https://doi.org/10.33778/kcsa.2022.22.3.101

설명 가능한 인공지능(XAI)을 활용한 침입탐지 신뢰성 강화 방안★

정 일 옥*, 최 우 빈**, 김 수 철***

요 약

다양한 분야에서 인공지능을 활용한 사례가 증가하면서 침입탐지 분야 또한 다양한 이슈를 인공지능을 통해 해결하려는 시도가 증가하고 있다. 하지만, 머신러닝을 통한 예측된 결과에 관한 이유를 설명하거나 추적할 수 없는 블랙박스 기반이 대부분으로 이를 활용해야 하는 보안 전문가에게 어려움을 주고 있다. 이러한 문제를 해결하고자 다양한 분야에서 머신러닝의 결정을 해석하고 이해하는데 도움이 되는 설명 가능한 AI(XAI)에 대한 연구가 증가하고 있다. 이에 본 논문에서는 머신러닝 기반의 침입탐지 예측 결과에 대한 신뢰성을 강화하기 위한 설명 가능한 AI를 제안한다. 먼저, XGBoost를 통해 침입탐지 모델을 구현하고, SHAP을 활용하여 모델에 대한 설명을 구현한다. 그리고 기존의 피처 중요도와 SHAP을 활용한 결과를 비교 분석하여 보안 전문가가 결정을 수행하는데 신뢰성을 제공한다. 본 실험을 위해 PKDD2007 데이터셋을 사용하였으며 기존의 피처 중요도와 SHAP Value에 대한 연관성을 분석하였으며, 이를 통해 SHAP 기반의 설명 가능한 AI가 보안 전문가들에게 침입탐지 모델의 예측 결과에 대한 신뢰성을 주는데 타당함을 검증하였다.

The Enhancement of intrusion detection reliability using Explainable Artificial Intelligence(XAI)

Jung Il Ok*, Choi Woo Bin**, Kim Su Chul***

ABSTRACT

As the cases of using artificial intelligence in various fields increase, attempts to solve various issues through artificial intelligence in the intrusion detection field are also increasing. However, the black box basis, which cannot explain or trace the reasons for the predicted results through machine learning, presents difficulties for security professionals who must us e it. To solve this problem, research on explainable AI(XAI), which helps interpret and understand decisions in machine le arning, is increasing in various fields. Therefore, in this paper, we propose an explanatory AI to enhance the reliability of machine learning-based intrusion detection prediction results. First, the intrusion detection model is implemented through XGBoost, and the description of the model is implemented using SHAP. And it provides reliability for security experts to make decisions by comparing and analyzing the existing feature importance and the results using SHAP. For this experim ent, PKDD2007 dataset was used, and the association between existing feature importance and SHAP Value was analyzed, and it was verified that SHAP-based explainable AI was valid to give security experts the reliability of the prediction res ults of intrusion detection models.

Key words : XAI, SHAP, intrusion detection, Explainable, machine learning

접수일(2022년 08월 31일), 수정일(2022년 09월 20일), (게재확정일(2022년 09월 28일)

★ 본 논문은 2022년 정부(국토교통부)의 재원으로 국토교통과학기술진흥원(KAIA)의 지원을 받아 연구가 수행된 연구임 (22TLRP-B152767-04, 자율협력주행 도로교통체계 통합보안 시스템 운영을 위한 기술 및 제도개발)

* 고려대학교/정보보호학과 (주저자, 교신저자)

** 경희대학교/응용수학과(공동저자)

*** 숭실대학교/IT정책경영학과(공동저자)

1. 서 론

인터넷에 대한 사용이 증가하면서 사이버위협 또한 증가하고 있다[1]. 다양한 연구를 통해 침입탐지(Intrusion Detection) 분야에서는 딥러닝(Deep Learning)과 머신러닝(Machine Learning)을 이용한 탐지 성능 향상이 검증되면서 이를 활용한 사례가 나날이 증가하고 있다[2]. 하지만, 사이버 보안 분야에서 머신러닝의 적용은 그렇게 쉽지만은 않다.

이는 사이버 보안 분야의 특성상 침입에 대한 오탐(False Positive)과 미탐(False Negative)에 민감하기 때문이다[3]. 잘못된 탐지에 대한 대응은 서비스 중지 등과 같은 가용성에 큰 영향을 줄 수 있으며, 지능화된 공격에 대한 미탐은 정보유출이나 시스템 파괴와 같은 큰 피해로 연결될 수 있기 때문이다. 또한, 전통적인 머신러닝이 블랙박스(Black-box) 기반의 모델이기 때문이다. 따라서 모델에 의한 예측이 유의미한 결과를 도출하더라도 이를 해석하고 판단해야 하는 분석가가 이해하지 못한다면 아무 소용이 없게 된다[4].

이러한 한계를 극복하기 위해서 많은 연구자가 설명 가능한 AI에 관한 연구에 관심을 가지게 된다. 2016년부터 DARPA를 통해 BAA-16-53(Explainable Artificial Intelligence)를 시작으로 설명 가능한 AI 연구가 시작되었으며[5], 아직은 초창기 단계지만, 다양한 분야에서의 요구로 인해 활발한 연구가 진행되고 있다.

최근 머신러닝 분야에서는 부스팅 방법을 통해 결합한 앙상블(ensemble) 방법인 XGBoost(eXtream Gradient Boosting)를 기반으로 모델이 좋은 성능을 나타내고 있다[6]. XGBoost는 의사결정나무에 비하여 높은 예측 성능을 보이지만, 여러 개의 의사결정나무를 혼합하는 방법이기 때문에 예측결과가 도출된 이유에 대하여 설명하기 힘든 블랙박스 모델이라는 한계가 있다. 이에 예측에 대한 유의미한 분석을 도출하기 위해 설명 가능한 AI(eXplainable AI, XAI) 기법 중 하나인 SHAP(SHapley Additive exPlanations)를 적용하고자 한다[7]. SHAP은 Shapley Value를 이용하여 예측에 영향을 미치는 변수들을 파악하는 방법으로 종속변수에 긍정적인 영향을 미치는 변수뿐만 아니라 부정적인 영향을 미치는 변수도 파악할 수 있는 기법이다. 또한 SHAP은 변수의 전역적인(Global) 영향력뿐 아니라 분석에 사용된 개개인에 대해서도 어떤 변수가 침입 의도에 긍정적인 영향을 미쳤는지, 부정적인 영향을 끼쳤는지 파악할 수 있는 지역적인(Local) 정보도 제공해 준다.

이에 본 연구에서는 XGBoost와 SHAP을 적용하여 더 우수한 침입탐지 모델을 구축하는 동시에 구축된 모델에 대한 유의미한 해석을 제공하는 것을 목표로 한다.

크게 설명 가능한 AI 기법을 적용하는 방법은 전통적인 머신러닝에 적용하는 방법과 딥러닝에 적용할 수 있는 방법으로 나눌 수 있는데, 본 연구에서는 전통적인 머신러닝에 설명 가능한 AI를 적용하는 것으로 범위를 제한하였다. 이는 도메인의 특성을 고려한 피처 추출 및 선택을 통해 설명 가능한 AI 모델의 적정성을 알아보고자 하였기 때문이다.

본 연구를 통해 구체적으로 이바지하는 바는 다음과 같다. 먼저, 기존의 블랙박스 기반의 기술과 잘 알려진 XAI 기술을 사용하여 사이버 보안 전문가에게 침입탐지 기반의 결정을 체계적으로 설명할 수 있는 능력을 제공하는 XAI 기반의 프레임워크를 제안한다. 둘째, 다양한 알고리즘 (Random Forest, MLP, Logistic Regression, Decision Tree, XGBoost[6])을 통해 모델을 성능을 예측하였으며, 가장 좋은 결과를 도출한 XGBoost 모델에 SHAP을 적용하였다. 셋째, 기존의 피처 중요도와 SHAP 변수 중요도를 비교 분석하여 SHAP의 효용성을 알아보았다.

본 논문의 구성은 다음과 같다. 2장에서는 설명 가능한 AI에 관한 연구와 침입탐지 분야에서의 XAI 관련 연구를 기술하고 3장에서는 본 논문에서 제안하는 기법에 관해 설명한다. 4장에서는 제안된 기법에 대한 실험 및 평가 결과에 대해 분석하였다. 5장에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련 연구

본 논문에서는 설명 가능 AI 기법과 침입탐지 분야에서 설명 가능한 AI 관련 연구를 살펴본다.

2.1 설명 가능한 AI (XAI)

XAI는 인공지능의 설명성 제공을 목적으로 한다. 인공지능은 자동화 및 높은 정확도를 통해 기존의 탐지기법에 비해 많은 관심을 받고 있지만, 블랙박스(Black-box) 특성으로 인해 현장에서는 적용하기 어려움을 가지고 있다[10]. 인공지능은 본질적으로 입력과 출력 간의 관계가 선형적이지 못해 입력이 출력에 어떤 영향을 주는지 명확하게 알기 어렵다. 이에 2001년에는 퍼뮤테이션 피처 중요도(Permutation Feature Importance) [11][12], 부분 의존성 플롯(Partial Dependence Plots:PDP)[13]을 통해 모델을 설명하기도 한다. 하지만, 딥러닝 등의 발전으로 피처에 대한 생성 또한, 분석가들이 하는 게 아니라 인공지능이 자체적으로 학습에 필요한 피처를 생성하거나 선택할 수 있기 때문에 인공지능 내부에 대한 해석이 더욱 어렵게 되었다. 이에 설명 가능한 이러한 블랙박스 기반의 모델을 들여다보는 방법으로 전통적으로 피처 중요도를 통해 보는 방법과 설명 가능한 AI 기법인 SHAP, LIME 등을 적용하는 방법, 시각화 등을 활용하기도 한다[14].

이러한 인공지능에 대한 해석과 설명을 제공할 수 있는 기술인 XAI는 높은 수준의 성능을 유지하면서도 더 설명 가능한 모델을 생성함으로써 분석가가 인공지능을 이해하고 신뢰할 수 있도록 해준다.

2.2 침입탐지 분야의 XAI

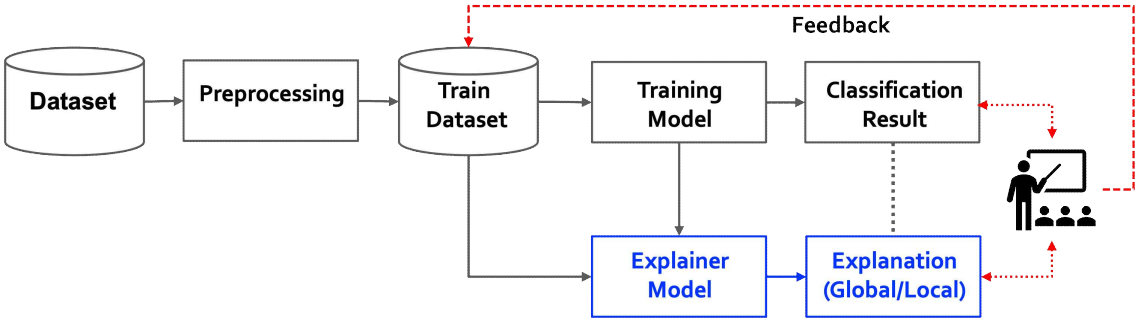
침입탐지 분야에 머신러닝(ML) 연구가 증가함에 따라 탐지 성능 향상과 이를 활용한 현장 적용이 증가하고 있다. 하지만, 많은 머신러닝 기반의 시스템이 블랙박스 기반의 모델로 예측된 결과에 대한 명확한 설명이 없으므로 기존의 탐지 패턴에 대한 설명이 명확한 시그니처 기반의 탐지기법보다 보안 전문가 관점에서 수용하기에 한계가 있다. 최근 침입탐지 분야에서도 기존의 블랙박스 기반의 모델에 대해 설명 가능한 AI를 적용하려는 연구가 많이 증가하고 있다[14]. Mane and Rao[15] 는 딥러닝 기반의 침입탐지시스템(IDS)에 대한 투명성을 최적화하기 위해 LIME, SHAP, ProtoDash 등을 활용한 XAI 프레임웍을 설계하였다. Wang et al.[16]은 침입탐지시스템(IDS)의 투명성과 설명 가능성을 추가하기 위해 SHAP 기반의 접근 방식을 설계하였으며, 설명 가능성을 향상하려는 방법으로 Local과 Global을 결합하였다. Wali and Khan[17]은 Random Forest 기반의 IDS에 대해 SHAP 기반의 설명 가능한 AI를 통해 예측 결과에 대한 설명을 시도하였다.

지금까지의 연구들이 기 정형화된 데이터셋(NSL-KDD[18], CICIDS[19])에 대해서 Random Forest와 딥러닝 등을 활용한 모델을 만들고 설명 가능한 AI를 통해 설명 가능성 등을 주장하였다면 본 논문에서는 침입탐지의 근거가 되는 패턴이 포함된 페이로드 기반의 데이터셋(PKDD2007)[20]을 선택하였으며, 부스팅 기반의 알고리즘(XGBoost)을 통해 모델을 생성하였다. 또한, 기존의 피처 중요도 기반과 설명 가능한 AI 기법 중 SHAP을 비교 분석한다.

3. 제안된 기법

3.1 제안 프로세스

전통적인 블랙박스 기반의 침입탐지 데이터셋에 설명 가능한 AI 알고리즘을 적용하여 모델에 대한 신뢰성을 향상 시키는 것이 본 프로세스의 목적이다. (그림 1)는 설명 가능한 AI를 적용한 프로세스를 나타내고 있다. 먼저, 수집된 데이터셋에 대해서 전처리 프로세스를 수행한다. 이렇게 전처리 된 데이터 중 학습에 필요한 데이터를 추출하여 Training Model을 생성한다. 이렇게 생성된 모델은 블랙박스 기반의 모델로 모델링 결과에 대한 피처 중요도를 통해 모델을 분석하고, 평가지표를 통해 평가를 수행한다. 이에 더불어 설명 가능한 AI 모델링을 통해 Explainer Model을 생성하게 된다. 보안 전문가는 분류 결과(Classificatio



(그림 1) 설명 가능한 AI 적용 프로세스

n Result)와 Explainer Model에서 도출된 Explanation 통해 분류 결과에 대한 신뢰성을 확보하게 된다. 또한, 잘못된 결과에 대해서는 피드백(Feedback)을 통해 기존의 모델을 더욱더 강화하게 된다.

3.2 피처 중요도 (Feature Importance)

데이터의 피처가 알고리즘에 대한 분류를 얼마나 정확히 하는지에 대한 영향도를 분석하는 기법이다. 피처 중요도는 특정 피처의 값을 임의의 값으로 치환했을 때 원래 데이터보다 예측 에러가 얼마나 더 커지는가를 측정하게 된다. 다음은 Fisher et al.(2019)[12] 가 제안한 피처 중요도 계산 방법이다.

1. 주어진 모델의 에러를 측정한다.

$$e^{original} = L(y, f) \tag{1}$$

2. X의 피처 k개(k=1,2,3...,p)에 대해서 피처 메트릭스 $X^{permutation}$ 를 만든 후 모델 에러($e^{permutation}$)를 측정한다.

$$e^{permutation} = L(X, f(X^{permutation})) \tag{2}$$

3. 피처 중요도 FI 를 구한다.

$$FI^k = e^{permutation} + e^{original} \tag{3}$$

3.3 SHAP(Shapley Additive exPlanations)

제안된 프로세스에 사용된 설명 가능한 AI 알고리즘은 SHAP(Shapley Additive exPlanations)을 사용하였다. SHAP은 SHapley Additive exPlanations의 약자로 모델 해석을 위한 잘 알려진 통합 프레임워크로 정의된다. SHAP은 최종 결정 예측에 대한 각 기능의 기여도를 계산하여 인스턴스의 예측을 설명한다. 기여는 음수 또는 양수일 수 있다. SHAP의 주요 장점은 선형 모델 분류기 대신 모든 모델 분류기에 적용할 수 있다는 것이다. 로컬 결정 해석만 검사하는 대신 SHAP은 기능의 입력값을 합산하고 모든 열 기능을 개별적으로 평균화하여 전역 해석을 검사한다[7].

$$\phi_i(v) = \sum_{S \in M \backslash i} \frac{|S|!\,(n - |Svert - 1)!}{n!} (v(S \cup i) - v(S)) \tag{4}$$

위 식에서 ϕ_i 는 i 번째 데이터에 대한 새플리(Shaply value)를 나타내며, n은 피처의 개수, S는 총 피처에서 i 번째 데이터를 제외한 모든 집합, $v(S)$ 는 i 번째 피처를 제외하고, 나머지 부분 집합이 결과에 공헌한 기여도, $v(S \cup i)$ 는 i 번째 피처를 포함한 전체 기여도를 나타낸다. 이처럼 SHAP은 모델 출력에 대해서 각 피처의 기여도를 분해한다.

3.4 XGBoost

부스팅(Boosting) 알고리즘은 약한 분류기(Weak Classifier)를 만들고 순차적으로 학습 후 학습 결과의 오차를 줄이는 방향으로 신규 분류기를 반복적으로 추가하는 앙상블 기술이다. 특히 기존의 Gradient Boosting 기법의 단점인 느린 학습 속도와 과적합 문제를 보완한 알고리즘이 XGBoost이다. XGBoost는 2016년 8월 Tianqi Chen[6]에 의해 소개되었다. XGBoost는 병렬 처리를 사용하기 때문에 학습과 분류가 빠르며, 평가 함수(Eval

uation function)를 포함해 다양한 최적화 옵션을 제공한다. 또한, Greedy Algorithm을 사용해 자동으로 Forest를 치기 때문에 과적합에 강하다. 하지만, 약한 분류기를 수없이 엮어놓기 때문에 다른 블랙박스 모델과 마찬가지로 특별한 후처리 없이 모델을 이해하기 어렵다. 따라서 XGBoost와 XAI와의 결합은 모델을 설명하는데 좋은 결합이 될 수 있다.

4. 실험 및 평가

이 세션에서는 침입탐지 데이터셋에 대해서 머신러닝 알고리즘을 통해 분류모델을 생성하고, 생성된 모델에 대해서 설명 가능한 AI 알고리즘인 SHAP을 적용하여 머신러닝 결과에 대한 해석 가능성을 알아본다.

우리는 본 실험을 통해 다음과 같은 질문을 해결하고자 한다. 블랙박스 기반의 침입탐지 모델의 예측 결과에 관해서 설명 가능한 AI를 통해 설명이 가능한가? 피처(Feature) 중요도와 설명 가능한 AI 알고리즘 SHAP 간의 관계는 어떻게 해석할 수 있는가?

이를 위해 본 실험에서 실험 데이터셋으로 페이로드가 포함된 웹어플리케이션 침입탐지 데이터셋인 PKDD2007[20]을 사용한다. 분류 알고리즘으로는 부스팅 기반의 알고리즘 XGBoost, 설명 가능한 AI 알고리즘으로는 SHAP을 사용하였다.

4.1 데이터셋

본 실험에 사용된 데이터셋 PKDD2007(ECML PKDD2007 Discovery Challenge Dataset)은 HTTP 요청 헤더와 페이로드가 포함된 웹 애플리케이션 침입탐지 데이터셋이다. HTTP 요청에 헤더와 페이로드가 모두 포함된 데이터셋을 선택한 이유는 보안 전문가들이 헤더와 페이로드를 통해 공격의 여부를 판단하기 때문이다. 데이터셋은 2007년 제18회 유럽 기계 회의(ECML)와 데이터베이스의 지식 발견 원칙 및 실무에 관한 제11차 유럽 회의(PKDD)에서 웹 트래픽 분석에 대한 챌린저(challenger)를 통해 만들어졌다. 챌린저(challenger) 목적으로 유효한 트래픽(Normal)과 7가지 유형의 공격(Attack)으로 분류된 요청을 포함하는 데이터셋이 제공되었다. 데이터셋에는 정상적으로 분류된 35,006건의 요청과 공격으로 분류된 15,110건의 요청이 포함된다. PKDD2007 데이터 세트는 실제 트래픽을 기록하여 생성된 후 정보 처리를 위해 처리되었다. 이 마스킹 프로세스는 모든 url, 매개 변수 이름 및 값을 임의로 생성된 문자열로 바꾸는 작업으로 구성되어 있다. 또한, 파일 형식은 XML 형식으로 되어 있으며, reqContext, class, request로 구성되어 있으며, request는 method, Protocol, URL, Query, Headers, Body로 구분되어 있다. <표 1>은 PKDD2007 데이터셋에서 Valid를 포함한 8가지 유형별 데이터 수를 나타낸다.

<표 1> PKDD2007 데이터셋

NO	공격유형	데이터 수
1	Valid	35,006
2	Cross-Site Scripting	1,825
3	SQL Injection	2,274
4	Path Traversal	2,295
5	LDAP Injection	2,279
6	XPATH Injection	2,279
7	Command execution	2,302
8	SSI attacks	1,856

4.2 데이터 전처리

전처리 단계는 수집된 데이터셋(PKDD2007)을 머신러닝에 적용하기 전 데이터를 정형화 및 전처리를 수행하는 단계이다. 이러한 전처리에서는 Normalization, Filed Selection, Feature Extractor & Selection, Sampling 등의 단계가 수행된다.

4.2.1 NORMALIZATION

해당 데이터셋은 비정형화된 XML 형식의 데이터로 구성되어 있다. 따라서 해당 데이터셋에 대해 Method, Version, http_url, http_query, http_body로 Normalization을 수행한다. 특히, 사용자 정보 등을 나타내는 값에 대해서는 Body 필드로 포

함한다. 그리고 Uri, Query, Body에 대해서 '\n' 문자를 제거하고, Uri Decoding을 적용한다.

4.2.2 FIELD SELECTION

각 데이터셋에서 실험에 사용할 필드에 대해 선택한다. Category Type은 Class, Method, Version으로 구분할 수 있으며, Text Type은 Uri, Query, Body를 선택한다.

4.2.3 FEATURE EXTRACTOR & SELECTION

분리된 http_url, http_query, http_body 필드에 대해서 <표 2>와 같이 공격유형에 따른 키워드, 빈도수, 형식 유무에 따라 피처를 추출하고 선택하였다.

<표 2> 피처 추출 및 설명 (일부)

피처	유형	설명
host_category	Category	host의 표현 식이 IP 형식인지 여부
weak_http_version	Category	http_version이 'HTTP/1.1'인지 여부
url_length,	numeric	http_uri 길이
url_query_length	numeric	http_query 길이
body_length	numeric	http_body 길이
path_length	numeric	http_uri의 path 길이
upperchar_freq	numeric	http_uri + http_query에서 대문자의 빈도수

4.3 실험 환경 및 평가 방법

본 실험 환경은 <표 3>과 같이 Ubuntu 20.04 LTS에서 Python 3.7을 사용하여 구현되었다. 사용된 고전적인 기계 학습 알고리즘은 Scikit-learn 0.20.4 을 사용하였다. 하드웨어 사양은 GPU는 NVidia Geforce RTX 2060 이었으며 128GB RAM, 8TB 하드 디스크, AMD Ryzen Threadripper 1900X 8-Core Processor 환경이다.

<표 3> Experimental environment

Project	Environment/Version
Operating System	Ubuntu 20.04 LTS
Framework	Scikit-learn 0.20.4
Backend	Engine Tensorflow
CPU	AMD Risen 1900X 8-Core
GPU	NVidia Geforce RTX 2060
Memory	128 G

본 실험의 목적은 전통적인 머신러닝을 통해 분류된 값에 관해서 설명 가능한 머신러닝을 통해 설명할 수 있는지 알아보는 데 있다. 이에 대한 평가지표로는 일반적으로 사용되는 Confusion Matrix[21]를 기반으로 기본적으로 사용되는 Accuracy(ACC), Precision, Recall, F1-score 사용하였다. Accuracy는 모든 샘플 중에서 정상과 공격이 올바르게 분류된 항목의 비율로 정의된다. Precision은 공격이라고 예측한 것 중 실제 공격이라고 분류한 비율을 말한다. Recall은 실제 공격 중 공격이라고 예측한 비율을 말한다. F1-score는 Precision과 Recall 간의 조화평균(harmonic mean)을 의미한다. 평가지표별 공식은 <표4> 와 같다.

<표 4> 평가지표별 공식

평가지표	공식
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
Precision	$TP / (TP + FP)$
Recall	$TP / (TP + FN)$
F1-score	$2 \times (Recall \times Precision) / (Recall + Precisio$

4.4 실험 결과

본 실험의 목적은 전통적인 머신러닝 결괏값에 관해서 설명할 수 있는 알고리즘인 SHAP 적용을 통해 얼마나 해석할 수 있는지를 알아보는 데 있다. 먼저, 학습 데이터에 대한 전통적인 머신러닝 알고리즘을 적용하여 모델을 생성하고 평가하였다.

4.4.1. 모델 성능

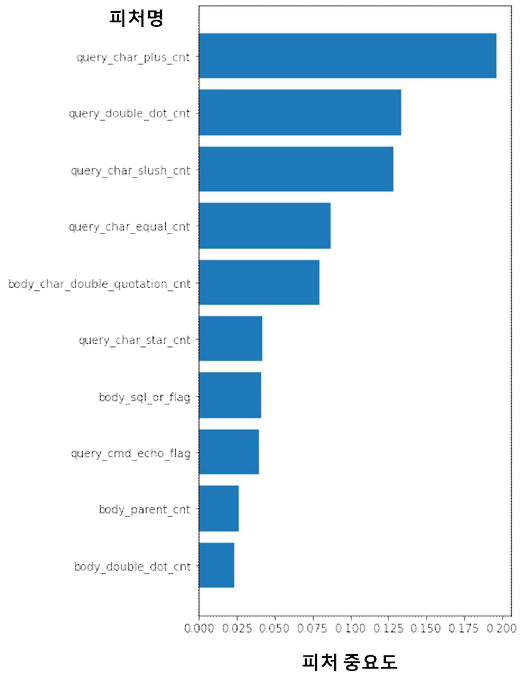
다음은 침입탐지 데이터셋에 대한 다양한 알고리즘을 통해 모델을 생성하였으며, 이에 대한 성능 평가표는 <표5>와 같다.

<표 5> 성능평가표

	ACC	Precision	Recall	F1-score
LR	0.8872	0.7694	0.7545	0.7538
MLP	0.9036	0.8000	0.7900	0.7909
XGBoost	0.9051	0.8273	0.7928	0.7974
DT	0.7858	0.7787	0.5212	0.5267
RF	0.8891	0.8198	0.7599	0.7625

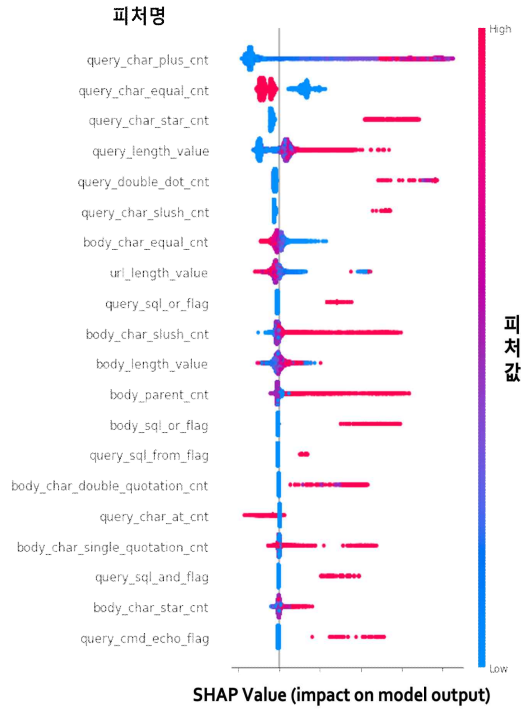
해당 <표 5>에서 보는 바와 F1-score 기준으로 XGBoost > MLP > RF(Random Forest) > LR (Logistic Regression) > DT(Decision Tree) 순으로 성능이 나타나고 있다. 또한, ACC(Accuracy) 기준으로는 XGBoost > MLP > RF > LR > DT 순으로 나타낸다. 이처럼 최근에는 Boosting 기반의 알고리즘(XGBoost, RF) 나 신경망(MLP) 기반의 알고리즘이 좋은 성능을 내고 있음을 알 수 있다. 이에 본 논문에서는 Boosting 기반의 알고리즘 중 XGBoost 알고리즘을 통해 만들어진 모델을 선정하여 피처 중요도와 설명 가능한 알고리즘 SHAP을 적용하였다.

4.4.2 피처(Feature) 중요도



(그림 2) 모델에 대한 피처 중요도

탐지 모델 결과를 분석하고 예측 결과에 중요하게 작용했던 피처(Feature)를 선정하였다. 피처 중요도는 데이터의 피처가 알고리즘의 정확한 분류에 얼마나 큰 영향을 미치는지를 분석하는 기법이다. 이에 탐지 모델 결과를 분석하고 결과를 예측하는데 많이 사용된다. 특정 피처의 값을 임의의 값으로 치환했을 때 원래 데이터보다 예측 에러가 얼마나 더 커지는가를 측정하는 것이다. (그림 2)는 모델에 대한 피처 중요도를 나타내고 있다. 가장 중요도가 높은 피처는 query_char_plus_cnt 이며, 해당 피처는 데이터 중 query 부분에 "+" 문자열 카운터가 많을 때 영향을 미치고 있다. 또한, query_double_dot_cnt, query_char_slush_cnt 등과 같이 query 부분에 관련된 문자열의 영향도가 크게 나타남을 알 수 있다. 하지만, 이러한 피처 중요도는 피처가 독립적일 잘 나타나며, 학습 데이터 전체를 기반으로 한다. 그러나 피처 중요도는 좋은 영향인지, 나쁜 영향인지 방향성이 없으며, 피처 간 의존성이 존재하면 신뢰하기가 어렵다.



(그림 3) SHAP의 Summary Plot

(그림 4) 개별 데이터(rn=60689)에 대한 SHAP 결과

4.4.3 SHAP 기반의 모델 분석

다음은 4.4.2에서 XGBoost 기반의 모델에서 SQL Injection에 대한 shap value를 구하고, Valid 데이터에 대한 SHAP의 Summary Plot을 사용하였다. Summary Plot은 Valid 데이터에 대한 AI 모델의 Feature Importance와 Feature Influence를 시각화하여 (그림 3)과 같은 plot을 제공한다. Summary Plot X축은 SHAP value를 의미하고 Y축은 SHAP value의 절댓값의 합의 순서대로 나열되어있고 Point의 색은 실제 feature value의 높고 낮음을 의미한다.

다음은 SHAP을 통해 XAI local 시각화에 대한 부분을 나타낸다. (그림 4)에서 보는 바와 같이 시각화는 Log odds 형태의 그래프와 확률값 형태 변환 그래프로 나타낼 수 있으며, 두 번째, 확률값 형태 변환 그래프에서 나타낸 것과 같이 개별 데이터인 60689번째 대한 예측 확률은 0.97이다. 해당 데이터에 대한 긍정적인 영향을 미치는 피처는 붉은색으로 body_sql_from_flag, body_char_slush_cnt 등이 크게 미치며, query_char_plus_cnt, query_char_equal_cnt 형태의 파란 색은 부정적인 영향을 미치고 있음을 나타낸다.

이처럼 피처 중요도와 SHAP을 비교해 보면 피처 중요도는 피처가 서로 의존적이라면 잘못 계산될 수 있으며, SHAP은 예측 결과에 대해서 각 데이터의 결정에 기준 피처가 어떤 영향을 미치는지 설명할 수 있음을 알 수 있다.

5. 결론

본 연구를 통해 침입탐지 분야의 블랙박스 머신러닝 모델에 설명 가능한 AI를 적용하여 예측 결과에 대한 타당성을 검증하려 하였다. 이를 위해 XGBoost를 통해 침입탐지 모델을 구현하고, SHAP을 활용하여 모델에 대한 설명을 추가하였다. 이러한 프로세스를 수행 후 기존의 피처 중요도와 SHAP을 활용한 결과를 비교 분석하여 설명 가능한 AI에서 나온 결과가 기존의 피처 중요도보다 침입탐지에 대한 타당성을 더욱더 정확하게 보여줌을 확인하였으며, 이는 결론적으로 보안 전문가가 정확한 판단을 수행하는데 투명성과 신뢰를 제공함을 확인할 수 있었다.

참고문헌

[1] “2021 국가정보보호백서”, KISA, 2021.5
[2] Capgemini Research Institute(2019), 'Reinventing Cybersecurity with Artificial Intelligence: The new frontier in digital security', 11 July 2019.
[3] Cisco(2019), 2019 연례 사이버보안 보고서.
[4] Barnard, Pieter & Marchetti, Nicola & Silva, Luiz. (2022). Robust Network Intrusion Detection through Explainable Artificial Intelligence (XAI). IEEE Networking Letters. 1-1. 10.1109/LNET.2022.3186589.
[5] D. Gunning and D. Aha, “DARPA’s Explainable

Artificial Intelligence (XAI) Program", AIMag, vol. 40, no. 2, pp. 44-58, Jun. 2019.

[6] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

[7] D. Fryer, I. Strümke and H. Nguyen, "Shapley Values for Feature Selection: The Good, the Bad, and the Axioms," in IEEE Access, vol. 9, pp. 144352-144360, 2021, doi: 10.1109/ACCESS.2021.3119110.

[8] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[9] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[10] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), 4793-4813.

[11] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[12] Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res., 20(177), 1-81.

[13] Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. R J., 9(1), 421.

[14] Srivastava, Gautam & Jhaveri, Rutvij & Bhattacharya, Sweta & Pandya, Sharnil & Rajeswari, & Reddy, Praveen & Yenduri, Gokul & Hall, Jon & Alazab, Mamoun & Gadekallu, Thippa. (2022). XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions.

[15] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," 2021, arXiv:2103.07110.

[16] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," IEEE Access, vol. 8, pp. 73127-73141, 2020.

[17] S. Wali and I. Khan. "Explainable AI and Random Forest Based Reliable Intrusion Detection System." Dec. 2021. [Online]. Available: https://www.techrxiv.org/articles/preprint/Explainable_AI_and_Random_Forest_Based_Reliable_Intrusion_Detection_system/ 17169080.

[18] Dhanabal, L., & Shantharajah, S. P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International journal of advanced research in computer and communication engineering, 4(6), 446-452.

[19] Gopalan, S. S., Ravikumar, D., Linekar, D., Raza, A., & Hasib, M. (2021, March). Balancing approaches towards ML for IDS: a survey for the CSE-CIC IDS dataset. In 2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA) (pp. 1-6). IEEE.

[20] Koronacki, J. N. K. J., Matwin, R. L. D. M. S., & Skowron, D. M. A. Knowledge Discovery in Databases: PKDD 2007.

[21] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. MAICS, 710(1), 120-127.

〔저자소개〕

정 일 옥 (Il-ok Jung)
2001년 2월 전남대학교 물리학과 학사
2008년 8월 고려대학교 컴퓨터공학과 석사
2021년 8월 고려대학교 정보보호학과 박사
email : okkida@korea.ac.kr

김 수 철 (Soo-chul Kim)
2008년 8월 고려대학교 컴퓨터공학과 석사
2022년 8월 숭실대학교 IT정책경영학과 박사 수료
email : kscfuture@naver.com

최 우 빈 (Woo-bin, Choi)
2018년 8월 경희대학교 응용수학과 학사
email : khuam1216@gmail.com