Contents lists available at Crown Academic Publishing

International Journal of Multidisciplinary Sciences and Advanced Technology

Journal home page: http://www.ijmsat.com

Formulation of Quick Response Code Dataset for Machine learning Analysis

S. O Subairu [1], J. K Alhassan [2], S.M Abdulhamid [3], J.A Ojeniyi [4]

[1] Cyber Security Science, Federal University of Technology Minna, Nigeria.
E-mail:-islam4life@futminna.edu.ng

[2] Computer Science, Federal University of Technology Minna, Nigeria.
E-mail:- jkalhassan@futminna.edu.ng

[3]Cyber Security Science, Federal University of Technology Minna, Nigeria.
E-mail:-shafii-abdulhamid@futminna.edu.ng

[4] Cyber Security Science, Federal University of Technology Minna, Nigeria.
E-mail:-ojeniyija@futminna.edu.ng

Corresponding Author: E:mail: islam4life@futminna.edu.ng

Original article

Received 6 March 2020, Accepted 17 March 2020, Available online 18 March 2020

ABSTRACT

Quick Response Code technology has made so easy many human digital transactions such as payment, authentication, advertisement, web navigation and others. This technology, despite being widely accepted because of its ease of creation, deployment and usage, has been recently a tool of personal identification theft in the hands of fraudster. Researchers in the area of application of machine learning to cyber security may find it difficulty sourcing QR code dataset. In order to fill this identified gap, a model was developed which incorporate data engineering principle to formulate QR code dataset in the form implementable on machine learning algorithm for analysis.

Keywords: Machine Learning, Quick Response Code, Dataset, cyber security.

1. Introduction

Machine Learning (ML) (Waller et al, 2013; Biamonte et al., 2017, Albzeirat et al., 2018) rely so much on data, It's the important factor which make possible the training of algorithm and this form the basis of the popularity of machine learning recently as it is being apply to many areas of scientific research. Dataset as the building block for effective machine learning analysis, sourcing it publicly would usually makes researcher work a bit easier. So many dataset has been formulated for different area of human discipline such as computing and other, which has led to great technological development.

2. Literature Review

Quick Response (QR) code is two dimensional matrix codes (Dabrowski, Krombholz, Ullrich, & Weippl, 2014) that has gained popularity in various human endeavor such as payment, authentication, web navigation, ticketing, access control (Yao & Shin, 2013). This technology is now being use for phishing and malware attacks (Tao, 2017).

Figure 1:QR Code Attack (Thompson & Lee, 2013)

There is a need for thorough research on this technology especially in the area of mitigation of attacks that could be easily launched using this technology as a vector. For researchers in the area of applying machine learning to security, will find it cumbersome when it comes to quick response code. Quick response code as a widely accepted technology for both consumers and provider of goods and services lacks publicly available dataset in the form that is implementable to machine learning algorithm. Formulation of this dataset will aid many researchers to effectively carryout research on this technology, hence this work aim to fill this research gap by formulating a quick response code dataset for machine learning analysis.

The quick response code dataset built was formulated for classes such as benign and phishing uniform resource locator [URL]. The URL used for the formulation of dataset was sourced from Phishtank, Openphish, Alto University Research, and University of New Brunswick URL databases. The total number of both benign and phishing URL gotten from these databases is fifty five thousand eight hundred, out of which twenty seven thousand, seven hundred and twenty (27,720), were selected base on features selected (Mohammad, Thabtah & McCluskey 2015)

3. Research and Methodology

In the formulation of quick response code dataset we developed a model which was used for the data collection, data engineering; which comprises of data wrangling, data cleansing and data preparation.

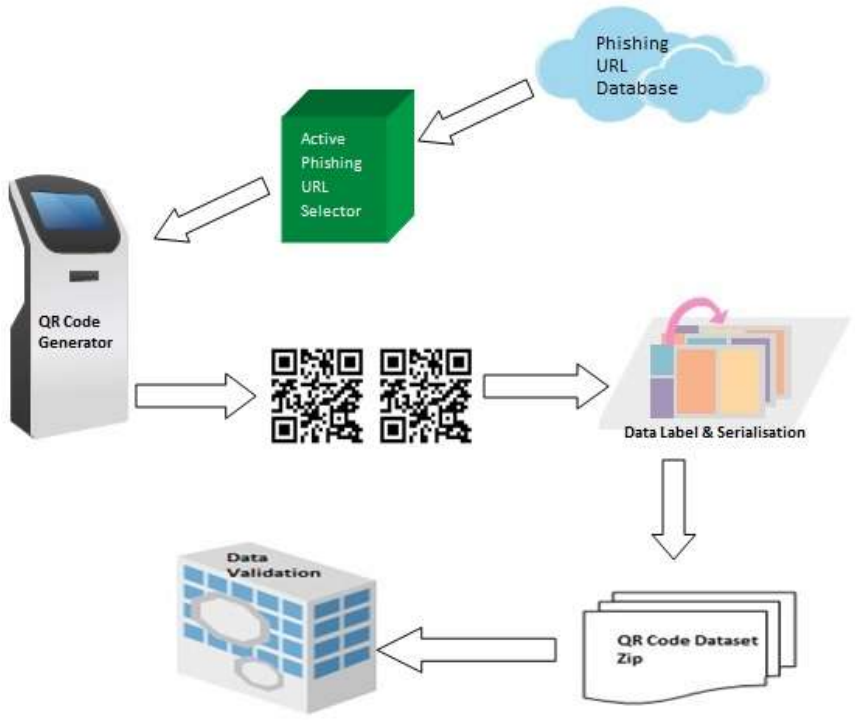

Figure 2: Dataset Formulation Process

As shown in figure 2, benign and phishing URL was downloaded from publicly available benign and phishing databases such as openphish and phishtank. Active phishing URL and benign URL was then examined, and selected based on features as documented in literatures such as (Mohammad *et al* 2015). Active phishing and benign URL was then fed into QR code generator; generated phishing and benign QR codes then renamed, resized and serialized using python script.

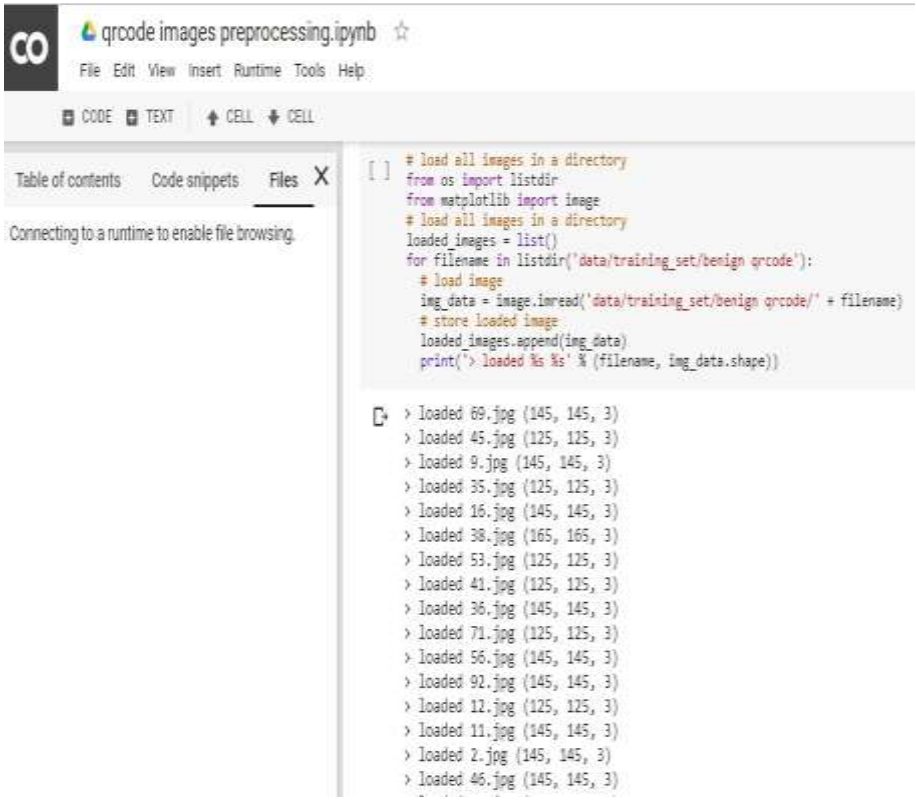

Figure 3: QR code dataset showing sizes variation

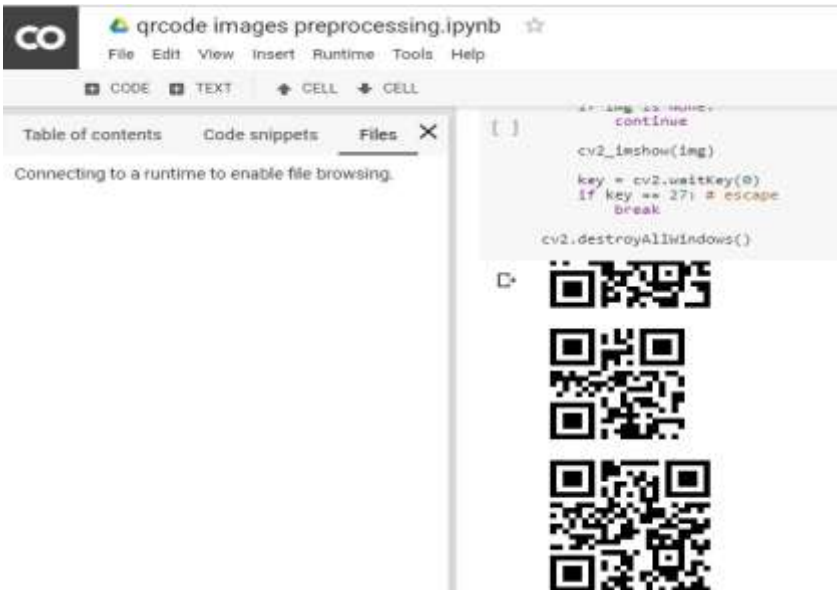

Figure 4: QR code dataset showing sizes variation

Both figure 3 and 4, shows the variation of QR code sizes after generation from selected URL.

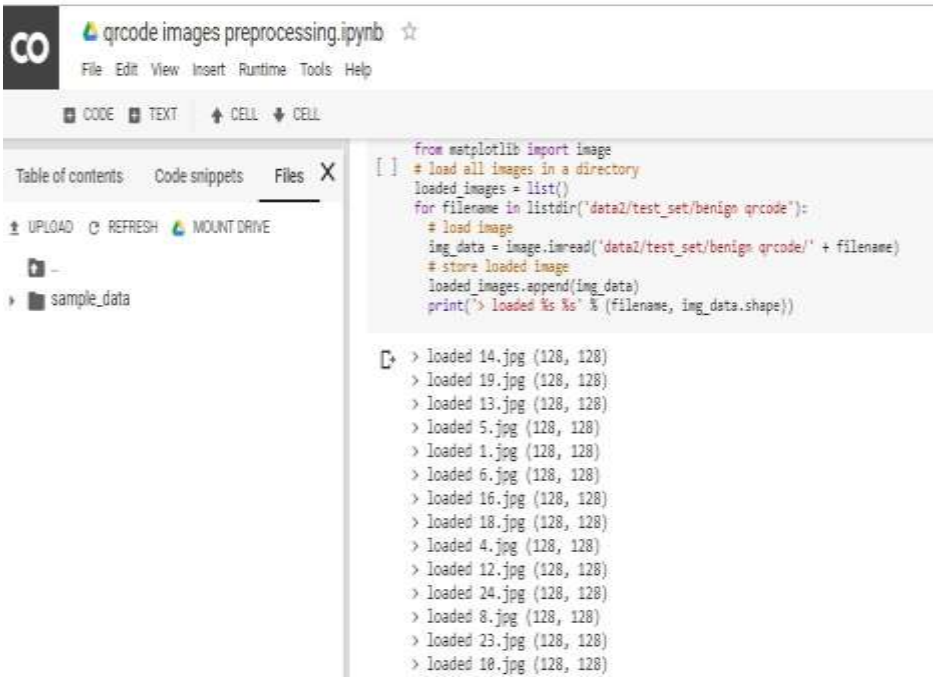

Figure 5: Dataset Resizing 128*128

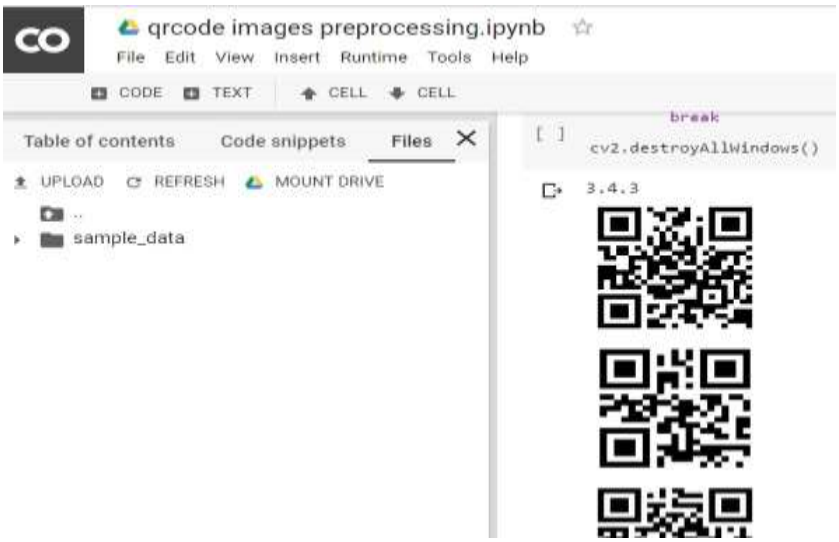

Figure 6: Dataset Resized

The resized QR code is shown vividly in figure 5 in dimension while figure 6 shows the pictorial view of the codes.

These processes was in line with data engineering; which is data wrangling being a process of identification of data, collection, merging, and preprocess a data sets in getting it ready for data cleansing.

Figure 7: Data Engineering

Data cleansing was then applied on our dataset to get rid of unwanted or incomplete data features. Likewise data that cannot be repair or having missing values are all removed during cleansing process so as to get the dataset correct both syntactically and semantically.

Data preprocessing as the final stage in the data engineering is to ensure that cleansed dataset are fully ready for machine learning algorithm preprocessing. This process could involve dataset normalization; which transform dataset input to an even distribution in acceptable range for the algorithm of machine learning. This process is usually done before feeding dataset into machine learning algorithm.

4. Result

The formulated Dataset has shown in figure 4, have been structured properly and ready to be implemented on a machine learning algorithm.

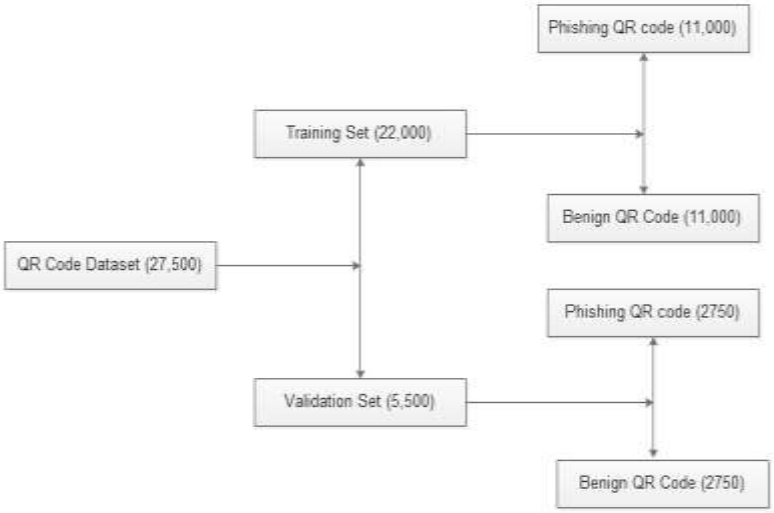

Figure 8: Quick Response Code Dataset Framework.

The quick response dataset has a total size of twenty seven thousand, five hundred (27,500), the training QR code is twenty two thousand (22,000), representing 80% of the entire dataset; which comprises of eleven thousand (11,000) each for both benign and phishing. The validation QR code is of the size of five thousand five hundred (5,500) representing 20% of the entire dataset; which comprises of two thousand seven hundred and fifty (2750) each for phishing and benign QR codes.

5. Conclusion

The QR code dataset formulated was first tested for normality, result shows it successfully passed the normality as shown in table 1 and figure 9. Then the dataset was equally tested on Google Colab and IBM Watson and result shown the dataset to be in order with data engineering concept, thus it is implementable on machine learning algorithm for various analysis.

Table 1: QRCODESTATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	11000	50.0	50.0	50.0
	1	11000	50.0	50.0	100.0
	Total	22000	100.0	100.0	

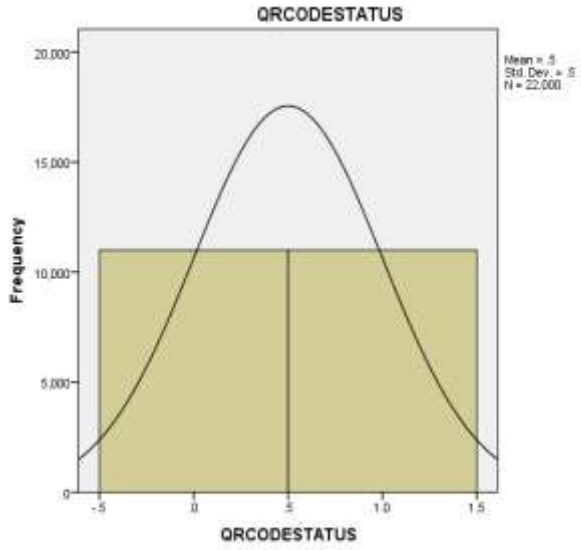

Figure 9: Graphical Normality test on Dataset

Preliminary usage of the formulated QR code dataset on the cloud shows promising result as shown in figures [9, 10 and 11]

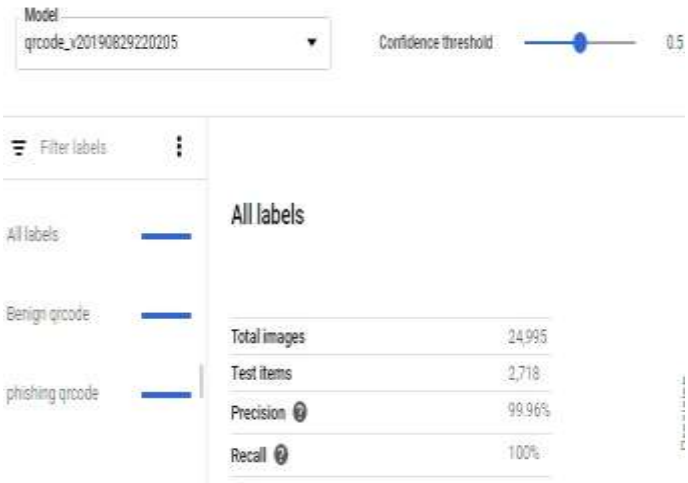

Figure 9 Showing QR code dataset labels

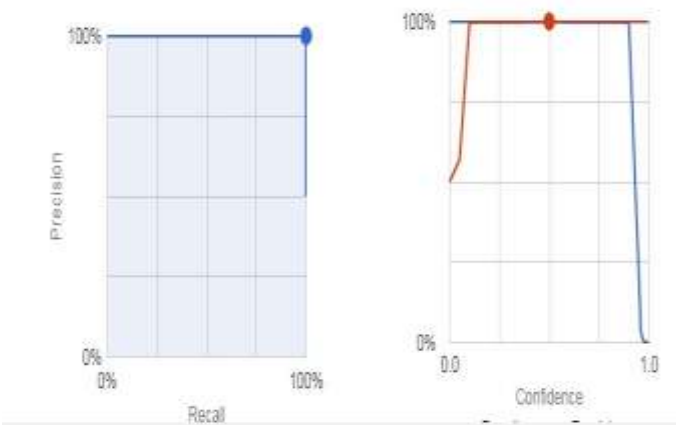

Figure 10 Showing QR code Model Recall & Confidence on QR code formulated dataset

Figure 11 Showing Confusion matrix of a model on QR code formulated

References

Albzeirat, M. K., Hussain, M. I., Ahmad, R., Al-Saraireh, F. M., & Ahmad, I. (2018). A novel mathematical logic for improvement using lean manufacturing practices. *Journal of Advanced Manufacturing Systems*, *17*(03), 391-413.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, *549*(7671), 195-202.

Dabrowski, A., Krombholz, K., Ullrich, J., & Weippl, E. R. (2014, November). QR inception: Barcode-in-barcode attacks. In *Proceedings of the 4th ACM workshop on security and privacy in smartphones & mobile devices* (pp. 3-10).

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review*, *17*, 1-24.

Tao, L. (2017). QR code Scams rise in China, Putting E-payment Security in spotlight. *South China Morning Post*.

Thompson, N., & Lee, K. (2013). Information security challenge of QR codes. *Journal of Digital Forensics, Security and Law*, *8*(2), 2.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77-84.

Yao, H., & Shin, D. (2013, May). Towards preventing qr code based attacks on android phone using security warnings. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security* (pp. 341-346).