Towards an explainable malicious URL classifier

Fabien Charmet[1], Tomohiro Morikawa[2], Takeshi Takahashi[1]
[1]National Institute of Information and Communications Technology
Tokyo, Japan
[2]University of Hyogo, Hyogo, Japan
email: fabien.charmet@nict.go.jp

1 Introduction

Artificial Intelligence (AI) is an indispensable tool for cybersecurity solutions, and various studies on the use of AI for attack detection have been reported until now. Using AI yields a better detection performance, however AI also suffers from the false positives generated in some cases. For example, SIEM appliances produce a tremendous amount of false positive alerts, which leads to the alert fatigue problem [1, 2]. To make the most out of AI in the cybersecurity field, techniques to efficiently deal with such false positives are needed.

Meanwhile, recent advances in the AI field have led researchers to investigate the capability for an AI model to explain the decisions and predictions it makes, also referred to as explainable AI (XAI). In the field of computer vision, Saliency maps [3] highlight the pixels that contributed the most. Likewise, in the cybersecurity field, multiple works on Intrusion Detection Systems (IDS) have been augmented with the LIME [4] or Shapley values [5] explanation methods. These methods determine which features in the input data contributed the most to classifying the sample as malicious or benign. However, the research is still in its early stages, and more applications should be able to benefit from XAI.

Among many cybersecurity use cases, in this paper, we work on the applicability of XAI for malicious URL detection. We present an ongoing work about the design of an explainable malicious URL classifier. Our work significantly diverges from the existing literature in that it tries to address the limitations of exploiting cybersecurity data while trying to classify it.

2 Our proposal

Most of the malicious URL classifiers use "traditional" AI algorithms, such as Support Vector Machines or Random Forests. To the best of our knowledge, only Hernandes et al. [6] have studied the application of LIME to Random Forest and SVM classifiers for malicious URL detection. We intend to push the design of our model one step further by extending the work from Pierazzi et al. [7], where

authors proposed a problem space attack that will generate malware samples using Adversarial Examples (AE) techniques, while evading the detection of state-of-the-art security classifiers. Authors also propose some formalization advice with regard to other topics (e.g., face recognition, Javascript code analysis etc.). We formulate the following research questions:

- How can we adapt the problem space attack from [7] so it is fit for malicious URLs instead of malware bytecode ?
- How can we prove that all theorems describing the security properties are maintained in our system ?
- Can we prove that these theorems are also applicable to a different type of AI models ?

Similarly to malware binaries, URLs can lose their "functional maliciousness" because of the modifications from AE techniques. This can be easily illustrated by removing the **.com** from **http://mymaliciouswebsite.com**, thus making the URL unreachable while still having it classified as malicious.

3 Methodology & Dataset

In order to answer the research questions we formulated, we divide our research into four steps:

First, we establish a baseline of URL classifiers. We will implement various models existing in the literature, while considering two types of features: natural and handcrafted. Natural features are simply the raw data representing the URL. For instance, we may convert each character into its one-hot encoding representation. Handcrafted features are NLP features and cybersecurity features that commonly appear in the literature.

Second, we investigate the feasibility of the problem space attack on our various classifiers. We will determine which type of features is the most fitting with regard to the formalization from [7].

Third, we evaluate the results of the maliciousness of the generated samples by using an external suite of security tools, such as VirusTotal [8].

Finally, we investigate the robustness of our classifier against Adversarial Examples techniques, in order to reinforce our classifier against various attacks. We aim to encompass the attack from [7] and other XAI attacks [9, 10].

The dataset we collected contains malicious and benign URLs. Malicious URLs have been collected by aggregating the output of multiple public blacklist providers. We have collected a total of 12 millions URLs using these blacklists. Benign URLs have been collected by using the top 1 million from Alexa [11]. We have preprocessed the dataset to extract 23 distinct NLP and cybersecurity features inspired from [12, 13]

References

[1] Muhamad Erza Aminanto et al. "Threat Alert Prioritization Using Isolation Forest and Stacked Auto Encoder With Day-Forward-Chaining Analysis". In: *IEEE Access* 8 (2020), pp. 217977–217986. DOI: 10.1109/ACCESS.2020.3041837.

[2] Ryosuke Ishibashi et al. "Generating Labeled Training Datasets Towards Unified Network Intrusion Detection Systems". In: *IEEE Access* 10 (2022), pp. 53972–53986. DOI: 10.1109/ACCESS.2022.3176098.

[3] Timor Kadir and Michael Brady. "Saliency, scale and image description". In: *International Journal of Computer Vision* 45.2 (2001), pp. 83–105.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[5] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[6] Paulo R Galego Hernandes et al. "Phishing Detection Using URL-based XAI Techniques". In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021, pp. 01–06.

[7] Fabio Pierazzi et al. "Intriguing Properties of Adversarial ML Attacks in the Problem Space". English. In: *2020 IEEE Symposium on Security and Privacy* (May 2020), pp. 1332–1349. ISSN: 2375-1207. DOI: 10.1109/SP40000.2020.00073.

[8] *VirusTotal*. 2022-05-27. URL: https://virustotal.com/.

[9] Aditya Kuppa and Nhien-An Le-Khac. "Adversarial xai methods in cybersecurity". In: *IEEE Transactions on Information Forensics and Security* 16 (2021), pp. 4924–4938.

[10] Aditya Kuppa and Nhien-An Le-Khac. "Black box attacks on explainable artificial intelligence (XAI) methods in cyber security". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.

[11] *Alexa Top 1M urls*. 2022-05-27. URL: http://s3-us-west-1.amazonaws.com/umbrella-static/index.html.

[12] Apoorva Joshi et al. "Using lexical features for malicious URL detection–a machine learning approach". In: *arXiv preprint arXiv:1910.06277* (2019).

[13] Ozgur Koray Sahingoz et al. "Machine learning based phishing detection from URLs". In: *Expert Systems with Applications* 117 (2019), pp. 345–357.