IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204348 |

DGA Detection Leveraging ML Techniques

Akash S Shinde[1],Brahmtej B Bargali[2], Sohan S P[3], Sushruth E S[4] , Dr. Gururaj T[5], Prof. Vaishnavi A I[6]

U.G. Student, Department of Computer Science Engineering, Bapuji Institute Engineering and Technology, Shamanur Rd, Davanagere, Karnataka, India[1]

U.G. Student, Department of Computer Science Engineering, Bapuji Institute Engineering and Technology, Shamanur Rd, Davanagere, Karnataka, India[2]

U.G. Student, Department of Computer Science Engineering, Bapuji Institute Engineering and Technology, Shamanur Rd, Davanagere, Karnataka, India[3]

U.G. Student, Department of Computer Science Engineering, Bapuji Institute Engineering and Technology, Shamanur Rd, Davanagere, Karnataka, India[4]

Associate Professor, Department of Computer Science Engineering, Bapuji Institute Engineering and Technology, Shamanur Rd, Davanagere, Karnataka, India[5]

Assistant Professor, Department of Computer Science Engineering, Bapuji Institute Engineering and Technology, Shamanur Rd, Davanagere, Karnataka, India[6]

ABSTRACT:This project presents a machine learning approach for detecting domain generation algorithm (DGA) traffic using n-grams and the Random Forest algorithm. The proposed method captures the frequency of character sequences in the domain names to identify malicious domains generated by DGAs. The system can also block suspicious hosts by filtering their packets.

KEYWORDS:N-gram, Random Forest, DGA(Domain Generated Algorithm) and Scapy.

I. INTRODUCTION

With the ever-increasing number of devices with access to the Internet including servers, home computers, phones, and others. This gives rise to many risks that can be exploited by malicious actors actively scanning the Internet for vulnerable hosts or trying to convince a user to download malware onto their hosts. Once installed on a host, the malware can be used to perform various unauthorized or harmful tasks, such as eavesdrop on communication, launch denial of service attacks, steal private information, send phishing emails, encrypt data, and request a ransom for its decryption, amongst many others. In this regard, one of the most critical parts of modern malware campaigns is the mal-ware's communication to a Command and Control (C&C) server to receive instructions on what task to perform.In the past, the malware had the location of the C&C server hard-coded into its binaries as domains or IP addresses. However, this strategy can easily be detected and prevented by malware analysts uncovering the domains or IP addresses and blacklisting them from being used. The term "in flux" refers to something that is constantly changing, the malware, in this case, evades detection and blacklisting by constantly having the domain registered to the C&C server IP address "in flux" to that generated by a DGA. Traditional methods to stop malware became much less effective with DGA present, leading researchers to leverage machine learning to train complex models that can detect patterns of DGA creation by processing large quantities of legitimate domains and DGA domains. The resulting models make decisions based purely on the domains that a host is querying and determines whether it is created using DGA in real-time.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204348 |

II. RELATED WORK

[1] Zhiqiang Wang, Hongyu Dong, Yaping Chi, Jianyi Zhang, Tao Yang, QixuLiu,"DGA and DNS Covert Channel Detection System based on Machine Learning", 2019In recent years, the application of covert channel to ensure network security has attractedmore and more attention from scholars and network attackers, but the research on its detection is relatively few. We propose a DGA and DNS covert channel detection system basedmachine learning, using improved TFIDF, specificity score and other algorithms to detectmalicious domain names.
[2] Ryan R Curtin, Andrew B. Gardner, Slawomir Grzonkowski, AlexyKleymenov,Alejandro Mosquera, "Detecting DGA Domains with Recurrent Neural Networks andSide Information", 2019Modern malware typically makes use of a domain generation algorithm (DGA) to avoidcommand and control domains or Ips being seized or sinkholed. This means that an infectedsystem may attempt to access many domains in an attempt to contact the command-and-control server. Therefore, the automatic detection of DGA domains is an important task,both for the sake of blocking malicious domains and identifying compromised hosts. However, many DGAs use English wordlists to generate plausibly clean-looking domain namesthis makes automatic detection difficult. In this work, we devise a notion of difficulty forDGA families called the smashword score; this measures how much a DGA family lookslike English words.

III. METHODOLOGY

1. Data Collection: The first step of the DGA detection process is to collect DNS traffic data from various sources. This data can be collected from DNS server logs, network packets, or other sources that provide DNS traffic information.
2. Data Preparation: After collecting the DNS traffic data, the next step is to prepare the data by filtering out irrelevant traffic and extracting relevant features. This step involves identifying the relevant features such as domain names, IP addresses, packet sizes, and protocol types that will be used in the DGA detection process.
3. Feature Extraction: Feature extraction involves generating n-grams from the domain names in the DNS traffic data. N-grams are contiguous sequences of n items from a given sample of text or speech. In this step, n-grams serve as features for the Random Forest model. The n-grams can be unigrams, bigrams, or trigrams, depending on the requirements of the model.
4. Model Training: In this step, the Random Forest model is trained on the extracted features. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The model is trained on a training dataset consisting of labeled data, i.e., data that is labeled as either DGA or non-DGA.
5. Model Testing: After training the model, it is tested on a separate dataset to evaluate its accuracy in detecting DGAs. This step involves testing the model's ability to correctly classify new data that it has not seen before. The testing dataset is also labeled data, and the model's performance is evaluated based on metrics such as accuracy, precision, and recall.
6. Live Traffic Analysis: Once the model is trained and tested, it can be used to analyze DNS traffic in real-time. This step involves using Scapy, a Python-based packet manipulation tool, to analyze DNS traffic in real-time. Scapy allows for the creation, manipulation, and decoding of network packets. The live DNS traffic data is analyzed to detect any DGAs present in the traffic.
7. DGA Detection: The final step of the DGA detection process is to apply the trained Random Forest model to the n-grams extracted from the live DNS traffic data using Scapy. If the model detects a DGA, appropriate action can be taken, such as blocking the domain or alerting the security team. This step helps in preventing potential cyber threats and protecting the network from malicious activities.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204348 |

III. EXPERIMENTAL RESULTS

Figures shows the results of DGA detection based on ML techniques using Ngram, Random Forest and Scapy algorithm. Fig(a) shows the retrieval of datasets and perform data processing. Fig(b) shows the training model.Fig(c)shows the loading of training model and scanning of the host network traffic using the network interface specified by the user.

```
(env) debian@debian:~/code/dga_prediction_model$ python initializer.py
----------------------------- MENU -----------------------------

    1. Get training data
    2. Train model
    3. Start capture network traffic
    4. Exit

----------------------------------------------------------------
----------------------------------------------------------------
Enter your choice [1-4]: 1
----------------------------------------------------------------
[*] Getting training data...
[*] Reduction to a unified form...
[*] Saving training data to disk...

Press Enter to continue.
```

Fig (a) shows the retrieval of datasets and perform data processing

```
(env) debian@debian:~/code/dga_prediction_model$ python initializer.py
----------------------------- MENU -----------------------------

    1. Get training data
    2. Train model
    3. Start capture network traffic
    4. Exit

----------------------------------------------------------------
----------------------------------------------------------------
Enter your choice [1-4]: 2
----------------------------------------------------------------
[*] Loading training dataset from disk...
[*] Calculating length for each domain...
[*] Creating model to definite matrix of ngram counts...
[*] Counting the ngram occurrences of legit domains...
[*] Training model...
[*] Saving model to disk...

Press Enter to continue.
```

Fig (b) shows the training model

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204348 |

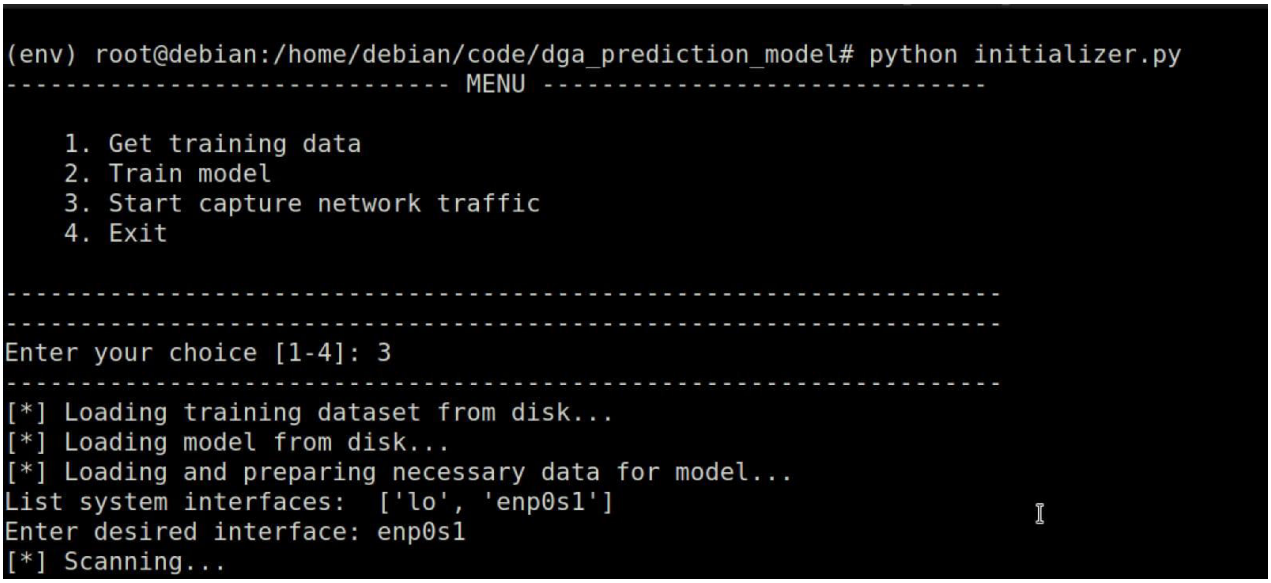

Fig (c)shows the loading of training model and scanning of the host network traffic using the network interface specified by the user

IV. CONCLUSION

In this project, we have demonstrated the effectiveness of a machine learning model based on frequency analysis of Ngrams and trained using a Random Forest algorithm in detecting DGA domains in DNS requests. The model was able to analyze the general active flow of DNS requests and identify suspicious domains that may be associated with DGA activity. Additionally, the model has the capability to block suspicious hosts by filtering their packets. Overall, our work demonstrates the potential of machine learning models to identify and mitigate threats in real-time and can be a valuable tool for network administrators looking to prevent DGA-related attacks on their systems.

REFERENCES

[1] Areej Alhogail, Isra Al-Turaiki, ”Improved Detection of Malicious Domain Names Using Gradient Boosted Machines and Feature Engineering”, 2022

[2] Nida Aslam, Irfan Ullah Khan, Samiha Mirza, AlanoudAlOwayed, Fatima M. Anis, Reef M. Aljuaid and RehamBaageel, ”Interpretable Machine Learning Models for Malicious Domains Detection Using Explainable Artificial Intelligence (XAI)”, 2022

[3] JuweriyaSolkar, Priyanka Bandagale, DaniyaSolkar, ”Detection and Identification of DGA Domains using Machine Learning”, 2021

[4] Carl-Simon Brandt, Jonathan Kleivard, Andreas Turesson, ”Convolutional, Adversarial and Random Forest-Based DGA Detection”, 2021

[5] Zheng Wang, ”Detecting Algorithmically Generated Domains Using a GCNN-LSTM Hybrid Neural Network”, 2021

[6] Al-Turaiki, I. and Altwaijry, N. A, ”Convolutional Neural Network for Improved Anomaly-Based Network Intrusion Detection.”, 2021

[7] lmashhadani, A. O., Kaiiali, M., Carlin, D., and Sezer, S. MaldomDetector, ”A system for detecting algorithmically generated domain names with machine learning. computers & Security”, 2020

[8] Shi, Y., Chen, G., and Li, J. ”Malicious Domain Name Detection Based on Extreme Machine Learning. Neural processing Letters”, 2018

[9] Chiba, D, Yagi, T., Akiyama, M., Shibahara, T., Yada, T., Mori, T., Goto, S., ”Discovering domain names abused in future”, 2016

[10] Mamum, M., Islam, S., Rathore, M., Lashkari, A., Stakhanova, N., and Ghorbani, ”Detecting malicious URLs using lexical analysis.”, 2016

[11] Choi, H., Zhu, B. B., and Lee, H. ”Detecting malicious web links and identifying their attack types. Proceedings of the 2nd USENIX conference on Web application develOpment”, 2011

[12] Ma, J., Saul, L., Savage, S., and Voelker, G. ”Learning to detect malicious URLs”, 2011

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details