Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.4, Aug. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.4.647

준 지도 이상 탐지 기법의 성능 향상을 위한 섭동을 활용한 초구 기반 비정상 데이터 증강 기법*

정 병 길,[1†] 권 준 형,[1] 민 동 준,[1] 이 상 근[2‡]
[1,2]고려대학교 (대학원생, 교수)

Abnormal Data Augmentation Method Using Perturbation Based on Hypersphere for Semi-Supervised Anomaly Detection*

Byeonggil Jung,[1†] Junhyung Kwon,[1] Dongjun Min,[1] Sangkyun Lee[2‡]
[1,2]Korea University (Graduate student, Professor)

요 약

최근 정상 데이터와 일부 비정상 데이터를 보유한 환경에서 딥러닝 기반 준 지도 학습 이상 탐지 기법이 매우 효과적으로 동작함이 알려져 있다. 하지만 사이버 보안 분야와 같이 실제 시스템에 대한 알려지지 않은 공격 등 비정상 데이터 확보가 어려운 환경에서는 비정상 데이터 부족이 발생할 가능성이 있다. 본 논문은 비정상 데이터가 정상 데이터보다 극히 작은 환경에서 준 지도 이상 탐지 기법에 적용 가능한 섭동을 활용한 초구 기반 비정상 데이터 증강 기법인 ADA-PH(Abnormal Data Augmentation Method using Perturbation based on Hypersphere)를 제안한다. ADA-PH는 정상 데이터를 잘 표현할 수 있는 초구의 중심으로부터 상대적으로 먼 거리에 위치한 샘플에 대해 적대적 섭동을 추가함으로써 비정상 데이터를 생성한다. 제안하는 기법은 비정상 데이터가 극소수로 존재하는 네트워크 침입 탐지 데이터셋에 대하여 데이터 증강을 수행하지 않았을 경우보다 평균적으로 23.63% 향상된 AUC가 도출되었고, 다른 증강 기법들과 비교했을 때 가장 높은 AUC가 또한 도출되었다. 또한, 실제 비정상 데이터에 유사한지에 대한 정량적 및 정성적 분석을 수행하였다.

ABSTRACT

Recent works demonstrate that the semi-supervised anomaly detection method functions quite well in the environment with normal data and some anomalous data. However, abnormal data shortages can occur in an environment where it is difficult to reserve anomalous data, such as an unknown attack in the cyber security fields. In this paper, we propose ADA-PH(Abnormal Data Augmentation Method using Perturbation based on Hypersphere), a novel anomalous data augmentation method that is applicable in an environment where abnormal data is insufficient to secure the performance of the semi-supervised anomaly detection method. ADA-PH generates abnormal data by perturbing samples located relatively far from the center of the hypersphere. With the network intrusion detection datasets where abnormal data is rare, ADA-PH shows 23.63% higher AUC performance than anomaly detection without data augmentation and even performs better than the other augmentation methods. Also, we further conduct quantitative and qualitative analysis on whether generated abnormal data is anomalous.

Keywords: Anomaly Detection, Data Augmentation, Network Intrusion Detection

Received(05. 09. 2022), Modified(06. 07. 2022), Accepted(06. 22. 2022)
* 본 연구는 국방과학연구소의 지원을 받아 수행되었습니다. (UD210010ED)
† 주저자, jbkcose@korea.ac.kr
‡ 교신저자, sangkyun@korea.ac.kr(Corresponding author)

I. 서 론

이상 탐지는 일반적인 상황에서 예상되는 패턴을 따르지 않는 비정상 데이터 포인트를 감지하는 방법론을 의미한다. 이상 탐지는 네트워크 침입 탐지[1], 시스템 모니터링[2], 질병 감지[3], 공장에서의 결함탐지[4] 등과 같이 다양한 분야에서 실용적으로 사용된다. 특히, 최근에는 사이버 보안 분야에서 사이버 공격 방식이 속도나 다양성 측면에서 크게 증대되고 있으므로, 이러한 공격 탐지를 위한 이상 탐지 방법론 연구의 중요성이 증대하고 있다.

기계 학습이 등장하기 이전에는 프로파일링 기반[5], 또는 규칙 기반[6] 이상 탐지와 같은 통계적인 기법이 많이 사용되었다. 하지만 기계 학습 연구가 진행됨에 따라 One-Class Support Vector Machine(OC-SVM)[7]이나 Support Vector Data Description(SVDD)[8] 등과 같은 기계 학습 기반의 이상 탐지 기법 연구 또한 많이 진행되었다. 또한, 최근에는 딥러닝 기술 기반의 이상 탐지 기법이 높은 성능을 보이면서, Deep Support Vector Data Description(DeepSVDD)[9] 혹은 Outlier exposure[10] 등의 관련 연구들이 제안되었다. 특히, Outlier exposure는 비정상 데이터를 학습에 사용함으로써 타 기법 대비 높은 성능을 도출할 수 있음을 보였으며, 이를 기반으로 비정상 데이터를 일부 학습에 참여시키는 준 지도 학습 기반의 이상 탐지 기법 연구가 활발히 진행되고 있다.

딥러닝 기반 이상 탐지 기법은 사이버-물리 시스템에 대한 침입 탐지[11], 스마트 그리드 공격 탐지[12] 등 사이버 보안 영역에서 유효한 성능을 보이면서 관련 연구도 증가하는 추세이다. 하지만, 이러한 연구들에서는 알려지지 않은 사이버 공격 또는 최신 공격 기법에 대한 정보들이 제한되어 있으므로 비정상 데이터 수집이 어려운 경우가 많다[13]. 비정상 데이터가 극소수로 존재할 경우, 준 지도 학습 기반의 이상 탐지 기법 학습에 어려움이 있으며, 더 나아가 이상 탐지를 위한 이상 점수에 대한 최적의 임계값을 찾기에 어려움이 있다.

이를 해결하기 위해, Generative Adversarial Networks(GAN)[14] 등의 생성 모델을 기반으로 비정상 데이터를 생성할 수 있는 기법들이 고려되었지만[15], 학습 데이터가 부족한 상황에서는 생성 모델 학습에 어려움이 있다. 또한, 생성된 데이터들이 실제 비정상 데이터에 근사하는지에 대한 보장이 없으며, 이를 기반으로 준 지도 이상 탐지 기법의 학습을 수행할 경우 성능 향상을 기대하기에 어려움이 있다. 따라서 극소수의 비정상 데이터를 보유한 상황에서도 효과적으로 비정상 데이터를 증강하기 위한 방안이 요구된다.

이에 따라, 본 연구에서는 준 지도 이상 탐지 기법에 적용 가능한 섭동을 활용한 초구 기반 비정상 데이터 생성 기법인 ADA-PH(Abnormal Data Augmentation Method using Perturbation based on Hypersphere)를 제안한다. 여기서 초구(Hypersphere)란 정상 데이터를 잘 표현할 수 있는 잠재 공간상의 데이터 군집을 의미한다[8]. 또한, 섭동(Perturbation)이란 적대적 예제 생성 시 데이터에 가하는 작은 노이즈(Noise)를 의미하며 일반적으로 딥러닝 모델의 추론 성능 저하를 유도하는 공격 기법 등에 많이 응용된다[16]. 본 기법에서는 이를 응용하여 신뢰도가 낮은 정상 데이터 샘플에 대해 초구의 중심으로부터 멀어지도록 섭동을 가하여 비정상 데이터를 생성한다. 이후, 이를 준 지도 학습 이상 탐지 기법에 활용함으로써 비정상 데이터가 매우 적은 환경에서도 효과적인 이상 탐지 수행을 기대할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 기존 연구 사례 및 제안하는 기법과 관련된 연구를 소개한다. 3장에서는 제안하는 기법인 ADA-PH를 소개하고, 4장에서는 네트워크 침입 탐지 데이터셋에 대한 이상 탐지 실험 결과를 보여준다. 마지막으로 5장에서는 결론과 한계점, 후속 연구 방향을 소개한다.

II. 관련 연구

본 연구는 준 지도 이상 탐지 기법의 성능 향상을 위한 비정상 데이터 증강 기법에 관한 것이다. 이에 따라, 본 장에서는 이와 관련된 기존의 이상 탐지 기법들과 종래 비정상 데이터 생성을 위한 연구들을 간략하게 소개한다.

2.1 이상 탐지 기법

종래에는 프로파일링 기반 이상 탐지[5]나 규칙 기반 이상 탐지[6]와 같이 통계적인 특성을 이용한 이상 탐지에 대한 연구가 많이 진행되었다. 이후, 기계 학습에 대한 연구가 활발해짐에 따라, 기계 학습

기반 이상 탐지 기법에 대한 연구가 많이 진행되었다. 대표적으로, Scholkopf 등[7]은 Support Vector Machine(SVM)을 단일 클래스 분류기로 확장한 기법인 One-Class Support Vector Machine(OC-SVM)을 제안하였다. 이후, Tax와 Duin[8]은 OC-SVM을 기반으로 정상 데이터를 초구 내부에 매핑시키는 기법인 Support Vector Data Description(SVDD)를 제안하였다. SVDD는 SVM과 같이 커널 함수를 활용하며, 잠재 공간에서 정상 데이터를 잘 표현할 수 있는 초구를 찾는 Minimum Enclosing Ball(MEB) 문제를 기반으로 이상 탐지를 수행한다. SVDD는 MEB 기반 이상 탐지 기법의 새로운 방향성을 제시하였으며, 최근 관련 연구가 활발히 진행되고 있다.

Ruff 등[9]은 기존 SVDD 기법의 커널 함수를 심층 신경망으로 대체한 Deep Support Vector Data Description(DeepSVDD)를 제안하였다. DeepSVDD는 정상 데이터셋을 기반으로 초구의 중심을 사전 학습한 뒤, SVDD와 마찬가지로 정상 데이터를 잘 표현할 수 있는 잠재 공간상의 초구를 학습한다. 추론 시에, 입력 데이터와 초구의 중심 간의 거리를 계산함으로써 이상 점수를 도출할 수 있다. 이에 따라, 데이터가 정상에 가까울수록 이상 점수는 작게 도출되며, 반대로 이상치에 가까울수록 크게 도출된다.

해당 연구들은 모두 정상 데이터를 기반으로 비지도 학습을 수행하는 이상 탐지 기법에 대한 연구를 다루고 있다. Hendrycks 등[10]은 비정상 데이터로 레이블이 된 일부 데이터를 학습에 활용할 때 성능 향상에 도움을 줄 수 있다는 연구 결과인 Outlier exposure를 제시하였다. 구체적으로, 일부 비정상 데이터를 사용함으로써 탐지 모델을 일반화하여 알려지지 않은 이상치를 정상과 잘 구분하도록 함으로써 이상 탐지 성능을 향상시킬 수 있다. 이후, Ruff 등[17]은 DeepSVDD에 Outlier exposure를 도입하여 준 지도 이상 탐지 기법으로 확장시킨 Deep Semi-supervised Anomaly Detection (DeepSAD)를 제안하였다. DeepSAD의 목적 함수는 DeepSVDD의 목적 함수를 기반으로 비정상 데이터에 대한 MEB 학습이 가능하도록 재구성되었다. 구체적으로, 비정상 데이터에 대한 손실 항이 추가되었고, 비정상 데이터와 정상 데이터에 대한 손실항 간의 균형을 위한 하이퍼 파라미터가 추가되었다.

이러한 준 지도 이상 탐지 기법들의 경우 학습에 사용된 비정상 데이터의 개수에 따라 성능 차이의 편차가 크므로[17], 높은 이상 탐지 성능 도출을 위해선 준 지도 학습에 유효한 비정상 데이터의 충분한 확보가 필요하다.

2.2 비정상 데이터 증강 기법

보유한 비정상 데이터가 부족한 상황에서 비정상 데이터의 증강을 통해 준 지도 이상 탐지 기법의 성능 향상을 기대할 수 있다. 대표적인 데이터 증강 기법으로 SMOTE[18] 기법이 있다. SMOTE는 소수 클래스의 데이터 샘플과 해당 샘플의 k-최근접 이웃 간의 보간을 통해 새로운 데이터를 생성하는 데이터 증강 기법이다. 하지만 비정상 데이터의 경우, 최근접 이웃 간의 보간점이 이상치라는 보장이 없다.

최근에는 이상 탐지 분야에서 심층 신경망을 기반으로 부족한 비정상 데이터를 증강하기 위한 다양한 연구가 진행된 바 있으며, 대부분 생성 모델을 기반으로 비정상 데이터를 증강한다. 생성적 적대 신경망(Generative Adversarial Network, GAN)[14]은 대표적인 생성 모델 중 하나이다. GAN은 일반적으로 생성자(Generator)와 식별자(Discriminator)로 구성된다. 생성자는 식별자가 실제 입력 데이터인지 생성 데이터인지 분간하기 어렵도록 데이터를 생성하는 것을 목표로 한다. 그리고 식별자는 실제 데이터와 생성 데이터를 잘 구분하는 것을 목표로 한다. 이를 기반으로, GAN의 학습은 생성자와 식별자의 손실 함수를 최소화함으로써 적대적으로 이루어진다. 결과적으로, 학습이 완료된 GAN 모델의 생성자를 활용하여 데이터를 생성하는 것이 가능하다. 이를 통해, 비정상 데이터로만 GAN 모델을 학습함으로써 준 지도 이상 탐지 기법에 활용될 비정상 데이터를 증강하는 것이 가능하다. 하지만, GAN 학습을 위한 비정상 데이터의 수가 극히 적을 경우, 생성된 데이터가 실제 비정상 데이터와 유사한지에 대한 보장이 없다[19].

Dai 등[20]은 학습 데이터의 분포로부터 신뢰도가 낮은 데이터를 생성할 수 있는 기법인 BadGAN을 제안하였다. BadGAN은 생성자가 학습 데이터 분포에서 벗어난 데이터를 생성하도록 학습을 수행한다. 이를 기반으로, 학습된 생성자를 통해 특징 공간에서 실제 데이터와 차이가 존재하는 보완 데이터(Complement sample)을 생성할 수 있다. 이를

Table 1. Comparison of data augmentation methods.

Methods	No requirement of abnormal data	Consideration of anomaly detection	Low training cost
SMOTE[18]	No	No	Yes
GAN[14]	No	No	No
BadGAN[19]	Yes	No	No
ADA-PH(ours)	Yes	Yes	Yes

기반으로, 최근에는 BadGAN을 활용하여 비정상 데이터 증강을 통해 이상 탐지 성능을 끌어올리거나[21], BadGAN을 데이터 클러스터링을 함께 사용하여 봇 탐지 성능 향상을 꾀한 시도가 있었다[22].

하지만 생성 모델 기반의 데이터 증강 방식은 데이터를 생성하기 위한 생성자의 학습이 필수적으로 요구되며, 이에 따라 모델이 학습 데이터셋의 분포를 충분히 반영할 수 있어야 한다[23]. 또한, 모델 학습 시 손실 함수의 min-max 문제의 최적 해를 찾기가 어려워 학습의 불안정성이 높다[24]. 따라서 본 연구에서는 생성 모델 기반의 데이터 증강 방식이 아닌, 적대적 예제(Adversarial example) 생성에 주로 사용되는 섭동(Perturbation)을 활용하여 데이터를 증강하고자 한다.

Goodfellow 등[16]은 적대적 예제를 통해 심층 신경망 모델의 오분류를 유도할 수 있는 기법인 Fast Gradient Sign Method(FGSM)을 제안하였다. 적대적 예제란 손실 값을 최대화할 목적으로 생성된 섭동이 입력 데이터에 가해진 데이터이다. 생성된 적대적 예제는 인지적인 관점에서 원본 데이터와 유사하지만, 심층 신경망의 오분류를 유도할 수 있다. 해당 연구를 기반으로, 효과적인 적대적 예제의 생성을 위해 Projected Gradient Descent(PGD)[25] 등의 다양한 적대적 공격(Adversarial attack) 기법들이 연구되었다.

최근에는 Goyal 등[26]이 적대적 예제를 활용하여 이상 탐지를 수행하는 기법인 Deep Robust One-Class Classification(DROCC)을 제안하였다. DROCC은 정상 데이터에 섭동을 가함으로써 정상 데이터의 매니폴드를 벗어난 적대적 예제를 생성한다. 이후, 생성한 적대적 예제를 비정상 클래스로 상정하고 이를 활용하여 이진 분류 모델을 학습한다. 학습된 이진 분류 모델은 정상 클래스의 결정 경계면을 보유하고 있으며, 이를 기반으로 이상 탐지를 수행할 수 있다. DROCC은 정상 데이터를 기반으로 섭동을 통해 효과적인 이상치를 생성할 수 있다는 점에서 이상 탐지 연구의 새로운 방향성을 제시하였다. 이상 탐지를 위한 이진 분류 모델을 학습하는 DROCC과 달리, 본 연구는 섭동을 사용하여 정상 데이터로부터 비정상 데이터를 생성함으로써 준 지도 학습 기반 이상 탐지 기법에 활용하고자 한다.

Table 1.은 사이버 보안 데이터셋에 대해 활용될 수 있는 종래 데이터 증강 기법들 및 ADA-PH와의 다양한 기준에서의 비교를 나타낸다. 구체적으로, 데이터 생성 시 비정상 데이터의 필요 여부, 이상 탐지 기법을 고려한 증강 기법인지에 대한 여부, 그리고 모델 학습 시 요구되는 연산량에 대해 비교하였다. ADA-PH의 경우, 학습 시 비정상 데이터가 요구되지 않으며, 생성 모델 기반 데이터 증강 기법 대비 연산량이 적다는 장점이 있다. 무엇보다, 이상 탐지 기법을 위해 고려된 데이터 증강 기법이라는 점에서 큰 차별성이 있다. DROCC의 경우 데이터 증강 기법이 아닌, 비지도 학습 기반 이상 탐지 기법이므로 비교 대상에서 제외되었다.

III. 제안 기법

본 장에서는 초구 기반의 섭동을 활용한 비정상 데이터 증강 기법인 ADA-PH을 구체적으로 설명한다. 또한 본 논문의 실험에서 사용될 준 지도 이상 탐지 기법 중 하나인 DeepSAD에 대해서도 설명한다.

3.1 ADA-PH: 섭동을 활용한 초구 기반 비정상 데이터 증강 기법

본 논문에서 제안하는 기법인 ADA-PH는 정상 데이터를 잘 표현할 수 있는 초구를 학습한 뒤, 초구로부터 거리가 먼 데이터에 섭동을 첨가하여 비정상 데이터를 생성할 수 있다. Fig. 1.은 ADA-PH 알고리즘에 대한 흐름도를 나타낸다.

심층 신경망 기반의 학습 파라미터 θ_e 및 θ_d를 가

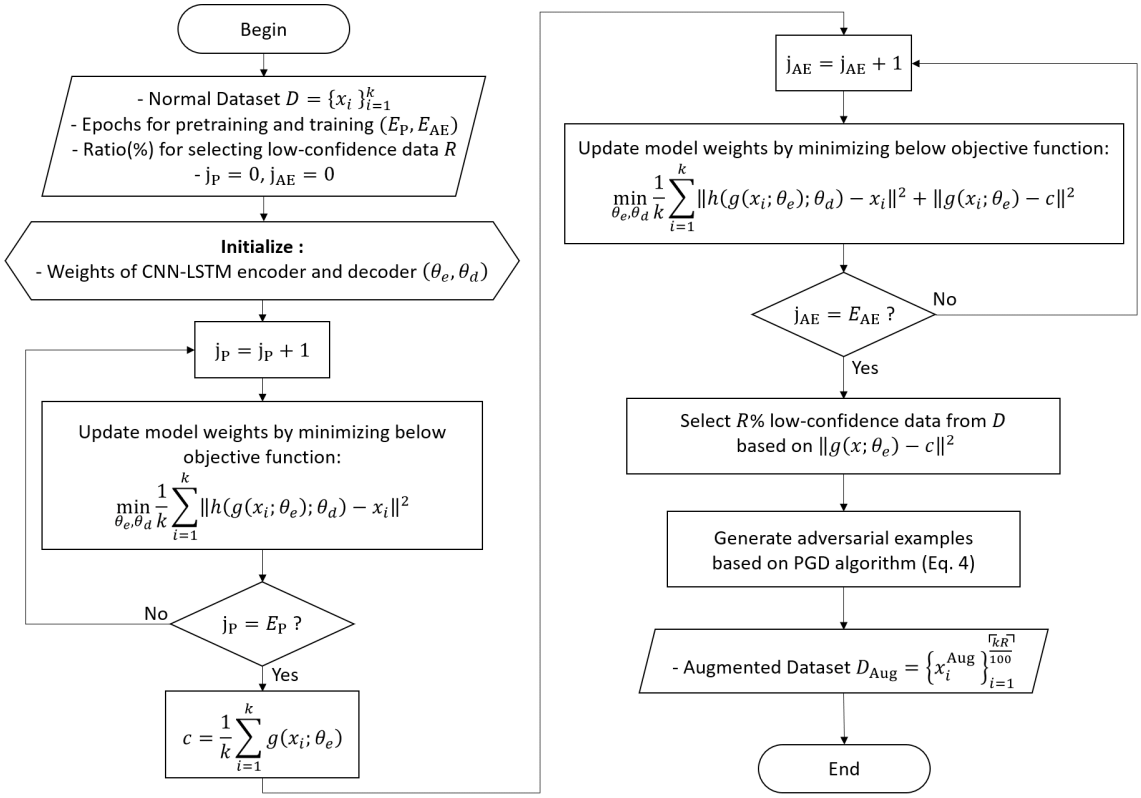

Fig. 1. Algorithm of ADA-PH.

진 인코더 g및 디코더 h로 구성된 임베딩 네트워크가 존재한다고 가정했을 때, k개의 정상 데이터 $x_1,\ldots,x_k$를 기반으로 정상 데이터의 표현을 학습하기 위한 사전 학습은 아래 수식 (1)을 최소화함으로써 수행될 수 있다.

$$\min_{\theta_e,\theta_d} \frac{1}{k}\sum_{i=0}^{k} \| h(g(x_i;\theta_e);\theta_d) - x_i \|^2 \quad (1)$$

사전 학습된 임베딩 네트워크를 기반으로, 초구를 학습하기 위한 초구의 중심 c는 아래 수식 (2)와 같이 임베딩된 정상 데이터들의 평균으로 계산할 수 있다.

$$c = \frac{1}{k}\sum_{i=0}^{k} g(x_i;\theta_e) \quad (2)$$

이후, 아래의 수식 (3)을 최소화함으로써 정상 데이터를 기반으로 MEB 학습을 수행한다.

$$\min_{\theta_e,\theta_d} \frac{1}{k}\sum_{i=0}^{k} \| h(g(x_i;\theta_e);\theta_d) - x_i \|^2 + \| g(x_i;\theta_e) - c \|^2 \quad (3)$$

수식 (3)의 좌항을 최소화함으로써 정상 데이터와 복원 데이터 간의 복원 오차를 최소화할 수 있으며, 우항을 최소화함으로써 임베딩된 정상 데이터와 초구의 중심 간의 거리를 최소화할 수 있다. 이상적인 학습이 이루어질 경우, 디코더를 통해 복원된 데이터는 정상 데이터의 특징과 근사하며, 동시에 임베딩된 데이터는 초구의 중심에 가까이 매핑될 것을 기대할 수 있다.

잠재 공간상의 정상 데이터와 초구 간의 거리인 $\| g(x;\theta_e) - c \|^2$을 기준으로, 거리가 상대적으로 먼 데이터를 신뢰도가 낮은 데이터로써 선별할 수 있다. 이후, PGD[24] 기법을 통해 선별된 데이터를 기반으로 반복적으로 섭동을 업데이트함으로써 최적의 섭동을 찾을 수 있다. 구체적으로, (0,1) 구간의 실수 값을 가지는 step size α만큼 씩 섭동을 반복적으

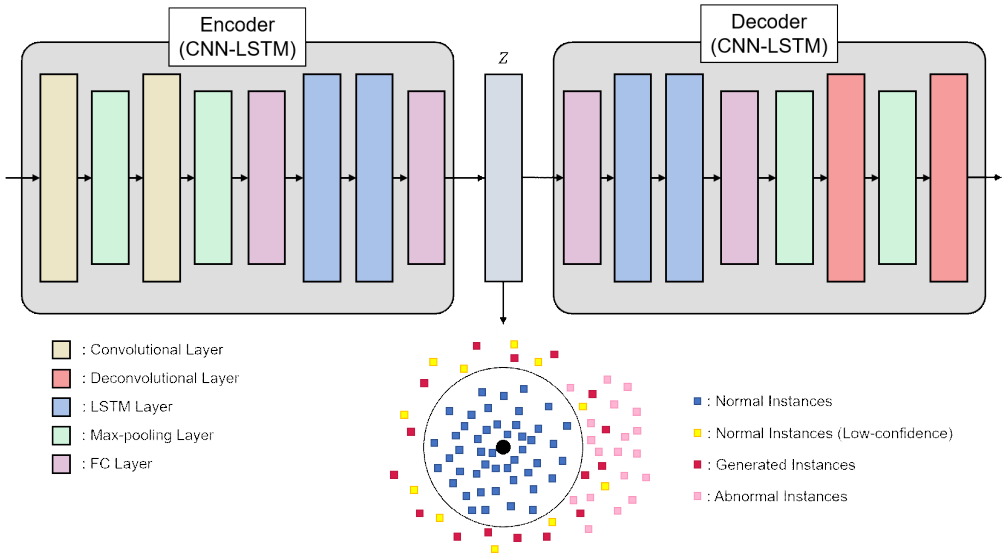

Fig. 2. Illustration of the CNN-LSTM model architecture and data generated from ADA-PH.

로 최적화하여 ℓ_2-norm ball에서 효과적인 최적의 섭동을 찾을 수 있다. 각 단계를 거칠 때마다, 찾은 최적의 섭동을 입력 데이터 x의 ϵ-neighbor에 사영시킨다. 이를 수식으로 표현하면 다음과 같다.

$$\begin{aligned} &x^0 = x, \\ &x^{t+1} = \Pi(x^t + \alpha sgn(\nabla_x \ell(x^t + \delta))), \\ &\| \delta \|_2 \leq \epsilon \end{aligned} \tag{4}$$

수식 (4)에서, δ는 입력 데이터 x에 가할 섭동을, ϵ는 섭동 δ의 ℓ_2-norm bound를, 그리고 Π는 ℓ_2-norm ball을 향한 사영 연산을 각각 의미한다. 또한, $sgn(.)$ 함수는 그래디언트의 부호를 나타내는 함수를 의미한다. 이 때, 섭동을 학습하기 위한 손실 함수 ℓ은 아래 수식 (5)와 같이 정의할 수 있다.

$$\ell(x) = \| g(x;\theta_e) - c \|^2 \tag{5}$$

손실 함수 ℓ는 잠재 공간상의 데이터와 초구의 중심과의 거리를 의미한다. 따라서 생성되는 섭동은 임베딩된 데이터로 하여금 초구의 중심을 기준으로 상대적으로 멀게 위치시키고자 한다.

앞서 선별된 데이터는 초구의 중심으로부터 거리가 먼 데이터이기에, 초구의 중심과 가까운 데이터에 비해 상대적으로 비정상 데이터에 근사하다. 따라서 선별된 데이터로부터 생성된 적대적 예제 또한 비정상 데이터에 근사할 것임을 기대할 수 있다. Fig. 2.는 생성된 데이터에 대한 묘사를 나타낸다.

3.2 증강된 비정상 데이터를 활용한 준 지도 이상 탐지 모델 학습

DeepSAD[17]는 준 지도 이상 탐지 기법 중에서 Tabular 데이터셋에 대해서 높은 성능이 도출된 기법 중 하나이다[27]. 따라서 본 연구에서는 ADA-PH을 포함한 비정상 데이터 증강 기법들의 성능을 검증하기 위한 준 지도 이상 탐지 기법으로 DeepSAD를 활용하였다. n개의 라벨링 되지 않은 데이터 $x_1,\ldots,x_n$ 와 정상($y=+1$) 혹은 비정상($y=-1$) 클래스를 가진 m 개의 라벨링 된 데이터 $(x_1,y_1),\ldots,(x_m,y_m)$가 존재한다고 가정했을 때, DeepSAD의 학습은 손실 함수인 아래 수식 (6)을 최소화함으로써 수행된다.

$$\begin{aligned} &\min_w \frac{1}{n+m}\sum_{i=1}^{n} \| \phi(x_i;w) - c \|^2 + \\ &\frac{\eta}{n+m}\sum_{j=1}^{m} (\| \phi(x_i;w) - c \|^2)^{y_i} + \frac{\lambda}{2}\sum_{l=1}^{L} \| W^l \|_2^2 \end{aligned} \tag{6}$$

이 때, ϕ는 $W^1, W^2, \ldots, W^L$의 학습 가중치를 가

진 L개의 은닉층을 보유한 심층 신경망, c는 사전 학습된 초구의 중심, 그리고 η는 양의 정수인 가중치 하이퍼파라미터를 각각 의미한다.

이상 탐지 환경에서 라벨링 되지 않은 대다수의 데이터는 정상 데이터라고 가정한다[17]. 따라서 수식 (6)을 기반으로 DeepSAD의 학습이 진행됨에 따라 정상 데이터를 잠재 공간상의 초구 내로, 비정상 데이터를 초구 바깥으로 매핑시킬 수 있음을 기대할 수 있다. 학습된 모델을 통해 잠재 공간상의 매핑된 데이터 ϕ와 초구의 중심 c 간의 거리를 기반으로 이상 점수 s는 아래 수식 (7)과 같이 계산할 수 있다.

$$s(x) = \| \phi(x;w) - c \| \quad (7)$$

수식 (7)에 따라, 잠재 공간상의 매핑된 데이터와 초구의 중심 간의 거리가 멀 수록 높은 이상 점수가 도출되며, 반대로 거리가 가까울 수록 낮은 이상 점수가 도출된다.

IV. 제안 기법

4.1 실험 환경 및 데이터셋

본 연구에서는 사이버 보안 분야에서 이상 탐지를 위한 비정상 데이터 증강 기법이 효과적인지를 검증하기 위해 네트워크 침입 탐지 데이터셋인 CICIDS-2017[28] 데이터셋과 TON-IoT[29] 데이터셋에 대해서 실험을 진행한다.

- CICIDS-2017: CICIDS-2017 데이터셋은 5일 동안 수집된 네트워크 트래픽 데이터들로 구성되어 있으며, 총 14종의 공격 타입으로 이루어져 있다. 전체 공격 데이터의 수는 655,364개이며 정상 데이터 수는 2,271,397개이다.
- TON-IoT: TON-IoT는 DoS, DDoS, 백도어 공격 등 9가지 유형의 공격 타입으로 구성되어 있으며 총 21,523,641개의 악성 데이터와 796,380개의 양성 데이터가 포함되어 있다.

본 연구에서는 비정상 데이터의 수가 극히 적은 환경을 설정하고 있으므로, 실험을 위해 각 데이터셋이 보유한 악성 데이터의 일부만을 사용하였다. 본 실험에서는 학습 데이터셋 내 비정상 데이터 수가 정상 데이터 수 대비 3%만을 포함하도록 설정하였다.

Table 2. The number of data points for the CICIDS-2017 dataset and the TON-IoT dataset.

		CICIDS-2017	TON-IoT
Training	Normal	84,030	29,318
	Abnormal	2,490	890
Validation	Normal	20,998	7,347
	Abnormal	632	206
Test	Normal	45,020	15,709
	Abnormal	1,330	475

학습에 사용된 각 데이터셋의 구체적인 정보는 Table 2.와 같다.

4.1.1 임베딩 네트워크 구조

본 연구에서는 네트워크 패킷 데이터를 사용하여 이상 탐지를 수행한다. 따라서 시계열 데이터인 네트워크 패킷 데이터의 임베딩을 위해 CNN-LSTM 모델을 사용한다[30]. 본 연구에서 사용된 CNN-LSTM 모델의 구조는 Fig. 2.에 묘사된 것과 같다.

4.1.2 학습 및 평가 방법

본 연구에서는 데이터를 생성할 수 있는 생성 모델들인 GAN, BadGAN을 비교군으로 ADA-PH와의 성능 비교를 수행하였다. Table 3.은 실험에서 사용된 각 기법들의 하이퍼파라미터를 나타낸다. BadGAN의 alpha는 complementary discriminator loss에 대한 weight를 나타낸다. 그리고 ADA-PH의 low confidence ratio는 사전에 설명한 바와 같이 정상

Table 3. Hyperparameters of each augmentation method.

Methods	Hyperparameters
GAN[14]	generator learning rate: 0.0004 discriminator learning rate: 0.0002 latent space dimension: 32
BadGAN [20]	generator learning rate: 0.00005 discriminator learning rate: 0.00005 alpha: 2.0
ADA-PH	low confidence ratio: 0.05 (5%) alpha: 0.01 epsilon: 2.0

데이터 중 섭동을 가하게 될 신뢰도가 낮은 데이터의 비율을 의미하며, alpha는 PGD 알고리즘의 step size α를 의미한다. 그리고 epsilon은 섭동의 크기에 대한 bound 파라미터 ϵ를 의미한다.

본 실험에서는 평가 데이터셋에 대한 각 모델들의 Area Under the Receiver Operating Characteristic Curve (AUROC, AUC)[31] 측정을 통해 이상 탐지 모델들의 성능 평가를 수행하였다. 여기서 ROC 곡선이란, 민감도(True Positive Rate, TPR)를 x축으로, 위양성률(False Positive Rate, FPR)를 y축으로 갖는 그래프를 의미하며, AUC는 ROC 곡선 아래 영역의 넓이를 나타내며 [0,1]의 실수값을 가질 수 있다. AUC 값이 0.5에 가까울수록 모델의 예측 성능이 무작위 추론에 가깝고, 1.0에 가까울수록 정확한 추론이 수행되었음을 의미한다.

4.2 실험 결과

본 연구에서 제안하는 비정상 데이터 증강 기법인 ADA-PH의 유효성을 입증하기 위하여 상기 언급한 비교군인 GAN, BadGAN과의 성능 비교를 수행하였다. 본 실험에서는 각 증강 기법들로부터 생성된 데이터를 활용하여 준 지도 이상 탐지 기법 중 하나인 DeepSAD를 기반으로 이상 탐지를 수행하였으며, 각 기법에 대한 이상 탐지 성능은 Table 4.의 AUC 점수와 Fig. 3.의 ROC 곡선으로 나타내었다.

Table 4.에서 ADA-PH의 AUC 수치는 CICIDS-2017 데이터셋과 TON-IoT 데이터셋에 대해 데이터 증강을 수행하지 않은 경우보다 24.3%, 13.2%만큼 각각 증가한다. 이는 ADA-PH가 이상 탐지 성능 향상에 도움이 된다는 것을 의미한다.

또한, 다른 생성 모델들인 GAN 및 BadGAN과 비교했을 때 ADA-PH의 AUC 수치가 평균적으로 26.5%, 8.5%만큼 각각 증가한다. 이는 ADA-PH를 기반으로 신뢰도가 낮은 정상 데이터에 섭동을 첨가하여 생성한 데이터가 GAN 기반의 다른 증강 기법들로부터 생성된 데이터보다 이상 탐지 성능 향상에 더 유효함을 보여준다. 특히, GAN의 경우, CICIDS-2017 데이터셋에 대해서 데이터 증강을 수행하지 않았을 때 보다 AUC 수치가 7.4% 감소

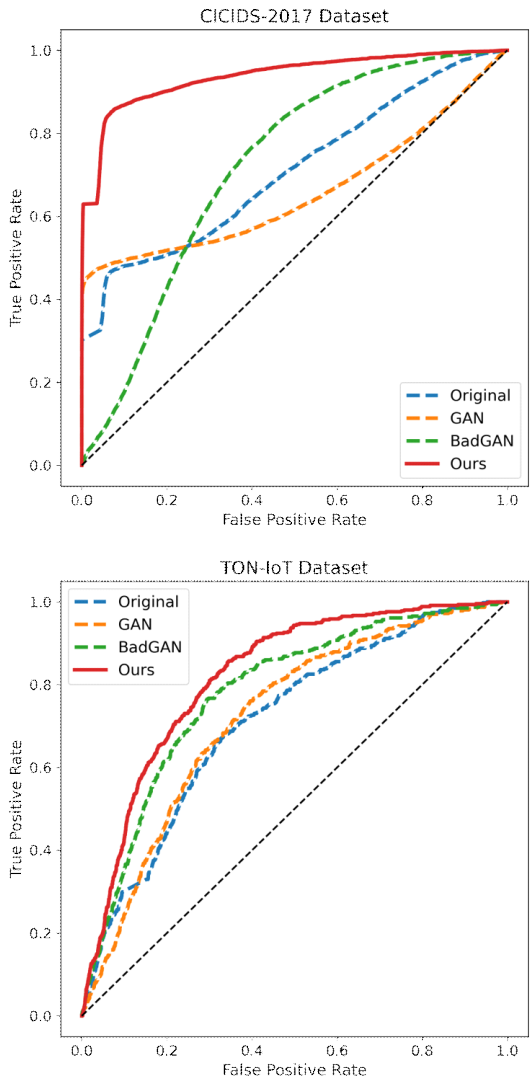

Fig. 3. ROC curve of different augmentation methods on CICIDS-2017 (top) and TON-IoT (bottom) datasets.

Table 4. AUC performance on each augmentation method for the CICIDS-2017 dataset and TON-IoT dataset.

Dataset	Original	GAN	BadGAN	**ADA-PH**
CICIDS-2017	0.7101	0.6575	0.7205	**0.9376**
TON-IoT	0.7118	0.7221	0.7797	**0.8203**

한다. 이는 GAN으로부터 생성된 데이터가 준 지도 학습 기반 이상 탐지 기법으로 하여금 성능 저해를 야기하는 데이터를 생성했다고 볼 수 있다.

4.2.1 t-SNE 및 KL-divergence 분석

본 장에서는 실험에서 활용된 증강 기법들인 GAN, BadGAN, 그리고 ADA-PH로부터 생성된 데이터들을 비정상 데이터로 상정할 수 있는지에 대한 분석 및 논의가 이루어진다. Table 5.는 CICIDS-2017 데이터셋에 대해 정상 데이터, 비정상 데이터, 그리고 생성 데이터의 조합을 기반으로 도출한 KL divergence[32] 수치이다. KL divergence 수치가 작을수록 두 데이터셋의 분포는 서로 근사하다. Table 5.에 따르면, ADA-PH는 정상 데이터의 분포와 생성 데이터의 분포 간의 KL divergence 및 비정상 데이터의 분포와 생성 데이터의 분포 간의 KL divergence 가 다른 기법들에 비해 평균적으로 3.5배 및 17.3배 낮다. 이는 ADA-PH가 다른 기법들보다 학습 데이터와 유사한 데이터를 생성할 수 있음을 보여준다. 또한, ADA-PH에서 비정상 데이터의 분포와 생성 데이터의 분포 간의 KL divergence 는 정상 데이터의 분포와 생성 데이터의 분포 간의 KL divergence 보다 2.2배 낮다. 이는 ADA-PH로부터 생성된 데이터가 정상 데이터보다 비정상 데이터와 유사하다는 것을 의미한다. 따라서 ADA-PH로부터 생성된 데이터는 GAN 및 BadGAN으로부터 생성된 데이터보다 비정상 데이터에 가깝다고 볼 수 있다.

Table 5. KL divergence between normal and abnormal (N and A), normal and generated (N and G), and abnormal and generated examples (A and G) for the CICIDS-2017 dataset.

Methods	N and A	N and G	A and G
GAN	19.090	150.490	385.167
BadGAN	19.090	385.149	812.853
ADA-PH	**19.090**	**76.552**	**34.577**

추가적으로, t-SNE[33] 알고리즘을 통해 각 증강 기법들로부터 생성된 데이터에 대한 정성적인 분석 또한 수행하였다. t-SNE 그래프 상에서 데이터가 서로 인접할수록 서로 유사한 분포에서 도출되었다고 볼 수 있다. Fig. 4.은 CICIDS-2017 데이터셋에 대해서 각 기법들로부터 생성된 데이터 및 정상 데이터와 비정상 데이터의 2차원 t-SNE 그래프를 보여준다. GAN과 BadGAN의 경우, t-SNE로부터 도출된 잠재 공간상에서의 생성 데이터는 정상 데이터와 비정상 데이터의 군집에서 멀리 위치해있음을 확인할 수 있다. 반면 ADA-PH의 경우, 다른 두 증강 기법들과 달리, 생성 데이터들이 신뢰도가 낮은 정상 데이터와 상대적으로 근접한 것을 확인할 수 있다. 이는 ADA-PH로부터 생성된 적대적 예제가 신뢰도가 낮은 정상 데이터로부터 섭동이 가해져 생성되었기 때문으로 볼 수 있다.

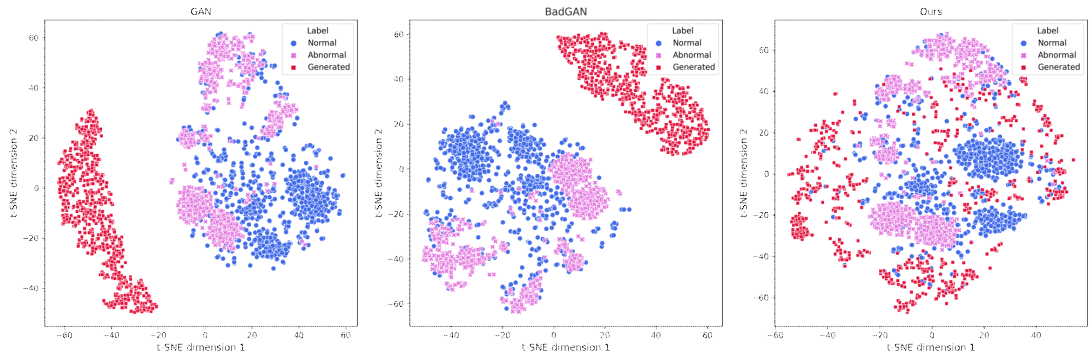

Fig. 4. 2-dimensional t-SNE of embedded data from GAN (left), BadGAN (middle), and ADA-PH (right) for the CICIDS-2017 dataset.

4.2.2 하이퍼파라미터 α 및 ϵ에 따른 성능 변화

Table 6.는 CICIDS-2017 데이터셋에 대해서 ADA-PH의 적대적 예제 생성에 사용된 PGD 기법의 하이퍼파라미터 α 및 ϵ에 따른 DeepSAD의 AUC 성능 변화를 보여준다. α의 경우, 0.01에서 가장 높은 AUC 수치가 도출되었고, 이는 두 번째로 높은 AUC 수치가 도출된 지점인 0.005의 경우보다 16.1% 높다. 그리고 ϵ의 경우, 2.0에서 가장 높은 AUC 수치가 도출되었고, 이는 두 번째로 높은 AUC 수치가 도출된 지점인 2.2의 경우보다 9.6% 높다. 따라서 우리는 최적의 α 및 ϵ 값으로 각각 0.01 및 2.0을 사용하였다.

Table 6. AUC scores of ADA-PH according to hyperparameters alpha(α) and epsilon(ϵ) for CICIDS-2017 dataset.

alpha(α)	AUC	epsilon(ϵ)	AUC
0.001	0.7420	1.6	0.7598
0.005	0.8073	1.8	0.8060
0.01	**0.9376**	**2.0**	**0.9376**
0.05	0.7837	2.2	0.8558
0.1	0.7798	2.4	0.6365

4.2.3 비정상 데이터 비율에 따른 성능 변화

Fig. 5.는 CICIDS-2017 데이터셋 및 TON-IoT 데이터셋에 대해 학습 데이터셋 내 비정상 데이터의 포함 비율에 따른 AUC 성능 변화 그래프를 각각 나타낸다. 우리는 학습 데이터셋 내 비정상 데이터의 포함 비율을 1% 부터 5%까지 증가시켜가며 각 증강 기법들의 AUC 성능 변화를 분석하였다. CICIDS-2017 데이터셋과 TON-IoT 데이터셋에 대해서 학습 데이터셋 내 비정상 데이터의 포함 비율이 1%에서 3%까지 증가함에 따라 다른 증강 기법들의 경우 AUC 수치가 평균적으로 각각 22.0%, 8.2% 증가하였다. 반면, ADA-PH의 경우 AUC 수치가 두 데이터셋에 대해 각각 52.8%, 9.6% 증가하였다. 이는 비정상 데이터가 매우 부족한 상황에서 ADA-PH로부터 생성된 데이터가 다른 증강 기법들과 비교했을 때 더욱 효과적임을 보여준다. 하지만, 두 데이터셋 모두 학습 데이터셋 내 비정상 데이터의 포함 비율이 4% 이상이 되는 지점부터 뚜렷한 성능 개선이 보이지 않는다. 이는 ADA-PH로부터 생성된 데이터가 비정상 데이터가 극히 일부인 상황에서 유효한 도움을 주지만, 비정상 데이터가 일정 수준으로 확보된 상황에서는 성능 개선에 한계가 있는 것으로 파악된다.

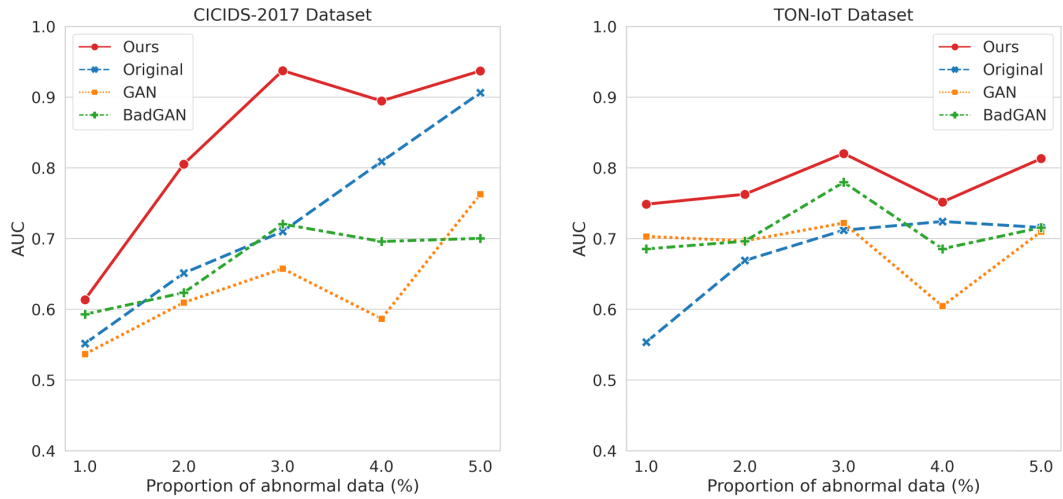

Fig. 5. AUC change of each data augmentation method according to the proportion of abnormal data in the training dataset for the CICIDS-2017 dataset (left) and the TON-IoT dataset (right).

V. 결 론

본 연구에서는 소수의 비정상 데이터가 존재하는 환경에서 준 지도 이상 탐지 기법의 성능 증대를 위한 효과적인 데이터 증강 기법인 ADA-PH를 제안하였다. ADA-PH를 통해 신뢰도가 낮은 정상 데이터에 섭동을 가해 적대적 예제를 생성할 수 있으며, 이를 준 지도 이상 탐지 기법의 학습을 위한 비정상 데이터로써 활용할 수 있다.

실험 결과에 따르면, ADA-PH로부터 생성된 비정상 데이터를 학습에 활용했을 경우 다른 증강 기법들의 경우보다 높은 AUC 수치가 도출된 것을 확인할 수 있었다. 또한, KL divergence와 t-SNE를 기반으로 수행한 분석을 통해, ADA-PH로부터 생성된 데이터의 분포가 다른 증강 기법들의 경우보다 실제 비정상 데이터의 분포와 가장 유사함을 확인할 수 있었다. 또한, 학습 데이터셋 내 비정상 데이터의 비율이 변화함에 따라서도 ADA-PH가 타 증강 기법들보다 가장 높은 AUC 수치가 도출되었으며, 이는 ADA-PH가 성능 확보의 측면에서 다른 증강 기법들보다 강건하다고 볼 수 있다. 하지만, 학습 데이터셋 내 비정상 데이터가 일정 비율 이상 확보된 환경의 경우, ADA-PH로부터 생성된 데이터는 성능 향상에 큰 도움을 주지 않는다. 따라서 이를 개선하기 위해, 우리는 군집화 알고리즘 등을 활용하여 비정상 데이터가 충분한 상황에서도 성능 개선에 유효한 데이터를 생성할 수 있도록 향후 연구를 수행할 예정이다.

본 논문에서 제안하는 기법은 소수의 비정상 데이터가 존재하는 상황에서의 네트워크 침입 탐지 및 시스템 이상 행동 탐지 등에 효과적으로 사용될 수 있음을 기대하며, 향후 실제 운용 환경에서 보다 유용하게 활용될 수 있도록 추가적인 연구를 진행할 필요가 있다.

References

[1] N. Shone, N. N. Tran, V. Dinh Phai and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 1, pp. 41-50, Jan. 2018.

[2] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised Anomaly Detection on Multivariate Time Series," Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 3395-3404, Aug. 2020.

[3] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," Journal of Medical Systems, vol. 42, no. 11, pp. 226, Oct. 2018.

[4] D. J. Atha and M. R. Jahanshahi, "Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection," Structural Health Monitoring, vol. 17, no. 5, pp. 1110 - 1128, Sep. 2018.

[5] D. Denning and P. G. Neumann, "Requirements and model for IDES-a real-time intrusion-detection expert system," SRI International Menlo Park, vol. 8, Aug. 1985.

[6] K. Ilgun, R. A. Kemmerer, and P. A. Porras, "State Transition Analysis: A Rule-Based Intrusion Detection Approach," IEEE Transactions on Software Engineering, vol. 21, no. 3, pp. 181 - 199, Mar. 1995.

[7] B. Scholkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support Vector Method for Novelty Detection," Proceedings of the 12th Neural Information Processing Systems, pp. 582 - 588, Dec. 1999.

[8] D. M. J. Tax and R. P. W. Duin, "Support Vector Data Description," Machine Learning, vol. 54, no. 1, pp. 45 - 66, Jan. 2004.

[9] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft,

"Deep One-Class Classification," Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 4393 - 4402, Jul. 2018.

[10] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep Anomaly Detection with Outlier Exposure," 7th International Conference on Learning Representations, May 2019.

[11] W. Yan, L. K. Mestha, and M. Abbaszadeh, "Attack Detection for Securing Cyber Physical Systems," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8471 - 8481, Oct. 2019.

[12] J. Wang, D. Shi, Y. Li, J. Chen, H. Ding, and X. Duan, "Distributed Framework for Detecting PMU Data Manipulation Attacks With Deep Autoencoders," IEEE Transactions on Smart Grid, vol. 10, no. 4, pp. 4401 - 4410, Jul. 2019.

[13] A. Liu, Y. Wang, and T. Li, "SFE-GACN: A novel unknown attack detection under insufficient data via intra categories generation in embedding space," Computers and Security, vol. 105, pp. 102262, Jun. 2021.

[14] I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," Proceedings of the 27th Neural Information Processing Systems, pp. 2672 - 2680, Dec 2014.

[15] J. Lee and K. Park, "GAN-based imbalanced data intrusion detection system," Personal and Ubiquitous Computing 25.1, pp. 121-128, Nov. 2021.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," Proceedings of the 3rd International Conference on Learning Representations, Mar. 2015.

[17] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K. R. Müller, and M. Kloft, "Deep Semi-Supervised Anomaly Detection," In 8th International Conference on Learning Representations, Apr. 2020.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence, vol. 16, pp. 321 - 357, Jun. 2002.

[19] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training Generative Adversarial Networks with Limited Data," Advances in Neural Information Processing Systems 33, pp. 12104-12114, Dec. 2020.

[20] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. Salakhutdinov, "Good Semi-supervised Learning That Requires a Bad GAN," Proceedings of the 30th Neural Information Processing Systems, pp. 6510 - 6520, Dec. 2017.

[21] P. Zheng, S. Yuan, X. Wu, J. Li, and A. Lu, "One-Class Adversarial Nets for Fraud Detection," Proceedings of the AAAI Conference on Artificial Intelligence 33, pp. 1286 - 1293, Jul. 2019.

[22] S. T. K. Jan, Q. Hao, T. Hu, J. Pu, S. Oswa, G. Wang, and B. Viswanath, "Throwing Darts in the Dark? Detecting Bots with Limited Data using Neural Data Augmentation," In IEEE Symposium on Security and Privacy (SP), pp. 1190 - 1206, Jul. 2020.

[23] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved

Techniques for Training GANs," Advances in Neural Information Processing Systems 29, Dec. 2016.

[24] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?," In International Conference on Machine Learning, pp. 3481-3490, Jul. 2018.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," Proceedings of the 6th International Conference on Learning Representations, Feb. 2018.

[26] S. Goyal, A. Raghunathan, M. Jain, H. V. Simhadri, and P. Jain, "DROCC: Deep Robust One-Class Classification," Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 3711 - 3721, Jul. 2020.

[27] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," ACM Computing Surveys, vol. 54(2), pp. 1-38, Mar. 2022.

[28] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," In Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp. 108-116, Jan. 2018

[29] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130-165150, Sep. 2020.

[30] Y. Zhang, X. Chen, L. Jin, X. Wang, and D. Guo, "Network Intrusion Detection: Based on Deep Hierarchical Network and Original Flow Data," IEEE Access, vol. 7, pp. 37004 - 37016, Mar. 2019.

[31] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145 - 1159, Jul. 1997.

[32] S. Kullback and R. A. Leibler, "On Information and Sufficiency," The Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79 - 86, Mar. 1951.

[33] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579 - 2605, Nov. 2008.

〈저자소개〉

정 병 길 (Byeonggil Jung) 학생회원
2020년 2월: 한양대학교(ERICA) 컴퓨터공학과 졸업
2020년 9월~현재: 고려대학교 정보보호대학원 석사과정
〈관심분야〉 이상탐지, 인공지능, 표현학습

권 준 형 (Junhyung Kwon) 학생회원
2018년 2월: 한양대학교(ERICA) 컴퓨터공학과 졸업
2020년 3월~현재: 고려대학교 정보보호대학원 석박사통합과정
〈관심분야〉 인공지능, 인공지능 보안, 이상탐지

민 동 준 (Dongjun Min) 학생회원
2018년 8월: 한양대학교(ERICA) 교통물류학과 졸업
2020년 3월~현재: 고려대학교 정보보호대학원 석사과정
〈관심분야〉 인공지능, 연합학습, 시계열 데이터 예측

이 상 근 (Sangkyun Lee) 정회원
2003년 2월: 서울대학교 컴퓨터공학 학사 졸업
2005년 6월: 서울대학교 전기컴퓨터공학 대학원 석사 졸업
2011년 7월: 미국 Wisconsin-Madison 대학 컴퓨터과학 석사, 박사 졸업
2011년 8월~2017년 2월: 독일 TU Dortmund 대학 PostDoc Fellow, Project Leader
2017년 3월~2020년 2월: 한양대학교(ERICA) 소프트웨어공학과 조교수
2020년 3월~현재: 고려대학교 정보보호대학원 조교수
〈관심분야〉 인공지능보안, 인공지능 모델 복제 공격과 방어 기술, 설명 가능한 인공지능, 인공지능 모델 압축