN	98.4
[35]	D1: 4,898 UCI D2: 7,044 PhishTank	D1: 6,157 UCI D2: 7,049 Several	D1: Random Forest D2: Random Forest	D1: 100 D2: 92.83
Our	36,173 PhishTank	51,316 PhishTank	DNN (Content) GCNN (URL)	98.37

VII. DISCUSSION

Due to the anonymity of the Internet, phishing is one of the most common cyberattacks in recent years. To establish a secure system, various dynamic phishing detection technologies are required in addition to traditional security mechanisms. This system should be resistant to new types of attack, with a strong emphasis on dynamic learning.

According to the machine learning or deep learning approach, URL-based, content-based, and URL&content-based phishing attack detection methods are used in the literature. Among these methods, there are limited studies on content-based and URL&content-based approaches. To increase the number of these studies, we think that one of the points that should be developed in the literature is the dataset. One of the shortcomings of existing datasets is that the quantity of data is not high and there is a small amount of phishing site content due to zero-day attacks. Another shortcoming of datasets is that they are not real-world datasets. We attribute this to the receipt of legitimate data from known indexing websites like Alexa, Google, etc. The contribution of existing datasets to the literature is indisputable, but as a contribution to the literature, the legitimate data in the real world in our dataset will provide quality to the study area.

URL-based detection systems are favored for real-time protection. However, this may result in a high false positive rate. As a result, this research proposes a new dataset and a two-stage phishing detection method that utilizes both URL-based and content-based detection schemas in a hybrid model. The proposed approach is effective with a high degree of accuracy and a low rate of false positives, as demonstrated by the experimental results.

When the test results of the URL-based approach and the content-based approach are compared, it has been determined that the accuracy rate of the content-based approach is lower. It can be expected that the results obtained in the experiments to be performed by combining these two approaches will be lower than the results obtained in the URL-based approach. However, with the UCDeM approach, we developed TshPhish in which the advantageous points of two models were combined. Therefore, better accuracy was obtained by conducting a study in which we used the strengths of the models. Consequently, this rate has been increased by supporting the content-based approach with the base accuracy rate of the URL-based model. Furthermore, considering the importance of FPR in phishing detection systems, it can be seen that our approach combines the strengths of the models.

VIII. CONCLUSIONS AND FUTURE WORKS

In this paper, first, a study on the collection of a relatively large URL and content dataset was carried out by collecting 36,173 phishing and 51,316 legitimate data from PhishTank. This dataset is a high-risk dataset whose links are reported as suspicious and risky to the PhishTank system.

Second, with this new dataset, URL-based, content-based, and a hybrid model of URL and content-based models were tested separately. All experiments were carried out with 5-fold cross-validation. As a result of our experimental work, in the URL-based model, a 97.68 % accuracy is achieved by using a new model of the GCNN model. The performance of the system was also measured using the content-based model, and 93.39 % accuracy was achieved using the DNN model. Finally, our main proposal, TshPhish, which uses the hybrid model of URL- and content-based models, was conducted, and in this model, 98.37 % accuracy was achieved. So, it can be said that the use of the hybrid model results in better efficiency in the detection of phishing attacks.

In future studies, our aim is to increase the amount of data in the high-risk URL and content dataset. In addition, the goal is to optimize the feature selection mechanism using evolutionary algorithms to increase the overall efficiency of the system.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] IBM Data Breach, IBM. [Online]. Available: https://www.ibm.com/tr-tr/security/data-breach

[2] Phishing Activity Trends Reports, Phishing Activity Trends Report 1st Quarter 2021, APWG. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2021.pdf

[3] Phishing Activity Trends Reports, Phishing Activity Trends Report 2nd Quarter 2021, APWG. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2021.pdf

[4] PhishTank. [Online]. Available: https://phishtank.org/what_is_phishing.php

[5] M. Korkmaz, O. K. Sahingoz, and B. Diri, "Detection of phishing websites by using machine learning-based URL analysis", in *Proc. of 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–7. DOI: 10.1109/ICCCNT49239.2020.9225561.

[6] E. Kocyigit, M. Korkmaz, O. K. Sahingoz, and B. Diri, "Real-time content-based cyber threat detection with machine learning", in *Intelligent Systems Design and Applications. ISDA 2020. Advances in Intelligent Systems and Computing*, vol. 1351. Springer, Cham, 2021, pp. 1394–1403. DOI: 10.1007/978-3-030-71187-0_129.

[7] K. S. Ray and R. Kusshwaha, "Detection of malicious URLs using deep learning approach", in *The "Essence" of Network Security: An End-to-End Panorama. Lecture Notes in Networks and Systems*, vol.

163. Springer, Singapore, 2021, pp. 189–212. DOI: 10.1007/978-981-15-9317-8_8.

[8] S. Sountharrajan, M. Nivashini, S. K. Shandilya, E. Suganya, A. B. Banu, and M. Karthiga, "Dynamic recognition of phishing URLs using deep learning techniques", in *Advances in Cyber Security Analytics and Decision Systems. EAI/Springer Innovations in Communication and Computing*. Springer, Cham, 2020, pp. 27–56. DOI: 10.1007/978-3-030-19353-9_3.

[9] S. G. Selvaganapathy, M. Nivaashini, and H. P. Natarajan, "Deep belief network based detection and categorization of malicious URLs", *Information Security Journal: A Global Perspective*, vol. 27, no. 3, pp. 145–161, 2018. DOI: 10.1080/19393555.2018.1456577.

[10] W. Yang, W. Zuo, and B. Cui, "Detecting malicious URLs via a keyword-based convolutional gated-recurrent-unit neural network", *IEEE Access*, vol. 7, pp. 29891–29900, 2019. DOI: 10.1109/ACCESS.2019.2895751.

[11] T. Rasymas and L. Dovydaitis, "Detection of phishing URLs by using deep learning approach and multiple features combinations", *Baltic Journal of Modern Computing*, vol. 8, no. 3, pp. 471–483, 2020. DOI: 10.22364/BJMC.2020.8.3.06.

[12] B. Wei, R. A. Hamad, L. Yang, X. He, H. Wang, B. Gao, and W. L. Woo, "A deep-learning-driven light-weight phishing detection sensor", *Sensors*, vol. 19, no. 19, p. 4258, 2019. DOI: 10.3390/s19194258.

[13] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J-P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL", *Electronics*, vol. 9, no. 9, p. 1514, 2020. DOI: 10.3390/electronics9091514.

[14] H.-h. Wang, L. Yu, S.-w. Tian, Y.-f. Peng, and X.-j. Pei, "Bidirectional LSTM malicious webpages detection algorithm based on convolutional neural network and independent recurrent neural network", *Applied Intelligence*, vol. 49, pp. 3016–3026, 2019. DOI: 10.1007/s10489-019-01433-4.

[15] X. Xiao, D. Zhang, G. Hu, Y. Jiang, and S. Xia, "CNN–MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites", *Neural Networks*, vol. 125, pp. 303–312, 2020. DOI: 10.1016/j.neunet.2020.02.013.

[16] P. Sun, P. Liu, Q. Li, C. Liu, X. Lu, R. Hao, and J. Chen, "DL-IDS: Extracting features using CNN-LSTM hybrid network for intrusion detection system", *Security and Communication Networks*, vol. 2020, 2020. DOI: 10.1155/2020/8890306.

[17] M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, "Efficient deep learning techniques for the detection of phishing websites", *Sādhanā*, vol. 45, art. no. 165, 2020. DOI: 10.1007/s12046-020-01392-4.

[18] W. Wang, F. Zhang, X. Luo, and S. Zhang, "PDRCNN: Precise phishing detection with recurrent convolutional neural networks", *Security and Communication Networks*, vol. 2019, art. ID 2595794, 2019. DOI: 10.1155/2019/2595794.

[19] X. Xiao, W. Xiao, D. Zhang, B. Zhang, G. Hu, Q. Li, and S. Xia, "Phishing websites detection via CNN and Multi-Head Self-Attention on imbalanced datasets", *Computers & Security*, vol. 108, art. 102372, 2021. DOI: 10.1016/j.cose.2021.102372.

[20] C. Wang, Z. Hu, R. Chiong, Y. Bao, and J. Wu, "Identification of phishing websites through hyperlink analysis and rule extraction", *The Electronic Library*, vol. 38, nos. 5/6, pp. 1073–1093, 2020. DOI: 10.1108/EL-01-2020-0016.

[21] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent phishing detection scheme using deep learning algorithms", *Journal of Enterprise Information Management*, 2020. DOI: 10.1108/JEIM-01-2020-0036.

[22] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, "Classifying phishing URLs using recurrent neural networks", in *Proc. of 2017 APWG Symposium on Electronic Crime Research (eCrime)*, 2017, pp. 1–8. DOI: 10.1109/ECRIME.2017.7945048.

[23] R. Yang, K. Zheng, B. Wu, C. Wu, and X. Wang, "Phishing website detection based on deep convolutional neural network and random forest ensemble learning", *Sensors*, vol. 21, no. 24, p. 8281, 2021. DOI: 10.3390/s21248281.

[24] L. Tang and Q. H. Mahmoud, "A deep learning-based framework for phishing website detection", *IEEE Access*, vol. 10, pp. 1509–1521, 2022. DOI: 10.1109/ACCESS.2021.3137636.

[25] J. Anitha and M. Kalaiarasu, "A new hybrid deep learning-based phishing detection system using MCS-DNN classifier", *Neural Computing and Applications*, vol. 34, no. 8, pp. 5867–5882, 2022. DOI: 10.1007/s00521-021-06717-w.

[26] H. Che, Q. Liu, L. Zou, H. Yang, D. Zhou, and F. Yu, "A content-based phishing email detection method", in *Proc. of 2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2017, pp. 415–422. DOI: 10.1109/QRS-C.2017.75.

[27] U. Ozker and O. K. Sahingoz, "Content based phishing detection with machine learning", in *Proc. of 2020 International Conference on Electrical Engineering (ICEE)*, 2020, pp. 1–6. DOI: 10.1109/ICEE49691.2020.9249892.

[28] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework", *Wireless Communication and Mobile Computing*, art. ID 4678746, 2018. DOI: 10.1155/2018/4678746.

[29] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information", *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 2015–2028, 2019. DOI: 10.1007/s12652-018-0798-z.

[30] A. A. Zuraiq and M. Alkasassbeh, "Review: Phishing detection approaches", in *Proc. of 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS)*, 2019, pp. 1–6. DOI: 10.1109/ICTCS.2019.8923069.

[31] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning", *IEEE Access*, vol. 7, pp. 15196–15209, 2019. DOI: 10.1109/ACCESS.2019.2892066.

[32] J. Feng, L. Zou, O. Ye, and J. Han, "Web2Vec: Phishing webpage detection method based on multidimensional features driven by deep learning", *IEEE Access*, vol. 8, pp. 221214–221224, 2020. DOI: 10.1109/ACCESS.2020.3043188.

[33] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, "OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network", *IEEE Access*, vol. 7, pp. 73271–73284, 2019. DOI: 10.1109/ACCESS.2019.2920655.

[34] C. Opara, Y. Chen, and B. Wei, "Look before You leap: Detecting phishing web pages by exploiting raw URL And HTML characteristics", 2020. ArXiv: 2011.04412.

[35] M. Aljabri and S. Mirza, "Phishing attacks detection using machine learning and deep learning models", in *Proc. of 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, 2022, pp. 175–180. DOI: 10.1109/CDMA54072.2022.00034.

[36] High-Risk URL and Content Dataset, kaggle. [Online]. Available: https://www.kaggle.com/datasets/08c5834d30c8bb5cc08d273d146e042512173f72adb8da5bbde0c8dd928fb879

[37] M. Korkmaz, E. Kocyigit, O. K. Sahingoz, and B. Diri, "Phishing web page detection using n-gram features extracted from URLs", in *Proc. of 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2021, pp. 1–6. DOI: 10.1109/HORA52670.2021.9461378.