# 도메인 이름 및 부가 정보를 활용한 기계학습 기반 보안 QR 코드 스캐너 개발

김보람, 권소연, 김유빈, 김은결, 이광재*

*상명대학교

{201921505, 202021241, 202021251, 202021252}@sangmyung.kr, begleam@smu.ac.kr*

# A Development of a Machine learning-based Security QR Code Scanner using Domain Name and its Side Information

Bo-Ram Kim, So-Yeon Kwan, Yu-Bin Kim, Eun-Gyeol Kim, Kwangjae Lee*

*Sangmyung Univ.

요 약

최근 코로나19 방역 패스로 인해 QR 코드 사용량이 급격히 증가하였고, QR 코드에 악성 URL을 삽입하여 정보를 탈취하는 공격 또한 늘어나고 있다. 본 논문은 QR 코드에 숨겨진 악성 URL을 탐지하는 보안 QR 코드 스캐너를 개발하였다. URL 탐지를 위해 URL의 도메인 이름으로 악성 여부를 판별하는 DGA 도메인을 탐지한다. DGA 도메인 탐지는 BiLSTM이 가장 성능이 좋았고, 탐지 정확도는 99.00%이다. DGA 도메인이 아닌 경우도 고려하여 WHOIS 정보로 악성 여부를 탐지한다. 이 탐지 또한 BiLSTM이 가장 성능이 좋았고, 탐지 정확도는 73.05%이다. 이 결과는 DGA 알고리즘으로 생성된 악성 URL은 거의 탐지할 수 있으며, 그렇지 않더라도 비교적 높게 탐지할 수 있다. QR 코드 스캔부터 결과 출력까지 0.65초로 빠르게 결과를 확인할 수 있었다.

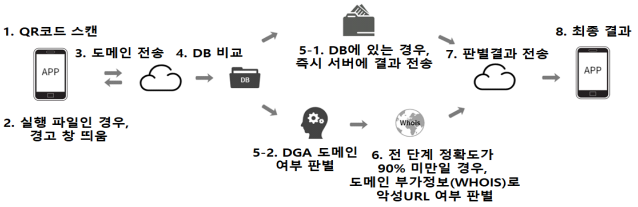
## Ⅰ. 서론

최근 발생한 코로나19의 확산은 우리 일상의 많은 것을 변화시켰다. 그 변화 중 하나는 바로 QR 코드의 생활화이다. URL, 이미지, 동영상 등 다양한 정보 저장이 가능한 QR 코드는 코로나19 방역을 위한 방역 패스로 활용되면서 이용량이 증가하였고, 결제 시스템이나 전자출입명부 시스템, 마케팅, 승차권 등 많은 분야에서 유용하게 사용되었다[1]. 그러나 QR 코드에 저장된 정보는 일반 사용자가 눈으로 파악하기 어려워 악성 URL이 숨겨진 QR 코드도 의심 없이 사용하게 만든다[2].

악성 URL은 피싱 공격이나 랜섬웨어 유포 등 심각한 문제를 일으킨다. 이를 피하고자 보안 분야에서도 악성 URL을 탐지 및 차단을 위한 연구가 이루어지고 있다. 기존 연구에는 URL 구성요소인 도메인 중 Domain Generation Algorithm(DGA)로 생성된 악성 도메인 탐지가 있으며, DGA는 도메인 생성 알고리즘으로 공격자가 Command & Control(C&C) 서버에서 공격 시 보안 시스템 회피를 위해 동적인 도메인을 생성하는 것이다. DGA로 생성된 도메인을 탐지하기 위한 기계학습은 개발되어 있으며, 높은 탐지 정확도를 가진다[3]. 하지만 DGA로 생성하지 않은 도메인은 찾지 못하는 단점이 있다. 본 논문에서는 기존 연구를 활용하여 DGA로 생성된 도메인도 판별하고, DGA로 생성되지 않은 도메인도 판별할 수 있도록 WHOIS에서 추출한 도메인 부가 정보를 기계학습에 이용하여 탐지하는 방식을 제안한다. 그리고 속도를 개선하기 위해서 이미 알려진 도메인은 DB로 비교한다.

## Ⅱ. 본론

제안하는 시스템은 QR 코드 내 악성 도메인 탐지를 위해 스마트폰 앱과 서버로 구성된다. 그림 1은 시스템의 구성도이며 사용자가 앱을 실행시켜 스캔하고자 하는 QR 코드를 화면에 맞춰 스캔하면 인식된 URL의 구성요소인 도메인만 서버로 전송된다. 다음으로 서버에서 전송된 도메인은 DB에 검색하여 DB에 포함된 도메인인지 확인한다. DB에 포함된 도메인이면 앱으로 결과를 전달하고, DB에 포함되지 않은 도메인은 기계학습으로 판별 과정을 거친다. DB에 포함되지 않은 도메인은 먼저 DGA 기계학습 모델로 DGA로 생성된 도메인인지 판별하고, DGA로 생성된 도메인은 결과를 앱으로 전달한다. DGA로 생성되지 않은 도메인은 WHOIS 기계학습 모델로 악성 여부를 판별한 결과를 앱으로 전달한다.




[그림 1] 제안하는 기계학습 기반 보안 QR 코드 스캐너의 구성도

### 1. DGA 기계학습

DGA 기계학습의 실험 데이터 셋은 Alexa Top Sites를 통해 수집한 non-DGA 도메인 데이터 50만 개와 netlab360에서 제공하는 DGA Domain Feed 데이터 50만 개를 사용한다[4]. 데이터셋 파일을 읽어 도메인들을 그림 2와 같이 영문자, 숫자, 특수문자에 인덱스 값을 부여하여 도메인과 인덱스를 매핑하였다. 영문자는 대소문자를 구분하지 않고 동일하게 취급하였다. 이후, 서로 다른 도메인 길이를 맞추기 위해 길이를 50으로 설정하고, 제로 패딩 작업을 진행했다. 그리고 여러 기계학습 모델들로 만들어진 학습 데이터를 훈련했다. 100만 개의 데이터셋을 학습 데이터 80만 개와 훈련 데이터 20만 개로 분류하고, 모델 훈련을 진행하였다. 본 논문에서 사용된 기계학습 모델은 LSTM, BiLSTM, GRU 알고리즘을 사용

```
index = {'-':'1', '.':'2', '1':'3', '0':'4', '3':'5', '2':'6', '5':'7', '4':'8',
         '7':'9', '6':'10', '9':'11', '8':'12', '_':'13', 'a':'14', 'c':'15',
         'b':'16', 'e':'17', 'd':'18', 'g':'19', 'f':'20', 'i':'21', 'h':'22',
         'k':'23', 'j':'24', 'm':'25', 'l':'26', 'o':'27', 'n':'28', 'q':'29',
         'p':'30', 's':'31', 'r':'32', 'u':'33', 't':'34', 'w':'35', 'v':'36',
         'y':'37', 'x':'38', 'z':'39'}
```

[그림 2] DGA 기계학습의 전처리에 사용되는 index

하였다. 더 성능 좋은 모델을 선정하기 위해서 정밀도, 재현율을 가지고 F1-score를 구해서 성능 평가를 진행했다. 성능 평가는 총 다섯 번의 반복을 진행하였고, 표 1과 같은 결과를 가진다. BiLSTM의 F1-score는 0.9921로 가장 높은 성능을, LSTM의 F1-score는 0.9738로 낮은 성능을 보여주었다. DGA 기계학습 모델 가운데 가장 정확도가 높은 BiLSTM 모델이 성능 평가에서도 5번의 훈련 모두 가장 높은 결과를 가졌기 때문에 DGA 악성 도메인을 판단하는 모델로 BiLSTM 모델을 선정하였다.

[표 1] DGA 기계학습 모델별 성능 평가 결과

| Run | LSTM | | | BiLSTM | | | GRU | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| 1 | 0.9922 | 0.9743 | 0.9832 | 0.9947 | 0.9896 | 0.9921 | 0.9841 | 0.9821 | 0.9831 |
| 2 | 0.9836 | 0.9726 | 0.9780 | 0.9869 | 0.9796 | 0.9832 | 0.9833 | 0.9202 | 0.9817 |
| 3 | 0.9841 | 0.9741 | 0.9791 | 0.9889 | 0.9810 | 0.9849 | 0.9881 | 0.9793 | 0.9837 |
| 4 | 0.9804 | 0.9746 | 0.9775 | 0.9936 | 0.9826 | 0.9831 | 0.9868 | 0.9761 | 0.9814 |
| 5 | 0.9739 | 0.9737 | 0.9738 | 0.9934 | 0.9777 | 0.9855 | 0.9918 | 0.9675 | 0.9795 |

## 2. WHOIS 기계학습

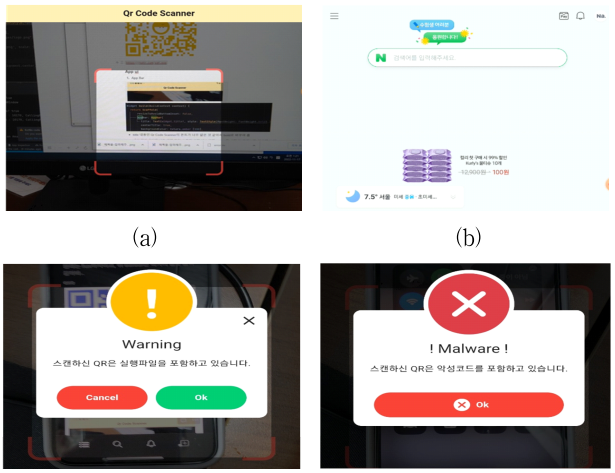
WHOIS 기계학습의 경우는 WHOIS를 통하여 얻은 도메인의 부가 정보를 데이터셋으로 제작하였다. 다양한 도메인 정보 중 악성 URL를 탐지하는 부가 정보로 주로 사용하는 도메인 등록일 날짜 정보와 등록자 이름, 등록기관을 사용하였다[5]. 날짜 정보는 int형식, 등록자 이름 및 등록기관은 띄어쓰기로 구분하여 각 단어를 숫자로 나타내주는 one hot encoding으로 변환하여 사용하였다. 구축한 모델 훈련 결과 LSTM 모델은 50.67%, BiLSTM 모델은 73.05%, GRU 모델은 49.52%의 정확도를 보였고, 표 2와 같은 결과를 가진다. 그리고 더 좋은 모델을 얻기 위해 성능 평가는 총 다섯 번의 반복을 진행했고, BiLSTM의 F1-score는 0.7698로 가장 높은 성능을, GRU의 F1-score는 0.6467로 가장 낮은 성능을 보여주었다. 결과적으로 BiLSTM 모델이 성능 평가에서 가장 높은 결과를 가졌기 때문에 WHOIS로 악성 도메인을 판단하는 모델로 BiLSTM 모델을 선정하였다.

[표 2] WHOIS 기계학습 모델별 성능 평가 결과

| Run | LSTM | | | BiLSTM | | | GRU | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| 1 | 0.5068 | 0.6362 | 0.6727 | 0.6986 | 0.8355 | 0.7609 | 0.4952 | 0.6897 | 0.6575 |
| 2 | 0.4942 | 0.6941 | 0.6615 | 0.6739 | 0.8489 | 0.7513 | 0.5018 | 0.6989 | 0.6467 |
| 3 | 0.4832 | 0.6823 | 0.6518 | 0.6688 | 0.8554 | 0.7507 | 0.5068 | 0.6887 | 0.6631 |
| 4 | 0.4957 | 0.6986 | 0.6629 | 0.6917 | 0.8677 | 0.7698 | 0.5003 | 0.6892 | 0.6597 |
| 5 | 0.4781 | 0.6802 | 0.6469 | 0.6884 | 0.8511 | 0.7611 | 0.5013 | 0.6935 | 0.6564 |

## 3. 실험 결과

보안 QR 코드 스캐너 앱의 실행 화면은 그림 3과 같다. 그림 3(a)을 보면 QR 코드를 스캔할 수 있는 구역이 존재하고 해당 구역 안에 QR 코드를 맞춰서 스캔할 수 있다. 정상 도메인일 경우 그 도메인을 실행시켜 인터넷 브라우저로 연결하고 최종적으로 그림 3(b)와 같은 결과가 나온다. 악성 도메인 중 큰 비중을 차지하는 것은 실행파일이다. 물론 정상적인 실행파일도 존재하겠지만 본 연구에서는 도메인 탐지에 중점을 두기 위해 실행파일을 스캔한 경우 그림 3(c)와 같이 경고성 알림창을 띄우는 것으로만 대처하였다. 악성 도메인을 포함하고 있는 QR 코드임을 판별한 결과를 받으면 그림 3(d)와 같은 결과를 표시한다.




(a) (b) (c) (d)

[그림 3] 보안 QR 코드 스캐너 실행 결과 (a) 앱 실행 화면, (b) 정상 결과 실행, (c) 실행파일 경고창, (d) 악성 코드 경고창

## Ⅲ. 결론

본 논문에서는 눈으로 판별이 어려운 QR 코드에 포함된 악성 URL을 피할 수 있도록 QR 코드 스캔 시 악성 도메인을 탐지하는 앱을 개발하였다. 이미 알려진 도메인은 검색 엔진이 빠른 mongoDB에 저장하여 판별하고, DGA로 생성한 도메인는 DGA 기계학습으로, 그렇지 않은 도메인은 WHOIS 기계학습으로 악성 여부를 판별하는 시스템을 추가로 구현하였다. DGA로 생성된 도메인 탐지 기계학습은 BiLSTM 모델을 선택하였고, 탐지 정확도는 99.00%를 얻었다. WHOIS를 통한 악성 판별 기계학습도 BiLSTM 모델을 선택하였고, 악성 도메인 판별 정확도는 73.05%를 얻었다. 앱과 서버를 연동 시 서버 수행 시간은 0.048초이고, QR 코드 스캔 후 실행 결과 화면 출력까지 0.65초가 소요되므로 빠른 속도로 URL의 판별 결과를 확인할 수 있었다. 개발한 앱을 활용한다면 파악이 어려운 QR 코드를 안전하게 이용할 수 있을 것으로 기대한다.

## 참 고 문 헌

[1] "Detailed Statistical Report: Worldwide QR Code Usage Before and After COVID-19." qrcode-tiger.com. https://www.qrcode-tiger.com/ko/qr-code-statistics-before-and-after-covid-19 (accessed Jan. 3, 2023).

[2] D.-W. Kim, Y.-T. Jo, and J.-M. Kim, "Cloud-based malware QR Code detection system," *J. Korea Inst. Inf. Commun. Eng.*, vol. 25, no. 9, pp. 1227-1233, 2021.

[3] K. Park, K. Kim, and Y. W, "Malicious Domain Detection through Machine Learning Malicious Domain Detection through Machine Learning", in *2017 Proc. Korea Softw. Conf.*, 2017, pp. 2022-2024.

[4] "DGA Datasets" netlab.360.com. https://netlab.360.com/zh (accessed Dec. 3, 2022).

[5] M. Kuyama, Y. Kakizaki, and R. Sasaki, "Method for detecting a malicious domain by using whois and dns features," in *3rd Int. Conf. Digit. Secur. Forensics*, Sep. 2016, vol. 74, pp. 74-80.