JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 12: 1786~1793, Dec. 2022

URL 주요특징을 고려한 악성URL 머신러닝 탐지모델 개발

김영준[1] · 이재우[2*]

Development of a Malicious URL Machine Learning Detection Model Reflecting the Main Feature of URLs

Youngjun Kim[1] · Jaewoo Lee[2*]

[1]Graduate Student, Department of Convergence Security, Chung-Ang University, Seoul, 06974 Korea
[2*]Assistant Professor, Department of Industrial Security, Chung-Ang University, Seoul, 06974 Korea

요 약

최근 코로나 19, 정치적 상황 등 사회적 현안을 악용한 스미싱, 해킹메일 공격이 지속되고 있다. 공격의 대부분은 악성 URL 접근을 유도하여 개인정보를 탈취하는 방식을 취하고 있는데, 이를 대비하기 위해 현재 머신러닝, 딥러닝 기술 연구가 활발하게 진행되고 있다. 하지만 기존 연구에서는 데이터 세트의 특징들이 단순하기 때문에 악성으로 판별할 근거가 부족하다고 판단하였다. 본 논문에서는 URL 데이터 분석을 통해 기존 연구에 반영된 URL 어휘적인 특징 이외에도 “URL Days”, “URL Words”, “URL Abnormal” 3종, 9개 주요특징을 추가 제안하였고, 4개의 머신러닝 알고리즘 적용을 통해 F1-Score, 정확도 지표로 측정하였다. 기존 연구와 비교 분석 시 평균 0.9%가 향상된 결과값과 F1-Score, 정확도에서 최고 98.5%가 측정됨에 따라 주요특징이 정확도 및 성능 향상에 기여하였다.

ABSTRACT

Cyber-attacks such as smishing and hacking mail exploiting COVID-19, political and social issues, have recently been continuous. Machine learning and deep learning technology research are conducted to prevent any damage due to cyber-attacks inducing malicious links to breach personal data. It has been concluded as a lack of basis to judge the attacks to be malicious in previous studies since the features of data set were excessively simple. In this paper, nine main features of three types, “URL Days”, “URL Word”, and “URL Abnormal”, were proposed in addition to lexical features of URL which have been reflected in previous research. F1-Score and accuracy index were measured through four different types of machine learning algorithms. An improvement of 0.9% in a result and the highest value, 98.5%, were examined in F1-Score and accuracy through comparatively analyzing an existing research. These outcomes proved the main features contribute to elevating the values in both accuracy and performance.

키워드 : 악성 URL, 피싱 URL, 머신러닝, 탐지모델

Keywords : Malicious URL, Phsing URL, Machine learning, Detection Model

Received 21 October 2022, **Revised** 27 October 2022, **Accepted** 1 November 2022

*** Corresponding Author** Jaewoo Lee(E-mail:jaewoolee@cau.ac.kr, Tel:+82-2-820-5935)
Assistant Professor, Department of Industrial Security, Chung-Ang University, Seoul, 06974 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.12.1786 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

과학기술정보통신부는 “최근 코로나19 지속, 정치적 상황 등 사회적 현안을 악용한 문자결제사기(스미싱), 해킹메일 유포를 통해 개인정보를 탈취하고 탈취 정보를 바탕으로 지능화된 사기전화(보이스피싱) 등 전기통신금융사기가 지속될 것으로 전망” 된다고 제시하였다[1]. 그뿐만 아니라 아래 그림 1과 같이 2020년부터 코로나 19 상황을 이용해 악성, 피싱 URL이 136,000개 이상 생성되었다고 한다[2]. 이러한 악성 URL은 사용자들의 접근 유도를 통해 사용자의 계정이나 개인정보 탈취, 악성코드 감염 등으로 악의적인 행위에 사용된다. 위와 같은 공격 행위는 개인뿐만 아니라 정부, 공공기관, 언론사 등 국가 주요 산업과 기반시설을 표적으로 지능화·고도화 되고 있다. 이처럼 공격대상이 개인에서 사회, 정부, 국가로 광범위해지는 상황에서 보안장비를 이용한 Rule-based 기반의 탐지방법[3]이나 사이버위협 정보 수집을 통해 만들어진 침해지표 기반의 사이버 위협 인텔리전스(Cyber Threat Intelligence, CTI) 정보[4]만으로는 악성 URL을 신속하게 탐지하고 대응하기에는 어려움이 존재한다.

위와 같은 문제를 해결하기 위해 최근에는 인공지능(AI)을 보안 분야와 접목한 보안 기술을 개발하고자 하는 기업이 많아지고 있으며, 정부에서도 인공지능 보안 기업을 육성하고자 노력하고 있다. 그중 한국인터넷진흥원을 통해 진행되고 있는 “제품·서비스 개발 지원사업은 5년간 AI 기반 보안 기업 100개사를 발굴하고 60개사를 육성”하고 있다[5]. 이처럼 인공지능 기술을 접목하여 악성 URL을 탐지 할 수 있는 모델 개발이 필요하다. 따라서, 본 논문에서는 악성 URL을 탐지하고 예측할 수 있도록 6종의 주요한 특징을 상세하게 제시하고, 4개의 머신러닝 알고리즘 적용을 통해 정확도와 성능을 높인 악성 URL 탐지모델을 제안하고자 한다.

본 논문의 공헌은 다음과 같다.

- 관련연구에서 공개한 특징 이외에 “URL Days”, “URL Words”, “URL Abnormal” 3종의 주요특징을 추가 도출하여 제안하였다.
- 4개의 머신러닝 알고리즘(Decision Tree, Random Forest, Extra Trees, Gradient Boosting)을 적용하여 악성 URL 모델을 개발하였다. 또한 해당 모델을 이용해 악성 URL을 신속하게 탐지하도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구되었던 머신러닝과 딥러닝 알고리즘을 이용한 악성 URL 탐지 방법에 관한 소개한다. 3장에서는 머신러닝 알고리즘을 이용한 악성 URL 탐지 모델 개발 구조를 설명하고, 4장에서는 머신러닝 기반의 악성 URL 탐지 모델 개발을 통해 성능 측정한 실험결과를 분석하였다. 5장에서는 향후 과제에 대한 의견 제시와 함께 결론을 구성하였다.

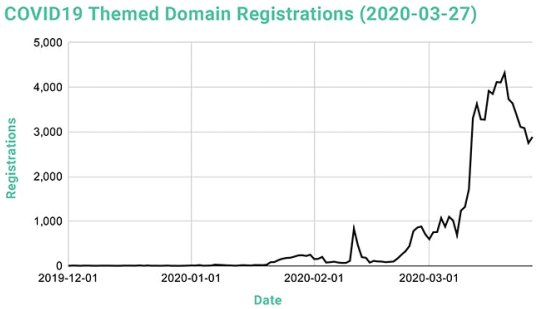

Fig. 1 Malware, Phishing URL Trends[2]

Ⅱ. 관련연구

머신러닝과 딥러닝 기반의 악성 URL 탐지모델과 관련된 연구는 대부분은 검색량, 어휘적 기반, 길이, 개수와 같은 단순한 특징을 기반으로 알고리즘을 학습시켜 모델을 개발하였다.

김종관 등[6]은 URL 검색량 기반의 특징 데이터를 추출하였고 머신러닝 알고리즘(SVM, KNN, LR)을 적용하여 악성 URL 탐지모델을 개발하였다. 학습데이터는 정상 URL 약 1,000,000건, 비정상 URL 14,767건을 이용하였으며, 정상·비정상 URL은 동일한 비율로 정상 200개, 비정상 200개로 구성하여, 검색량 데이터만 포함된 데이터 세트를 생성하였다. 머신러닝 모델별 탐지정확도는 80% 수준의 성능을 보여준 연구 사례였다.

강홍구 등[7]은 URL 어휘적 기반의 특징을 이용한 여러 머신러닝 알고리즘별로 모델을 생성하고 악성 URL를 예측하는 시스템을 고안하였다. 시스템은 특징 추출 모듈, 백터 생성 모듈, 모델 생성 모듈, 악성 URl 예측 모듈로 구성하였다. 학습데이터는 정상 URL 436,722건, 악성 URL 158,081건으로 총 594,803건의 URL 데이터를 이용하였다. 어휘력 특징으로는 URL만으로 추출

가능한 길이(length), 개수(count), 존재(existence) 유형으로 학습데이터를 생성하였고, 5개의 머신러닝 알고리즘(DT, RF, GBM, XGB, SVM)에 적용과 알고리즘 간의 조합으로 예측 정확도를 높여 90% 수준의 성능을 도출한 연구 사례였다.

Hevapathige, Asela 등[8]은 Super Learner 앙상블을 사용한 악성 URL 분류 모델을 개발하는 연구를 진행하였다. 학습데이터로는 정상 URL 476,062건, 악성 URL 272,802건으로 총 748,864건의 학습데이터를 이용하였다. 특징으로는 총 4종으로 길이(Length), 개수(Count), 참・거짓(Boolean), 계산(Calculation)의 특징 추출을 진행하였다. 그 이후에는 6개의 머신러닝 알고리즘(ADABoost, Bagging, Gradient Boosting, Extra Trees, Random Forest, Histogram, XGBoost)을 이용해 앙상블 학습으로 머신러닝 알고리즘을 적용한 정확도보다 높은 결과로 95% 수준의 성능을 보여준 연구 사례였다.

Chen, Yu 등[9]은 CNN 알고리즘을 이용한 악성 URL를 탐지하는 모델을 연구하였다. 학습데이터로는 정상 URL 5,000건, 악성 URL 5,000건으로 비율을 동일하게 하여 총 10,000건의 데이터를 사용하였다. 특징으로는 Whois 정보, URL 길이, 특수문자 수, 키워드 및 기타 텍스트를 추출하였다. 딥러닝 신경망 모델인 CNN을 적용해 80% 수준의 정확도를 보여준 연구 사례였다.

앞서 소개한 연구사례는 악성 URL 탐지모델 정확도가 80% 이상 측정되었다. 하지만 검색량, 어휘적 기반, 길이, 개수와 같은 단순한 특징은 실제 URL의 악성을 검증하기 위한 근거로는 미흡하다고 생각하였다.

본 논문에서는 URL정보를 이용하여 악성으로 판단할 수 있는 주요특징을 추가적으로 도출하고, 기존 연구보다 정확도가 높은 모델 개발을 목표로 진행하였다.

Ⅲ. 주요특징 기반 악성URL 모델 제안

관련연구에서 제안된 특징으로는 URL의 길이, 문자와 숫자 그리고 특수문자 개수, 문자열 포함 여부로 단순하게 구성되었기 때문에, 악성으로 판별할 근거로 부족하다고 판단하였다.

본 논문에서는 선행연구에서 제시된 특징들 이외에도 악성으로 판단 할 수 있는 주요 특징들을 추가로 도출하였다. 그림2와 같이 “URL Days”, “URL Words”, “URL Abnormal” 3종으로 제시하였다. 또한 관련연구보다 정확도 및 성능 향상 검증을 위해 4개의 머신러닝 알고리즘 적용 이후 모델의 성능을 측정하기 위해 정확도, 재현율, 정밀도, F1-Score 지표를 이용한 검증방식을 제안하였다.

3.1에서는 학습데이터의 정보와 출처를 소개한다. 3.2에서는 전처리 과정에서 신규 도출된 주요특징을 3.2.1부터 3.2.3까지 악성 판단 근거를 포함하여 제시하고, 3.2.4부터 3.2.6까지는 기존 연구에서 사용된 특징을

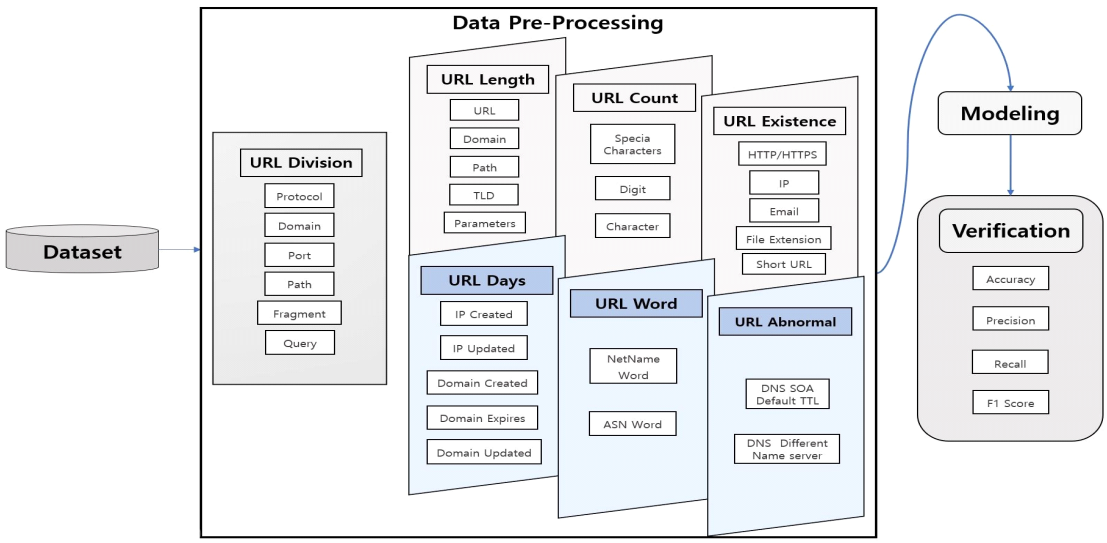

Fig. 2 A framework for developing a malicious URL detection model that adds three type features(blue areas)

간략하게 소개한다. 3.3은 4개의 머신러닝 알고리즘 정보와 함께 3.4에서 성능 검증 방식을 제시한다.

3.1. 데이터 수집

본 논문에서 수집된 학습데이터는 뉴브런즈윅 대학교에서 제공된 URL 데이터(ISCX-URL2016[10])를 총 48,025개(정상 35,378개, 악성 12,647개)를 수집하였으며, PhishTank[11], URLhaus[12], openphish[13]에서 악성 URL 124,837개, DMOZ-ODP[14]에서 제공된 정상 URL 100,000개를 추가로 수집하였다.

3.2. 데이터 전처리

머신러닝 알고리즘을 적용하기 전 데이터 세트를 생성하기 위해 전처리 및 주요특징 도출이 이루어져야 한다. 먼저, 전처리로는 결측치, 틀린값, 이상치 데이터를 파악하여 수정・삭제를 진행하고, 주요특징으로는 URL 구조 및 데이터 분석으로 총 6종의 43개를 도출하여 데이터 세트를 생성했다.

주요특징을 도출하기 위해 사전에 URL 구조를 파악하여야한다. 먼저, 아래 그림 3과 같이 6개의 구조로 분류하고, 데이터세트를 생성하기 위해 URL을 Protocol, Domain, Port, Path, Parameters, Fragment로 분할하여 각 항목을 분석하였다[15].

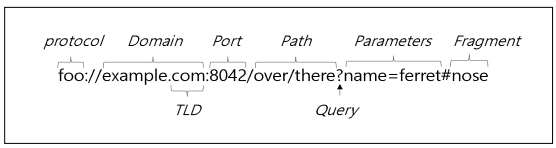

Fig. 3 URL Structure

3.2.1. 일수(Days) 기반 특징 추출

일수 기반의 특징은 Python Whois 모듈을 통해 아래 그림 4와 같이 도메인의 생성일자, 갱신일자, 만료일자를 추출하였다. 현재일 기준으로 기간이 어느 정도 지났고 남았는지를 일수로 도출하여 주요특징으로 포함하였다. 생성 또는 갱신일자가 최신일 경우 악의적으로 사용할 목적으로 생성되었을 가능성이 높기 때문이고, 만료일자의 경우도 한시적으로 URL을 사용하기 때문에 서비스 기간이 대부분 짧은 것으로 분석하였다. 이처럼 생성일자와 갱신일자 주기가 최신이고 만료일자가 짧은 이유로는 악성 도메인이 보안장비에서 차단되고 다시 공격시도를 하기 위해 다른 신규 도메인을 생성하기 때문이다.

도메인을 이용해 확보된 IP도 생성일자와 갱신일자가 최신일 때도 악용될 목적으로 생성되었을 가능성이 높다. 그렇기에 IP의 생성 또는 갱신일자를 주요특징에 포함하였고 그림 5와 같이 추출하였다.

```
Input : Domain
Output: Update, Creation, Expiration Day

domain_info = whois.whois(domain)
update = domain_info.updated_date
creation = domain_info.creation_date
expiration = domain_info.expiration_date

today = datetime.datetime.now()
Update_day = today – update
  return Domain_Update_day
Creation_day = today – creation
  return Domain_Creation_day
Expiration_day = expiration – today
  return Domain_Expiration_day
```

Fig. 4 Domain Whois Requset Source Code

```
Input : IP
Output: Update, Creation Day

ip_info = IPWhois(IP)
ip_info = ip_info . lookup_whois()

create = str(w['nets'][0]['created'])
update = str(w['nets'][0]['updated'])

today = datetime.datetime.now()
Update_day = today – update
  return IP_Update_day
Creation_day = today – creation
  return IP_Creation_day
```

Fig. 5 IP Whois Requset Source Code

3.2.2. 단어(Words) 기반 특징 추출

단어 기반 특징은 Netname을 추출하기 위해 총 세 단계의 과정을 진행하였다. 먼저, 첫 번째로 악성 도메인에서 Python dns.resolver 모듈을 이용하여 IP를 획득한다. 두 번째로 획득한 IP는 IP Whois 모듈을 통해 Netname을 추출한다. 마지막으로 Netname 목록을 그룹화하여 비교 대상 항목 지정하였다. 악성 도메인을 통해 추출된 Netname 분석 시 공격자가 도메인을 등록할 때 동일한 Netname을 사용하는 것을 알 수 있었다. 그렇기 때문에 도메인의 Netname 조회 시 악성 도메인과 동일한 Nename을 사용할 경우 악성 도메인 가능성이 높

다고 판단하였다. 또한 IP Whois 모듈을 이용해 그림 6과 같이 ASN(Autonomous System Numbers)도 추가로 추출하였다. ASN도 Netname과 마찬가지로 IP 서브넷을 식별하기 위해 고유하게 부여된 번호이다. 그렇기에 ASN을 조직 단위로 사용하기 때문에 악성으로 식별이 가능하다.

```
Input : IP Netname, ASN
Output: Malware Netname, ASN = 0, Other = 1
IP_info = IPWhois(IP)
IP_info = ip_info . lookup_whois()
Netname = str(w['nets'][0]['name'])
ASN = str(w['nets'][0]['postal_code'])

Match = re.search('OVH|HGBLOCK|LACNIC|BG-SUNET|...|
                   IS-HAFNARFJ', Netname)

Match = re.search(20006|94063|80045|90012|85034|...|
                   60666|75202', asn)
if Match:
  return 0
else:
  return 1
```

Fig. 6 IP Netname, ASN Match Source Code

3.2.3. 비정상적인(Abnormal) 특징 추출

비정상적인 특징은 DNS의 SOA 레코드 질의를 통해 그림 7과 같이 DNS TTL(Time To Live) 정보를 추출하였다. 정상 URL은 DDNS 사용 등으로 인해 TTL 주기가 비교적 짧다. 그리고 악성 URL의 경우는 DNS 질의가 되지 않거나 TTL 주기가 대부분 “86400”으로 설정된 것으로 분석하였다.

DNS SOA 레코드 질의 데이터 중 Name Server도 주요 특징으로 포함하였다. 정상 도메인에 대해 DNS 질의 시 Name Server 명칭에 도메인이 포함되어 구성되어 있다. 하지만 악성 도메인에 대한 질의 결과 확인 시 Name Server가 도메인 명칭과 상이하였다.

```
Input : domain
Output: DNS SOA Recode TTL
res = dns.resolver.query(domain, 'SOA')
SOA_records = []
  for val in res:
    SOA_records.append(val.to_text())
return DNS_SOA
```

Fig. 7 DNS SOA Recode Request, Source Code

3.2.4. 길이(Length) 기반 특징 추출

길이 기반의 특징은 Domain, Path, Parameter, TLD의 문자열 길이로 추출하였다. 악성 URL은 길이가 대부분 길게 사용되기 때문에 아래 그림 8처럼 구조별 문자열 길이를 측정하였다[16].

```
Input : URL, Domain, Path, TLD, Parameters Data
Output: URL Structure Length
return len(URL)
domain = res.parsed_url.netloc
return len(domain)
Path = res.parsed_url.path
return len(path)
TLD = res.tld
return len(TLD)
Parameters = res.parsed_url.query
return len(Parameters)
```

Fig. 8 URL Structure Length Source Code

3.2.5. 개수(Count) 기반 특징 추출

개수 기반 특징은 URL 내 포함되는 특수문자 18종(‘%’, ‘$’, ‘=’, ‘@’, ‘?’, ‘&’, ‘#’, ‘.’, ‘_’, ‘-’, ‘;’, ‘{’, ‘}’, ‘[’, ‘]’, ‘|’, ‘+’, ‘*’)과 문자, 숫자로 추출하였다. 그림 9와 같이 count 함수를 이용해 특수문자 18종이 몇 개가 존재하는 지 개수로 추출하였다. 특수문자 중 ‘.’은 악성 URL을 접근시키기 위한 서브도메인으로 정상 URL를 이용한다. 이외에도 SQL 인젝션 취약점 공격 시 ‘=’, ‘*’, ‘;’ 등 다양한 특수문자가 SQL 쿼리 작성에 사용된다. 이처럼 특수문자를 과도하게 사용할 경우 정상 URL로 판별하기 어렵기에 특수문자 개수를 주요특징으로 도출하였다.

```
Input : URL
Output: Special Characters Count
feature = ['@', '?', '-', '=', '.', '#', '%', '+', '$', '!', '*', ',', '//']
for i in range(len(Match)):
  dataset[a] = dataset['url'].apply(lambda i: i.count(a))
```

Fig. 9 Special Characters Count Source Code

3.2.6. 존재여부(Existence) 기반 특징 추출

존재여부 기반 특징 선별은 HTTPS 적용 여부로 판별하였다. HTTPS는 웹 사이트에서 전송되는 모든 데이터를 암호화 또는 접속하는 사이트가 신뢰되는 사이트인지 엄격한 인증과정을 거치는 부분이다. 그래서 악성 URL은 비교적 HTTPS를 거의 사용하지 않기 때문에 HTTPS 존재 여부를 주요 특징으로 포함하였다.

```
Input : Protocol
Output: HTTPS = 1, Other = 0
```

```
Protocol = urlparse(URL).scheme
Match = str(protocol)
  if match=="https":
    return 1
  else:
    return 0
```

Fig. 10 Secure Site Existence Source Code

이외에도 URL 내 IP 또는 E-mail, 파일 확장자 15종(“.php”, “.html”, “.htm”, “.hwp”, “.hwpx”, “.pptx”, “.docx”, “.iso”, “.js”, “.lnk”, “.vbs”, “.xls”, “.xml”, “.zip”, “.xlsx”)의 존재 여부 체크를 통해 주요 특징 항목으로 추출하였다. IP와 Email이 URL에 포함이 될 경우 피싱 URL로 악용된 사례가 다수 존재하고, 파일 확장자의 경우는 웹쉘 또는 웹해킹 취약점으로 악용되기 때문에 포함하였다.

단축 URL(Short URL)은 복잡하고 긴 주소를 짧게 줄여주고 실제 접근되는 URL 확인이 어려운 부분을 악용하고 있어 주요 특징으로 포함하였다[16].

```
Input : URL
Output:  Short URL = 1, Other = 0
```

```
match = re.search('bit₩.ly | ... | me2₩.kr', url)
 if match:
   return 1
 else:
   return 0
```

Fig. 11 Short URL Existence Source Code

3.3. 머신러닝 모델링

본 논문에서는 URL 주요특징 데이터를 효과적으로 이용하기 위해 트리 기반 알고리즘을 이용하였다.

트리 기반 알고리즘은 분기별로 주요특징에 대한 조건식을 거쳐 악성과 정상을 구분 할 수 있는 최적의 기준을 찾을 수 있기에 사용하였다. 이러한 머신러닝 알고리즘으로는 Decision Tree, Random Forest, Extra Trees, Gradient Boosting로 4개를 선정하였다.

3.4. 모델 성능 검증

본 논문에서는 악성 URL 탐지모델의 성능을 측정하기위해 Accuracy(정확도), Recall(재현율), Precision(정밀도), F1-Score 지표를 이용하였다. Accuracy(정확도)는 실제 데이터에서 예측 데이터가 얼마나 같을지를 판단하는 지표이며, 직관적으로 모델 예측 성능을 측정할 때 주로 사용된다. Precision(정밀도), Recall(재현율)는 Positive 데이터의 예측 성능에 더 초점을 맞춘 평가 지표이다. 정밀도는 예측 값이 Positive인 대상 중 예측과 실제 값이 Positive로 일치한 데이터의 비율을 말한다. 그리고 재현율은 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻하고 있다. F1-Score는 정밀도와 재현율을 결합한 지표로 사용된다. 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가지기 때문에 정밀도와 재현율의 편향 차이를 줄일 수 있어 정확한 성능 검증을 가능하게 한다[17].

Ⅳ. 실험 결과 및 분석

URL 주요특징을 고려한 악성 URL 머신러닝 탐지 모델은 Jupyter Notebook 6.3.0 환경에서 Python 3.8.8 언어를 사용하여 실험을 하였다.

4.1. 실험 데이터 세트

실험으로 활용될 데이터 세트는 수집된 정상과 악성 URL 비중을 1:1로 비율로 측정하여, 약 100,000개 URL 데이터를 최대한 균형 있게 구성하였다. 또한 위의 표1과 같이 6종의 43개의 주요특징을 도출해 데이터 세트를 생성하였다.

Table. 1 Dataset

Category	Main Features
URL Length	URL Length
	Domain Length
	Path Length
	TLD Length
	Parameters Length
URL Count	Specia Characters Count
	Digit Count
	Character Count
URL Existence	HTTP/HTTPS Existence
	IP Existence
	Email Existence
	File Existence
	Short URL Existence

Category	Main Features
URL Days(new)	IP Created/Updated Days
	Domain Created/Updated/Expires Days
URL Words(new)	NetName Words
	ASN Words
URL Abnormal(new)	DNS SOA Record Default TTL
	Name server Different URL

4.2. 성능 측정 실험

생성된 데이터 세트를 8:2 비율로 각각 학습 세트와 테스트 세트로 분류하였으며, Random_state 값은 2로 설정하였다. 4개의 머신러닝 알고리즘(Decision Tree, Random Forest, Extra Trees, Gradient Boosting)을 적용시켜 테스트 세트로 정확도를 측정하여, 아래 표 2와 같은 결과를 도출하였다.

Precision(정밀도), Recall(재현율), Accuracy(정확도), F1-Score 4개의 지표 모두 95%이상 높은 수치로 측정되었으며, 4개의 머신러닝 알고리즘 중 Random Forest가 F1-Score, 정확도가 98.5%로 가장 좋은 결과를 나타냈다.

Table. 2 Accuracy Result

Algorithm	Performance			
	Precision	Recall	F1-Score	Accuracy
DT	97%	97%	96.8%	96.9%
RF	99%	98%	98.5%	98.5%
ET	98%	95%	96.4%	96.5%
GB	98%	98%	98%	98%

4.3. 실험결과 분석

관련연구에서 사용되었던 어휘적 특징을 포함한 데이터 세트[7]와 본 논문에서 제안한 주요특징을 추가한 데이터 세트를 이용해 4개의 머신러닝 알고리즘을 적용하고, F1-Score, Accuracy(정확도) 지표를 통해 비교 분석하였다.

표 3과 같이 4개의 알고리즘 대상 F1-Score, Accuracy(정확도) 지표를 이용해 성능 측정 시 관련연구에서 제시된 데이터세트보다 평균 0.9% 향상된 결과 값을 보여주었으며, URL 데이터 분석을 통한 주요특징 도출의 중요성이 실험결과를 통해 검증하였다.

Table. 3 Performance Compare Result

Algorithm	Performance			
	AS-IS		TO-BE	
	F1-Score	Accuracy	F1-Score	Accuracy
DT	96.1%	96.3%	96.8%	96.9%
RF	97.2%	97.3%	98.5%	98.5%
ET	95.6%	95.9%	96.4%	96.5%
GB	96.9%	97%	98%	98%

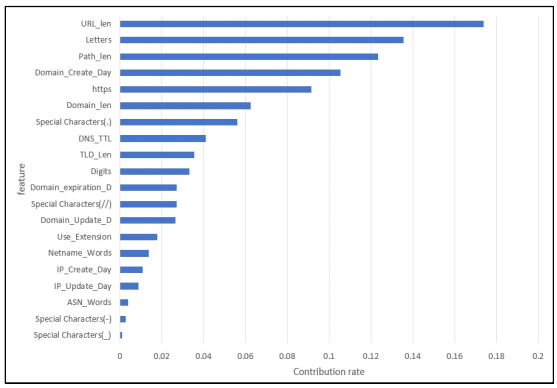

Fig. 12 Random Forest Algorithms Bar chart according to Main Feature contribution

위 그림12와 같이 정확도의 기여도에 대한 주요특징을 막대차트로 비교하였다. 신규로 생성한 주요특징에서 "Domain Create Day" 13%, "DNS TTL" 4%, "Domain Expiration Day" 2%, "Domain Update Day" 2%로 정확도를 높이는 데 기여하였다. "Domain Create Day"의 정상 URL 평균값이 4,938일이며, 악성 URL은 평균값이 597일로 생성일자가 크게 차이가 나는 것을 확인 할 수 있었다.

Ⅴ. 결 론

본 논문에서는 URL의 주요특징을 고려한 악성URL 머신러닝 탐지모델 개발을 제안하였다. 악성 URL 탐지 모델 개발하기 위해 학습데이터 수집, 주요특징 도출 과정을 설명하고 실험결과를 통해 검증하였다. 주요특징을 도출하는 과정에서 URL 구조 분석을 통해 단순히 어휘적인 특징을 추출한 게 아닌 악성으로 판단될 수 있는 3종의 주요 특징(“URL Days”, “URL Words”, “URL

Abnormal")을 추가 도출하였다. 기존의 악성 URL 탐지 모델 연구들과 비교하여 분석한 결과 F1-Score와 정확도에서 평균 0.9% 향상된 것을 확인할 수 있었고, Random Forest 알고리즘에서 최고 98.5%가 측정되어 주요특징의 중요성을 실험결과로 통해 검증하였다.

제안한 모델에서 학습데이터에 중점을 두어 주요특징을 구성만 하더라도 기존에 연구되었던 모델보다 뛰어난 성능을 보여주고 실험 결과로 증명하였다.

향후에도 악성 URL 탐지 정확도의 유효성을 높이기 위해 정부, 기관, 단체에서 수집된 URL 데이터를 전달받아 지속적인 주요특징의 데이터 세트 생성과 새로운 주요특징 도출 연구가 활발하게 이뤄져야 한다.

References

[1] N. S. Kim, "Ministry of Science and ICT, '21 cyber threat analysis and '22 viewpoint analysis," Ministry of Science and ICT, 2021. [Internet]. Available: https://doc.msit.go.kr/SynapDocViewServer/viewer/doc.html?key=7d38743144ff45fb8688b4f2255dfc13&convType=html&convLocale=ko_KR&contextPath=/SynapDocViewServer/.

[2] Spotting and blacklisting malicious COVID-19-themed sites [Internet]. Available: https://www.helpnetsecurity.com/2020/04/07/covid-19-malicious-sites/.

[3] Y. B. Kwon and I. S. Kim, "A Study on Anomaly Signal Detection and Management Model using Big Data," *The Journal of The Institute of Internet, Broadcasting and Communication*, vol. 16, no. 6, pp. 287 - 294, Dec. 2016.

[4] S. G. Lee, D. W. Kim, B. J. Kim, T. W. Lee, S. W. Han, and J. K. Lee, "Comprehensive Analysis Strategy in Cyber Threat Intelligence Environment," *Review of KIISC*, vol. 31, no. 5, pp. 33-38, Oct. 2021.

[5] Leading the domestic security market with AI technology [Internet]. Available: http://www.itdaily.kr/news/articleView.html?idxno=206661.

[6] J. K. Kim, M. H. Jang, S. N. Lim, and M. S. Kim, "A Study on the Detection Method of Malicious URLs based on the Internet Search Engines using the Machine Learning," *The Transactions of The Korean Institute of Electrical Engineers*, vol. 70, no. 1, pp. 114-120, Jan. 2021.

[7] H. K. Kang, S. S. Shin, D. Y. Kim, and S. T. Park, "Design and Implementation of Malicious URL Prediction System based on Multiple Machine Learning Algorithms," *Journal of Korea Multimedia Society*, vol. 23, no. 11, pp. 1396 - 1405, Nov. 2020.

[8] A. Hevapathige and K. Rathnayake, "Super Learner for Malicious URL Detection," in *Proceedings of 2022 2nd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, pp. 114-119, 2022.

[9] Y. Chen, Y. Zhou, Q. Dong, and Q. Li, "A Malicious URL Detection Method Based on CNN," in *Proceedings of 2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, Shenyang, China, pp. 23-28, 2020.

[10] University of new brunswick ISCX-URL2016 URL dataset [Internet]. Available: https://www.unb.ca/cic/datasets/url-2016.html.

[11] Phishing URLs provided by Phishing Tank [Internet]. Available: http://data.phishtank.com/data/online-valid.csv.

[12] Malicious URLs provided by URLhaus [Internet]. Available: https://urlhaus.abuse.ch/.

[13] Phishing websites provided by OpenPhish [Internet]. Available: https://openphish.com/.

[14] Multinational Open Content Directory on World Wide Web Links by DMOZ [Internet]. Available: https://www.dmoz-odp.org.

[15] The Internet Society, "Rfc3986: Uniform resource identifier (uri): Generic syntax," 2005. [Online]. Available: https://tools.ietf.org/html/rfc3986.

[16] J. S. Park, "Based on URL pattern analysis Preventive measures against harmful sites," M. S. thesis, Konkuk University, 2019.

[17] C. M. Kwon, *Python Machine Learning Perfect Guide*, Gyeonggi, Korea, Wikibook, 2019.

김영준(Youngjun Kim)

2015년 2월 한서대학교 항공소프트웨어공학과 학사
2021년 3월 ~ 현재 중앙대학교 융합보안학과 석사과정
※관심분야 : 인공지능, 침해사고 대응, 악성코드

이재우(Jaewoo Lee)

2006년 2월 서울대학교 컴퓨터공학부 학사
2008년 2월 서울대학교 컴퓨터공학부 석사
2017년 8월 University of Pennsylvania, Ph.D in Computer and Information Science
2018년 3월 ~ 현재 중앙대학교 산업보안학과 조교수
※관심분야 : Cyber Physical System Security