ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 5, 291-296 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Ensemble Learning of Deep URL Features based on Convolutional Neural Network for Phishing Attack Detection

Seok-Jun Bu[*,1], Hae-Jung Kim[2]

[1]*Deptartment of Computer Science, Yonsei University, Seoul 03722, Korea*

[2]*Department of Computer Science, Kyungil University, Daegu 38428, Korea*

ARTICLE INFO

Article history:
Received: 20 July, 2021
Accepted: 10 October, 2021
Online: 14 October, 2021

Keywords:
Phishing detection
Deep learning
Ensemble learning
Convolutional neural network
Recurrent neural network

ABSTRACT

The deep learning-based URL classification approach using massive observations has been verified especially in the field of phishing attack detection. Various improvements have been achieved through the modeling of character and word sequence of URL based on convolutional and recurrent neural networks, and it has been proven that an ensemble approach of each model has the best performance. However, existing ensemble methods have limitations in effectively fusing the nonlinear correlation between heterogeneous features extracted from characters and the sequence of sub-domains. In this paper, we propose a convolutional network-based ensemble learning approach to systematically fuse syntactic and semantic features for phishing URL detection. By learning the weights that integrating the heterogeneous features extracted from the URL, an ensemble rule that guarantees the best performance was obtained. A total of 45,000 benign URLs and 15,000 phishing URLs were collected and 10-fold cross-validation was conducted for quantitative validation. The obtained classification accuracy of 0.9804 indicates that the proposed method outperforms the existing machine learning algorithms and provides plausible solution for phishing URL detection. We demonstrated the superiority of the proposed method by receiver-operating characteristic (ROC) curve analysis and the case analysis and confirmed that the accuracy improved by 1.93% compared to the latest deep model.

1. Introduction

Network security based on information technology for protecting personal information and system resources from various types of threats may be defined through policies and methods. Various methods for network administrator have been developed to protect networks and cyber assets including detection mechanism against active attacks [1]. However, few studies have been conducted to analyze the characteristics of phishing attacks, which steal entire input information from users. Phishing attack in its broadest sense can be defined as a scalable act of deception whereby impersonation is used by an attacker to obtain the information from an individual [2]. Considering that the most common form of online phishing attack is malicious hyperlinks embedded in messages, the recent technological trend in which personal connections are reinforced due to the explosive growth of social media services is particularly vulnerable [3].

Existing security systems primarily conduct rule-based detection mechanism using phishing databases to identify malicious URLs [4]. However, phishing URLs based on web applications have zero-day exploit characteristics that frequently involve novel attack instances, as URLs can be generated very conveniently in such applications. For this reason, phishing URLs hardly detected by predefined databases or simple detection rules [2, 5, 6].

Meanwhile, previous study based on ensemble of the convolutional neural network (CNN) and recurrent neural network (RNN) for the modeling the character and word-level features found that classification of malicious URLs was improved [7,8].

In Figure 1, we visualize the phishing URLs into feature space generated by the t-SNE dimension reduction method. Blue and red dots represent normal URLs and phishing URL instances, respectively. The Euclidean distance was determined based on the similarity of character combinations constituting the URL, and a cluster of short and regular URLs was mainly formed at the

*Corresponding Author: Seok-Jun Bu, Yonsei University, sjbuhan@yonsei.ac.kr

bottom. On the other hand, in the center, instances where it is difficult to distinguish between normal and phishing URLs due to subdomains are intricately confused.

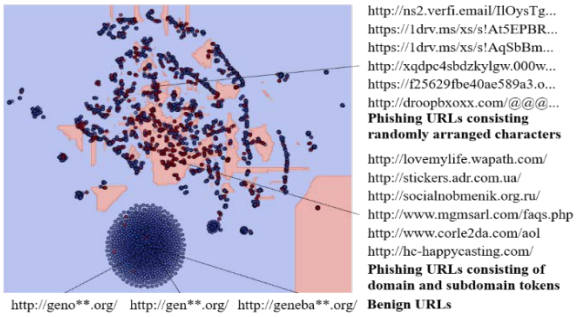

Figure 1: The feature space of phishing URLs and the necessity of ensemble learning

Phishing URL features can be distinguished into a syntactic feature consisting of a sequence of randomly arranged characters and a semantic feature consisting of a sequence of domain and subdomain words. In the existing deep learning-based ensemble approach, a simple rule-based ensemble that averages the output of syntactic-semantic convolution and recurrent networks at log-scale was applied, but it showed the limitation to effectively model the complex nonlinear correlation of features and resulted the degradation of accuracy and recall.

Taken together, we propose an ensemble learning network based on CNN that can systematically utilize the syntactics and semantics of URL features using CNN and RNN. The proposed ensemble network is a deep learning algorithm that can extract effective features for phishing URL classification using filter operations that can be trained using data. The joint learning of ensemble rule based on deep representations of URLs provides plausible solution for phishing detection. We collected a total of 45,000 benign URLs and 15,000 phishing URLs and the proposed method was validated through 10-fold cross-validation, and chi-squared test. The analytic results indicated the best performance among the machine learning-based phishing detector. To the best of our knowledge, this is the first attempt that convolutional neural network is incorporated to learn the ensemble rule for phishing detection. The main findings of this research can be summarized as follows:

- The convolutional neural network works well for learning the ensemble rule of fusing heterogeneous features of URL representations, resulting the best accuracy and recall for phishing detection.
- We categorized the features of URLs into character and word levels, and demonstrated the convolutional and recurrent neural networks to effectively model each feature.

The remainder of this paper is organized as follows. In Section 2, we review the previous URL modeling methods based on machine learning and clarify the contributions of this paper by discussing the differences between them. In Section 3, we illustrate how the heterogeneous URL features are extracted by the deep learning and fused with convolutional neural network. The performance of the model is evaluated in Section 4 through various experiments, including the 10-fold cross-validation and ROC curve analysis. Finally, section 5 concludes the paper with some discussion of future directions.

2. Related Works

Previous studies on phishing URL classification can be classified into the following categories as summarized in Table 1: those on phishing URL detection based on the blacklist, which were mainly performed before 2010; those on modeling of words extracted from the text based on traditional machine learning; and those on text feature extraction through the latest deep learning algorithms.

The author proposed a system that extracts lexical features from the text according to ex-pert-defined rules, constructs a blacklist on known phishing URLs, and detects new phishing URLs through a simple comparison algorithm [9]. However, this method has the limitation of detecting new phishing URLs in terms of generalization performance. To confirm the validity of the machine learning method in the field of phishing URL classification, the authors applied fundamental machine learning methods including naive Bayes classifiers to the word combination found in URLs and classified phishing URLs that were not included in training datasets [10]. The authors enhanced the performance of phishing URL classification systems based on machine learning by applying a support vector machine (SVM), which is widely known to perform more complex nonlinear mapping [11]. Verma significantly increased phishing URL classification accuracy through the implementation of a random forest algorithm that was designed to perform effective modeling of hierarchical elements of lexical features in the URLs [12].

The researchers extracted semantic features from phishing URLs using a word-to-vector model capable of embedding word vectors based on their statistical meaning using deep learning algorithms. Furthermore, they applied long short-term memory (LSTM) and gated re-current unit (GRU) deep learning algorithms specialized for time series modeling, including gate operations, to enhance the phishing URL classification performance of existing modeling methods [8,13]. It was proposed a convolution-recurrent network to effectively model semantics extracted from the word-to-vector model [14].

The majority of the current research in deep learning-based phishing detection focuses mainly on optimizing the operation of the neural network [16]. In particular, the comparative study in [17] proves the superiority of the ensemble approach based on CNN variations. This motivates our decision to consider the ensemble learning approach proposed in this paper. The proposed method deviates from existing work in that it implements and learns the ensemble rule with convolutional operation based on CNN to consider the heterogeneous URL features.

Table 1: Related works on phishing URL detection with respect of URL features and modeling method.

URL Features	Method	Author
Bag-of-words	Naive Bayes	Prakash [9]
Lexical Features	Matching Rules	Ma [10]
Bag-of-words	SVM	Le [11]
Lexical Features	Random Forest	Verma [12]
Word embeddings	LSTM	Bahnsen [8]
Word embeddings	GRU	Zhao [13]
Lexical Features	Generative adversarial network (GAN)	Anand [15]
Word embeddings	CNN-LSTM	Yang [14]

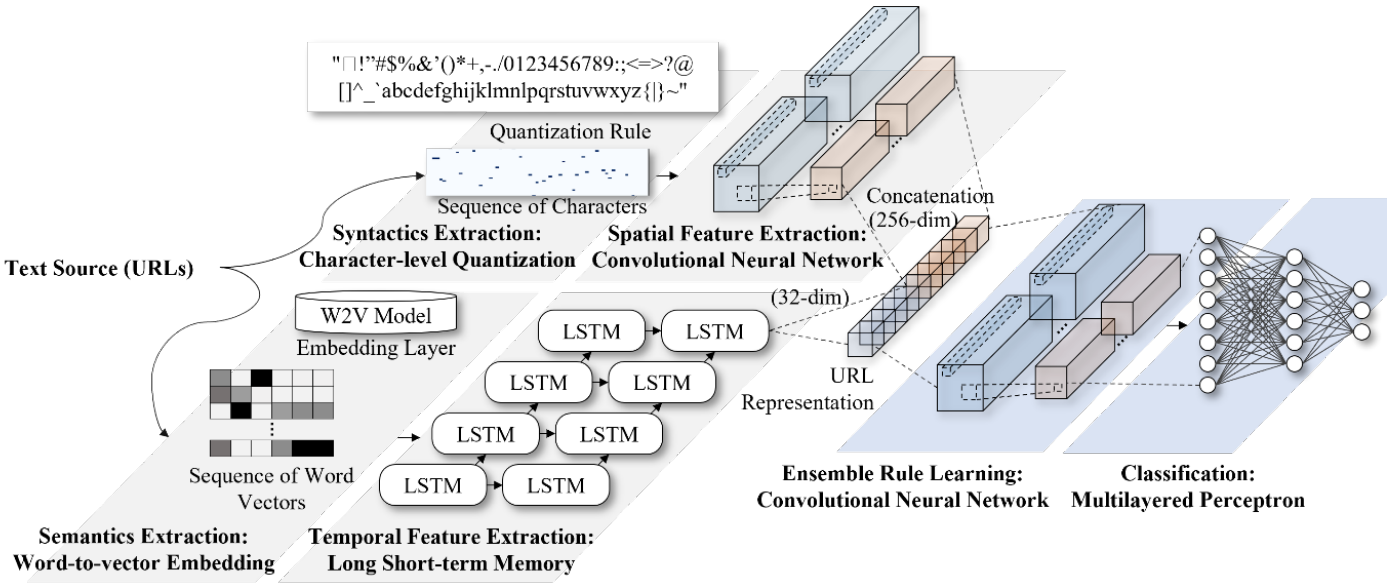

Figure 2: Proposed CNN-based ensemble learning method for phishing URL detection

3. Proposed Method

In this section, we describe the combination of the convolutional neural network and recurrent neural network to extract URL features and its ensemble learning method. Figure 2 visualize the diagrams of deep learning-based phishing classifier that extract the syntactic and semantics from URLs, as well as the proposed CNN-based ensemble learning network for the late-fusion of URL features.

3.1. Deep Learning-based Phishing URL Feature Extraction

Two types of deep learning algorithms and individual preprocessing steps were applied to conduct the modeling of syntactic and semantic features of phishing URLs. First, an integer was assigned to each character, and modeling of a low-level signal obtained through this process was performed by the CNN to model the syntactic features of random characters, including enumerated special characters, which are frequently observed in phishing URLs. Second, each word was embedded based on the word-to-vector model, and the modeling of a sequence of words obtained through this process was performed by the LSTM to model the semantic features of domains and sub-domains composing the internal URLs.

In detail, a preprocessing step for each character was performed to replace the characters with their unique Unicode values based on UTF-8 encoding, and an integer sequence of up to 100 characters was extracted in consideration of the average length of URL characters in the datasets collected. In total, 139 types of characters were used, and a vector in the dimension of n ×100 ×139 based on n of observations were inputted into the character-level CNN.

The convolution operation $\phi_c^l(\cdot)$ in Equation 1 applies a parameterized filter to the input vector and extracts syntactics from sequence of characters in URL. A filter size $m \times m$ is applied to the ith row and jth column nodes of the lth layer.

$$\phi_c^l(x_{ij}) = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab} x_{(i+a)(j+b)} \quad (1)$$

The pooling operation $\phi_p^l(\cdot)$ in the lth pooling layer performs the extraction of representative value and defined as Equation 2, with the pooling distance τ in the region $k \times k$ among the input vectors, and outputs the maximum activation value from the region.

$$\phi_p^l(x_{ij}) = \max_{\tau \in R} x_{ij \times \tau} \quad (2)$$

The learning of the convolution operation is the process of optimizing the weight of the filter w that extracts the syntactics while preserving the spatial correlation between characters, and the pooling operation is based on extraction of emphasized features.

Meanwhile, representative features of URLs include semantics that can be derived from a sequence of words such as domains and sub-domains. Phishing URL classification accuracy can be enhanced through the parallel utilization of deep learning algorithms for additional modeling of a sequence of subdomains [18].

The modeling of semantics of phishing URLs was carried out through word embedding based on the word-to-vector model and LSTM deep learning algorithm application for time series modeling. Moreover, 20 words that appeared in sub-domains were additionally extracted since phishing URLs generally included various sub-domains. Each word was replaced as vectors in 32 dimensions using the word-to-vector model, and URLs formed as n ×20 ×32 sized vector according to n observations were input in the phishing word-level LSTM.

The LSTM network is a type of RNN in which three types of nonlinear gates are implemented. The LSTM $\phi_L^l(\cdot)$ performs the time-series modeling of sequence of domain and subdomains.

$$\phi_L^l(x_{ij}) = o_t \odot \tanh(c_t) \quad (3)$$

The input gate(i), forget gate (f), output gate (o), and LSTM cell state (c) were defined based on the input domain sequence of

x = (x(t),…, x(t-ω)) with word sequence length ω, as shown in Equation 4. b, σ and ⊙ refer to the bias added to each neural network, the sigmoid activation function of neural networks, and Hadamard multiplication, respectively.

$$\begin{aligned} i_t &= \sigma(W_{ix}x(t) + W_{im}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x(t) + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x(t) + W_{ho}h_{t-1} + b_o) \\ g_t &= tanh(W_{xh}x(t) + W_{hc}h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \end{aligned} \quad (4)$$

3.2. Ensemble Learning based on the Convolutional Network

The proposed ensemble network utilizes the deep URL representations with a size of (n × 256) and (n × 32) with n observations from the intermediate layer of CNN and LSTM in Section 3.1. Contrary to the existing CNN-LSTM ensemble-based phishing URL detector, the model is optimized to weight the outputs from multi-level URL representations.

The character-level and word-level representations of phishing URL derived from the character-level CNN and word-level RNN were concatenated to form a vector of size (n × 288). The proposed fusion neural network was trained to minimize errors that might occur in the process of mapping the input vector to the benign or phishing URLs.

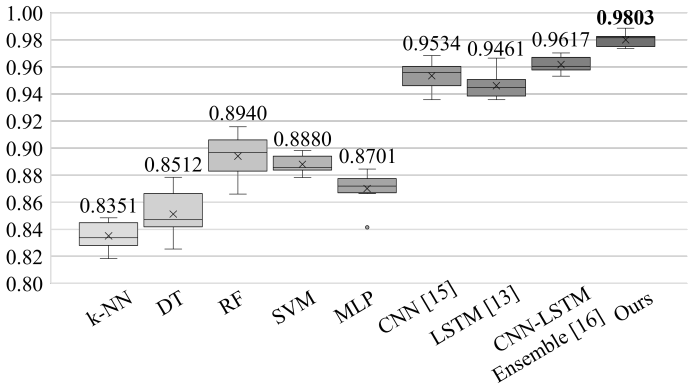

Figure 3: Results of 10-fold cross validation comparing the proposed ensemble learning with other existing methods (k-NN: k-nearest neighbor; DT: decision tree; RF: random forest; SVM: support vector machine; MLP: multilayered perceptron; CNN: convolutional neural network; LSTM: long short-term memory).

As the proposed method contains a convolution layer in order to systematically fuse level-based features, effective features are selected from input vectors from 288 dimensions and output the predictive label $\hat{y}$, as shown in Equation 5.

$$p(\hat{y}_i|x_i) = argmax \frac{exp(\phi^{l-1}(x_i)w^l + b^l)}{\sum exp(\phi^{l-1}(x_i)w^l + b^l)} \quad (5)$$

At this stage, the Softmax function, which is an activation function of the neural network, was applied to facilitate the encoding of the output vector at the probability of [0,1] range and to promote the differentiation process that operates during the optimization of the loss function. The entire mapping results obtained from inputs to outputs in the character-level and semantic-level neural networks, including the proposed ensemble network, is differentiable and can be learnt by the massive URL observations.

The entire weights of neural networks are tuned by applying a backpropagation algorithm based on gradient descent to the cross-entropy loss function L_{CE} shown in Equation 6.

$$L_{CE} = -\sum_i y_i log(\hat{y}_i) \quad (6)$$

The proposed ensemble network, which fuses features according to deep URL representations, performs the optimization of ensemble rules in consideration of joint learning of CNN and LSTM.

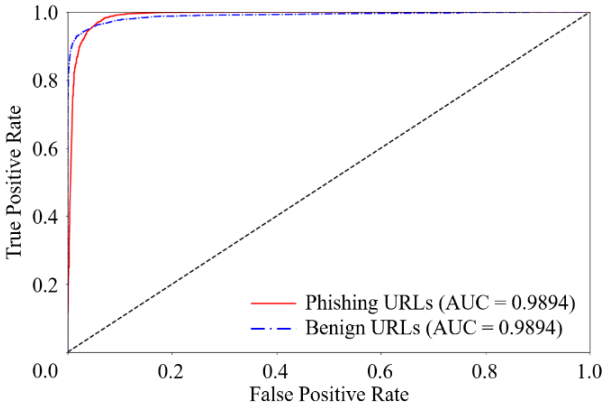

Figure 4: ROC curve and area under the curve (AUC) of classification result based on ensemble learning approach.

4. Experimental Results

4.1. Phishing URL Dataset

A total of 45,000 of URLs were collected by crawling method and 15,000 of phishing URLs were collected from Phishtank [19,20] where provides a blacklist of phishing URLs. Benign URLs were mainly collected from open directory project(ODP), which is the URL database to categorize URLs. The number of observations was intentionally adjusted because of the data imbalance issue, to reflect the conditions in which the number of phishing URLs was much lower than that of benign URLs. We noted the length of the URL is one of the critical phishing features, given that the average lengths of phishing and benign URLs are 75.74 and 35.83, respectively.

4.2. Phishing URL Classification Performance

Figure 3 shows the result of 10-fold cross validation on the proposed ensemble network and other machine learning-based models to verify the phishing URL classification performance of the proposed method. The average accuracy of the random forest algorithm, CNN, and RNN were achieved 0.8940, 0.9534, and 0.9461, respectively, with the CNN and RNN exhibiting significantly higher accuracy performance than the random forest algorithm.

The comparative result based on the ensemble of CNN [7] and RNN [8] achieved 0.9641. Based on the performance improvement, we confirmed the significance of proposed ensemble learning approach in consideration of character and word-level URL modeling. Regarding the proposed ensemble network designed to fuse Multi-level URL features, the best classification accuracy was achieved as 0.9803.

Table 2: Qualitative evaluation of complementarity based on the case analysis of CNN and LSTM (0: benign, 1: phishing).

Category		URL (accessed date: 19-10-2020)	CNN	LSTM
CNN Advantages	Phishing	https://1drv.ms/xs/s!AhtvzT3KrwqMZzLMKnTc8clHnRA?wdFormId=%7BA0F7982D%2D71A4%2D4DE0%2DB4C4%2DC16A0F044	0.9874	0.7385
	Benign	http://market.security***.net	0.0031	0.8441
LSTM Advantages	Phishing	http://bitcoin24-wallet.site	0.0722	0.9837
	Benign	http://www.knightfeatu***.com/kfweb/content/features/kffeatures/puzzlesandcrosswords/kf/sudoku/sudoku_classic/sudoku_classic.html	0.8384	0.0073
Misclassified	Benign	http://archives.seattletimes.nwsou***.com/cgi/bin/texis.cgi/web/vortex/display?slug=will&date=199903	0.8815	0.8764
	Phishing	http://tesla-present.site/ethereum/	0.0584	0.0354

Table 3: A confusion matrix for phishing URL classification based on the ensemble network

		Predicted (w/o ensemble learning)		
		Benign	Phishing	**Recall**
Actual	Benign	9035 (8902)	74 (207)	**0.9919 (0.9773)**
	Phishing	115 (266)	2776 (2625)	**0.9602 (0.9080)**
	Precision	**0.9874 (0.9710)**	**0.9740 (0.9269)**	Accuracy: 0.9843 (0.9606)

Since it is essential to minimize false negatives and improve recall in the field of phishing URL detection, the ROC curve and AUC are described in Figure 4. Table 3 summarizes the classification results based on the model that exhibited the best accuracy. Considering the false negatives and the recall of phishing instance of 0.9602, it is inferred that additional modeling should be carried out mainly focusing on the generalization strength of the model.

4.3. Discussion

Table 2 indicates the advantages of each model based on practical classification cases, to aid in the classification of the performance of deep learning models according to multi-level URL representations. The two upper rows show the robustness against random character enumeration of CNN. The character-level CNN classified URL as phishing instance with a probability of 0.9874, considering that a phishing URL feature is hidden in the sequence of random characters. On the other hand, in the second case, the word-based LSTM misclassified benign URL as phishing with a probability of 0.8441 because the benign words are included in the URL.

The word-level LSTM supplements the entire system by reflecting a sub-domain that was not used by the character-level CNN in the form of words. The LSTM was able to classify certain words such as ‘security’ and ‘bitcoin’ based on the massive observations that such words are frequently used in phishing URLs. The CNN, however, misclassified benign URLs as phishing based on the number of sub-domains.

5. Concluding Remarks

5.1. Conclusion

This study introduced a character-level CNN and word-level RNN for phishing URL representation and proposed an ensemble network that can effectively fuse the syntactics and semantics of phishing URLs extracted from each model. The proposed ensemble network, implemented as convolutional neural network, provide the plausible solution of parameterization and optimization of the ensemble rule. Specifically, it exhibited a classification accuracy of 0.9804, which is the highest compared to other machine learning methods including deep learning models.

5.2. Future Work

Further studies should be performed to guarantee the generalization strength of model, in consideration of zero-day attack characteristics of phishing attacks. The latest deep learning algorithms, such as one-shot learning, should be thoroughly examined. Meanwhile, the scope of this study is limited to the modeling of syntactics and semantics of URL and optimizing the ensemble rule. In this regard, a symbolic AI approach to fully exploit and utilize the domain knowledge is promising. In addition, the neural-symbolic integration approach that can calibrate the deep learning model should be additionally considered in the future studies.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1063085).

References

[1] H.-J. Kim, Image-based malware classification using convolutional neural network, Springer: 1352–1357, 2017, doi:10.1007/978-981-10-7605-3_215.
[2] S.-J. Bu, S.-B. Cho, “Deep Character-Level Anomaly Detection Based on a Convolutional Autoencoder for Zero-Day Phishing URL Detection,” Electronics, **10**(12), 1492, 2021, doi:10.3390/electronics10121492.
[3] V. Suganya, “A review on phishing attacks and various anti phishing techniques,” International Journal of Computer Applications, **139**(1), 20–23, 2016, doi:10.5120/ijca2016909084.
[4] K.L. Chiew, K.S.C. Yong, C.L. Tan, “A survey of phishing attacks: Their types, vectors and technical approaches,” Expert Systems with Applications, **106**, 1–20, 2018, doi:10.1016/j.eswa.2018.03.050.

[5] I. Qabajeh, F. Thabtah, F. Chiclana, "A recent review of conventional vs. automated cybersecurity anti-phishing techniques," Computer Science Review, **29**, 44–55, 2018, doi:10.1016/j.cosrev.2018.05.003.

[6] S.-J. Bu, S.-B. Cho, "Integrating Deep Learning with First-Order Logic Programmed Constraints for Zero-Day Phishing Attack Detection," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: 2685–2689, 2021, doi: 10.1109/ICASSP39728.2021.9414850.

[7] H. Le, Q. Pham, D. Sahoo, S.C.H. Hoi, "URLNet: Learning a URL representation with deep learning for malicious URL detection," ArXiv Preprint ArXiv:1802.03162, 2018, doi:10.475/123_4.

[8] A.C. Bahnsen, E.C. Bohorquez, S. Villegas, J. Vargas, F.A. González, "Classifying phishing URLs using recurrent neural networks," in 2017 APWG symposium on electronic crime research (eCrime), IEEE: 1–8, 2017, doi:10.1109/ECRIME.2017.7945048.

[9] P. Prakash, M. Kumar, R.R. Kompella, M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in 2010 Proceedings IEEE INFOCOM, IEEE: 1–5, 2010, doi:10.1109/INFCOM.2010.5462216.

[10] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond blacklists: learning to detect malicious web sites from suspicious URLs," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 1245–1254, 2009, doi:10.1145/1557019.1557153.

[11] A. Le, A. Markopoulou, M. Faloutsos, "Phishdef: Url names say it all," in 2011 Proceedings IEEE INFOCOM, IEEE: 191–195, 2011, doi:10.1109/INFCOM.2011.5934995.

[12] R. Verma, K. Dyer, "On the character of phishing URLs: Accurate and robust statistical learning classifiers," in Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, 111–122, 2015, doi:10.1145/2699026.2699115.

[13] J. Zhao, N. Wang, Q. Ma, Z. Cheng, "Classifying malicious URLs using gated recurrent neural networks," in International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Springer: 385–394, 2018, doi:10.1007/978-3-319-93554-6_36.

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

[15] A. Anand, K. Gorde, J.R.A. Moniz, N. Park, T. Chakraborty, B.-T. Chu, "Phishing URL detection with oversampling based on text generative adversarial networks," in 2018 IEEE International Conference on Big Data (Big Data), IEEE: 1168–1177, 2018, doi:10.1109/BigData.2018.8622547.

[16] F. Tajaddodianfar, J.W. Stokes, A. Gururajan, "Texception: A character/word-level deep learning model for phishing URL detection," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: 2857–2861, 2020, doi:10.1109/ICASSP40776.2020.9053670.

[17] D. Vasan, M. Alazab, S. Wassan, B. Safaei, Q. Zheng, "Image-Based malware classification using ensemble of CNN architectures (IMCEC)," Computers & Security, **92**, 101748, 2020, doi:10.1016/j.cose.2020.101748.

[18] Q. Li, M. Cheng, J. Wang, B. Sun, "LSTM based phishing detection for big email data," IEEE Transactions on Big Data, 2020, doi:10.1109/TBDATA.2020.2978915.

[19] L.L.C. OpenDNS, "PhishTank: An anti-phishing site," Online: https://Www. Phishtank. Com, 2016 (accessed: 1 Oct. 2021).

[20] Q. Cui, G.-V. Jourdan, G. V Bochmann, R. Couturier, I.-V. Onut, "Tracking phishing attacks over time," in Proceedings of the 26th International Conference on World Wide Web, 667–676, 2017, doi:10.1145/3038912.3052654.