Journal of the Korea Academia-Industrial cooperation Society
Vol. 23, No. 1 pp. 24-32, 2022

https://doi.org/10.5762/KAIS.2022.23.1.24
ISSN 1975-4701 / eISSN 2288-4688

딥러닝을 활용한 산업제어 시스템에 대한 이상징후 탐지

황종배[1], 장세창[1], 장재원[1], 김기현[2], 하재철[3*]
[1]호서대학교 정보보호학과, [2](주) 앤앤에스피, [3]호서대학교 컴퓨터공학부

Anomaly detection on ICS(Industrial Control System) using deep learning method

Jong-Bae Hwang[1], Se-Chang Jang[1], Jae-Won Jang[1], Ki-Hyun Kim[2], Jae-Cheol Ha[3*]
[1]Department of Information Security, Hoseo University
[2]NNSP
[3]Division of Computer Engineering, Hoseo University

요 약 4차 산업혁명 시대가 도래하면서 스마트 공장을 비롯한 산업제어 시스템에서의 제어 공정과 관련한 사이버 보안이 매우 중요해지고 있다. 본 논문에서는 사이버 공격으로 산업제어 시스템에서 발생할 수 있는 이상징후를 딥러닝 기법을 이용하여 탐지하는 모델을 개발하였다. 이상징후 탐지를 위해 수자원 시스템에 대한 공개 데이터 셋인 SWaT(Secure Water Treatment)를 사용하였다. 또한, 산업제어 시스템에서 발생하는 신호가 시계열 데이터인 점을 고려하여 RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory), GRU(Gated Recurrent Unit) 딥러닝 모델을 개발하였다. 이상징후 탐지 성능을 평가하기 위해서는 기존 정밀도(Precision)와 재현율(Recall) 평가 지표뿐만 아니라 이를 보완한 RP(Range-based Precision), RR(Range-based Recall), TaP(Time-Series Aware Precision), TaR(Time-Series Aware Recall)을 구현하여 비교하였다. 이상징후 탐지 실험 결과, 기존 모델과 비교하여 TaR 관련 지표에서는 LSTM이, TaP 관련 지표에서는 GRU가 우수한 성능을 나타냄을 확인하였다.

Abstract With the advent of the era of the 4th Industrial Revolution, cyber security related to control processes in industrial control systems, including smart factories, is becoming very important. In this paper, we developed deep learning models to detect abnormal signatures that may occur in the industrial control system due to cyber-attacks. Our anomaly detection research adopted the SWaT (Secure Water Treatment) data set used in process simulation of the water resource systems. Considering that the signals generated by the industrial control system are time-series data, we developed RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit) deep learning models. To evaluate the anomaly detection performance of these models, we implemented existing Precision and Recall evaluation indicators as well as RP (Range-based Precision), RR (Range-based Recall), TaP (Time-Series Aware Precision), and TaR (Time-Series Aware Recall). From the results of the anomaly detection experiment, it was confirmed that the RNN model in the TaR shows excellent performance compared to the existing model, and the GRU model is supreme in the TaP.

Keywords : Industrial Control System, Anomaly Detection, Deep Learning, Recurrent Neural Network, TaP, TaR

이 논문은 산업통상자원부 '산업혁신인재성장지원사업'의 재원으로 한국산업기술진흥원(KIAT)의 지원을 받아 수행된 연구임.
(2021년 차세대 디스플레이 공정・장비・소재 전문인력 양성사업, 과제번호 : P0012453)
*Corresponding Author : Jae-Cheol Ha(Hoseo Univ.)
email: jcha@hoseo.edu
Received September 27, 2021 Revised October 27, 2021
Accepted January 7, 2022 Published January 31, 2022

1. 서론

4차 산업혁명 시대가 도래하면서 스마트 공장을 비롯하여 산업제어 시스템에서의 제어 공정과 관련한 보안이 매우 중요해지고 있다. 산업제어 시스템은 전력, 가스, 상하수도, 원자력, 운송, 제조 등의 산업 현장을 모니터링하고 제어하는데 사용되는 시스템을 의미한다. 산업제어 시스템은 물리적 장치의 공정 상태를 계측, 모니터링하고 장치의 동작을 직접 제어하기도 한다.

과거에는 국가 기반 시설, 국방 관련 시설 및 산업기반시설 등이 폐쇄망으로 운영되어 단순한 조작 실수나 내부망 사용자의 공격과 같은 피해 사례가 대부분이었으며 그 오동작의 피해는 크지 않았다. 그러나 4차 산업혁명 시대가 도래하면서 공정 부분의 자동화가 진행되었고 이로 인해 물리적 공격뿐만 아니라 원격에서의 사이버 공격도 가능해졌다. 최근 사이버 공격에 의한 침해사고는 전력, 수자원, 교통, 원자력, 국방, 스마트 공장 등 다양한 분야에 걸쳐 발생하고 있다.

산업제어 시스템이 사이버 공격을 받게 되면 물리적 피해를 유발하게 되며 특히 사람의 안전과 환경에 영향을 끼칠 수 있다. 따라서 스마트 팩토리와 같이 네트워크와 물리적 시설이 공존하는 사이버 물리 시스템(Cyber Physical System)에 대한 안전성과 관련한 보안 문제가 대두되고 있다. 그러나 산업제어 시스템에 관한 보안 연구는 관련 시설에 접근하기도 어렵고 자체 보안상의 이유로 공정 과정의 데이터를 외부에 공개하지 않아 실질적인 공격 실험과 대응 기법의 검증이 쉽지는 않다.

본 논문에서는 산업제어 시스템에서 발생할 수 있는 사이버 공격상의 이상징후를 딥러닝 기법을 이용하여 탐지하고자 한다. 이상징후 탐지 실험에서는 2015년 SUTD(Singapore University of Technology and Design)의 iTrust 연구소에서 공개한 수자원 시스템에 대한 공개 데이터 셋인 SWaT를 사용하였다. 또한, 산업제어 시스템에서의 신호가 대부분 시계열 데이터인 점을 고려하여 순환 신경망을 이용한 딥러닝 모델을 개발하였다.

본 논문의 구성은 다음과 같다. 2장에서는 순환 신경망인 RNN, LSTM, GRU와 실험에 사용한 SWaT 테스트베드 데이터 셋을 설명한다. 3장에서는 딥러닝 모델의 성능을 비교하기 위한 평가 기법을 설명한다. 4장에서는 이상징후 탐지 실험 과정과 결과를 분석하며, 5장에서 결론을 맺는다.

2. 딥러닝 기법과 학습 데이터

2.1 순환 신경망

순환 신경망 RNN은 인공 신경망의 한 종류로서 유닛 간의 연결이 순환적 구조를 갖는 특징을 가지고 있다. RNN은 순방향 신경망인 다층 퍼셉트론, 합성곱 신경망과 달리 내부의 메모리를 이용해 시퀀스 형태의 입력을 처리할 수 있다[1,2]. 이러한 순환 구조는 신경망 내부에 상태를 저장할 수 있어 음성 인식과 같은 시변적 동적 특징을 모델링하는데 효과적이다.

기본적인 RNN 알고리즘의 구조는 Fig. 1과 같이 표현할 수 있다. Fig. 1의 왼쪽을 보면 입력 값 x_t를 받아 출력 값 y_t를 만들고, 해당 출력을 다시 입력으로 받아 사용하는 순환적 구조를 보인다. 그림의 오른쪽은 각 타임 스텝(time step) t마다 펼쳐서 타임 스텝별 입력, 출력 그리고 가중치를 나타낸 것이다. RNN의 각 계층에서 수행하는 연산은 Eq. (1)과 같다.

$$h_t = \tanh(h_{t-1} W_h + x_t W_x + b) \tag{1}$$

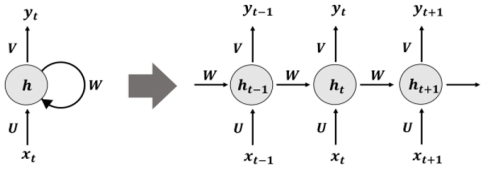

Fig. 1. Structure of RNN algorithm

그러나 RNN은 시퀀스 데이터의 장기 의존 관계성(long-term dependency)을 학습하기에는 한계가 있다. 즉, 신경망 하나를 통과할 때마다 기울기 값이 조금씩 작아져 신경망의 길이가 길어짐에 따라 이전 타임 스텝까지 역전파 되기 전에 0이 되어 소멸하게 되는 기울기 소실(gradient vanishing)이 발생하거나, 역으로 기울기가 너무 커지는 기울기 폭발(gradient exploding) 문제가 일어나기 때문이다.

이를 해결하기 위해서 RNN 계층의 신경망 구성에 특수 목적의 게이트를 추가하는 방법을 사용하는데 대표적으로 LSTM 알고리즘과 GRU 알고리즘이 있다. LSTM 알고리즘은 Fig. 2와 같이 은닉층에 메모리 셀, 입력 게이트, 망각 게이트, 출력 게이트를 추가한 구조이다.

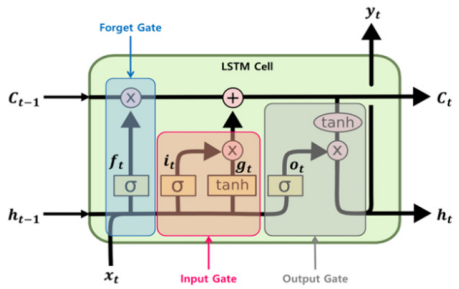

Fig. 2. Structure of LSTM algorithm

추가된 4가지 중 망각 게이트가 중요하다. 망각 게이트는 LSTM 알고리즘의 첫 단계로 어떤 기억 정보를 버릴 것인지 정한다. LSTM 알고리즘은 불필요한 기억 정보를 지우는 과정을 통해 기울기 값이 급격하게 사라지거나 증가하는 문제를 방지할 수 있다[3].

GRU 알고리즘은 2014년에 처음으로 제안되었으며, 여기에서는 출력 게이트가 존재하지 않는다. LSTM 알고리즘에서 두 상태 벡터인 c_t와 h_t가 하나의 벡터 h_t로 합쳐졌으며 Z_t는 망각 게이트와 입력 게이트를 모두 제어한다. LSTM 알고리즘보다 적은 수의 매개변수를 가짐에도 불구하고 음성 인식 분야에서 LSTM 알고리즘과 유사한 성능을 가진다[4].

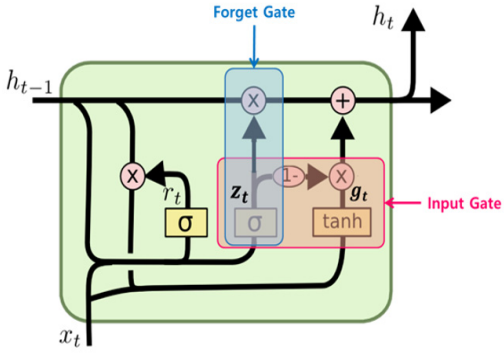

Fig. 3. Structure of GRU algorithm

2.2 SWaT 테스트베드

2015년, J. Goh 등은 SUTD의 iTrust 연구소에서 수처리 시스템에 대한 SWaT 테스트베드를 구성하고 총 36가지의 공격을 수행한 데이터 셋을 공개하였다[5]. SWaT 테스트베드의 각 프로세스별 구성도는 Fig. 4와 같은데 총 6단계로 구성되어 있다.

Fig. 4에서 P1은 원수를 탱크에 공급하여 저장하는 단계이다. P2는 원수의 질을 조정하며 P3에서 여과막을 이용하여 불순물을 제거한다. P4는 자외선을 이용하여 잔여 염소를 제거하며, P5는 P4에서 전달받은 물을 역삼투 시스템을 통해 불필요한 무기물을 걸러낸다. 마지막 프로세스인 P6에서 물 분배 시스템에서 최종적으로 생산된 물을 방류한다.

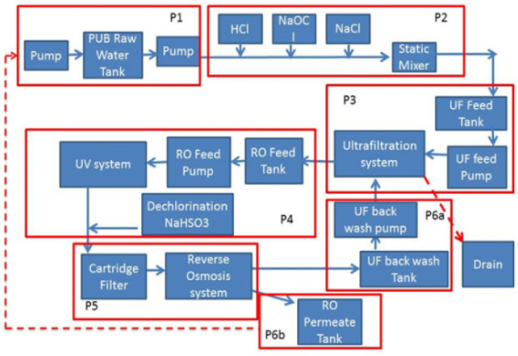

Fig. 4. SWaT architecture

SWaT 테스트베드의 데이터 셋에 포함된 공격은 크게 4가지로 분류할 수 있다. 먼저 SSSP(Single Stage Single Point) 공격 유형은 하나의 단계에 대하여 하나의 데이터를 조작한 공격이다. SSMP(Single Stage Multi Point) 공격 유형은 하나의 단계에 대하여 다수의 데이터를 조작한 공격이다. MSSP(Multi Stage Single Point)는 다수의 단계에 단일 데이터를 조작한 공격이며, MSMP(Multi Stage Multi Point)는 다수의 단계에 다수의 데이터를 조작한 공격이다. Table 1은 총 36가지 공격을 유형별로 정리한 것이다.

Table 1. Number of attacks per category

Attack Category	Number of attacks
SSSP	26
SSMP	4
MSSP	2
MSMP	4

3. 이상징후 탐지 성능 평가

3.1 정밀도와 재현율

연구자들은 오래전부터 기계학습 모델을 평가하기 위한 평가 지표로 정밀도와 재현율을 사용했다. 본 논문에

서도 설계한 딥러닝 모델을 평가하기 위해 기본적으로 정밀도와 재현율을 사용하였다.

정밀도는 예측을 정상으로 한 대상 중에 실제값이 정상으로 일치한 데이터의 비율이다. 재현율은 실제값이 정상인 대상 중에 예측값이 정상으로 일치한 데이터의 비율이다. 평가 기법인 정밀도와 재현율을 수식으로 나타낸 것이 Eq. (2)와 Eq. (3)이다. Eq. (2)와 (3)의 TP(True Positive)는 실제 참인데 분류 모델이 예측을 참이라고 판단된 경우이고, FP(False Positive)는 실제 거짓인데 분류 모델이 예측을 참이라고 판단한 경우이다. FN(False Negative)은 실제 참인데 예측이 거짓으로 판단한 경우이다.

$$\mathrm{Precision} = \frac{TP}{(TP+FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{3}$$

스팸메일 분류나 암 환자 분류와 같이 포인트 기반 데이터 분류 모델의 경우 위의 방식으로 평가할 수 있다. 그러나 사이버 공격에 의한 이상징후 탐지를 위해 시계열 데이터를 분석해 보면 긴 시간 동안 수행되는 공격도 있고, 비교적 짧은 시간 동안 수행되는 공격이 존재한다. 이처럼 특정 신호가 일정한 범위를 가지는 데이터를 사용하여 학습한 모델의 경우 기존 정밀도나 재현율과 같은 평가 기법을 사용하기에는 한계가 있다.

3.2 범위 기반 이상징후 탐지 능력 평가

그림 Fig. 5는 범위 기반 이상징후 탐지가 필요한 예를 나타낸 것이다. 이상징후가 포함된 데이터를 분석해 보면 Fig. 5의 $a1$ 과 같이 이상징후가 분포하고, 범위 기반 이상징후를 탐지하기 위한 학습모델이 Method 1, 2와 같이 예측을 하였다. 그림에서 보면 학습기법 Method 1은 $a1$을 정확하게 탐지하였다. 그러나 Method 2의 경우 $a1$을 정확하게 탐지하였다고 볼 수 없다. Method 2가 예측한 $p2$는 $a1$의 일부분을 정확하게 예측했으나 $p3$의 일부분은 이상징후가 아님에도 불구하고 이상징후라고 탐지하였다.

이와 같은 학습모델 Method 2에 대해 기존의 평가 방법인 재현율과 정밀도는 $p3$에 대해 잘못 예측한 부분을 제외시키는 문제점을 안고 있었다. 이와 같은 이유로 2018년 N. Tatbul 등은 포인트 기반 평가 방법의 단점을 보완하기 위해 범위 기반 정밀도와 범위 기반 재현율을 제안하였다[6].

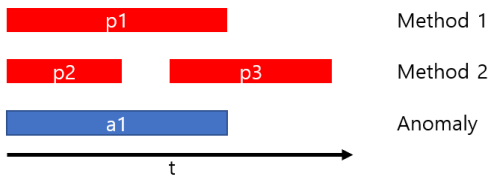

Fig. 5. Example of anomaly and predictions produced by two method

Table 2 는 범위 기반 정밀도, 재현율을 계산하는데 사용된 변수의 정의이다.

Table 2. Notation for range-based evaluation

Notation	Description
R, R_i	set of real anomaly ranges, the i^{th} real anomaly range
P, P_j	set of predicted anomaly ranges, the j^{th} predicted anomaly range
N, N_r, N_p	number of all points, number of real anomaly ranges, number of predicted anomaly ranges
α	relative weight of existence reward
$\gamma(), \omega(), \delta()$	overlap cardinality function, overlap size function, positional bias function

범위 기반 이상징후 탐지 방법에서 재현율 RR은 Eq. (4)와 같이 Eq. (5)에서 구한 재현율 값을 모두 더하여 N_r로 나눔으로써 계산하게 된다. 여기서 Eq. (5)는 i 번째 이상징후에 대하여 모델이 예측했는지 재현율을 구하는 식이다. Eq. (5)에서 α의 범위는 $0 \le \alpha \le 1$이다.

$$Recall_T(R,P) = \frac{\sum_{i=1}^{N_r}(Recall_T(R_i,P))}{N_r} \tag{4}$$

$$Recall_T(R_i,P) = \alpha \times ExistenceReward(R_i,P) + (1-\alpha) \times OverlapReward(R_i,P) \tag{5}$$

Eq. (5)에 나타낸 $ExistenceReward()$함수는 실제 범위와 예측한 범위 겹치는 범위가 있다면 1을, 겹치는 범위가 없다면 0을 반환하는 함수로서 Eq. (6)과 같이 정의하였다. 또한, Eq. (7)에 표기한 $CardinalityFactor()$함수는 R_i가 하나의 예측된 P_j와 겹치면 1을 반환하고, x개와 겹칠 경우 $1/x$을 반환하는 함수이다. 여기서 $\gamma(), \omega(), \delta()$의 범위는 각각 $0 \le \gamma() \le 1, 0 \le \omega() \le 1, \delta \ge 1$이다.

$$ExistenceReward(R_i,P) = \begin{cases} 1, If \sum_{j=1}^{N_p} | R_i \cap P_j | \geq 1 \\ 0, Otherwise \end{cases} \quad (6)$$

$$OverlapReward(R_i,P) = CardinalityFactor(R_i,P) \times \sum_{j=1}^{N_p} \omega(R_i, R_i \cap P_j, \delta) \quad (7)$$

$$CardinalityFactor(R_i,P) = \begin{cases} 1, If\ R_i\ overlaps\ with\ at most\ one\ P_j \in P \\ \gamma(R_i,P), Otherwise \end{cases} \quad (8)$$

범위 기반 이상징후 탐지 방법에서 정밀도 RP을 나타낸 식은 아래 Eq. (9)와 (10)과 같다. 여기서 사용된 $ExistenceReward()$, $OverlapReward()$함수는 Eq. (6)과 (7)에 나타낸 것과 같으며 R_i 대신 P_j를 인수로 사용하면 된다.

$$\mathrm{Precision}_T(R,P) = \frac{\sum_{j=1}^{N_P} \mathrm{Precision}_T(R,P_j)}{N_p} \quad (9)$$

$$\mathrm{Precision}_T(R,P_j) = \alpha \times ExistenceReward(R,P_j) + (1-\alpha) \times OverlapReward(R,P_j) \quad (10)$$

3.3 시계열 인지 이상징후 탐지 능력 평가

상기한 RR, RP는 범위 기반 평가 기법이지만 산업제어 시스템의 특성을 완전히 고려한 평가 방법은 아니다. 추가로 사이버 공격이 끝난 직후 공격의 영향으로 비정상적으로 구동하는 것처럼 탐지될 수 있는데 이에 대한 고려가 필요하다.

W. S. Hwang 등은 공격은 끝났지만, 공격의 여파로 비정상적으로 보이는 범위를 별도의 변수인 모호한 인스턴스를 설정하고 이를 고려한 새로운 평가 방법인 TaP와 TaR을 발표하였다[7]. RR과 RP의 경우 제어시스템에서 수집된 실제 데이트 셋에서 모호한 인스턴스를 간과하는 문제점이 있다. 즉, 공격이 실행된 이후 해당 공격의 영향이 제어시스템에 얼마나 오래 남아 있는지 추정하기 어렵다.

다음 Fig. 6은 $a1$이후의 $a'1$의 범위가 정상임에도 불구하고 이상징후로 탐지될 수 있다는 것을 보여주는 예제이다. TaP와 TaR에서는 다양한 이상징후를 탐지했을 때 높은 점수를 할당하고 $a'1$과 같이 모호한 인스턴스 범위를 탐지했을 때도 긍정적인 점수를 제공한다. 또한, Method 1의 $p1$과 같이 부분적으로 탐지했을 경우 부분적인 점수를 제공한다.

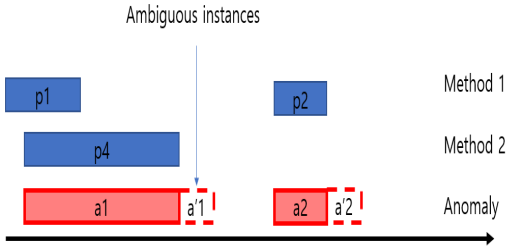

Fig. 6. Inaccurate evaluation of precision and recall in time-series data

TaP와 TaR에서 a는 이상징후 인스턴스 집합, $a = \{t, t+1, \cdots, t+l-1\}$이다. 여기서 t는 첫 번째 이상징후 인스턴스 순서를 의미하며, l은 길이로 인스턴스 수이다. A는 이상징후 집합으로, $A = \{a_1, a_2, \cdots, a_n\}$이다. 여기서 n은 이상징후 항목 수를 나타낸다. p는 의심되는 인스턴스 집합으로, $p = \{t', t'+1, \cdots, t'+l'-1\}$이다. P는 모델이 예측한 집합으로, $P = \{p_1, p_2, \cdots, p_m\}$이다. 여기서 m은 예측 수이다.

이상징후 a에 영향을 받은 모호한 인스턴스를 $a' = \{t+l, t+l+1, \cdots, t+l+\delta-1\}$로 나타내며 a이후의 δ개의 인스턴스를 포함한다. 모호한 인스턴스 집합은 $A' = \{a'_1, a'_2, \cdots, a'_n\}$으로 나타내며 n은 이상 항목 수이다.

시계열 인지 이상징후 탐지 평가 방법 중 TaR 역시 기본적으로는 범위 기반 평가 지표이며 다음 Eq. (11)과 같이 정의한다. 여기서 α는 TaR^d와 TaR^p의 비율을 제어하며 0과 1 사이 값을 가진다.

$$TaR = \alpha \times TaR^d + (1-\alpha) \times TaR^p \quad (11)$$

Eq. (12)는 TaR^d를 나타낸 것으로 여기서 $O()$는 중첩 점수를 계산하는 함수로 예측 p가 이상징후 a를 탐지할 가능성을 나타낸다.

$$TaR^d = \frac{|A^d(\theta)|}{|A|} \quad (12)$$
$$where\ A^d(\theta) = \{a | a \in A \text{ and } \frac{\sum_{p \in P} O(a,p)}{|a|} \geq \theta\}$$

$$O(a,p) = |a \cap p| + S(a',p) \quad (13)$$

다음 Eq. (14)에서 정의한 $S()$함수는 모호한 인스턴스 a'을 탐지한 p에 대한 부분점수를 부여하는 함수이다. 여기서 $t_{a'}$은 모호한 인스턴스 a'의 첫 번째 인덱스이다. 모호한 인스턴스 a'은 이상징후 a에서 멀어질수록 이상징후일 가능성이 작다. 이 개념을 반영하기 위해 시그모이드 함수(sigmoid function)의 역함수인 로짓 함수(logit function)를 사용하였다.

$$S(a',p) = \sum_{i \in (a' \cap p)} \frac{1}{1+e^{i'}} \quad (14)$$
$$where\ i' = -6 + \frac{12(i - t_{a'} - 1)}{\delta - 1}$$

$$TaR^p = \frac{1}{|A|} \times \sum_{a \in A} \min\left(1, \frac{\sum_{p \in P} O(a,p)}{|a|}\right) \quad (15)$$

시계열 인지 이상징후 탐지 평가 방법 중 TaP를 정의한 것이 Eq. (16)이다. 또한, 아래 Eq. (17)에 사용된 $O()$는 위의 Eq. (13)과 같다.

$$TaP = \alpha \times TaP^d + (1-\alpha) \times TaP^p \quad (16)$$

$$TaP^d = \frac{|P^c(\theta)|}{|P|} where\ P^c(\theta) = \quad (17)$$
$$\{p | p \in P \text{ and } \frac{\sum_{a \in A} O(a,p)}{|p|} \geq \theta\}$$

$$TaP^p = \frac{1}{|P|} \times \sum_{p \in P} \left(\frac{\sum_{a \in A} O(a,p)}{|p|}\right) \quad (18)$$

4. 이상징후 탐지 실험 및 비교분석

상기한 내용과 같이 산업제어 시스템에서 발생하는 신호는 시계열 데이터이다. 본 논문에서는 이러한 특성을 고려하여 RNN, LSTM, GRU 딥러닝 모델을 개발하였다. 또한, 이상징후 데이터는 범위 기반 데이터이므로 3.2절에서 설명한 TaP이나 TaR과 같은 범위 기반 평가 기법이 더 적합함을 확인하였다.따라서 산업제어 시스템에 적합한 딥러닝 모델과 평가 방법을 찾기 위한 실험을 진행하였다.

4.1 실험 과정

SWaT 데이터 셋에는 2개의 정상 파일과 1개의 공격 파일이 있다. 데이터 셋은 총 11일 동안 중지 없이 캡처한 데이터로 7일은 정상 동작 데이터이며, 4일은 공격이 포함된 데이터이다. 테스트베드가 가동 후 물탱크가 완전히 비어 있는 상태에서 원수를 공급하여 시스템의 정상 동작까지 30분이 소요된다.

SWaT 데이터 셋 중에서 2개의 정상 파일은 version0, version1으로 구분되며, version0는 해당 30분이 포함된 데이터 셋이며, version1은 해당 30분이 제거된 데이터 셋이다. 해당 데이터 셋은 51개의 센서와 액추에이터에 대하여 매초 기록된, 총 946,722개의 정보로 구성되어 있다.

SWaT 데이터에 대한 이상징후 실험 과정을 나타낸 것이 Fig. 7이다. SWaT 데이터 셋의 경우 Level 0에 있는 센서 및 액추에이터의 값으로 구성되어 있다. 따라서 데이터의 범위가 서로 달라 Normalization 단계에서 Min-MaxScaler를 사용하여 모든 데이터를 0~1 사이의 값으로 만들게 된다. 각 센서 및 액추에이터의 데이터 분포를 확인하여 학습에 불필요하다고 판단되는 P102, P201, P202, P204, P206, P401, P403, P404, P502, P601, P603 총 11개의 특성(feature)을 제거하였다.

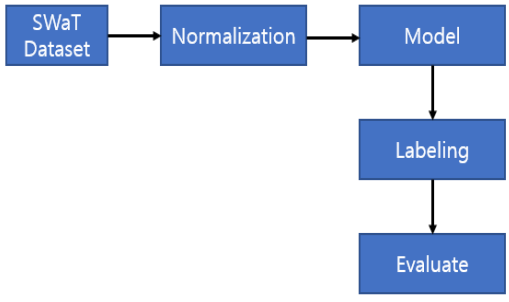

Fig. 7. Anomaly detection process for SWaT data set

SWaT 데이터 셋은 시계열 데이터이기 때문에 해당 실험에서는 순환 신경망인 RNN, LSTM, GRU 모델을 설계하여 구현하였다. 순환 신경망의 경우 학습단계에서 사용하는 데이터 셋은 시퀀스 데이터이므로 해당이 데이터를 특정 타임 스텝씩 저장하여 새로운 데이터 셋을 만들어야 한다. 본 논문에서는 타임 스텝을 30, 45, 60, 75, 90, 100으로 설정하여 결과를 비교 분석하였다.

새로 구축한 데이터 셋을 사용하여 RNN, LSTM, GRU 모델을 학습한다. 3개의 모델 모두 옵티마이저(optimizer)는 Adam을 사용하였으며, 손실 함수는 MSE(Mean Squared Error)를 사용하였다. 해당 모델이 예측하는 것은 정상, 비정상을 예측하는 것이 아니다. 학습에 사용되는 입력의 다음 타임 스텝을 예측한다. 따라서 모델이 예측한 것을 바탕으로 정상, 비정상 레이블링(labeling)을 진행한다. 그림 Fig. 7의 Labling 과정을 통해 만든 예측한 레이블과 실제 결과 레이블을 Evaluate 과정에서 평가한다.

해당 과정에서 사용한 평가 기법은 정밀도, 재현율, RP, RR, TaR, TaP를 사용하였다. RP, RR을 계산할 때 $\alpha = 0.5$로 설정했으며 δ는 각각 Flat bias와 Front-End bias를 사용하였다. TaP, TaR의 경우 $\alpha = 0.5$, $\delta = 180$으로 설정했으며 TaP를 계산할 때 θ는 0.7, TaR을 계산할 때 θ는 0.1로 설정하여 계산하였다.

4.2 이상징후 탐지 분석 결과

본 논문에서는 산업제어 시스템에서 순환 신경망 모델을 사용하여 이상징후를 탐지할 때 어떤 모델이 적합한지 비교 분석하고자 한다. 또한, 기존 평가 기법인 재현율, 정밀도는 범위 기반 이상징후를 평가하는데 부적절하다고 판단하여 최근 제안된 RP, RR, TaR, TaP를 사용하여 모델을 평가하고자 한다. 실험용 장비의 제원은 다음과 같으며 개발 언어는 Python 3.8을 사용하였다.

CPU : Intel(R) i5-8400 (2.80GHz)

RAM : 32.0GB

GPU : NVIDIA GeForce RTX 2060(6.0GB)

본 논문에서는 3개의 모델의 성능을 비교할 뿐 아니라 타임 스텝에 따라 어떻게 다른지 비교분석 하였다. 타입 스텝을 30, 45, 60, 75, 90, 100으로 설정하여 3가지 모델 전부 비교하였다. 정밀도와 재현율은 서로 보완적인 지표로 분류 모델의 성능을 평가하는데 높은 수치를 얻는 것이 가장 좋은 성능을 의미한다. 이상징후 탐지에서는 재현율이 정밀도보다 상대적으로 중요한 지표다. 따라서 본 논문에서는 재현율이 우선적으로 높으면서 정밀도가 높은 타임 스텝과 딥러닝 모델을 찾고자 한다.

Table 3는 RNN 모델의 결과이다. RNN 모델의 경우 SWaT 데이터 셋을 사용하여 이상징후를 탐지하는 모델로 적합하지 않음을 알 수 있다. 타임 스텝이 30, 75일 때를 제외하면 모두 같은 값이 나왔다. RNN 모델의 경우 6가지 평가 기법을 비교하였을 때 타입 스텝이 75일 때 가장 좋은 성능을 보임을 확인할 수 있다.

Table 3. RNN based anomaly detection results

	RNN					
time step	Precision	RP	TaP	Recall	RR	TaR
30	0.16	0.06	0.10	0.93	0.08	0.54
45	0.12	0.50	0.05	1.00	1.00	0.5
60	0.12	0.50	0.05	1.00	1.00	0.5
75	0.16	0.06	0.08	0.94	0.76	0.62
90	0.12	0.50	0.05	1.00	1.00	0.50
100	0.12	0.50	0.05	1.00	1.00	0.50

Table 4은 LSTM 모델의 결과이다. LSTM 모델의 경우 모든 타입 스텝의 결과가 달리 나왔다. 그 중에서 타임 스텝이 60일 때 가장 좋은 성능을 보였다. 정밀도의 경우 타임 스텝이 커질수록 낮은 결과가 나왔으며 재현율의 경우 점점 높아지는 결과를 보였다. TaP의 결과는 비슷한 성능을 보였지만 TaR의 경우 타임 스텝에 따라 비교적 큰 차이가 나는 것을 확인할 수 있다.

Table 4. LSTM based anomaly detection results

	LSTM					
time step	Precision	RP	TaP	Recall	RR	TaR
30	0.88	0.08	0.21	0.66	0.46	0.28
45	0.81	0.07	0.19	0.67	0.47	0.37
60	0.33	0.07	0.10	0.75	0.58	0.65
75	0.42	0.06	0.10	0.73	0.52	0.51
90	0.22	0.08	0.09	0.76	0.60	0.66
100	0.39	0.06	0.12	0.76	0.57	0.57

Table 5는 GRU 모델의 결과이다. 타임 스텝이 60일 때 다른 모델과 달리 TaP 결과가 TaR 결과보다 높게 나오는 것을 확인할 수 있으며 가장 좋은 성능을 보인다. GRU를 제외한 2가지 모델은 TaP, TaR의 값이 많이 차이가 났다. 이를 통해 GRU가 전체적으로 가장 성능이 좋은 이상징후 탐지 모델임을 확인할 수 있다.

Table 5. GRU based anomaly detection results

	GRU					
time step	Precision	RP	TaP	Recall	RR	TaR
30	0.57	0.14	0.30	0.04	0.38	0.22
45	0.79	0.06	0.18	0.68	0.45	0.35
60	0.95	0.12	0.47	0.64	0.39	0.32
75	0.56	0.14	0.33	0.06	0.44	0.28
90	0.55	0.13	0.32	0.07	0.44	0.37
100	0.42	0.08	0.19	0.08	0.47	0.38

위의 Table 3, 4, 5를 통해 설계한 3가지 모델 중 범위 기반 이상징후를 탐지할 때 RNN 모델은 적합하지 않다는 것을 확인할 수 있다. 또한, 타임 스텝에 따라 정밀도와 재현율이 다르기에 적절한 타임 스텝을 설정하여 학습에 사용할 데이터 셋을 구축해야 함을 알 수 있다.

Table 6의 경우 기존 기계학습 모델 연구와 이번 실험을 통해 확인한 결과를 비교한 것이다. 먼저 iForest 모델의 경우 기존 평가 방법인 정밀도와 재현율이 다른 모델과 비슷한 수치를 보이지만 범위 기반 평가 방법은 가장 낮은 수치를 보인다. OCSVM 모델은 iForest 모델보다 TaP, TaR값이 높은 것을 확인할 수 있다.

본 논문에서 설계한 RNN, LSTM, GRU 3가지 모두 iForest 모델보다 좋은 성능을 보이는 것을 확인할 수 있다. OCSVM 모델보다 RNN, LSTM 모델의 TaR값이 높지만 TaP 값은 낮으며 OCSVM 모델처럼 두 값의 차이가 크다. 하지만 제안한 모델 중 GRU는 TaR값은 비교적 낮지만 TaP 값이 다른 4가지 모델보다 높으며 TaP, TaR 두 값이 높은 것을 확인할 수 있다.

실험 결과, 본 논문에서 제안한 GRU 모델이 기존 평가 기법인 정밀도 지표에서 0.95로 가장 높았으며, 재현율 지표에서 RNN 모델이 0.94로 기존의 OCSVM, iForest 모델보다 높은 결과를 보였다. 또한, 범위 기반 평가 기법인 TaP 지표에서 기존의 연구된 OCSVM 모델보다 GRU 모델이 0.47로 높은 수치를 보였다. TaR 지표 역시 본 논문에서 제안한 LSTM 모델이 0.65로 가장 높은 수치를 보였다.

결론적으로 이상징후 탐지 실험 결과, 기존 모델과 비교하여 TaR 관련 지표에서는 LSTM이, TaP 관련 지표에서는 GRU가 우수한 성능을 나타냄을 확인하였다. 이상징후 탐지 모델을 설계할 때 재현율, 정밀도 모두 중요하며 두 지표의 차이가 작을수록 좋은 모델이다. 본 논문에서 구현한 모델 중 GRU 모델이 TaR의 값은 LSTM 모델과 비교했을 때 낮지만 두 지표의 차이가 크지 않음으로 가장 적합한 모델이라고 할 수 있다.

Table 6. Detection Results using SWaT dataset

	Precision	RP	TaP	Recall	RR	TaR
OCSVM[8]	0.17	0.14	0.17	0.85	0.61	0.55
iForest[9]	0.30	0.04	0.05	0.74	0.52	0.04
RNN	0.16	0.06	0.08	0.94	0.76	0.62
LSTM	0.33	0.07	0.10	0.75	0.58	0.65
GRU	0.95	0.12	0.47	0.64	0.39	0.32

5. 결론

산업제어 시스템을 대상으로 하는 사이버 공격의 경우 물리적 피해로 직접 이어질 수 있기에 공정 과정에서의 보안 기능 구현이 매우 중요하다. 본 논문에서는 SWaT 테스트베드 데이터 셋을 이용하여 순환 신경망 3가지를 설계하여 이상징후를 탐지에 적합한 모델을 제안하였다. 또한, 탐지 성능을 평가하기 위해 기존 평가 방법인 정밀도와 재현율뿐만 아니라 범위 기반 평가와 시계열 인지 기반 평가를 하여 제안 모델의 성능을 비교하였다.

이상징후 탐지 실험 결과, 기존 기계학습 모델과 비교하여 TaR 관련 지표에서는 LSTM이, TaP 관련 지표에서는 GRU가 우수한 성능을 나타냄을 확인하였다. 향후 전력 시설이나 및 스마트 공장과 같은 산업제어 시스템에 대한 사이버 공격이 증가할 것으로 예상되므로 이상징후 탐지를 위한 각 시스템의 공정 특성에 맞는 특징 추출 및 탐지율 향상을 위한 연구가 필요할 것이다.

References

[1] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmiduber, "A novel connectionist system for improved unconstrained handwriting recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5), 2009
DOI: https://doi.org/10.1109/TPAMI.2008.137

[2] H. Sak, A. Senior and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), pp. 338-342, 2014

[3] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks," Journal of machine learning research, 3, pp. 115-143, 2002
DOI: https://doi.org/10.1162/153244303768966139

[4] J. C. Heck, and F. M. Salem, "Simplified minimal gated unit varations for recurrent neural networks," 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1593-1596, 2017
DOI: https://doi.org/10.1109/MWSCAS.2017.8053242

[5] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A Dataset to Support Research in the Design of Secure Water Treatment Systems," International conference on critical information infrastructures security, pp. 88-99, 2016
DOI: https://doi.org/10.1007/978-3-319-71368-7_8

[6] N. Tatbul, T. Lee, S. Zdonik, M. Alam, and J.

Gottschlich, "Precision and Recall for Time Series", arXiv preprint arXiv:1803.03639, 2018

[7] W. S. Hwang, J. H. Yun, J. Kim, and H. C. Kim, "Time-Series Aware Precision and Recall for Anomaly Detection: Considering Variety of Detection Result and Addressing Ambiguous Labeling," Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2241-2244, 2019
DOI: https://doi.org/10.1145/3357384.3358118

[8] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Neural computation, 13(7), pp. 1443-1471, 2001
DOI: https://doi.org/10.1162/089976601750264965

[9] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," 2008 eighth ieee international conference on data mining, pp. 413-422, 2008
DOI: https://doi.org/10.1109/ICDM.2008.17

황 종 배(Jong-Bae Hwang) [준회원]

- 2021년 2월 : 호서대학교 컴퓨터정보공학부 (공학사)
- 2021년 3월 ~ 현재 : 호서대학교 정보보호학과 석사과정

〈관심분야〉
산업제어 시스템 보안, 인공지능 보안, 네트워크 보안

장 세 창(Se-Chang Jang) [준회원]

- 2016년 3월 ~ 현재 : 호서대학교 정보보호학과 학사과정

〈관심분야〉
양자 암호, 네트워크 보안, 부채널 분석

장 재 원(Jae-Won Jang) [준회원]

- 2016년 3월 ~ 현재 : 호서대학교 정보보호학과 학사과정

〈관심분야〉
인공지능보안, 양자내성암호, 부채널분석

김 기 현(Ki-Hyun Kim) [정회원]

- 1993년 2월 : 경북대학교 전자공학과 (공학사)
- 1995년 2월 : 경북대학교 일반대학원 전자공학과 (공학석사)
- 2011년 8월 : 충북대학교 일반대학원 컴퓨터공학과 (공학박사)
- 2014년 3월 ~ 현재 : ㈜앤앤에스피 연구소장

〈관심분야〉
시스템 및 네트워크 보안, 보안 관제, 제어망 보안

하 재 철(Jae-Cheol Ha) [종신회원]

- 1989년 2월 : 경북대학교 전자공학과 (공학사)
- 1993년 8월 : 경북대학교 일반대학원 전자공학과 (공학석사)
- 1998년 2월 : 경북대학교 일반대학원 전자공학과 (공학박사)
- 1998년 3월 ~ 2007년 2월 : 나사렛대학교 정보통신학과 교수
- 2007년 3월 ~ 현재 : 호서대학교 컴퓨터정보공학부 교수
- 2013년 1월 ~ 현재 : 한국정보보호학회 상임부회장
- 2009년 1월 ~ 현재 : 한국산학기술학회 이사

〈관심분야〉
암호학, 네트워크 보안, 부채널 분석, 인공지능 보안