Transactions of The Nigerian Association of Mathematical Physics
Journal homepage: https://nampjournals.org.ng

A SYSTEM FOR THE DETECTION OF MALICIOUS DOMAIN NAMES USING IMPROVED DEEP-LEARNING MODEL

Egwali, Annie O. and Ekhator, Roland O.

Department of Computer Science, Faculty of Physical Sciences. University of Benin, Benin City, Nigeria.

ARTICLE INFO

Article history:
Received xxxxx
Revised xxxxx
Accepted xxxxx
Available onlinexxxxx

Keywords:
Malicious Domain Detection,
DNS,
Cyber Security,
Deep Learning,
Domain Generation.

ABSTRACT

The tremendous growth of innovative technologies used for online services in the global economic space brings vulnerabilities to security breaches. The upsurge of these vulnerabilities created a level playing field for cyber-attacks to flourish, with assailants constantly adapting new nefarious methods to compromise information and deceive naïve users of the cyberspace. Despite the amazing and numerous anti-phishing approaches and solutions, the increasing incidences caused by malicious domain name system attacks such as spam, phishing and malware could be attributed to the dynamism in the approaches used by cybercriminals to counterfeit the techniques. To address these issues, many cyber security researchers have switched their focus to machine learning-based methodologies for malicious DNS detection. In this paper, we introduced the usage of machine-based model to detect the dynamism of malicious DNS by exploring Machine Learning, Ensemble learning and Deep-Learning. A customized web Crawler was implemented to extract URL attribute for model extraction. Furthermore, a Cross validation approach was used towards the classification and regression metrics (statistical approach) to evaluate their performance to an accuracy of 89.9%. Our experiment is based on both active and passive DNS analysis.

1. Introduction

The tremendous growth of innovative technologies used for online services and businesses in the global digital space brings along vulnerabilities to network security. A vulnerable network paves the way for hackers to manipulate and violate system infrastructure [1]. The upsurge of these vulnerabilities has created a level ground for cyber attacks to flourish. Often times domains that have previously committed suspicious acts are Black listed while the white lists include well-known and trustworthy domain names ([2]; [3]; [4]). This scenario is only used for broadly distributed infections rather than targeted ones. Attackers implant malicious programs (code) through the network vulnerabilities into host which grants them access to remotely control the host ([5]; [6]; [7]). The affected host then issue resolution requests, using a large number of nonexistent domain names randomly generated by the domain generation algorithm (DGA), a program that generates a large list of domain names and provide malware with new domains to evade security countermeasures in a short time ([8]; [9]).especially botnets ([10]; [11]). Some identified problems are limitations of conventional approaches used by many for malicious domain detection.

*Corresponding author. Egwali A.O.
E-mail address:annie.egwali@uniben.edu
https://www.doi.org/10.60787/tnamp-19-63-70

According to [12], these approaches are blacklisting of domain names, Malicious domain name (MDN) attack is a persisting problem in domain name system (DNS). Several researchers assert the fact that more work should be done and direction should be focused on the attacks analysis of the network traffic, dissection of the webpage content, DNS traffic analysis, and analysis of salient lexical features. The non-consideration of the domain name and the DGA data for computing the maliciousness of the URL results in a lack of precision. Hence, they advised effective mechanisms for malicious domain detection to help improve the precision of malicious URL detection using algorithm features selection and dynamic machine learning models that can act passively and actively in detecting malicious DNS; making use of Deep Learning algorithms to detect malicious accounts based on domain names to blacklisting associated Ips.

This paper aim therefore is to design a malicious name detection model using a data-centric methodology, to implement and train the proposed model using a DGA dataset and to validate the model performance using classification metrics (Accuracy, Precision, F1-score and Recall) or regression metrics (RMSE, MSE), to operationalize best performance model in a web browser.

2. RELATED WORKS

The DNS has been increasingly used by attackers to maintain and manage their materials infrastructure. Generally, malicious detection systems employ various approaches to investigate and establish the protected DNS environment. The verification of malicious DNS using multiple features can also be described in different dimension, in order to distinguish legitimate and malicious domains. A detection technology based on the passive domain name analysis method was proposed in [5] where a technique called EXPOSURE was used to train and monitor the DNS traffic of a commercial ISP. Several types of features were extracted by the researchers including the domain name lifetime, period similarity number of accesses; number of IPs parsed, whether IP is shared by other domain names, digital symbol ratio and length of longest meaningful substring. However, a classifier was constructed using decision tree algorithm.

A botnet detection algorithm basedon DNS traffic features using Power Spectral Density (PSD) testing technology which detects MDNs by analyzing malicious behavior within large volumes of DNS trafficwas proposed in [13]. A DGA classifier that leveraged long short-term memory (LSTM) networks for real-time production of DGA's without the need for contextual information or manually created features was also developed [14]. The experimental result showed that the method was significantly better than some state-of-the-art techniques. To obtained the data for the study of Local DNS records, [15] investigated the performance of various DL algorithms such as RRN, LSTM, and other approaches in which LSTM performed best in identifying malicious DNS requests. AnMDNs detection algorithm based on Algorithmically generated domain was developed using cluster correlation that identifies the names generated by a domain generation algorithm or its variants [16]. Various features such as TTL, the distribution of IP Addresses; WHOIS features, and historical information from the domain names in each cluster were extracted and the Support Vector Machine (SVM) algorithm was used to identify the MDNs.

Using Deep Learning (DL) approaches for the recognition of fraudulent domain names [17] extracted textual characteristics from domain names and passing them to LSTM and bidirectional LSTM. In another instance, [12] introduced selected features such as blacklist domain names features, DNS based features, web-based features and lexical features to identify malicious domain through features extracted from domain names. In a study, [18] aim to use a new metric to evaluate real unbalanced traffic data. Their experimental result shows that the level of the precision model and the value of the area under the curve (AUC) reach a certain maximum height. In [19], a system was developed that detected the feasibility of MDN account in relation to block chain so as to know whether it is malicious or not. They also use numerous features such as DN string based, DNS query based, DNS graph based and temporal aspect based extracted from domain names.

3. METHODOLOGY

Our proposed approach is based on three features which we extract from domain name online repository categorizing them into three groups; lexical-based, DNS statistical-based and third party-based features.

3.1 Lexical-based features

The lexical features ensure that MDNs can be detected using various features. In this work, we have extracted fourteen features from each domain details as represented in table 1:

Table 1: List of DNS-based features

Feature	Feature name	Description
0	**DNS_ID**	**Identifier of the DNS**
Lexical		
1	Subdomain	Has sub-domain or not
2	TLD	Top-level domain
3	SLD	Second-level domain
4	Len	Length of domain and subdomain
5	Numeric percentage	Counts the number of digits in the domain and subdomain
6	Character distribution	Counts the number of each letter in the domain
7	Entropy	Entropy of letter distribution
8	1-gram	1-gram of the domain in letter level
9	2-gram	2-gram of the domain in letter level
10	3-gram	3-gram of the domain in letter level
11	Longest word	Longest meaningful word in SLD
12	Distance from bad words	Computes average distance from bad words
13	Typos	Typosquatting
14	Obfuscation	Max value for URL obfuscation

3.2 DNS Statistical-Based Features

Statistical-based features were extracted based on the arrangement of DNS information in a distinct casement. However, these types of features are statistical information evaluated from the line section of the DNS feedback. Seven features were extracted from each domain as shown in in table 2.

Table 2: List of DNS statistical-based features

DNS statistical		
1	Unique country	The number of distinct country names in the window
2	Unique ASN	The number of distinct ASN values in the window
3	Unique TTL	The number of distinct TTL values in the window
4	Unique IP	The number of distinct IP values in the window
5	Unique domain	The number of distinct domain values in the window
6	TTL means	The average TTL in the window
7	TTL variance	The variance of TTL in the window

3.3 Third-party Features

Twelve features of the third party were extracted from two sources; WHOIS and Alexa rank. They contain the biographical properties of a domain as shown the table 3.

Table 3: Third-party Features Extracted

1	Domain name	Name of the domain
2	Registrar	Registrar of the domain
3	Registrant name	The name of the domain has been registered
4	Creation date time	The date and time the domain created
5	Emails	The emails associated with a domain
6	Domain age	The age of a domain
7	Organization	What organization it is linked to
8	State	The state the main branch is
9	Country	The country where the main branch is
10	Name server count	The total number of name servers linked to the domain
11	Alexa rank	The rank of the domain by Alexa
12	Status	The class of the DNS-Benign or Malicious

The software methodology adopted in this study is the Cross Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM is both an industry-proven methodology and a process model. As an industry methodology, it provides a concrete description of typical project stages tasks associated with each stage as well as the details of the interrelationship between the various tasks. As a process model, it provides a sketch that shows the data mining. In this study, CRISP-DM undergoes five stages include; (i) Business and data understanding, (ii) Data preparation, (iii) Modeling, (iv) Evaluation, and (v) Deployment. Figure 1 shows the proposed architecture, adopting the CRISP-DM methodology.

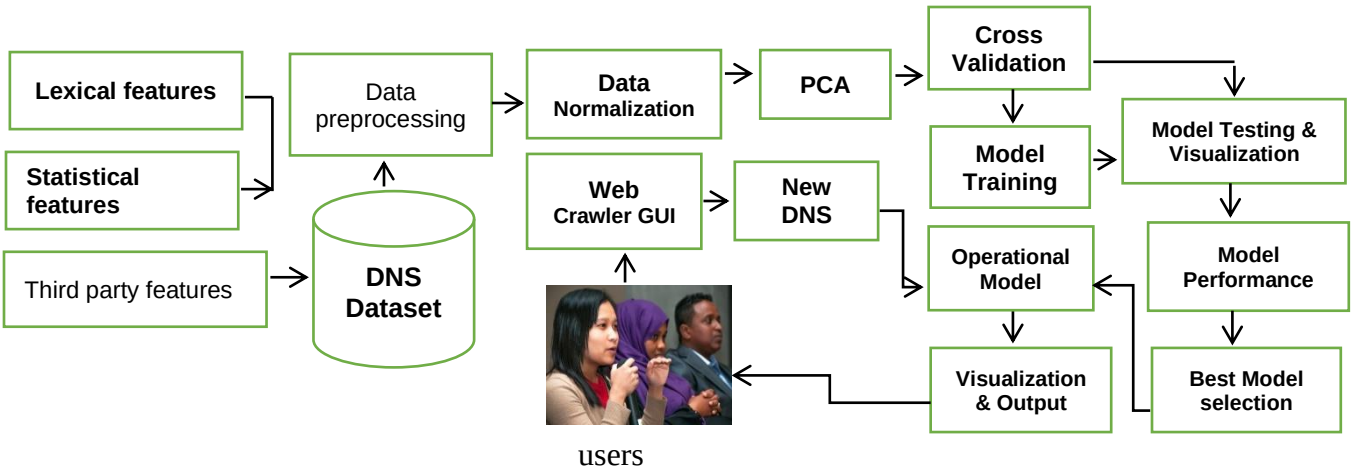

Figure 1: Proposed Malicious DNS prediction system

3.4 Data Preparation (DNS Dataset Description)

The CICBellDNS2021 was extracted at the Canadian Institute for Cyber security. The dataset compromises malware, spam, phishing, and benign URLs stored in separate comma-separated value (CSV) files (See table 4). The Correlation Matrix Analysis and Principal Component Analysis were then introduced to determine the relationship that exist among features in the dataset so that the most highly correlated features with predictor will be considered relevant for model building.

Table 4: Statistics of the domains dataset

Category	Original domains	Domains processed	Packets processed	Size in megabytes
Malware	26,895	9,432	182,266	2.7
Spam	8,254	1,976	61,046	2.4
Phishing	16,307	12,586	95,492	2.8
Benign	988,667	500,000	6,907,719	266

4. RESULTS AND DISCUSSION

Figure 2 shows the datasets consisting of 16 attributes. Some features were not correlated, these features were dropped to improve model performance. Hence, a total of 16 predictor features and one target feature (status) were used in the dataset for the model-building phase. For the model to perform better, the epoc value was increased from 5 to 10 as shown in figure 2. Figure 3 shows that all 16 features extracted from the 16 features of the data preprocessing significantly contribute to the variance in the dataset, hence they were used for model training with the help of Epoc values to determine the number of iterative or training that the model will perform.

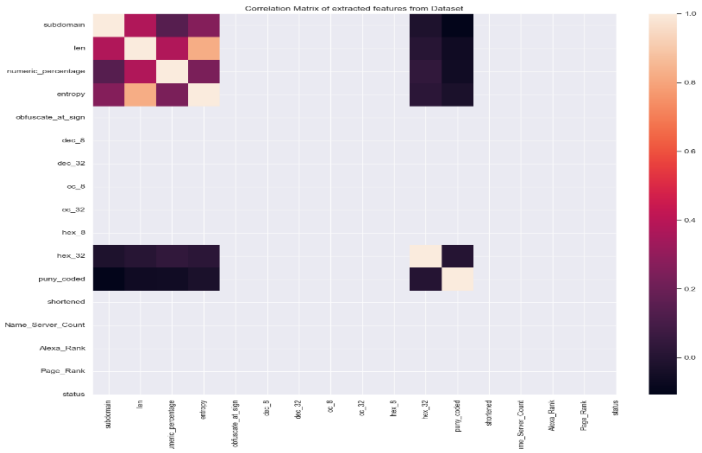

Figure 2: Correlation Matrix for dataset

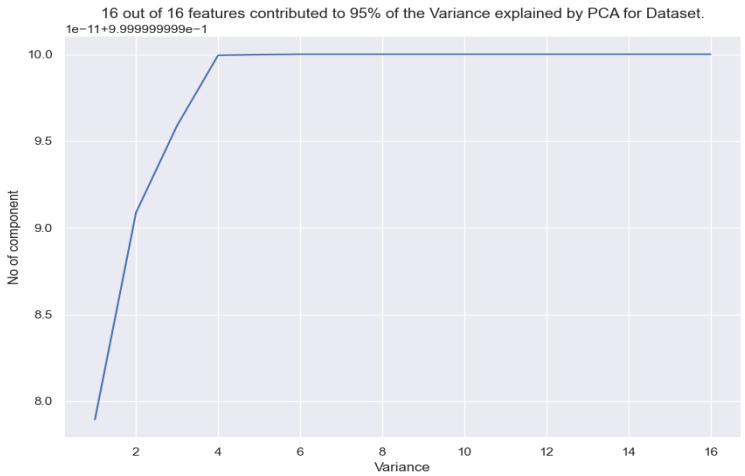

Figure 3: PCA diagram at 0.95% for dataset

4.1 Model Evaluation and Performance

Using the testing dataset, the trained models were evaluated. The models predict the labels of the test samples, which are then compared to the true labels to determine their performance. The evaluation of various models were conducted using several performance and validity metrics. We only focus on 5-fold cross validation classification and regression metrics. In table 5 shows the mean of the model's 5-fold cross validation classification metrics. In the EL models categories, RF had the highest accuracy of 89.9%. In ML models, DT had the highest accuracy of 86.9%. In the DL models, GRU had the highest accuracy score of 77.2%. Hence, RF (Ensemble Learning) had the highest accuracy score of 89.9% at 5-fold cross validation approach and was considered appropriate for the best detection model for predicting malicious DNS.

Table 5: Model Regression Metric using 5-fold cross validation approach

Algorithm	R^2	MSE	RMSE
DT	0.48	0.13	0.36
LogR	-0.38	0.34	0.59
SVM	-0.38	0.34	0.59
ANN	-0.36	0.34	0.58
KNN	0.41	0.15	0.38
RF	0.60	0.10	0.32
Xgboost	0.34	0.16	0.41
MLP	-0.08	0.27	0.52
DNN	-0.10	0.27	0.52
GRU	0.08	0.23	0.48
LSTM	-0.05	0.26	0.51

For the root mean square error, we choose the best model by identifying the least value in the evaluation metrics. Hence, in the EL models categories; RF had the lowest RMSE of 0.60. In ML models, DT had the lowest RMSE of 0.48. In the DL models, GRU had the lowest RMSE of 0.08. Hence, GRU has the least error rate and it is considered best for the detection of malicious DNS in terms of reduce error rate of detecting malicious DNS.

5. CONCLUSION AND FUTURE WORK

This study demonstrates the development of three modeling approaches for detecting malicious DNS with the ability to quantify the uncertainty in the prediction or detection. The modeling approach consists of various classifiers which were evaluated in order to select for the best performance. Also, the models were evaluated using 5-fold cross-validation to ensure that they were exposed to all data and to detect potential over-fitting a procedure frequently used within. The models obtained produce a good fit of the experimental data with an accuracy of 89.9%.

In our future work, we hope to optimize the best selected models for real-time malicious DNS detection by implementing it as a web browser plugin in and also to use another combine models for accuracy performance.

REFERENCES

[1] Li, K., Yu, X., & Wang, J. (2021) A Review: How to Detect Malicious Domains. In *Advances in Artificial Intelligence and Security: 7th International Conference, ICAIS Dublin, Ireland, July 19-23, 2021, Proceedings, Part III 7* (pp. 152-162). (2021) Springer International Publishing.

[2] Hamroun, C., Amamou, A., Haddadou, K., Haroun, H., &Pujolle, G.(2022.) A review on lexical based MDN detection methods. *6th Cyber Security in Networking Conference (CSNet)* (pp. 1-7). IEEE.

[3] Egwali A. O. and Alile S. O. (2020). A Casual Network Based System For Predicting Multi-Stage Attack with Malicious IP. International Journal of Academic Multidisciplinary Research 4 (5), 1-8.

[4] Alile S. O. and Egwali A. O. (2020). A Bayesian Belief Network Model For Detecting Multi-stage Attacks With Malicious IP Addresses. I.J. Wireless and Microwave Technologies, 2, 30-41

[5] Bilge L., Kirda, Kruegel E., Balduzzi C. (2011). EXPOSURE. Finding Malicious Domains Using Passive DNS Analysis. In Proceedings of the 18th Network and Distributed System Security Symposium, San Diego, CA, USA, 6 February 2011; Internet Society: Reston, VA, USA, PP. 1-17.

[6] Hong Zhao, Zhaobin Chang, GuangbinBao, and Xiangyan Zeng.(2019). MDNs detection algorithm based on N-Gram. Journal of Computer Networks and Communications, Volume 2019, Article ID 4612474, page 9 Publish 3 February 2019. Guest Editor: Saman S. Chaeikar.

[7] Halgamuge, M. N. (2022). Estimation of the success probability of a malicious attacker on blockchain-based edge network. *Computer Networks*, *219*, 109402.

[8] Ren, F., Jiang, Z., Wang, X., & Liu, J. (2020). A DGA domain names detection modeling method based on integrating an attention mechanism and deep neural network. *Cybersecurity*, *3*(1), 1-13.

[9] Ayub, M. A., Smith, S., Siraj, A., & Tinker, P. (2021, June). Domain Generating Algorithm based Malicious Domains Detection. In *2021 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)* (pp. 77-82). IEEE.

[10] Sachenko, B., Lysenko, S., Bobrovnikova, K., Savenko, O., &Markowsky, G. (2021, September). Detection DNS tunneling botnets. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (Vol. 1, pp. 64-69). IEEE.

[11] Ma D. J., Zhang S., Kong F., Fu Z.. (2021). MDN Detection Based on Doc2vec and Hybrid network. In IOP conference series: and Environmental Science. IOP publishing: Bristo, UK, Volume 693 [Cross Ref]

[12] Palaniappan, G., Sangeetha, S., Rajendran, B., Goyal, S., &Bindhumadhava, B. S. (2020). Malicious domain detection using machine learning on domain name features, host-based features and web-based features. *Procedia Computer Science*, *171*, 654-661.

[13] Kwon J., Lee J., Lee H., and perring A. (2016). "PsyBog:Psy Bog: a scalable botnet detection method for large scale DNS Traffic", Computer Networks vol.97, pp.48-73.

[14] Woodbridge J., Anderson H.S., Ahuja A., Grant D. (2016). Predicting Domain Generation Algorithms with Long Short-Term Memory Networks.arXi,v arXiv:1611.00791.

[15] Vinayakumar R., Soman K.P., Poornachandran P. (2018). Detecting MDNs using deep learning approaches at scale.J. Intell. Fuzzy System. 34, 1355-1367

[16] Zang x., Gong j., and Hu X.. (2018). “Detecting MDN based on AGD”, *Journal of International Technology*, vol.39, no.7. pp. 15-25.

[17] BharathiB.,andBhauvana J. (2019). Domain name detection and classification using deep neural networks. In international symposium on Security in Computing and Communication; Springer: Singapore.

[18] SuliangLuo., Gang Han., An Li., JianlangPeng. (2022).“Detecting MDNs from domain generation algorithms using bi-directional LSTM network” Proc. SPIE 12455, International Conference on Signal Processing and Communication Security (ICSPCS) Dalian China. doi: 10.1117/12.2655176

[19] Sachan, R. K., Agarwal, R., & Shukla, S. K. (2023). Identifying malicious accounts in blockchains using domain names and associated temporal properties. *Blockchain: Research and Applications*, 100136.