# Not all DGAs are Born the Same – Improving Lexicographic based Detection of DGA Domains through AI/ML

L. Torrealba Aravena*†, P. Casas*, J. Bustos-Jiménez†, G. Capdehourat‡, M. Findrik§
†NIC Labs, Universidad de Chile, *AIT Austrian Institute of Technology
‡Universidad de la República & Plan Ceibal, §cyan Security Group

***Abstract*—Timely identification of DNS queries to Domain Generation Algorithm (DGA) domains is crucial to limit malware propagation and its potential impact, particularly to prevent coordinated activities of botnets. We explore an approach for swift detection of DGA-generated domains by analyzing lexicographic features exclusively derived from the domain name as observed in a DNS query. We propose a reputation-based scoring system for domain names, based on the co-occurrence frequency of $n$-grams with respect to a list of well-known benign domains or whitelist. We further extract meaningful features from domain names and employ machine learning techniques to enhance detection performance. Experimental results on detecting 25 different families of DGA domains reveal that combining reputation scores with other basic lexicographic features largely outperforms current state of the art approaches.**

***Index Terms*—DGA Detection, n-grams, Lexicographic Analysis, DNS, Machine Learning.**

## I. Introduction

Domain Generation Algorithms (DGAs) have become prevalent in malware to establish and maintain a Command and Control (C&C) infrastructure. Botnets heavily rely on C&C servers to coordinate bots, i.e., compromised machines. To evade detection, botnets often employ DGAs that generate a diverse set of (quasi) random domain names based on a seed parameter, sometimes relying on pre-defined dictionaries [1]. By employing a shared algorithm, botmasters can register the C&C server on the network for a short duration with a randomly selected DGA domain name, allowing it to hide behind different domain names at different times. Detecting and neutralizing the C&C server domain name is therefore a key strategy to combat botnets.

We propose an approach to detection of DGA-generated domains by analyzing lexicographic features [2] derived exclusively from the domain name, as observed in the monitored DNS queries. This approach allows for efficient computation and large-scale monitoring, as the features are derived solely from processing the domain name, eliminating the need for external information sources. Moreover, by not accessing the content of a domain itself, privacy preservation for end-users is significantly enhanced.

Lexicographic analysis, and in particular $n$-grams, have been largely explored for classification of domain names. In the realm of machine learning for domain name analysis, various approaches have been investigated [3], [4], [5], including Random Forest models, XGBoost classifiers, and CNNs. These models leverage a combination of lexicographic-based, content-based, and other relevant features to achieve high accuracy in identifying malicious domains. The usage of $n$-grams in self-explainable approaches, without relying on learning models, is also part of the state of the art [6], [7].

Following state of the art [7], our method employs a reputation-based scoring system for domain names, utilizing the co-occurrence frequency of $n$-grams as compared to a whitelist of well-known benign domains. To identify DGA domains, we segment the domain name into $n$-grams of varying lengths and use them to calculate a reputation or similarity score against the list of $n$-grams obtained from the whitelisted domains. The resulting reputation score is finally compared against a predefined detection threshold for DGA/non-DGA binary classification. Additionally, we extract meaningful lexicographic features from domain names and leverage machine learning techniques to enhance the effectiveness of detection. Lexicographic features encompass various characteristics, including domain length, the randomness of the characters, and the frequency of $n$-grams, the latter by using the computed reputation scores.

We evaluate the performance of the different proposals on a publicly available dataset of domains names, consisting of a list of well-known domains (considered as benign) and a list of DGA-generated domains, using 25 different DGA families [1]. Results show that: (i) $n$-gram-based reputation scores can discriminate between DGA and non-DGA domains for different DGA algorithms; (ii) while the proposed reputation score significantly improves detection over state of the art [7], if fails to properly detect DGAs when these are generated through dictionary-based approaches; (iii) detection performance can be highly improved by using even simple learning models, combining reputation scores with other lexicographic features. This paper builds upon *PHISHWEB*, our recent work on web phishing detection [8], with a particular focus on DGAs and learning-based detection.

## II. DGA Detection with RVP and RF3

The proposed DGA detector is a variation of a previously proposed algorithm for detection of malicious domains [7], which is based on the computation of a *reputation score* or value RV for the domain under analysis. In a nutshell, a domain name is segmented in $n$-grams of different length $n$, and a score is computed based on the occurrence of these $n$-grams on a set of well-known benign names, serving as a

978-3-903176-58-4 

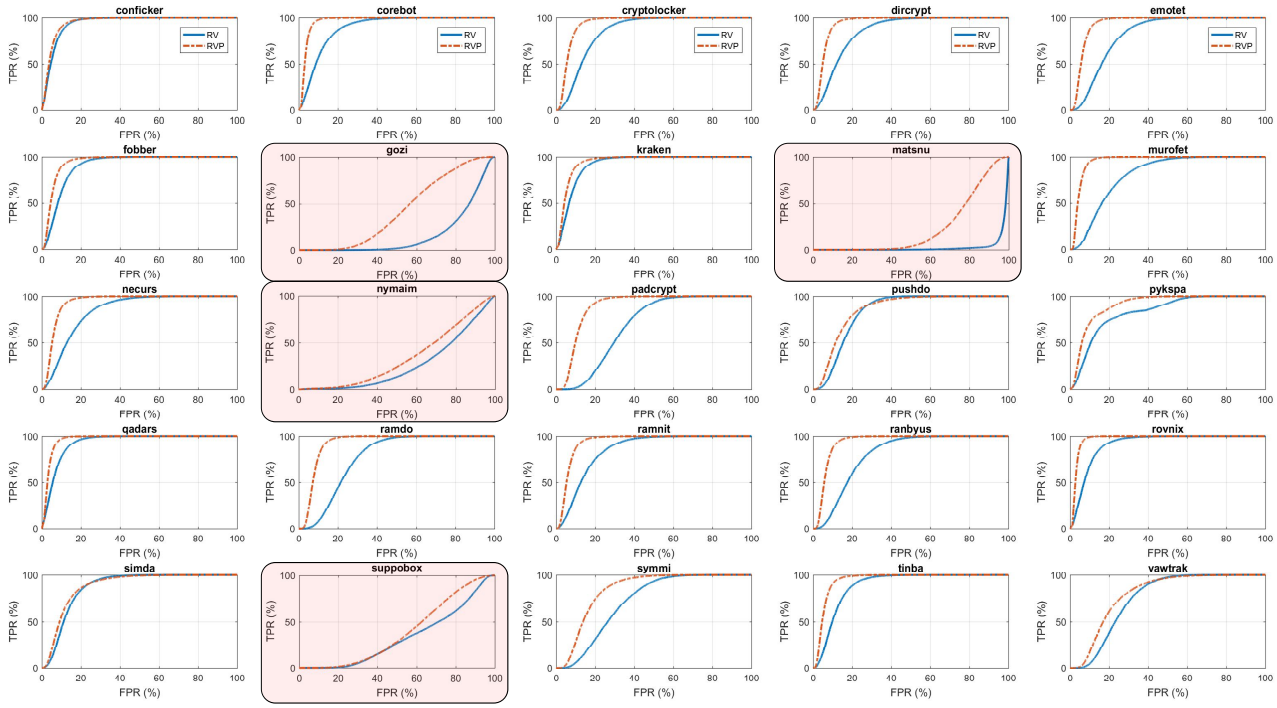


Figure 1. Individual DGA-algorithm detection performance for RV and RVP scores, reported as ROC curves. RVP clearly outperforms RV for all DGA families, but fails to properly detect dictionary-based DGAs: `gozzi`, `suppobox`, `nymaim`, and `matsnu`.

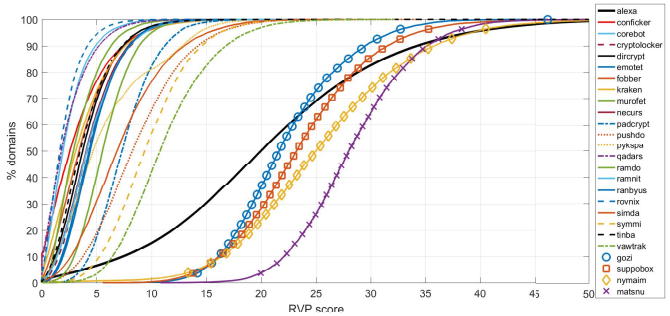


Figure 2. RVP scores for benign domains (`alexa`) and dictionary-based DGAs strongly overlap, severely impacting detection performance.

reputation list. We refer to our version of the reputation score as Reputation Value *PHISHWEB* or RVP. For a certain domain $d$, RVP is defined as follows:

$$\text{RVP}(d) = \sum_{i=1}^{m} \frac{n}{\lambda_d} \times W_n(i) = \sum_{i=1}^{m} \frac{n}{\lambda_d} \times log_2 \left( \frac{N_n(i)}{n} \right)$$

where $m$ is the total number of $n$-grams derived from $d$, for $n = 3$ to 7, $\lambda_d$ is the length of $d$, and $N_n(i)$ is the total number of occurrences of $n$-gram $i$ in the reputation list of domains. The idea of the reputation value is to reflect how similar are the sub-strings of domain $d$ to the list of *benign* sub-strings in this list. The higher the value of RVP, the higher the chances of $d$ being a benign domain. Detection is achieved by simple thresholding on RVP. As we show in the results, RVP drastically improves detection performance over the former score RV [7], which is strongly biased by the length of a domain, limiting its usefulness in the practice.

While RVP provides highly accurate DGA detection performance, results can be further boosted by combining RVP with a small set of simple lexicographic features, through a machine learning driven approach. In particular, we consider a 3-tuple to characterize a domain $d$, including its RVP score RVP($d$), its length $\lambda_d$, and a measure of the randomness of its characters $H(d)$, and use it as input to a standard random forest (RF) model, trained for binary classification. $H(\cdot)$ corresponds to the empirical entropy of the characters composing the domain name. We refer to this detector as RF3.

## III. Detection Results: RV vs. RVP vs. RF3

We use a publicly available DGA benchmark [1] for evaluation purposes, consisting of domains generated by 25 different DGA families from the Netlab Opendata Project repository (`https://data.netlab.360.com/dga/`), and using Alexa as an authoritative source for benign domain names. The dataset contains top-337.500 Alexa domains as whitelist, and 13.500 DGA domains per different family, resulting in a total of 675.000 domains, 50/50 balanced.

Fig. 1 compares the performance of RV [7] and our RVP reputation scores, for each individual DGA algorithm $i$ – i.e., `alexa` vs $\texttt{DGA}_i$. Firstly, RVP clearly outperforms RV for all DGA families, significantly reducing false alarms for higher detection rates. Still, RVP fails to properly detect so-called dictionary-based DGAs, referring to DGAs which use pre-defined lists of words which are similar to those used in standard domains. The list includes `gozzi`, `suppobox`, `nymaim`, and `matsnu`. Fig. 2 clearly shows how the empirical

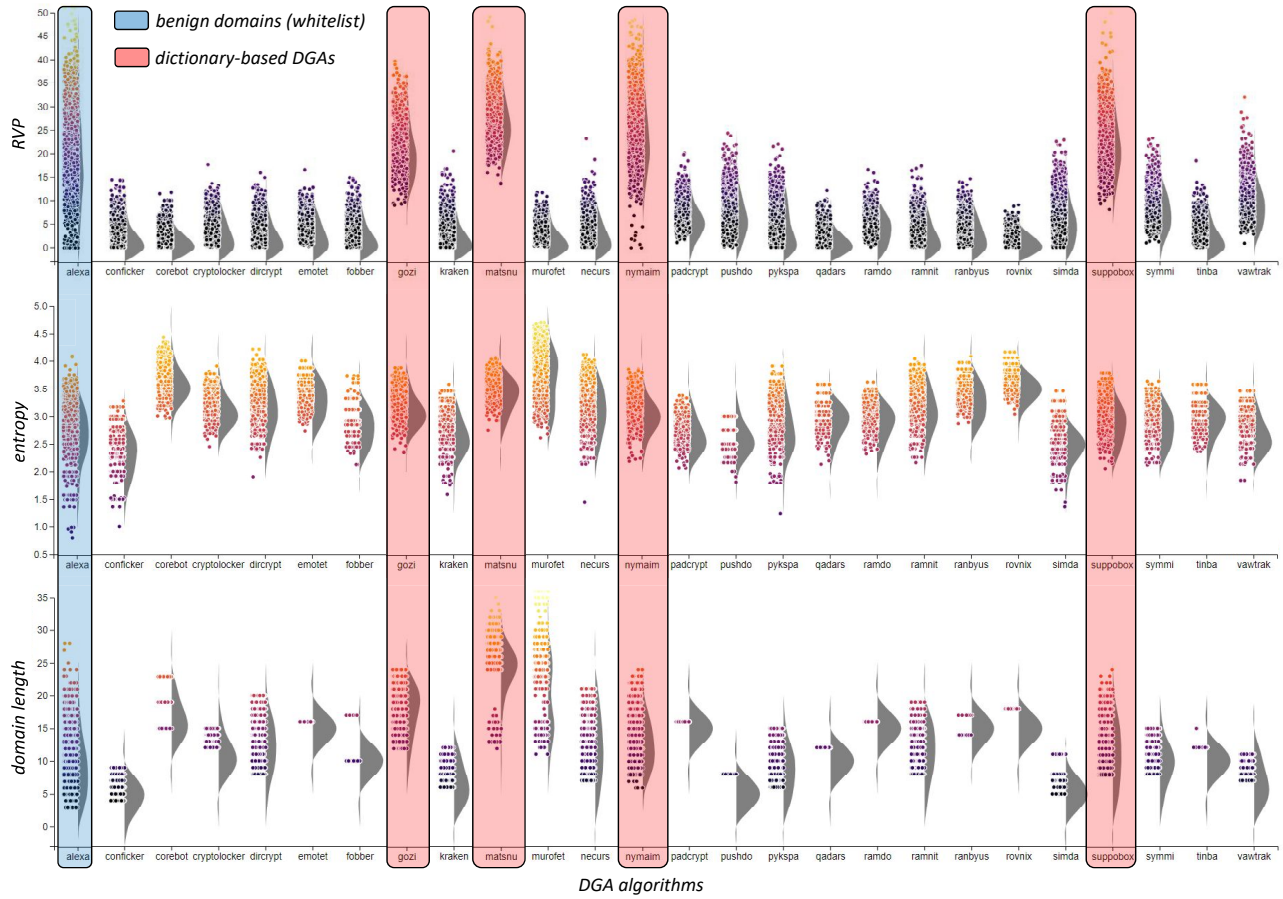


Figure 3. Empirical distribution of the input features used by RF3, for benign domains (`alexa`) and per DGA type. For ease of comparison, dictionary-based DGA algorithms are marked with boxes. DGAs such as `emotet`, `padcrypt`, or `pushdo` generate domains of constant length.

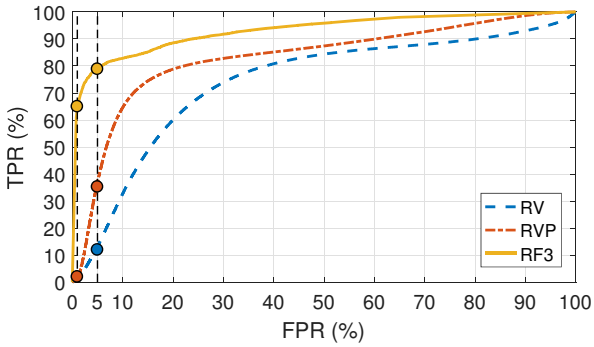


Figure 4. DGA detection performance. For the same false alarms' rate, RF3 largely outperforms RVP, particularly for FP ratios below 1%.

distributions of RVP values for benign domains (`alexa`) and dictionary-based DGAs strongly overlap, severely impacting detection performance.

Luckily, results can be significantly improved with RF3, combining RVP($d$) scores with domains length $\lambda_d$ and empirical entropy $H(d)$. Fig. 3 depicts the empirical distribution of the input features used by RF3, for benign domains (`alexa`) and per DGA type. Dictionary-based DGA algorithms are marked with boxes. Interestingly, DGAs such as `emotet`, `padcrypt`, `pushdo`, and others generate domains of constant length, and entropy values are higher for `corebot`, `murofet`, and `rovnix`. We train RF3 through standard five-fold cross validation, and evaluate the normalized importance of each of the input features on the classification output: while RVP clearly stands out, with a normalized impurity reduction of 72%, both $\lambda_d$ and $H(d)$ capture almost 30% of the impurity decrease, evidencing how relevant they are to improve the classification power of RF3 - the same conclusions are drawn through feature permutation based importance assessment. In particular, see how both $\lambda_d$ and $H(d)$ introduce higher heterogeneity to the description of a domain name $d$ as compared to RVP – cf. Fig. 3.

Finally, Fig. 4 reports the overall DGA detection performance for RV, RVP, and RF3 in the complete dataset. RF3 largely outperforms both RV and RVP, particularly for false alarm rates below 5%. For a FPR of 5%, RF3 detects 80% of the DGA domains, falling to 35% and 12% for RVP and RV, respectively. For a more applicable FPR of 1%, RF3 detects 65% of the DGA domains, whereas RVP and RV detect only 2% of them. These results confirm that a simple approach such as RF3 can significantly boost the descriptive properties obtained with RVP, combining reputation scores with other lexicographic features for better detection performance.

## ACKNOWLEDGMENT

This work has been partially supported by the Austrian FFG ICT-of-the-Future project *DynAISEC – Adaptive AI/ML for Dynamic Cybersecurity Systems* – ID 887504.

## REFERENCES

[1] A. Cucchiarelli, C. Morbidoni, L. Spalazzi, and M. Baldi, "Algorithmically Generated Malicious Domain Names Detection based on N-grams Features," *Expert Systems with Applications*, vol. 170, p. 114551, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420311957

[2] C. M. R. d. Silva, E. L. Feitosa, and V. C. Garcia, "Heuristic-Based Strategy for Phishing Prediction: A Survey of URL-Based Approach," *Comput. Secur.*, vol. 88, no. C, jan 2020. [Online]. Available: https://doi.org/10.1016/j.cose.2019.101613

[3] M. Korczynski, M. Wullink, S. Tajalizadehkhoob, G. C. M. Moura, A. Noroozian, D. Bagley, and C. Hesselman, "Cybercrime after the sunrise: A statistical analysis of dns abuse in new gtlds," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ser. ASIACCS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 609–623. [Online]. Available: https://doi.org/10.1145/3196494.3196548

[4] P. Mowar and M. Jain, "Fishing out the Phishing Websites," in *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 2021, pp. 1–6.

[5] M. Korkmaz, E. Kocyigit, O. K. Sahingoz, and B. Diri, "Phishing Web Page Detection Using N-gram Features Extracted From URLs," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2021, pp. 1–6.

[6] H. Zhao, Z. Chang, W. Wang, and X. Zeng, "Malicious Domain Names Detection Algorithm Based on Lexical Analysis and Feature Quantification," *IEEE Access*, vol. 7, pp. 128 990–128 999, 2019.

[7] H. Zhao, Z. Chang, G. Bao, and X. Zeng, "Malicious Domain Names Detection Algorithm Based on N-Gram," *J. Comput. Networks Commun.*, vol. 2019, pp. 4 612 474:1–4 612 474:9, 2019.

[8] L. Torrealba Aravena, J. Bustos-Jiménez, and P. Casas, "PHISHWEB: A Progressive, Multi-Layered System for Phishing Websites Detection," in *Proceedings of the 22nd ACM Internet Measurement Conference*, ser. IMC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 764–765. [Online]. Available: https://doi.org/10.1145/3517745.3563028