JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 6, pp. 47-54, June 2023
https://doi.org/10.9708/jksci.2023.28.06.047

Intrusion Detection System based on Packet Payload Analysis using LightGBM

Gun-Nam Kim*, Han-Seok Kim*, Soo-Jin Lee*

*Graduate Student, Dept. of Defense Science, Korea National Defense University, Nonsan, Korea
*Graduate Student, Dept. of Defense Science, Korea National Defense University, Nonsan, Korea
*Professor, Dept. of Defense Science, Korea National Defense University, Nonsan, Korea

[Abstract]

Most studies on machine learning-based intrusion detection systems use metadata. However, since metadata is information generated by analyzing packets, it is difficult to ensure real-time intrusion detection in a real network environment. Therefore, in this paper, we proposed a machine learning-based intrusion detection system that can quickly detect network intrusions by directly analyzing the payload of packets. The UNSW-NB15 Dataset and the LightGBM model were used to verify the detection performance of the proposed technique. We first used the 'Payload-Byte' technique to label PCAP files in the Dataset, then conducted learning with the LightGBM model and analyzed detection performance. Experimental results showed that our approach can achieve a significant improvement in the binary classification with accuracy of 99.33% and F1-score of 98.73%. However, Multi-class classification showed similar detection performance to previous studies with accuracy of 85.63% and F1-score of 85.68%.

▸ **Key words**: Machine Learning, IDS, Packet Payload, LightGBM, UNSW-NB15

[요 약]

기계학습 기반의 침입탐지시스템에 대한 연구는 대부분 메타데이터를 활용한다. 그러나 메타데이터는 패킷을 분석해서 생성되는 정보이기 때문에, 실제 네트워크 환경에서 실시간 침입탐지를 보장하기는 어렵다. 이에 본 논문에서는 패킷의 페이로드(payload)를 직접 분석하여 신속하게 네트워크 침입을 탐지할 수 있는 기계학습 기반의 침입탐지시스템을 제안하였다. 제안하는 기법의 성능을 검증하기 위해 UNSW-NB15 데이터세트와 LightGBM 모델을 활용하였다. 먼저 'Payload-Byte' 기법을 활용하여 데이터세트 내의 PCAP 파일에 대한 라벨링을 실시한 후 LightGBM 모델로 학습을 실시하고 탐지성능을 분석하였다. 실험 결과 이진 분류는 정확도 99.33%, F1-score 98.73%를 달성하여 탐지성능이 크게 향상됨을 확인하였다. 그러나 다중 분류는 정확도 85.63%, F1-score 85.68%로 선행연구와 유사한 탐지성능을 보였다.

▸ **주제어**: 기계학습, 침입탐지시스템, 패킷 페이로드, LightGBM, UNSW-NB15

• First Author: Gun-Nam Kim, Co-Author: Han-Seok Kim, Corresponding Author: Soo-Jin Lee
*Gun-Nam Kim (rjsska2918@gmail.com), Dept. of Defense Science, Korea National Defense University
*Han-Seok Kim (14.10083a@gmail.com), Dept. of Defense Science, Korea National Defense University
*Soo-Jin Lee (cyberkma@gmail.com), Dept. of Defense Science, Korea National Defense University
• Received: 2023. 05. 09, Revised: 2023. 05. 31, Accepted: 2023. 06. 08.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

4차 산업혁명 시대가 되면서 인터넷의 활용은 기하급수적으로 증가하고 있다. 과학기술정보통신부의 발표 자료에 따르면, 2018년 12월과 비교하였을 때 최근 무선 통신량은 218% 증가하였으며, 초고속 인터넷 회선 수는 11.7% 증가하였다.(23년 2월 기준)[1] 이렇듯 통신량이 증가한 만큼 네트워크를 통해 시스템에 침입해 개인정보를 탈취하거나 시스템을 마비시키는 등의 사이버 위협도 함께 증가하고 있어 이러한 위협에 대응하기 위한 수단 중 하나인 침입탐지시스템(Intrusion Detection Systems)의 중요성은 나날이 증가하고 있다.

침입탐지시스템은 각종 네트워크 공격으로부터 기관이나 개인의 내부 자원을 보호하는 시스템으로서, 탐지 방법에 따라, 사전에 입력된 공격 패턴과 일치하면 공격으로 판단하는 오용 탐지(misuse detection)와 정상적인 패턴에서 벗어난 행위를 공격으로 인식하는 이상행위 탐지(anomaly detection)로 구분된다. 그러나 공격방식이 점점 더 다양화되고 교묘해지면서 정상 패턴이나 공격 패턴을 사전에 정의하여 적용하는 것이 어려워지고 있는 상황이다.

이러한 문제점을 해결하기 위해 최근에는 기계학습을 활용해 침입탐지시스템을 구축하려는 연구가 활발히 진행되고 있다. 그러나 대부분의 연구는 패킷의 내용 자체가 아니라 패킷을 분석하여 생성하는 정보인 메타데이터(metadata)를 활용하여 모델을 학습시킨다. 이러한 메타데이터 기반의 침입탐지 모델은 실제 네트워크로 유입되는 패킷에 대한 분석을 통해 메타데이터가 생성되어야만 동작할 수 있다. 즉, 기존 접근방법은 구축된 침입탐지시스템들은 실제 네트워크 환경에서 동작할 때 메타데이터 생성을 위한 패킷 분석 시간이 추가로 필요하여 실시간 침입탐지를 보장하는 것이 제한될 수 있다. 그리고 통신량이 급증하고 있는 현재 네트워크 환경에서는 그러한 문제가 더욱 심각해질 수 있다.

이에 본 연구에서는 침입탐지의 실시간성을 보장하면서 탐지성능도 개선하기 위해 네트워크로 유입되는 패킷의 페이로드를 특성 공간에 두고 이를 직접 학습하여 침입을 탐지하는 기계학습 기반 침입탐지시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 모델 구축 및 성능평가 실험에 사용된 UNSW-NB15 데이터세트에 대해 설명하고, 선행연구를 정리한다. 3장에서는 제안하는 침입탐지시스템의 구축절차와 실험 방법을 설명하고, 실험 결과를 분석한다. 마지막으로 4장에서 결론을 맺는다.

II. Preliminaries

1. UNSW-NB15 Dataset

본 논문에서 제안하는 침입탐지시스템의 구축 및 성능평가에 사용된 데이터세트는 UNSW-NB15이다[2]. 호주 사이버보안센터(Australian Center for Cyber Security)의 UNSW Cyber Range Lab에서 'IXIA PerfectStorm'을 사용하여 네트워크 패킷을 수집한 후, 'Argus'와 'Bro-IDS Tool'을 이용하여 정상 및 비정상 트래픽을 분류한 공공 데이터세트로서, PCAP 및 CSV 파일 형태로 제공되고 있다[3]. 본 연구는 패킷 페이로드 분석을 기반으로 한 침입탐지시스템 구축이 주목적이기 때문에 페이로드 확인이 가능한 PCAP 형식의 데이터세트만을 사용하였다.

UNSW-NB15 데이터세트는 175,341개의 학습(Train) 데이터와 82,232개의 테스트(Test) 데이터로 구성되어 있으며, 정상 데이터와 비정상 데이터의 개수와 유형은 Table 1에서 보는 바와 같다. 학습 특성은 총 45개이며, 비정상 데이터는 9개의 공격유형으로 분류되어 있다[4]. 다만, UNSW-NB15 데이터세트의 PCAP 파일 크기가 매우 크고, 이 중에는 중복된 패킷과 페이로드가 없는 데이터가 포함되어 있어 실험의 효율성을 위해 이 부분을 우선 제거하였다. 그리고 정상 데이터와 비정상 데이터의 불균형 비율을 줄이기 위해 언더샘플링(Undersampling)하였다. 실제 실험에 사용한 데이터의 개수와 유형은 Table 2에서 보는 바와 같다.

Table 1. Configuration of UNSW-NB15 Dataset

Category	Train	Test
Analysis	2,000	677
Backdoor	1,746	583
DoS	12,264	4,089
Exploits	33,393	11,132
Fuzzers	18,185	6,062
Generic	40,000	18,871
Normal	56,000	37,005
Reconnaissance	10,492	3,496
Shellcode	1,133	378
Worms	130	44
Total	175,343	82,337

Table 2. Configuration of Experimental Dataset

Category	Total	Train	Test
Analysis	1,208	966	242
Backdoor	1,239	991	248
DoS	3,397	2718	679
Exploits	13,992	11,193	2,799
Fuzzers	12,722	10.178	2,544
Generic	17,580	14,064	3,516
Normal	21,000	16,800	4,200
Reconnaissance	7,562	6,050	1,512
Shellcode	1,088	870	218
Worms	93	74	19
Total	79,881	63,904	15,977

2. Related works

패킷(packet)은 컴퓨터에서 네트워크가 전달하는 데이터의 형식화된 블록이며, IP 패킷은 헤더(header)와 페이로드(payload)로 구성된다. 페이로드에는 송신자가 전송하고자 하는 내용이 포함되며, 이러한 페이로드에 비정상 데이터를 삽입하는 방식으로 네트워크 침입이 이루어질 수 있기 때문에 패킷 분석 기반의 침입탐지시스템을 구축하려는 연구가 지속적으로 수행되고 있다.

Jung 등[5]은 패킷의 페이로드를 분석하여 특정 패턴을 검출하는 4가지 방식을 비교, 분석하여 어떤 방식이 가장 효율적인 방식이 될 수 있는지를 제시하였다. 실험 결과 99%의 확률로 특정 패턴을 검출하는 것은 가능하였으나, 사전에 패턴이 정해져 있어야 한다는 점과 이진 분류만이 가능하다는 부분에서 한계가 있다.

Wang과 Stolfo[6]는 정상적 활동과 구분되는 패턴을 탐지하기 위해서 페이로드 특성(payload feature)의 분포를 활용하여 분류하는 방법을 제안하였다. 그러나 페이로드 특성은 문자들의 분포와 관련이 있어 의미 있는 정보를 반영하는데 한계를 가진다.

Mahoney와 Chan[7]은 포트 번호, TCP(Transmission Control Protocol) 플래그 및 기타 패킷 헤더 정보와 같은 메타데이터를 사용하여 침입을 탐지하는 PHAD(Packet Header Anomaly Detection)를 제안하였다. 추가로 제안한 NETAD(Network Anomaly Detection)는 패킷 헤더의 처음 48바이트를 사용하여 각 인터넷 프로토콜에 따라 다른 침입탐지 모델을 생성하였으나, 오경보율(false alarm rate)이 높다.

전통적인 방법의 한계를 해결하기 위해서 최근에는 CNN(Convolutional Neural Network), RNN (Recurrent Neural Network), LSTM(Long Short Term Memory) 등 기계학습(딥러닝)을 기반으로 한 침입탐지 연구가 활발하게 진행되고 있다[8-10].

Kim 등[11]은 패킷의 개수만을 이용하여 네트워크 기반 공격을 탐지할 수 있는 알고리즘과 이를 이용한 침입탐지 시스템을 제안하였다. Lee[12]은 딥러닝을 활용하여 DDoS 공격 중 TCP SYN Flood 공격을 패킷 카운팅으로 탐지하는 시스템을 제안하였다. LSTM 모델을 활용하여 학습시킨 결과 96%의 탐지율을 달성하기는 했지만, 단일 공격 패턴에 대한 이진 분류만 실시하였으며, 다양한 네트워크 기반 공격에 대해서는 성능을 검증하지 않았다.

본 연구와 동일한 UNSW-NB15 데이터세트를 이용하여 네트워크 기반 침입탐지시스템 네트워크 기반 침입탐지시스템을 구축하고자 하는 시도도 진행되었다. Jing과Chen 등[13]은 SVM(Support Vector Machine)을 적용하여 이진 분류는 85.99%, 다중 분류는 75.77%의 정확도를 달성하였다. Kabir 등[14]은 이진 분류에서 SVM(Support Vector Machine) 알고리즘을 사용하여 정상/비정상 트래픽을 분류했을 때 가장 높은 정확도(82.11%)를 달성하였으며, 다중 분류에서는 SVM으로 1차 분류된 결과를 Decision Trees에 적용하여 86%의 정확도를 달성하였다. Meghdouri 등[15]은 주성분분석(Principal Component Analysis, PCA)을 사용하여 데이터를 전처리 후 실험을 진행하였는데, 적용 모델 중 RF(Random Forest)가 Precision 84.9%, Recall 85.1%로 가장 높은 탐지성능을 보였다.

III. The Proposed Scheme

1. Experimental Preparation

1.1 Dataset Labeling

UNSW-NB15의 패킷 페이로드를 활용하기 위해서는 PCAP 형식의 파일을 통해 페이로드를 직접 확인해야 하지만 PCAP 파일은 라벨링이 되어 있지 않아 모델 학습에 직접 활용하는 것은 불가능하다. 이러한 라벨링 문제를 해결하기 위해 본 연구에서는 Farrukh 등이 제안한 'Payload-Byte' 기법[16]을 통해 PCAP 파일에 대한 라벨링을 수행하였다.

Payload-Byte 기법은 Fig. 1에서 보는 바와 같이 하나의 데이터세트에 대한 메타데이터와 PCAP 파일이 존재할 경우, 메타데이터를 기반으로 PCAP 파일 내에서 해당 패킷을 찾아 메타데이터와 같은 클래스로 분류한다. 탐색 된 패킷은 1바이트당 하나의 특성으로 분류되며 패킷 페이로드의 최대 크기가 사실상 1,500바이트이기 때문에

Payload-Byte 기법으로 변환된 데이터세트는 1,500개의 특성을 가지게 된다.

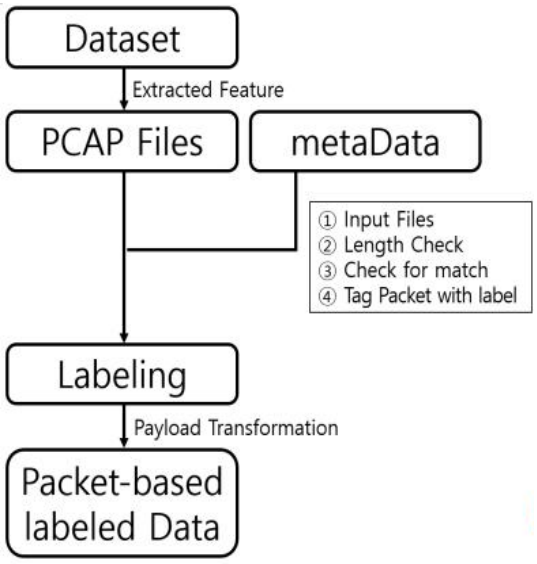

Fig. 1. Concept of Payload-Byte

1.2 LightGBM

LightGBM은 Gradient Boosting 기반의 기계학습 모델로서, 적은 메모리를 사용하여 실행 속도가 빠르면서도 높은 정확도를 가지기 때문에 최근 다양한 연구에 활용되고 있다. 데이터세트 샘플을 줄이는 GOSS(Gradient-based One-Side Sampling)와 데이터세트의 특성 수를 줄이는 EFB(Exclusive Feature Bundling) 알고리즘을 활용하기 때문에 기존 Gradient Boosting 기반 모델인 XGBoost와 비교했을 때 속도와 정확도 측면에서 모두 높은 성능을 보인다. 또한, EFB 알고리즘을 통해 자체적으로 특성 차원을 줄일 수 있으므로 특성 차원이 큰 데이터세트를 활용할 때도 모델의 학습이나 분류 측면에서 다른 프레임워크에 비해 더 좋은 성능을 보인다는 장점을 가진다[17].

본 연구에서 사용한 UNSW-NB15 데이터세트의 특성은 총 45개이지만, Payload-Byte 기법을 적용하여 새롭게 생성된 데이터세트는 1,500개의 특성을 가지기 때문에 특성 공간이 매우 크다고 볼 수 있다. 그리고 실시간 침입탐지를 보장하기 위해서는 모델의 실행 속도 역시 빨라야 하므로 본 연구에서는 LIghtGBM 모델을 가장 적합한 학습 모델로 판단하였다.

2. Experimental Design

본 연구에서 적용한 데이터세트는 메타데이터가 아닌 PCAP 형식이며, 라벨링이 되어 있지 않은 상태이기 때문에 그대로 학습에 활용할 수는 없다. 따라서 앞서 언급한 바와 같이 Payload-Byte 기법을 이용하여 라벨링 작업을 수행해야 하며, 라벨링 작업이 수행되는 과정에서 16진수로 표현된 페이로드는 자동으로 10진수로 변환된다. 그리고 10진수로 변환된 페이로드를 이용하여 1바이트당 하나의 특성을 생성한다. 이때, 페이로드가 1,500바이트보다 작은 경우에 실제 페이로드 이외의 영역은 특성 공간이 null로 채워진다.

모델 학습에 필수적인 라벨링을 위해 사용한 Payload-Byte 기법 내에서 16진수를 10진수 형태로 자동 변환하기는 했지만, Python에서는 16진수 앞에 '0x'를 붙이면 정수로 인식시킬 수 있기 때문에 10진수 변환 작업이 불필요하다. 따라서 라벨링이 되어 있는 PCAP 형식의 데이터세트만 확보할 수 있다면 패킷 페이로드를 그대로 학습시킬 수 있게 되며, 학습된 모델이 침입탐지를 수행할 때도 패킷 페이로드를 그대로 활용할 수 있어 실시간 침입탐지가 보장될 수 있다.

한편, PCAP 파일 내의 데이터 중 가장 큰 페이로드의 크기는 1,380바이트이기 때문에, Payload-Byte 기법을 통해 변환된 데이터세트 내의 모든 데이터는 최소 120개 이상의 특성이 null 값을 가진다. 따라서 null 특성이 모델 학습이나 성능에 어떤 영향을 미칠 수 있는지 확인하기 위해 성능 평가 실험은 2단계로 구분하여 진행하였다. 초도 실험은 null 값을 포함하여 1,500개의 특성을 모두 학습시킨 경우와 120개의 특성을 삭제한 후 학습시킨 경우의 탐지성능을 비교하는 형태로 진행하였으며 10회 반복 실험을 통해 확인한 결과 탐지성능이 거의 동일하게 나타났다. 이에 본 실험에서는 특성을 삭제하지 않고, 변환된 데이터세트를 원본 형태 그대로 LightGBM 분류 모델에 적용하여 성능평가를 수행하였고, 세부적인 실험 과정은 Fig 2에서 보는 바와 같다.

실험은 Windows 11 Pro 기반에 11세대 Gen Intel i5-11300H CPU와 40GB RAM이 탑재된 데스크탑 환경에서 진행하였고, 사용된 개발언어는 Python 3.10이다.

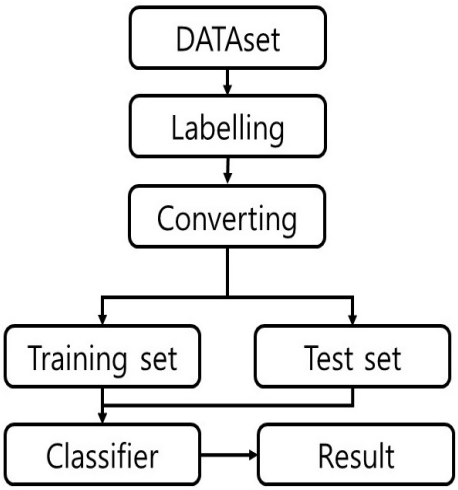

Fig. 2. Experimental Design

3. Parameter Setting

반복 실험을 통해 LightGBM 모델이 가장 우수한 성능을 달성하는 최적의 하이퍼파라미터를 선정하였으며, 실험에 적용한 하이퍼파라미터는 Table 3에서 보는 바와 같다.

Table 3. Hyperparameters of LightGBM Model

Category	Hyperparameters
Binary Classification	n_estimators= 1000, boost_from_average= False, num_leaves= 31, learning_rate= 0.1, max_depth= -1
Multi-class Classification	n_estimators= 181, boost_from_average= False, num_leaves= 100, learning_rate= 0.1, max_depth= 8, bagging_seed= 16

4. Results and Analysis

변환된 원본 형태의 데이터세트를 그대로 LightGBM 분류 모델에 적용하여 실험한 결과, Table 4에서 보는 바와 같이 이진 분류에서 정확도는 99.33%, F1-score는 98.73%, Precision과 Recall은 각각 98.16%, 99.31%를 달성하였으며, 동일 데이터세트를 활용하여 이진 분류를 시도했던 선행연구[12-14] 대비 탐지성능이 향상됨을 확인할 수 있었다.

다중 분류의 경우 F1-score를 측정하는 방법은 다양할 수 있으나, 본 연구에서는 데이터가 불균형할 경우 이를 반영할 수 있는 Micro F1-score를 활용하였다. 다중 분류에서는 Table 5에서 보는 바와 같이 정확도 85.63%, F1-score 85.68%, Precision과 Recall은 각각 87.88%, 85.61%로서, 선행연구[13-15]와 거의 유사한 탐지성능을 보였다. 그러나 탐지 속도 측면에서 본 연구는 Fig 3에서 보는 바와 같이 메타데이터를 생성하는 단계가 생략되기 때문에 선행연구들보다 탐지 속도가 더 빨라진다는 점에서 성과가 있다고 볼 수 있다. 실제 메타데이터 생성 과정은 캡처된 패킷을 분석하고 이해하기 쉬운 형태로 변환되는 작업을 거친 후 모델 학습에 사용되며 이 과정에는 일정 시간이 소요된다. 또한, Li 등[18]은 패킷 캡처 및 처리 기술에 대한 비교 연구를 진행하였는데 실험 결과에 따르면 어떤 패킷 캡처 및 처리 도구를 사용하더라도 패킷을 분석하는 과정이 존재하고 이러한 과정에서 약간의 패킷 손실이 발생한다는 사실을 알 수 있다. 이를 통해 본 연구에서 제안하는 침입탐지 방법은 메타데이터 생성에 필요한 절차가 생략되고, 생성 과정에서 발생하는 시간이 줄어 탐지 속도가 향상된다고 볼 수 있다. 다중 분류의 혼동행렬은 Fig. 4에서 확인할 수 있다.

Table 4. Experimental Results(Binary Classification)

Category	Proposed	[12]	[13]	[14]
Main Approach	LightGBM	LSTM	SVM	SVM +DT
Accuracy (%)	99.33	96	85.99	82.11
F1-score (%)	98.73	-	-	-
Precision (%)	98.16	-	-	-
Recall (%)	99.31	-	-	-

Table 5. Experimental Results(Multi-class Classification)

Category	Proposed	[13]	[14]	[15]
Main Approach	LightGBM	SVM	SVM + DT	PCA +RF
Accuracy (%)	85.63	75.77	86	-
F1-score (%)	85.68	-	86	84.9
Precision (%)	87.88	-	-	851.
Recall (%)	85.61	-	-	84.9

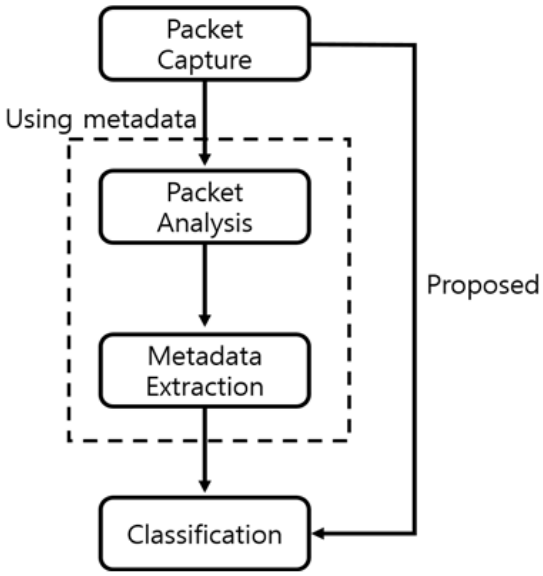

Fig. 3. Comparison of Detection Process

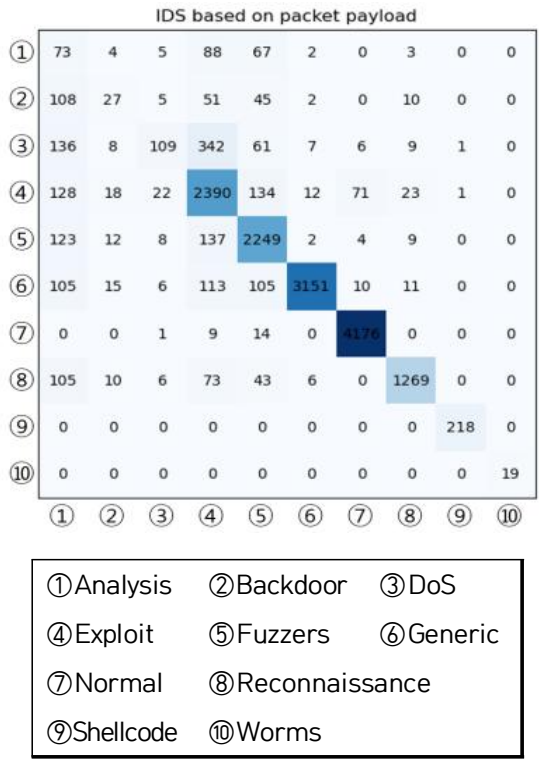

Fig. 4. Confusion Matrix of Multi-class Classification

IV. Conclusions

최근 증가하는 통신량만큼 네트워크를 통한 공격 또한 빠르게 증가하고 있다. 따라서 본 연구에서는 네트워크 침입탐지시스템에서 가장 중요한 요소들인 실시간성과 성능을 동시에 보장하기 위해 패킷의 페이로드를 특성 공간에 두고 이를 학습하여 침입을 탐지하는 기계학습 기반 침입탐지시스템을 제안하였다.

네트워크 침입탐지를 다루고 있는 선행연구들은 대부분 패킷의 내용 자체가 아닌 메타데이터를 기반으로 침입을 탐지하는 접근방법을 채택하였다. 메타데이터는 구조화된 형식에 포함된 '데이터에 대한 데이터', '정보에 대한 정보'로서, 데이터베이스에 수록된 데이터 집합을 기술하는 정보를 담고 있는 데이터를 말한다[19]. 그러나 메타데이터는 표준화된 형식이 존재하는 것이 아니며, 특정 연구나 데이터세트에 따라 달라지는 등 다양하게 정의될 수 있다. 반면 패킷의 페이로드는 실제 전송되는 데이터 자체로 모든 실험의 데이터에서 같은 형식으로 나타내진다. 이는 패킷 분석 과정을 생략하여 실험 및 탐지가 더 빨라진다는 장점 외에도 다양한 데이터세트와 모델을 활용한 연구 간 정확한 탐지성능 비교가 가능하다는 것을 의미한다.

제안된 접근방법에 대한 성능평가 실험은 PCAP 형식의 파일을 제공하는 UNSW-NB15 데이터세트를 사용하였으며, Payload-Byte 기법을 적용하여 PCAP 파일에 대한 라벨링을 수행하였다. 데이터를 라벨링 한 이후에는 16진수 형태로 표현된 각 패킷의 페이로드를 10진수로 변환하고, 1,500개의 특성을 가지도록 변환된 데이터를 LightGBM 모델에 적용하여 학습 및 탐지성능 분석을 수행하였다.

변환된 원본 형태의 데이터세트 그대로를 LightGBM 분류 모델에 적용하여 실험한 결과, 이진 분류에서는 정확도 99.33%, F1-score 98.73%, Precision 98.16%, Recall 99.31%를 달성하여 선행연구 대비 탐지성능이 향상됨을 확인하였다. 한편, 다중 분류에서는 정확도 85.63%, F1-score 85.68%, Precision 87.88%, Recall 85.61%로 나타나 선행연구와 거의 유사한 수준임을 확인하였다. 그러나 본 연구에서 제안된 접근방법은 패킷 페이로드를 직접 분석하여 침입을 탐지하기 때문에 메타데이터를 생성해야만 하는 기존 접근 방법들보다 실시간성 측면에서 큰 장점을 가진다.

본 연구에서는 학습에 사용될 데이터세트의 특성이 1,500개로 특성 공간이 크다는 점, 그리고 빠른 실행속도 및 높은 탐지성능을 가진다는 점을 고려하여 LightGBM 모델만을 적용하였다. 그러나 향후에는 보다 다양한 분류 모델을 적용하고 데이터세트도 다양화하면서 제안된 접근방법의 탐지성능을 분석할 예정이다.

REFERENCES

[1] Ministry of Science and ICT, “Statistics Information”, https://www.msit.go.kr/bbs/list.do?sCode=user&mPid=74&mId=99

[2] N. Moustafa and J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1-6, Nov 2015. DOI: 10.1109/milcis.2015.7348942

[3] Australian Center for Cyber Security, “UNSW-NB15 Data Set,” https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

[4] N. Moustafa and J. Slay, “The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set,” Information Security Journal: A Global Perspective, vol. 25, no. 1–3, pp. 18–31, Jan 2016. DOI: 10.1080/19393555.2015.1125974

[5] K. H. Jung, B. H. Lee, D. Yang, "Performance Analysis of Detection Algorithms for the Specific Pattern in Packet Payloads," Journal of the Korea Institute of Information and Communication Engineering 22, 5, pp.794-804, 2018. DOI: https://doi.org/10.6109/jkiice.2018.22.4.794

[6] K. Wang and S. J. Stolfo, “Anomalous Payload-Based Network Intrusion Detection,” Recent Advances in Intrusion Detection: 7th International Symposium, pp. 203–222, Sep 2004. DOI: 10.1007/978-3-540-30143-1_11

[7] M. V. Mahoney and P. K. Chan, “Learning nonstationary models of normal network traffic for detecting novel attacks,” Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 376-385, Jul 2002. DOI: 10.1145/775047.775102

[8] W. Wang et al., “HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection,” IEEE Access, vol. 6, pp. 1792–1806, 2018. DOI: 10.1109/access.2017.2780250

[9] B. A. Pratomo, P. Burnap, and G. Theodorakopoulos, “Unsupervised Approach for Detecting Low Rate Attacks on Network Traffic with Autoencoder,” 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), pp. 1-8, Jun 2018. DOI: 10.1109/cybersecpods.2018.8560678.

[10] H. Liu, B. Lang, M. Liu, and H. Yan, “CNN and RNN based payload classification methods for attack detection,” Knowledge-Based Systems, vol. 163, pp. 332–341, Jan 2019. DOI: 10.1016/j.knosys.2018.08.036.

[11] T. W. Kim, J. l. Jung, J. Y. Lee, "DoS/DDoS attacks Detection Algorithm and System using Packet Counting," Journal of the Korea Society for Simulation, vol.19, no.4, pp. 151-159, 2010.

[12] B. H Lee, "Deep Learning LSTM Model based TCP SYN Flood Detection System," Thesis for the Degree of Master of Agriculture. Graduate School, KNU. Korea. 2020.

[13] D. Jing and H.-B. Chen, “SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset,” 2019 IEEE 13th International Conference on ASIC (ASICON), pp. 1-4, Oct 2019. DOI: 10.1109/asicon47005.2019.8983598.

[14] M. H. Kabir, M. S. Rajib, A. S. M. T. Rahman, Md. M. Rahman, and S. K. Dey, “Network Intrusion Detection Using UNSW-NB15 Dataset: Stacking Machine Learning Based Approach,” 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), pp. 1-6, Feb 2022. DOI: 10.1109/icaeee54957.2022.9836404

[15] F. Meghdouri, T. Zseby, and F. Iglesias, “Analysis of Lightweight Feature Vectors for Attack Detection in Network Traffic,” Applied Sciences, vol. 8, no. 11, pp 2196, Nov 2018. DOI: 10.3390/app8112196.

[16] Y. A. Farrukh, I. Khan, S. Wali, D. Bierbrauer, and N. Bastian, “Payload-Byte: A Tool for Extracting and Labeling Packet Capture Files of Modern Network Intrusion Detection Datasets,”, pp 58-67, Sep. 2022. DOI: 10.36227/techrxiv.20714221.v1.

[17] KE, Guolin, et al, “Lightgbm: A highly efficient gradient boosting decision tree,” Advances in neural information processing systems 30, pp.3149-3157, Dec 2017. DOI: 10.5555/3294996.3295074

[18] J. Li, C. Wu, J. Ye, J. Ding, Q. Fu, Q, and J. Huang, “The comparison and verification of some efficient packet capture and processing technologies,.” IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 967-973, 2019. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00177.

[19] T. W. Nam, D. G. Oh, “Study on the Semantic Extension of the Concept of Metadata, Journal of the Korean Society for Library and Information Science, vol. 44, no. 4, pp. 373–393, Nov 2010. DOI: https://doi.org/10.4275/KSLIS.2010.44.4.373

Authors

Gun-Nam Kim received B.S. degree in 2016 from the Department of Chinese Studies, Chang-Won National University. He is currently a graduate student in the Department of Defense Science, Korea National Defense University. His research interests include Machine Learning, Intrusion Detection System and Cyber security Strategy.

Han-Seok Kim received B.S. degree in 2014 from the Department of International Relations, Korea Military Academy. He is currently a graduate student in the Department of Defense Science, Korea National Defense University. His research interests include Machine Learning and Intrusion Detection System.

Soo-Jin Lee received B.S., M.S. and Ph.D. degrees in Computer Science from Korea Military Academy, Yonsei University and Korea Advanced Institute of Science and Technology(KAIST) in 1992, 1996 and 2006. He is currently a professor of the Department of Defense Science, Korea National Defense University from 2006. His research interests include National Cybersecurity Policy, Intrusion Detection System, Mobile Network Security, Machine Learning, Encryption theory and applications.