(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2025 년 5 월 30 일 (30.05.2025)

WIPO | PCT

(10) 국제공개번호
WO 2025/110422 A1

(51) 국제특허분류:
H04L 9/40 (2022.01) *G06N 3/0455* (2023.01)

(21) 국제출원번호: PCT/KR2024/012559

(22) 국제출원일: 2024 년 8 월 22 일 (22.08.2024)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2023-0161868 2023 년 11 월 21 일 (21.11.2023) KR

(71) 출원인: 이화여자대학교 산학협력단 (**EWHA UNIVERSITY-INDUSTRY COLLABORATION FOUNDATION**) [KR/KR]; 03760 서울특별시 서대문구 이화여대길 52 (KR).

(72) 발명자: 박형곤 (**PARK, Hyunggon**); 10416 경기도 고양시 일산동구 일산로 241, 110동 204호 (KR). 이주홍 (**RHEEY, Joohong**); 03484 서울특별시 은평구 가좌로 9길 19-1, 401호 (KR).

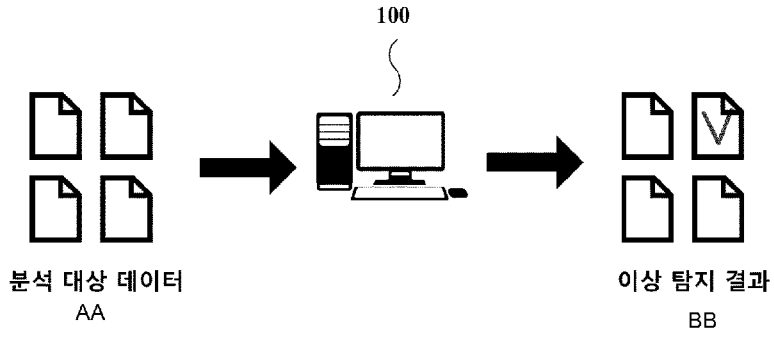
(74) 대리인: 특허법인(유한)아이시스 (**ISIS IP LAW LLC**); 06162 서울특별시 강남구 선릉로 94길 14, 12층 (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) **Title:** ANOMALY DETECTION METHOD ROBUST AGAINST ADVERSARIAL ATTACKS

(54) **발명의 명칭:** 적대적 공격에 강건한 이상 탐지 방법



(57) **Abstract:** An anomaly detection method robust against adversarial attacks includes the steps in which: an anomaly detection device acquires analysis target data; the anomaly detection device calculates a first outlier score for the analysis target data; the anomaly detection device pre-detects whether an anomaly is present in the analysis target data on the basis of the first outlier score; the anomaly detection device calculates a second outlier score for the analysis target data when whether an anomaly is present in the analysis target data cannot be determined as a result of the pre-detection; and the anomaly detection device post-detects whether an anomaly is present in the analysis target data on the basis of the second outlier score. (Representative drawing) FIG. 1

(57) **요약서:** 적대적 공격에 강건한 이상 탐지 방법은 이상 탐지 장치가 분석 대상 데이터를 획득하는 단계; 상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제1 이상치 점수를 계산하는 단계; 상기 이상 탐지 장치가 상기 제1 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 선탐지 하는 단계; 상기 선탐지 결과 상기 분석 대상 데이터에 이상이 있는지 판단 불가능 경우, 상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제2 이상치 점수를 계산하는 단계; 및 상기 이상 탐지 장치가 상기 제2 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 후탐지 하는 단계; 를 포함한다. (대표도) 도 1

WO 2025/110422 A1

[다음 쪽 계속]

규칙 4.17에 의한 선언서:

— 신규성을 해치지 아니하는 개시 또는 신규성 상실의 예외에 관한 선언 (규칙 4.17(v))

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 적대적 공격에 강건한 이상 탐지 방법

기술분야

[1] 이하 설명하는 기술은 이상 탐지 방법에 대한 것이다.

배경기술

[2] 이상 탐지(Anomaly Detection) 기술은 정상(Normal)이 아닌 이상(Abnormal) 상황을 탐지하는 기술을 의미한다. 예를 들어 이상 탐지 기술은 시스템이나 네트워크에서 일어나는 이상 행위나 패턴을 탐지하는 것을 포함할 수 있다. 또는 이상 탐지 기술은 기계 장치가 고장이 나는 상황을 탐지하는 것을 포함할 수 있다. 최근에는 기계학습(Machine Learning, ML)의 한 분야인 인공신경망(Artificial Neural Network, ANN)에 기반한 인공지능 기술이 발전하기 시작하면서, 인공지능에 기반한 인공신경망 기반의 이상 탐지 방법 등이 개발되고 있다. 예를 들어 네트워크 트래픽 데이터를 인공신경망 기반의 모델에 입력한 뒤, 인공신경망 모델의 출력값을 기반으로 이상 행위가 있었는지 판단하는 기술들이 있다.

[3] [선행기술문헌]

[4] [특허문헌]

[5] (특허문헌 1) 한국 공개특허 10-2023-0031117

발명의 상세한 설명

기술적 과제

[6] 인공신경망 기반의 이상 탐지 모델을 활용함에 있어 문제되는 것 중 하나는 적대적 공격(Adversarial Attacks)이다. 적대적 공격은 입력 데이터에 미세한 변화를 가져와 이상 탐지 모델을 속이는 공격이다. 예를 들어, 적대적 공격은 사람 눈으로는 볼 수 없는 미세한 노이즈(Noise)을 입력데이터에 가하여 모델이 잘못된 판단을 하도록 유도하는 것을 의미할 수 있다. 이상 탐지 모델은 적대적 공격이 수행된 데이터에 대하여 정확한 판단을 할 수 없어, 이상 탐지 모델의 신뢰도를 하락시킨다.

[7] 이하 설명하는 기술은 적대적 공격에 강건한 이상 탐지 방법을 개시하고자 한다.

과제 해결 수단

[8] 적대적 공격에 강건한 이상 탐지 방법은 이상 탐지 장치가 분석 대상 데이터를 획득하는 단계; 상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제1 이상치 점수를 계산하는 단계; 상기 이상 탐지 장치가 상기 제1 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 선탐지 하는 단계; 상기 선탐지 결과 상기 분석 대상 데이터에 이상이 있는지 판단 불가능 경우, 상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제2 이상치 점수를 계산하는 단계; 및 상기

이상 탐지 장치가 상기 제2 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 후탐지 하는 단계; 를 포함한다.

발명의 효과

[9] 이하 설명하는 기술을 이용하면 이상을 탐지할 수 있다. 예를 들어 이하 설명하는 기술을 이용해 네트워크 트래픽 데이터로부터 이상 트래픽을 검출하여 사이버 공격을 탐지할 수 있다.

[10] 이하 설명하는 기술을 이용하면 적대적 공격이 수행된 데이터에서도 이상을 탐지할 수 있다. 즉 이하 설명하는 기술을 이용하면 적대적 공격에 강건한 이상 탐지 방법을 수행할 수 있다. 예를 들어 이하 설명하는 기술을 이용하면 사이버 공격을 위하여 적대적 공격이 수행된 네트워크 트래픽 데이터로부터 이상 트래픽을 검출하여 사이버 공격을 탐지할 수 있다.

도면의 간단한 설명

[11] 도1은 이상 탐지 장치(100)가 적대적 공격에 강건한 이상 탐지 방법을 수행하는 전체적인 과정이다.

[12] 도2는 적대적 공격에 강건한 이상 탐지 방법의 실시예 중 하나의 순서도(200)이다.

[13] 도3은 적대적 공격에 강건한 이상 탐지 방법이 수행되는 예시 중 하나이다.

[14] 도4는 이상 탐지 모델에 입력되는 분석 대상 데이터의 예 중 하나이다.

[15] 도5는 정상(Normal) 데이터, 이상(Abnormal) 데이터 및 적대적(Adversarial) 데이터에 대한 이상치 점수이다.

[16] 도6은 이상 탐지 장치(300)의 실시예 중 하나의 구성이다.

발명의 실시를 위한 형태

[17] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있다. 명세서의 도면에 이하 설명하는 기술의 특정 실시 형태가 기재될 수 있다. 그러나, 이는 이하 설명하는 기술의 설명을 위한 것이며 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니다. 따라서 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경 물, 균등 물 내지 대체 물이 이하 설명하는 기술에 포함하는 것으로 이해되어야 한다.

[18] 다양한 구성요소들을 설명하기 위해서 제1, 제2, A, B 등의 용어가 사용될 수 있다. 하지만 상기 용어는 단지 하나의 구성요소를 다른 구성요소들과 구별하기 위해서 사용될 뿐, 상기 용어로 해당 구성요소들을 한정하려고 하는 것이 아니다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. “및/또는” 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[19] 이하 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 기

재된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계 동작 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.

[20] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성 부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전 부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.

[21] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백 하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[22] 이하 이상 탐지 장치가 적대적 공격에 강건한 이상 탐지 방법을 수행하는 전체적인 과정을 살펴본다.

[23] 도1은 이상 탐지 장치(100)가 적대적 공격에 강건한 이상 탐지 방법을 수행하는 전체적인 과정이다.

[24] 이상 탐지 장치(100)는 물리적으로 다양한 형태로 구현될 수 있다. 예를 들어 이상 탐지 장치(100)는 PC, 노트북, 스마트기기, 서버 또는 데이터처리 전용 칩셋 등의 형태를 가질 수 있다.

[25] 이상 탐지 장치(100)는 적대적 공격에 강건한 이상 탐지 방법을 수행할 수 있다. 이상 탐지 장치(100)는 분석 대상 데이터를 획득하여, 이상 탐지 결과를 출력하는 장치일 수 있다.

[26] 구체적으로 이상 탐지 장치(100)는 분석 대상 데이터를 획득할 수 있다. 이상 탐지 장치(100)는 분석 대상 데이터에 대한 제1 이상치 점수를 계산할 수 있다. 이상 탐지 장치(100)는 제1 이상치 점수를 기반으로 분석 대상 데이터에 이상이 있는지 여부를 선탐지 할 수 있다. 선탐지 결과, 분석 대상 데이터에 이상이 있는지 판단 불가능한 경우, 이상 탐지 장치(100)는 분석 대상 데이터에 대한 제2 이상치 점수를 계산할 수 있다. 이상 탐지 장치(100)는 제2 이상치 점수를 기반으로 분석 대상 데이터에 이상이 있는지 여부를 후탐지 할 수 있다.

[27] 이하 적대적 공격에 강건한 이상 탐지 방법에 대해 구체적으로 설명한다.

[28] 도2는 적대적 공격에 강건한 이상 탐지 방법의 실시예 중 하나의 순서도(200)이다. 도3은 적대적 공격에 강건한 이상 탐지 방법이 수행되는 예시 중 하나이다.

[29] 이상 탐지 장치는 분석 대상 데이터를 획득할 수 있다(210).

[30] 분석 대상 데이터는 이하 설명하는 기술을 통해 이상을 탐지할 때 필요한 데이터일 수 있다. 즉 분석 대상 데이터는 이상이 있는지 탐지하기 위한 분석 대상이 되는 데이터일 수 있다.

[31] 예를 들어 분석 대상 데이터는 네트워크 트래픽을 측정한 데이터일 수 있다. 네트워크가 정상적일 때 측정한 분석 대상 데이터는 정상(Normal) 데이터가 될 수 있다. 반대로 네트워크에 이상이 생겼을 때(ex 사이버 공격, 트래픽 쏠림 현상 발생) 측정한 분석 대상 데이터는 이상(Abnormal) 데이터가 될 수 있다. 또 다른 예로, 분석 대상 데이터는 기계 작동음 데이터일 수 있다. 기계가 정상적으로 작동할 때 측정한 분석 대상 데이터는 정상 데이터가 될 수 있다 반대로 기계에 이상이 생겼을 때 측정한 분석 대상 데이터는 이상 데이터가 될 수 있다.

[32] 분석 대상 데이터는 적대적 공격(Adversarial Attack)이 수행된 데이터를 포함할 수 있다. 적대적 공격은 입력 데이터에 미세한 변화를 가져와 모델을 속이는 공격이다. 예를 들어 적대적 공격이 수행된 이상 데이터를 모델에 입력하면, 모델은 이를 정상 데이터로 판단할 수 있다. 따라서 적대적 공격에 강건한(Robustness) 모델을 구축할 필요가 있다. 이하 설명하는 기술은 적대적 공격이 수행된 데이터에서도 이상을 잘 탐지할 수 있다.

[33] 분석 대상 데이터는 복수개의 특징 값을 가질 수 있다. 분석 대상 데이터는 복수개의 데이터일 수 있다. 예를 들어 분석 대상 데이터는 d개의 특징 값을 가지는 n개의 데이터일 수 있다. 즉 분석 대상 데이터(**X**)는 $X \in \mathbb{R}^{n \times D}$일 수 있다.

[34] 이상 탐지 장치는 분석 대상 데이터에 대한 제1 이상치 점수를 계산할 수 있다(220).

[35] 제1 이상치 점수는 분석 대상 데이터가 정상 데이터 인지 이상 데이터인지 판단하는데 필요한 점수일 수 있다. 예를 들어 제1 이상치 점수가 미리 설정된 값 이상이라면 분석 대상 데이터를 이상 데이터로 판단할 수 있다. 반대로 제1 이상치 점수가 미리 설정된 값 이하라면 분석 대상 데이터를 정상 데이터로 판단할 수 있다.

[36] 제1 이상치 점수는 분석 대상 데이터를 이상 탐지 모델(Anomaly Detection Model)에 입력 한 뒤, 이상 탐지 모델이 연산하는 값을 기반으로 계산된 것일 수 있다.

[37] 이상 탐지 모델은 기계학습(Machine Learning, ML) 기반의 모델일 수 있다. 기계 학습은 컴퓨팅 장치가 데이터를 통해 학습하여 특정 대상 혹은 조건을 이해할 수 있게 하거나 데이터의 패턴을 찾아내 분류하는 기술적 방식일 수 있다. 상기 기계 학습 기반의 모델은 결정 트리(Decision Tree), RF(random forest), KNN(K-nearest neighbor), 나이브 베이즈(Naive Bayes), SVM(support vector machine) 및 ANN(artificial neural network) 등이 포함될 수 있다.

[38] 이상 탐지 모델은 인공신경망(Artificial Neural Network, ANN) 기반의 모델일 수 있다. ANN은 DNN(Deep Neural Network)가 될 수 있다. DNN

은 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), RBM(Restricted Boltzmann Machine), DBN(Deep Belief Network), GAN(Generative Adversarial Network) 및 RL(Relation Networks) 등이 포함될 수도 있다.

[39] 이상 탐지 모델은 오토인코더(Autoencoder) 기반의 모델일 수 있다. 오토인코더 기반의 모델은 입력 받은 데이터으로부터 잠재 공간(Latent Space)을 형성한 뒤, 잠재 공간으로부터 입력 데이터를 재구성할 수 있다.

[40] 더 나아가 이상 탐지 모델은 변이형 오토인코더(Variational Autoencoder, VAE) 기반의 모델일 수 있다. 변이형 오토인코더의 인코더는 입력 데이터로부터 평균 벡터와 분산 벡터를 추출할 수 있다. 변이형 오토인코더의 인코더는 평균 벡터와 분산 벡터에 대응하는 다변량 정규분포로부터 샘플링을 통하여 잠재 변수(Latent variable)가 있는 잠재 공간을 형성할 수 있다. 변이형 오토인코더의 디코더는 잠재 공간 내 잠재 변수를 활용하여 입력 데이터와 유사한 분포를 따르는 출력 데이터를 생성할 수 있다.

[41] 이상 탐지 모델은 정상 데이터만을 기반으로 학습된 모델일 수 있다. 따라서 이상 탐지 모델은 정상 데이터는 잘 처리할 수 있다. 반대로 이상 탐지 모델은 이상 데이터는 잘 처리하지 못할 수 있다. 예를 들어 이상 탐지 모델이 변이형 오토인코더 모델인 경우를 본다. 정상 데이터만을 기반으로 학습된 변이형 오토인코더 모델은 정상 데이터를 입력 받으면 정상 데이터와 유사한 형태의 데이터를 재구성한다. 반대로 정상 데이터만을 기반으로 학습된 변이형 오토인코더 모델은 이상 데이터를 입력 받으면 이상 데이터를 잘 재구성하지 못한다. 이상 데이터로는 학습된 적이 없으며, 정상 데이터를 재구성하는 방법만을 학습하였기 때문이다. 따라서 이상 탐지 모델에 입력되는 분석 대상 데이터와 이상 탐지 모델의 출력값인 재구성 데이터를 비교하여 제1 이상치 점수를 계산할 수 있다.

[42] 수학식1은 제1 이상치 점수를 계산할 때 이용되는 식 중 하나이다. 수학식1은 평균 절대 오차(Mean Absolute Error, MAE) 기반의 제1 이상치 점수를 계산할 때 이용되는 식 중 하나이다.

[43] [수학식 1]

[44]

$$e(x_i) = \frac{1}{d}\sum_{k=1}^{d}\left|x_i^{(k)} - \hat{x}_i^{(k)}\right|$$

[45] 수학식1에서 x_i는 i번째 분석 대상 데이터(입력 데이터)이다. 수학식1에서 x_i는 $[x^{(1)}_1, x^{(2)}_2, x^{(3)}_3, ..., x^{(d)}_i]$ 일 수 있다. 수학식1에서 $\hat{x}_i$는 i번째 분석 대상 데이터를 이상 탐지 모델에 입력하였을때의 출력값이다. 수학식1에서 $\hat{x}_i$는 $[\hat{x}^{(1)}_1, \hat{x}^{(2)}_2, \hat{x}^{(3)}_3$

, ..., $\hat{x}_i^{(d)}$] 일 수 있다. 수학식1에서 $e(x_i)$는 i번째 분석 대상 데이터에 대한 제1 이상치 점수이다. 수학식1에서 d는 분석 대상 데이터의 특징 값의 개수(차원 수)이다.

[46] 이상 탐지 장치는 제1 이상치 점수를 기반으로 분석 대상 데이터에 이상이 있는지 여부를 선탐지 할 수 있다(230).

[47] 이상 탐지 모델은 적대적 공격이 수행된 분석 대상 데이터에 취약할 수 있다. 예를 들어 이상 탐지 모델에 적대적 공격이 수행되어 정상 데이터 형태와 비슷한 형태를 가지는 분석 대상 데이터가 입력 되면, 이상 탐지 모델은 입력 데이터를 정상 데이터와 비슷하게 재구성할 수 있다. 이 경우 이상 데이터를 정상 데이터로 잘못 판단할 수 있다. 따라서 제1 이상치 점수만으로는 정상 데이터와 적대적 공격이 수행된 이상 데이터를 구분하기 힘들 수 있다. 이러한 문제점을 해결하기 위해서 이하 설명하는 기술은 이상 탐지를 선탐지 및 후탐지로 나누어 수행한다.

[48] 선탐지 하는 것은 이상이 있는지 여부 판단이 쉬운 또는 이상이 있는지 여부가 확실한 분석 대상 데이터를 기반으로 이상의 유무를 탐지하는 과정일 수 있다. 구체적으로 선탐지 하는 것은 제1 이상치 점수만을 이용하였을 때 이상이 있는지 확실히 판단이 가능한 경우인지 확인하고, 이상이 있는지 확실히 판단이 가능한 경우라면 분석 대상 데이터를 정상 데이터 또는 이상 데이터로 구분하는 과정일 수 있다. 따라서 선탐지 결과 이상이 있는지 확실히 판단 불가능 하다면, 이상이 있는지 여부는 후탐지 과정을 통해 판단된다. 선탐지 과정에서 판단 불가능한 분석 대상 데이터를 불확실(Uncertain) 데이터라고 부를 수 있다. 후탐지 과정에 대해선 이하에서 구체적으로 설명한다.

[49] 선탐지 하는 것은 제1 이상치 점수를 제1 기준값 및 제2 기준값과 비교하여 수행되는 과정을 포함할 수 있다. 제1 기준값은 제2 기준값보다 클 수 있다(제1 기준값 > 제2 기준값).

[50] 일 실시예로 선탐지 하는 것은 제1 이상치 점수가 제1 기준값보다 초과(또는 이상)인 경우 이상이 있는 것으로 탐지하는 과정을 포함할 수 있다. 또는 선탐지 하는 것은 제1 이상치 점수가 제2 기준값보다 미만(또는 이하)인 경우 정상인 것으로 탐지하는 과정을 포함할 수 있다. 또는 선탐지 하는 것은 제1 이상치 점수가 제1 기준값보다 이하(또는 미만)이면서 제2 기준값보다 이상(또는 초과)라면 이상이 있는지 판단 불가능한 경우로 판단하는 과정을 포함할 수 있다.

[51] 수학식2는 선탐지 과정에서 이용되는 식 중 하나이다.

[52] [수학식 2]

[53]

$$t = \begin{cases} 0, & e(x_i) \le \delta_l \\ 2, & \delta_l < e(x_i) \le \delta_u \\ 1, & e(x_i) > \delta_u \end{cases}$$

[54] 수학식2에서 t는 이상 여부이다. 수학식2에서 0은 정상인 경우이다. 수학식2에서 1은 이상이 있는 경우이다. 수학식2에서 2는 이상이 있는지 판단 불가능한 경우이다. 수학식2에서 δ_u는 제1 기준값이다. 수학식2에서 δ_l은 제2 기준값이다. 수학식2에서 x_i는 i번째 분석 대상 데이터(입력 데이터)이다. 수학식2에서 $e(x_i)$는 i번째 분석 대상 데이터에 대한 제1 이상치 점수이다.

[55] 선탐지 결과 분석 대상 데이터에 이상이 있는지 판단 불가능한 경우, 이상 탐지 장치는 분석 대상 데이터 대한 제2 이상치 점수를 계산할 수 있다(240).

[56] 선탐지 결과 이상이 있는지 판단 불가능 경우는 전술한 바와 같이 제1 이상치 점수만을 이용하였을 때 이상이 있는지 확실히 판단이 불가능한 경우일 수 있다. 즉 선탐지 결과 이상이 있는지 판단이 불가능한 경우는, 제1 이상치 점수가 제1 기준값 이하(미만)이면서 제2 기준값 이상(초과)인 경우를 포함할 수 있다. 따라서 선탐지 결과 이상이 있는지 판단 가능하다면 제2 이상치 점수는 계산할 필요가 없다.

[57] 제2 이상치 점수는 불확실 데이터가 정상 데이터인지 이상 데이터인지 판단하는데 필요한 점수일 수 있다. 예를 들어 제2 이상치 점수가 미리 설정된 값 이상이라면 불확실 데이터를 이상 데이터로 판단할 수 있다. 반대로 제2 이상치 점수가 미리 설정된 값 이하라면 불확실 데이터를 정상 데이터로 판단할 수 있다.

[58] 제2 이상치 점수는 섀플리 값(Shapley Value)을 기반으로 계산된 것일 수 있다.

[59] 일 실시예로 제2 이상치 점수는 섀플리 값에 기반을 둔 특징 중요도 값을 기반으로 계산된 것일 수 있다. 특징 중요도 값은 분석 대상 데이터에 포함된 각 특징 값이 이상 탐지에 영향을 주는 정도를 포함할 수 있다. 구체적으로 특징 중요도 값은 분석 대상 데이터에 포함된 각 특징 값이 이상 탐지 모델에 영향을 주는 정도를 포함할 수 있다. 예를 들어 이상탐지 모델이 변이형 오토인코더라면, 특징 중요도 값은 변이형 오토인코더의 기반의 이상 탐지 모델이 잠재 변수를 만드는데 각 특징 값이 얼마나 기여하였는지를 의미할 수 있다.

[60] 수학식3은 특징 중요도 값을 계산하는데 이용되는 식 중 하나이다.

[61] [수학식 3]

[62]

$$\emptyset^{(j)} = \sum_{P \subseteq N \setminus \{j\}} \frac{|P|!\,(|N| - |P| - 1)!}{|N|!} \left(v(P \cup \{j\}) - v(P)\right)$$

[63] 수학식3에서 Φ^j는 n개의 입력 데이터에 대한 j번째 특징 벡터($f^{(j)}$)에 대한 특징 중요도 값이다. 특징 벡터($f^{(j)}$)는 벡터 형태로 [$x^{(j)}_1$, $x^{(j)}_2$, $x^{(j)}_3$, ..., $x^{(d)}_n$]일 수 있다 (1<=j<=d). 즉 특징 벡터는 도4와 같이 d개의 특징 값을 가지는 n개의 입력 데이터에서 j번째 특징 값을 추출하여 만든 벡터일 수 있다. 수학식3에서 N는 특징 집합이다. 수학식3에서 P ($\subset$ N {j})는 j번째 특징을 배제한 부분집합 일 수 있다. 수학식3에서 v (.)는 가치함수일 수 있다.

[64] 제2 이상치 점수는 분석 대상 데이터와 정상 데이터를 비교하여 계산된 것 일 수 있다. 일 실시예로 제2 이상치 점수는 분석 대상 데이터에 포함된 특징 값에 대한 특징 중요도 값과 정상 데이터에 포함된 특징 값에 대한 특징 중요도 값의 차이를 기반으로 계산된 것일 수 있다.

[65] 수학식4는 제2 이상치 점수를 계산하는데 이용되는 식 중 하나이다.

[66] [수학식 4]

[67]

$$e_s(\emptyset_i) = \frac{1}{d}\sum_{j=1}^{d}\left|\bar{\emptyset}^{(j)} - \emptyset_i^{(j)}\right|$$

[68] 수학식4에서 Φ_i는 i번째 분석 대상 데이터(x_i)에 대한 특징 중요도 값이다. 수학식4에서 Φ_i는 [$\Phi^{(1)}_1$, $\Phi^{(2)}_2$, $\Phi^{(3)}_3$, …$\Phi^{(d)}_i$] 일 수 있다. 수학식4에서 $\Phi^{(j)}_i$ 는 i번째 분석 대상 데이터의 j번째 특징 값에 대한 특징 중요도 값이다. 수학식4에서 $\bar{\Phi}^{(j)}$는 정상 데이터의 j번째 특징 값에 대한 특징 중요도 평균 값이다. 수학식4에서 $e_s(\Phi_i)$는 i번째 분석 대상 데이터에 대한 제2 이상치 점수이다. 수학식4에서 d는 특징 중요도 값의 차원수로, 분석 대상 데이터의 차원수와 같다.

[69] 이상 탐지 장치는 제2 이상치 점수를 기반으로 분석 대상 데이터에 이상이 있는지 여부를 후탐지 할 수 있다(250).

[70] 후탐지 하는 것은 선탐지 과정에서 이상 여부를 판단하지 못한 불확실 데이터에 대하여 이상이 있는지 판단하는 과정을 포함할 수 있다.

[71] 후탐지 하는 것은 제2 이상치 점수와 제3 기준값을 비교하여 수행되는 과정을 포함할 수 있다.

[72] 일 실시예로 후탐지 하는 것은 제2 이상치 점수가 제3 기준값보다 초과(또는 이상)인 경우 이상이 있는 것으로 탐지하는 과정을 포함할 수 있다. 또는 후탐지 하는 것은 제2 이상치 점수가 제3 기준값보다 미만(또는 이하)인 경우 정상인 것으로 탐지하는 과정을 포함할 수 있다.

[73] 수학식5는 후탐지 과정에서 이용되는 식 중 하나이다.

[74] [수학식 5]

[75]

$$t_s = \begin{cases} 0, & e_{s(\Phi_i)} \leq \delta_s \\ 1, & e_{s(\Phi_i)} > \delta_s \end{cases}$$

[76] 수학식5에서 t_s는 이상 여부이다. 수학식5에서 0은 정상인 경우이다. 수학식5에서 1은 이상이 있는 경우이다. 수학식5에서 δ_s는 제3 기준값이다. 수학식5에서$e_s(\Phi_i)$는 i번째 분석 대상 데이터에 대한 제2 이상치 점수이다.

[77] 이하 이상 탐지 모델을 구축하고, 구축된 이상 탐지 모델을 이용해 이상을 탐지한 실험결과에 대해 살펴 본다.

[78] 실험에 이용한 데이터 셋은 NSL-KDD데이터 셋이다. NSL-KDD는 네트워크 침입 탐지를 위해 사용되는 데이터 셋이다.

[79] 실험에 이용된 이상 탐지 모델은 변이형 오토인코더이다. 변이형 오토인코더는 d (d=40)개의 특징 값을 가지는 n개의 분석 대상 데이터($X \in \mathbb{R}^{n \times d}$)를 입력 받을 수 있다. 변이형 오토인코더는 m(m=20)개의 은닉층의 노드를 가질 수 있다. 변이형 오토인코더는 분석 대상 데이터로(**X**)부터 평균 벡터와 분산 벡터를 추출할 수 있다. 변이형 오토인코더는 평균 벡터와 분산 벡터에 대응하는 다변량 정규 분포로부터 샘플링을 통해 잠재 변수(latent variable, $Z \in \mathbb{R}^{m \times d}$)를 생성할 수 있다. 변이형 오토인코더는 생성된 잠재 변수(**Z**)를 활용하여 분석 대상 대상 데이터(**X**)와 유사한 분포를 따르는 출력 데이터($\hat{X} \in \mathbb{R}^{n \times d}$) 생성할 수 있다.

[80] 적대적 공격이 수행된 데이터를 구축하기 위하여 FGSM (Fast Gradient Sign Method)를 이용할 수 있다. FGSM (Fast Gradient Sign Method)는 데이터를 한 단계만 갱신 시켜 적대적 예제를 빠르게 생성하는 방법일 수 있다.

[81] 도5는 정상(Normal) 데이터, 이상(Abnormal) 데이터 및 적대적(Adversarial) 데이터에 대한 이상치 점수이다. 도5(a)는 제1 이상치 점수이다. 도5(b)는 제2 이상치 점수이다. 적대적 데이터는 이상 데이터에 사람은 구별할 수 없는 노이즈를 추가한 형태의 적대적 공격이 수행된 데이터이다.

[82] 도5(a)에서 볼 수 있듯이, 정상 데이터와 비정상 데이터의 제1 이상치 점수 분포는 다르다. 따라서 제1 이상치 점수를 기반으로 정상 데이터와 비정상 데이터를 구별할 수 있다. 반면에 정상 데이터와 적대적 데이터의 제1 이상치 점수의 분포는 매우 유사하다. 제1 이상치 점수를 기반으로 정상 데이터와 적대적 데이터를 구별하기는 어렵다.

[83] 도5(b)와 볼 수 있듯이, 정상 데이터와 비정상데이터의 제2 이상치 점수 분포는 다르다. 따라서 제2 이상치 점수를 기반으로 정상 데이터와 비정상데이터를 구별할 수 있다.

[84] 표6은 이상 탐지 모델의 성능 평가 결과이다.

[85] [표1]

	Accuracy	Precision	Recall	F1- score
Baseline	0.7874	0.8004	0.9075	0.8506
Proposed	**0.9343**	**0.9371**	**0.9663**	**0.9515**

[86] 기존 방법(Baseline)의 이상 탐지 정확도(Accuracy)는 0.7874, 정밀도(Precision)은 0.8004, 재현율(Recall)은 0.9075 및 F1-score는 0.8506이다.

[87] 반면에 적대적 공격에 강건한 계층적 이상 탐지 방법(Proposed)의 정확도는 0.9343, 정밀도는 0.9371, 재현율은 0.9663 및 F1-score는 0.9515인 것을 확인할 수 있다.

[88] 전술한 계층적 이상 탐지 방법의 성능이 기존 이상 탐지 방법보다 향상된 것을 확인할 수 있다.

[89] 이하 이상 탐지 장치에 대해 설명한다.

[90] 도6은 이상 탐지 장치(300)의 실시예 중 하나의 구성이다.

[91] 도6의 이상 탐지 장치(300)는 도1에서 설명한 이상 탐지 장치(100)에 해당할 수 있다. 즉 도6의 이상 탐지 장치(300)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는 장치일 수 있다.

[92] 이상 탐지 장치(300)는 입력장치(310), 저장장치(320), 연산장치(330), 출력장치(340), 인터페이스 장치(350) 및 통신장치(360)를 포함할 수 있다.

[93] 입력장치(310)는 일정한 명령 또는 데이터를 입력 받는 인터페이스 장치(키보드, 마우스, 터치스크린 등)를 포함할 수도 있다. 입력장치(310)는 별도의 저장장치(USB, CD, 하드디스크 등)를 통하여 정보를 입력 받는 구성을 포함할 수도 있다. 입력장치(310)는 입력 받는 데이터를 별도의 측정장치를 통하여 입력 받거나, 별도의 DB를 통하여 입력 받을 수도 있다. 입력장치(310)는 통신장치(360)을 통해 데이터를 유선 또는 무선 통신으로 입력 받을 수도 있다.

[94] 입력장치(310)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 정보를 입력 받을 수 있다. 입력장치(310)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 모델을 입력 받을 수 있다. 입력장치(310)는 분석 대상 데이터를 입력 받을 수 있다. 입력장치(310)는 이상 탐지 모델을 입력 받을 수 있다.

[95] 저장장치(320)는 일정한 정보를 저장하는 장치가 될 수도있다. 저장장치(320)는 입력장치(310)를 통해 입력 받은 정보를 저장할 수 있다. 저장장치(320)는 연산장치(330)가 연산하는 과정에서 생성되는 정보를 저장할 수 있다. 즉 저장장치(320)는 메모리를 포함할 수 있다.

[96] 저장장치(320)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 정보를 저장할 수 있다. 저장장치(320)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 모델을 저장할 수 있다. 저장장치(320)는 분석 대상 데이터를 저장할 수 있다. 저장장치(320)는 이상 탐지 모델을 저장할 수 있다.

[97] 연산장치(330)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다. 연산장치(330)는 이상 탐지 장치(300)를 제어하는 제어신호를 생성할 수 있다. 연산장치(330)는 이상 탐지 장치(300)에 포함된 입력장치(310), 저장장치(320), 출력장치(340), 인터페이스 장치(350) 및 통신장치(360)을 제어하는 제어신호를 생성할 수 있다.

[98] 연산장치(330)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 연산을 할 수 있다. 연산장치(330)는 분석 대상 데이터에 대한 제1 이상치 점수를 계산할 수 있다. 연산장치(330)는 제1 이상치 점수를 기반으로 분석 대상 데이터에 이상이 있는지 여부를 선탐지할 수 있다. 선탐지 결과 분석 대상 데이터에 이상이 있는지 판단 불가능한 경우, 연산장치(330)는 분석 대상 데이터에 대한 제2 이상치 점수를 계산할 수 있다. 연산장치(330)는 제2 이상치 점수를 기반으로 분석 대상 데이터에 이상이 있는지 여부를 후탐지할 수 있다.

[99] 출력장치(340)는 일정한 정보를 출력하는 장치가 될 수도 있다. 출력장치(340)는 데이터 과정에 필요한 인터페이스, 입력된 데이터, 분석결과 등을 출력할 수도 있다. 출력장치(340)는 디스플레이, 문서를 출력하는 장치, 스피커등과 같이 물리적으로 다양한 형태로 구현될 수도 있다. 출력장치(340)는 저장장치(330)에 저장된 정보를 출력할 수 있다. 출력장치(340)는 연산장치(330)가 연산하는 과정에서 생성된 정보를 출력할 수 있다. 출력장치(340)는 연산장치(330)가 연산한 결과를 출력할 수 있다. 출력장치(340)는 이상 탐지 결과를 출력할 수 있다.

[100] 인터페이스 장치(350)는 외부로부터 일정한 명령 및 데이터를 입력 받는 장치일 수 있다. 인터페이스 장치(350)는 이상 탐지 장치(300)를 제어하기 위한 제어신호를 입력 받을 수 있다. 인터페이스 장치(350)는 이상 탐지 장치(300)가 분석한 결과를 출력할 수 있다. 인터페이스 장치(350)는 물리적으로 연결된 입력 장치 또는 외부 저장장치로부터 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 정보를 입력 받을 수 있다.

[101] 통신장치(360)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미할 수 있다. 통신장치(360)는 Wi-Fi(Wireless Fidelity), Wi-Fi Direct, 블루투스(Bluetooth), UWB(Ultra Wide Band) 또는 NFC(Near Field Communication), USB(Universal Serial Bus), 혹은 HDMI(High Definition Multimedia Interface), LAN(Local Area Network) 등과 같은 네트워크 통신을 수행할 수 있다. 통신장치(360)는 이상 탐지 장치(300)를 제어하는데 필요한 제어 신호를 수신할 수 있다. 통신장치(360)는 이상 탐지 장치(300)가 분석한 결과를 전송할 수 있다. 통신장치(360)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 정보를 수신받을 수 있다. 통신장치(360)는 전술한 적대적 공격에 강건한 이상 탐지 방법을 수행하는데 필요한 모델을 수신 받을 수 있다.

[102] 전술한 적대적 공격에 강건한 이상 탐지 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다.

[103] 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.

[104] 상기 일시적 판독 가능 매체는 스태틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램 (Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(Synclink DRAM, SLDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

[105] 상기 비일시적 판독 가능 매체는 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM (read-only memory), PROM (programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[106] 본 실시예 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형예와 구체적인 실시예는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

청구범위

[청구항 1] 이상 탐지 장치가 분석 대상 데이터를 획득하는 단계;
상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제1 이상치 점수를 계산하는 단계;
상기 이상 탐지 장치가 상기 제1 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 선탐지 하는 단계;
상기 선탐지 결과 상기 분석 대상 데이터에 이상이 있는지 판단 불가능 경우, 상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제2 이상치 점수를 계산하는 단계; 및
상기 이상 탐지 장치가 상기 제2 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 후탐지 하는 단계; 를 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 2] 제1항에 있어서,
상기 분석 대상 데이터는 적대적 공격(Adversarial Attack)이 수행된 데이터를 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 3] 제1항에 있어서,
상기 분석 대상 데이터는 복수개의 특징 값을 가지는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 4] 제1항에 있어서,
상기 제1 이상치 점수는 상기 분석 대상 데이터를 이상 탐지 모델(Anomaly Detection Model)에 입력 한 뒤, 상기 이상 탐지 모델이 연산하는 값을 기반으로 계산된 것인, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 5] 제4항에 있어서,
상기 이상 탐지 모델은 오토인코더(Autoencoder) 또는 변이형 오토인코더(Variational Autoencoder, VAE) 기반의 모델인, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 6] 제1항에 있어서,
상기 선탐지 하는 것은 상기 제1 이상치 점수를 제1 기준값 및 제2 기준값과 비교하여 수행되는 과정을 포함하고,
상기 제1 기준값은 상기 제2 기준값보다 큰, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 7] 제6항에 있어서,
상기 선탐지 하는 것은 상기 제1 이상치 점수가 상기 제1 기준값보다 초과 또는 이상인 경우 이상이 있는 것으로 탐지하는 과정을 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 8] 제6항에 있어서,

상기 선탐지 하는 것은 상기 제1 이상치 점수가 상기 제2 기준값보다 미만 또는 이하인 경우 정상인 것으로 탐지하는 과정을 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 9] 제6항에 있어서,
상기 선탐지 결과 이상이 있는지 판단이 불가능한 경우는 상기 제1 이상치 점수가 상기 제1 기준값보다 미만 또는 이하이면서 상기 제2 기준값보다 초과 또는 이상인 경우를 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 10] 제1항에 있어서,
상기 제2 이상치 점수는 섀플리 값(Shapley Value)을 기반으로 계산된 것인, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 11] 제10항에 있어서,
상기 제2 이상치 점수는 상기 섀플리 값을 기반으로 둔 특징 중요도 값을 기반으로 계산된 것이고,
상기 특징 중요도 값은 상기 분석 대상 데이터에 포함된 각 특징 값이 이상 탐지에 영향을 주는 정도를 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 12] 제11항에 있어서,
상기 특징 중요도 값은 하기 수학식3을 이용하여 계산되는 것인, 적대적 공격에 강건한 이상 탐지 방법.

[수학식 3]

$$\emptyset^{(j)} = \sum_{P \subseteq N \setminus \{j\}} \frac{|P|!\,(|N| - |P| - 1)!}{|N|!} \left(v(P \cup \{j\}) - v(P)\right)$$

상기 수학식3에서 Φ^j는 n개의 입력 데이터에 대한 j번째 특징 벡터($f^{(j)}$)에 대한 특징 중요도 값이다. 상기 특징 벡터($f^{(j)}$)는 벡터 형태로 [$x^{(j)}_1$, $x^{(j)}_2$, $x^{(j)}_3$, ..., $x^{(d)}_n$]이다(1<=j<=d). 상기 특징 벡터는 d개의 특징 값을 가지는 n개의 입력 데이터에서 j번째 특징 값을 추출하여 만든 벡터이다. 상기 수학식3에서 N는 특징 집합이다. 수학식3에서 P ($\subseteq$ N {j})는 j번째 특징을 배제한 부분집합 일 수 있다. 상기 수학식3에서 v (.)는 가치함수이다.

[청구항 13] 제1항에 있어서,
상기 제2 이상치 점수는 상기 분석 대상 데이터와 정상 데이터를 비교하여 계산되는 것인, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 14] 제13항에 있어서,
상기 제2 이상치 점수는 상기 분석 대상 데이터에 포함된 특징 값에 대한 특징 중요도 값과 상기 정상 데이터에 포함된 특징 값에 대한 특징 중요도

값의 차이를 기반으로 계산된 것인, 적대적 공격에 강건한 이상 탐지 방법.

[청구항 15] 제14항에 있어서,
상기 제2 이상치 점수는 하기 수학식4를 이용하여 계산되는 것인, 적대적 공격에 강건한 이상 탐지 방법.
[수학식 4]

$$e_s(\Phi_i) = \frac{1}{d}\sum_{j=1}^{d}\left|\overline{\Phi}^{(j)} - \Phi_i^{(j)}\right|$$

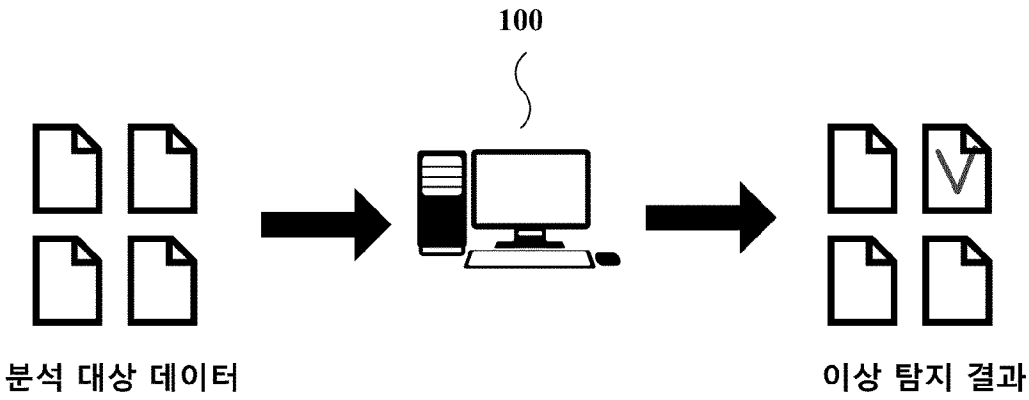
상기 수학식4에서 Φ_i는 i번째 분석 대상 데이터(x_i)에 대한 특징 중요도 값이다. 상기 수학식4에서 Φ_i는 $[\Phi^{(1)}{}_1, \Phi^{(2)}{}_2, \Phi^{(3)}{}_3, \ldots\Phi^{(d)}{}_i]$ 일 수 있다. 상기 수학식4에서 $\Phi^{(j)}{}_i$ 는 i번째 분석 대상 데이터의 j번째 특징 값에 대한 특징 중요도 값이다. 상기 수학식4에서 $\overline{\Phi}^{(j)}$는 정상 데이터의 j번째 특징 값에 대한 특징 중요도 평균 값이다. 상기 수학식4에서 $e_s(\Phi_i)$는 i번째 분석 대상 데이터에 대한 제2 이상치 점수이다. 상기 수학식4에서 d는 특징 중요도 값의 차원수로, 상기 분석 대상 데이터의 차원수와 같다.

[청구항 16] 제1항에 있어서,
상기 후탐지 하는 것은 상기 제2 이상치 점수를 제3 기준값과 비교하여 수행되는 과정을 포함하는, 적대적 공격에 강건한 이상 탐지 방법.

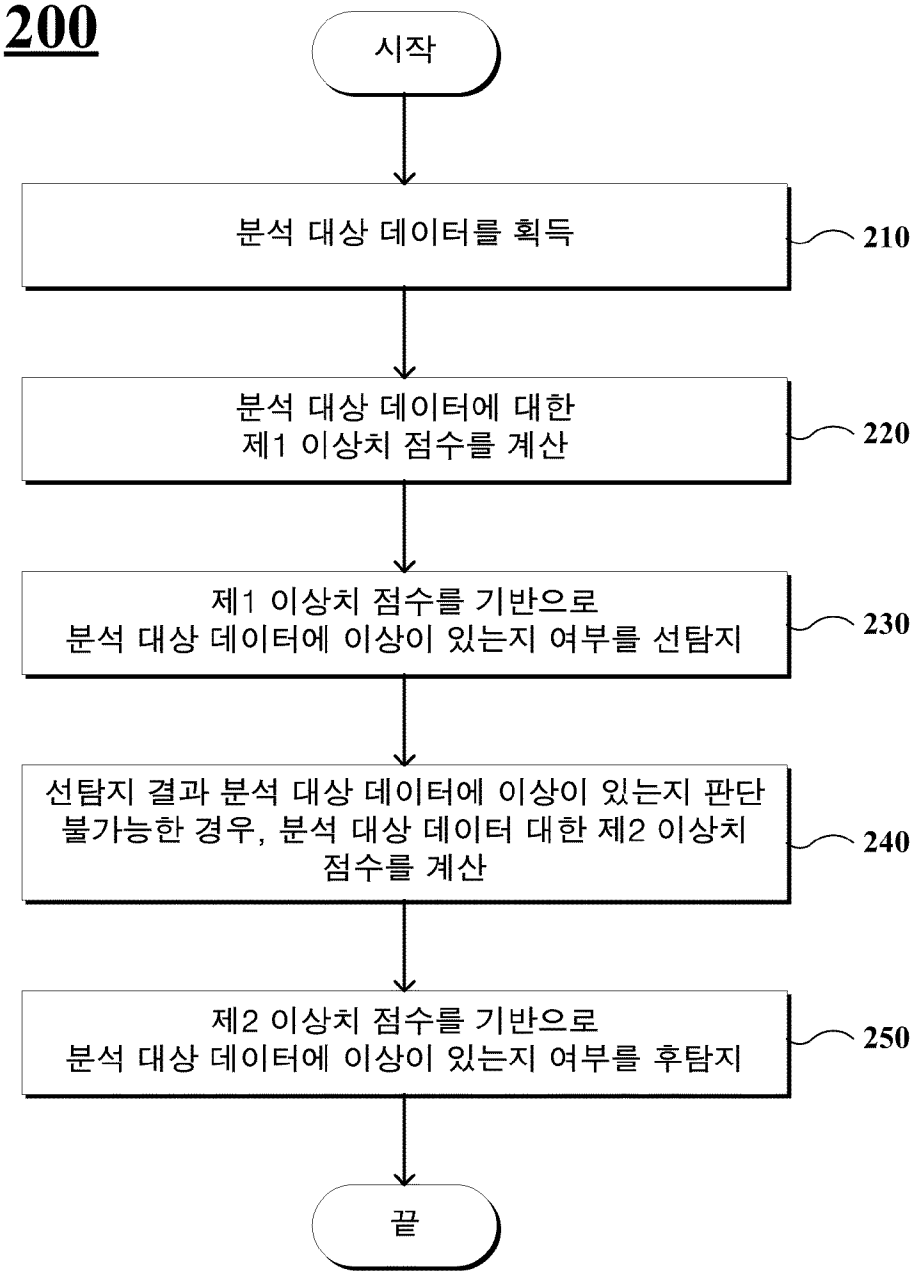
[청구항 17] 분석 대상 데이터를 입력 받는 입력장치;
상기 분석 대상 데이터에 대한 제1 이상치 점수를 계산하고, 상기 제1 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 선탐지 하고, 상기 선탐지 결과 상기 분석 대상 데이터에 이상이 있는지 판단 불가능한 경우, 상기 이상 탐지 장치가 상기 분석 대상 데이터에 대한 제2 이상치 점수를 계산하고 상기 제2 이상치 점수를 기반으로 상기 분석 대상 데이터에 이상이 있는지 여부를 후탐지 하는 연산장치; 및
상기 분석 대상 데이터를 저장하는 저장장치;를 포함하는, 이상 탐지 장치.

[청구항 18] 제1항에 기재된 적대적 공격에 강건한 이상 탐지 방법을 실행하기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

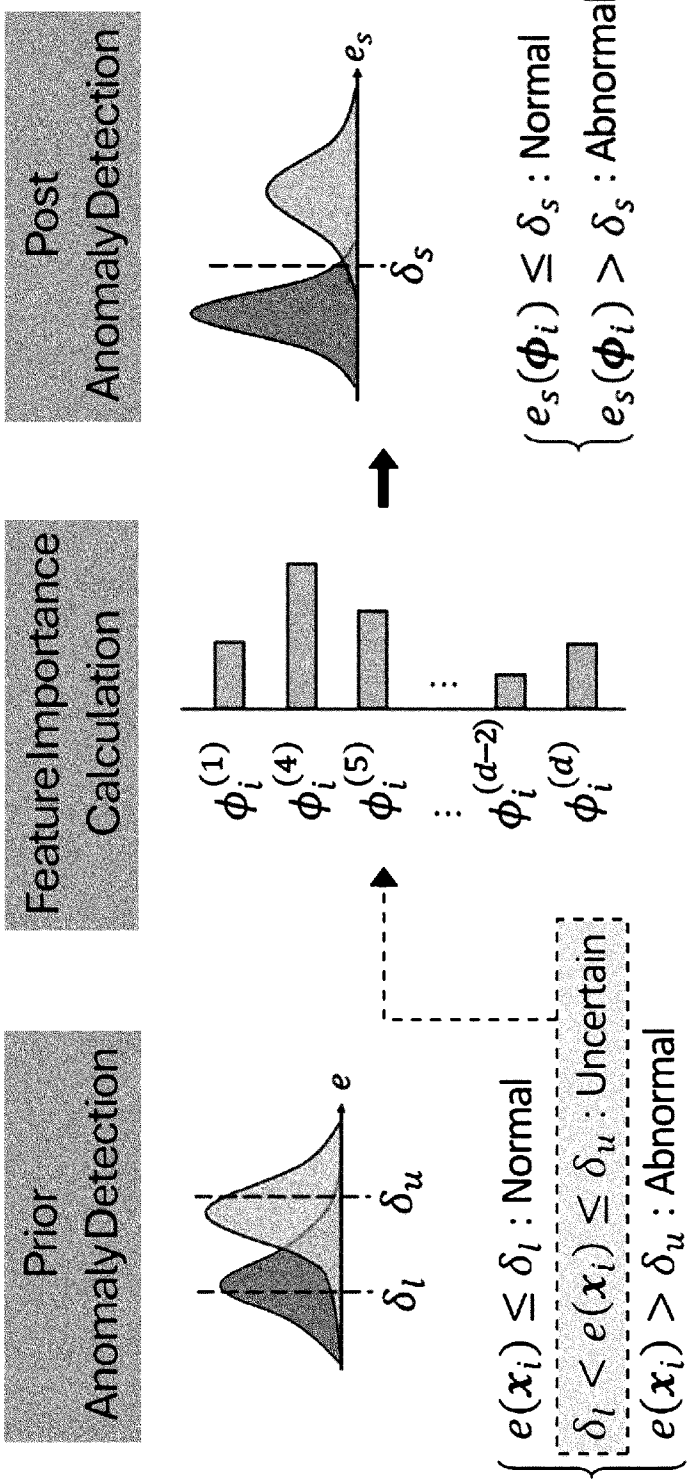
[도1]


100
분석 대상 데이터
이상 탐지 결과

[도2]


200
시작
분석 대상 데이터를 획득
210
분석 대상 데이터에 대한
제1 이상치 점수를 계산
220
제1 이상치 점수를 기반으로
분석 대상 데이터에 이상이 있는지 여부를 선탐지
230
선탐지 결과 분석 대상 데이터에 이상이 있는지 판단 불가능한 경우, 분석 대상 데이터 대한 제2 이상치 점수를 계산
240
제2 이상치 점수를 기반으로
분석 대상 데이터에 이상이 있는지 여부를 후탐지
250
끝

[도3]

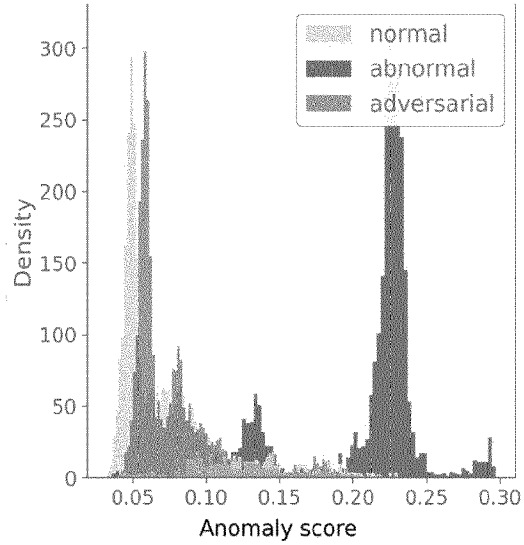

Prior Anomaly Detection
Feature Importance Calculation
Post Anomaly Detection
e
δ_l
δ_u
$e(x_i) \leq \delta_l$: Normal
$\delta_l < e(x_i) \leq \delta_u$: Uncertain
$e(x_i) > \delta_u$: Abnormal
$\phi_i^{(1)}$ $\phi_i^{(4)}$ $\phi_i^{(5)}$ ... $\phi_i^{(d-2)}$ $\phi_i^{(d)}$
e_s
δ_s
$e_s(\phi_i) \leq \delta_s$: Normal
$e_s(\phi_i) > \delta_s$: Abnormal

[도4]

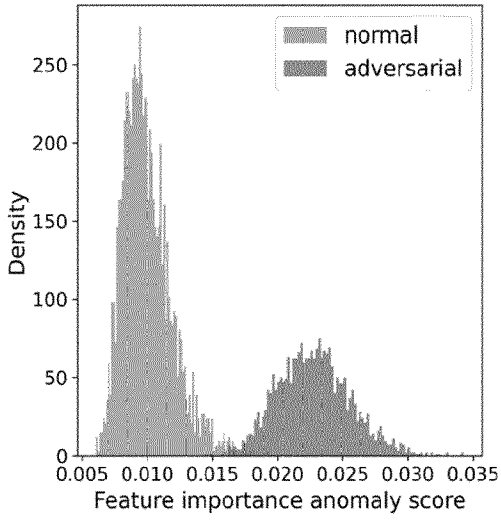
$x_1^{(1)}$	$x_1^{(2)}$	⋯	$x_1^{(d)}$
$x_2^{(1)}$	$x_2^{(2)}$	⋯	$x_2^{(d)}$
⋮	⋮	⋱	⋮
$x_n^{(1)}$	$x_n^{(2)}$	⋯	$x_n^{(d)}$

$f^{(1)}$

[도5]

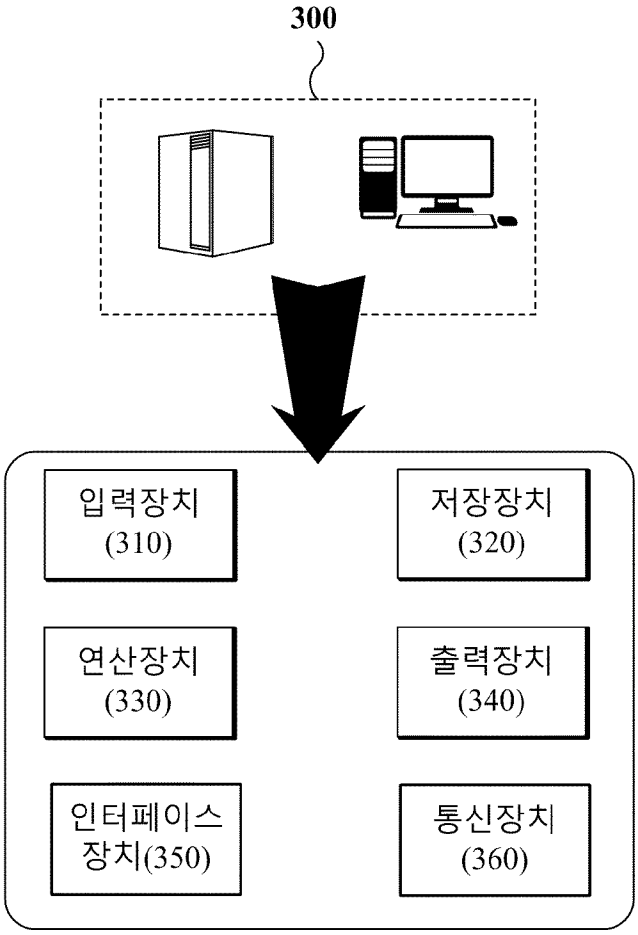


(a)



(b)

[도6]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2024/012559

A. CLASSIFICATION OF SUBJECT MATTER

H04L 9/40(2022.01)i; **G06N 3/0455**(2023.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L 9/40(2022.01); G06N 3/04(2006.01); G06N 3/08(2006.01); G06V 10/764(2022.01); G06V 10/82(2022.01); G10L 15/10(2006.01); G10L 15/16(2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above
Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & keywords: 적대적 공격(Adversarial Attack), 이상탐지(anomaly detection), 강건성(robustness), 특징(feature), 오토인코더(autoencoder), 새플리값(shapley value)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KR 10-2023-0018310 A (AGENCY FOR DEFENSE DEVELOPMENT) 07 February 2023 (2023-02-07) See claims 1-3.	1-4,17-18
Y		5,10-11,13-14,16
A		6-9,12,15
Y	KR 10-2023-0014248 A (KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY) 30 January 2023 (2023-01-30) See claims 1-2.	5,10-11,13-14,16
A	JP 2023-518331 A (MITSUBISHI ELECTRIC CORP.) 28 April 2023 (2023-04-28) See paragraphs [0096]-[0106]; and figures 2-3.	1-18
A	JP 2022-080285 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 27 May 2022 (2022-05-27) See claims 1-5; and figure 5.	1-18

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 November 2024	**29 November 2024**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2019-0136893 A (CARNEGIE MELLON UNIVERSITY et al.) 10 December 2019 (2019-12-10) See paragraphs [0123]-[0128]; and figure 3.	1-18
X	김민주 등. 적대적 공격에 강건한 특징 중요도 기반 계층적 네트워크 이상탐지 방법. The 4th Korea Artificial Intelligence Conference. 20 September 2023 (KIM, Minju et al. Robust Feature Importance-based Hierarchical Network Anomaly Detection Method against Adversarial Attacks). [Retrieved on 13 November 2024]. Retrieved from <URL: https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE11554740>. See entire document. * This document is a document declaring 'exceptions to lack of novelty' by the applicant.	1-18

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2024/012559

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR 10-2023-0018310 A	07 February 2023	None	
KR 10-2023-0014248 A	30 January 2023	KR 10-2592120 B1	20 October 2023
JP 2023-518331 A	28 April 2023	CN 115443463 A	06 December 2022
		EP 3935544 A1	12 January 2022
		EP 3935544 B1	10 August 2022
		JP 7378659 B2	13 November 2023
		US 11462211 B2	04 October 2022
		US 2021-0319784 A1	14 October 2021
		WO 2021-205746 A1	14 October 2021
JP 2022-080285 A	27 May 2022	CN 114519185 A	20 May 2022
		US 12045713 B2	23 July 2024
		US 2022-0156563 A1	19 May 2022
KR 10-2019-0136893 A	10 December 2019	CN 110554602 A	10 December 2019
		CN 110554602 B	01 October 2024
		EP 3576021 A1	04 December 2019
		EP 3576021 B1	30 October 2024
		US 11386328 B2	12 July 2022
		US 11676025 B2	13 June 2023
		US 2019-0370660 A1	05 December 2019
		US 2020-0026996 A1	23 January 2020

국 제 조 사 보 고 서

국제출원번호
PCT/KR2024/012559

A. 발명이 속하는 기술분류(국제특허분류(IPC))

H04L 9/40(2022.01)i; **G06N 3/0455**(2023.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

H04L 9/40(2022.01); G06N 3/04(2006.01); G06N 3/08(2006.01); G06V 10/764(2022.01); G06V 10/82(2022.01); G10L 15/10(2006.01); G10L 15/16(2006.01)

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 적대적 공격(Adversarial Attack), 이상탐지(anomaly detection), 강건성(robustness), 특징(feature), 오토인코더(autoencoder), 섀플리값(shapley value)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	KR 10-2023-0018310 A (국방과학연구소) 2023.02.07 청구항 1-3	1-4,17-18
Y		5,10-11,13-14,16
A		6-9,12,15
Y	KR 10-2023-0014248 A (한국과학기술원) 2023.01.30 청구항 1-2	5,10-11,13-14,16
A	JP 2023-518331 A (MITSUBISHI ELECTRIC CORP.) 2023.04.28 단락 [0096]-[0106]; 및 도면 2-3	1-18
A	JP 2022-080285 A (INTERNATIONAL BUSINESS MASCHINES CORPORATION) 2022.05.27 청구항 1-5; 및 도면 5	1-18
A	KR 10-2019-0136893 A (카네기 멜론 유니버시티 등) 2019.12.10 단락 [0123]-[0128]; 및 도면 3	1-18

☑ 추가 문헌이 C(계속)에 기재되어 있습니다.

☑ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

"A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌
"D" 본 국제출원에서 출원인이 인용한 문헌
"E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌
"L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌
"O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌
"P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌
"T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌
"X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.
"Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.
"&" 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일	국제조사보고서 발송일
2024년11월29일(29.11.2024)	**2024년11월29일(29.11.2024)**

ISA/KR의 명칭 및 우편주소

대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)

팩스 번호 **+82-42-481-8578**

심사관

양정록

전화번호 **+82-42-481-5709**

국 제 조 사 보 고 서

국제출원번호
PCT/KR2024/012559

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	김민주 등, `적대적 공격에 강건한 특징 중요도 기반 계층적 네트워크 이상탐지 방법`, The 4th Korea Artificial Intelligence Conference, 2023.09.20, [검색일: 2024.11.13], 출처: <URL: https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE11554740> 전문 * 본 문헌은 출원인에 의해 `신규성 상실 예외`가 선언된 문헌임.	1-18

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2024/012559

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2023-0018310 A	2023/02/07	없음	
KR 10-2023-0014248 A	2023/01/30	KR 10-2592120 B1	2023/10/20
JP 2023-518331 A	2023/04/28	CN 115443463 A	2022/12/06
		EP 3935544 A1	2022/01/12
		EP 3935544 B1	2022/08/10
		JP 7378659 B2	2023/11/13
		US 11462211 B2	2022/10/04
		US 2021-0319784 A1	2021/10/14
		WO 2021-205746 A1	2021/10/14
JP 2022-080285 A	2022/05/27	CN 114519185 A	2022/05/20
		US 12045713 B2	2024/07/23
		US 2022-0156563 A1	2022/05/19
KR 10-2019-0136893 A	2019/12/10	CN 110554602 A	2019/12/10
		CN 110554602 B	2024/10/01
		EP 3576021 A1	2019/12/04
		EP 3576021 B1	2024/10/30
		US 11386328 B2	2022/07/12
		US 11676025 B2	2023/06/13
		US 2019-0370660 A1	2019/12/05
		US 2020-0026996 A1	2020/01/23