Title (Use Title Case)

1st Author Name a, 2nd Author Name b, 3rd Author Name c

a 1st Author Affiliation, City, Country, e-mail address

b 2nd Author Affiliation, City, Country, e-mail address

c 3rd Author Affiliation, City, Country, e-mail address

**Abstract**

A concise and factual abstract is required, with a restriction of **150 words**. The abstract should state briefly the purpose of the research, the principal results and major conclusions. An abstract is often presented separately from the article, so it must be able to stand alone. For this reason, References should be avoided, but if essential, then cite the author(s) and year(s). Also, non-standard or uncommon abbreviations should be avoided, but if essential they must be defined at their first mention in the abstract itself.

**Keywords**

Immediately after the abstract, provide a maximum of 6 keywords, using American spelling and avoiding general and plural terms and multiple concepts (avoid, for example, 'and', 'of'). Be sparing with abbreviations: only abbreviations firmly established in the field may be eligible. These keywords will be used for indexing purposes.

1. Introduction

 환경에 대한 관심이 높아지고 있으며 인체에 해로울 수 있는 실내공기오염의 중요성이 강조되고 있다. 특히 실내공기오염(IAP)는 실외공기오염(OAP)보다 위험하다. 왜냐하면 실내에 머무는 시간이 길고 밀폐된 공간이기에 오염물질이 집중적으로 몸에 해로운 영향을 주고, 폐에 전달되는 과정이 짧기 때문이다. 또한 심혈관, 폐암, 기관지 천식 등 호흡기 질환과 같은 심각한 위험에 노출된다.[1]. 그렇기 때문에 실내공기오염(IAP)은 인간의 건강에 심각한 위협이 되고 매년 수백만명이 사망한다. 따라서 많은 연구자들이 이러한 문제를 해소하기 위해 주요 발생 원인과 실내 공기질(IAQ)의 제어에 대한 접근방식이 고안되고 있다.[2] 또한 이를 해결하기 위한 방안 중 하나로 실내환경에서 해로운 영향을 끼치는 인자들인 TVOC, CO2와 같은 주요 인자들의 측정에 있어 이상치의 중요성이 강조되고 있다[3]. 각 특성마다 이상치의 범위를 정의하는데 있어서 다양한 방법이 연구되고 있지만 환경 센서에 적용할 수 있는 이상치를 판별하기 위한 방법론을 연구해야한다. 더욱이 강조되는 상황은 인자의 정보를 알지 못하는 경우의 이상치를 판단하는데 있어 새로운 방법이 필요하다.

이상 탐지는 주어진 집합에서 비정상적인 점이나 패턴을 찾는 기술이다. 비정상에는 점 이상, 집단적 이상(collective outlier) 및 상황적 이상(contextual outlier)이 있다. 점 이상치는 다른 데이터와 비교하여 차이가 큰 데이터이다. 집단적 이상치는 데이터를 모았을 때 비정상으로 보이는 데이터를 의미한다. 이 중 상황적 이상치는 상황에 따라 정상일 수도 있고 비정상일 수도 있는 데이터에 대해서 상황 및 환경에 따라 장기적인 데이터로 인해 비정상적으로 보이는 데이터 집단을 말한다. 실시간 탐지에서는 모델 예측과 센서 측정 간의 차이를 비정상 측정의 분류로 본다.[4] - 각 이상치마다 탐지하는 방식이 다른지 확인필요. 그 중 주로 연구되는 분야와 우리의 방식을 연결하면 좋을 것 같음 그리고 그 연구분야가 어떻게 진행되고 있는지를 좀 더 깊게 파악하는 게 좋을 듯

실내환경에서는 구조와 사물에 따라 주요 환경 인자들의 수치가 달라질 수 있으므로 상황적 이상치에 대한 탐지가 필요하다. 실시간으로 여러 유형의 센서 데이터에서 상황별 이상값을 찾는 알고리즘에 대한 연구가 활발히 진행되고 있다.[5] 실내환경의 상황적 이상치의 경우, 범위가 달라질 수 있기 때문에 학습시키기 위한 조건이 까다로워진다. 다시 말해서 정상 데이터로 훈련시키기 위해 데이터를 수집해야 하지만 노이즈가 있는 경우를 배제하지 못한다. 그래서 대부분의 데이터가 노이즈를 포함한 정상 샘플이라는 가정을 하여 라벨의 취득 없이 학습을 시키는 경우인 비지도학습 이상치 탐색인 경우의 연구가 활발히 진행되고 있다. 라벨링이 되어있지 않은 경우 이상치를 어떻게 정의하느냐에 따라 정상치와 이상치로 구분될 수 있는 범위가 달라진다. 그렇기에 기준점이 적용되는 분야, 연구 등에 따라 다양하게 정의가 내려질 수 있다. 네트워크 이상 탐지의 경우, 네트워크 트래픽에서 예상되는 정상 패턴을 따르지 않는 예외적인 패턴을 이상치로 정의한다.[6] 자율 주행 분야의 경우, 고장을 진단하기 위해 Resnet 분류를 활용한 고장을 진단하는 모델로 동적 특성과 다른 경우를 고장으로 결함 진단을 정의한다.[7]  자연 이미지 데이터에서 다중 클래스 분류에서 이상치 탐지 방법을 평가할 때 클래스 레이블을 이미 사용할 수 있는 기존 분류 데이터 세트를 조정한다.[8]

시계열에 대한 연구에 있어서 베어링 결함 진단 문제[9]나 다변량 이상치 탐지에 대한 많은 연구는 [10]를 참조할 수 있다. 일변량 시계열에서는 범위를 넘는 점이나 하위 시퀀스가 나타나는 경우에 이상이 있는 것으로 간주할 수 있다.[11] 이와 달리 다변량 시계열에서는 각 시계열의 비정상적인 값과 하위 시퀀스의 이상 여부 뿐만 아니라 이러한 변수 간의 관계를 조사한다.[12] 본 논문에서의 환경 센서의 경우 다변량 시계열의 이상 탐지 문제를 다룬다.

(위 파란색 문단과 병합하여 해당 연구를 왜 진행하는지 기존 연구를 통해 제시. 영석이와 논의한 결과 산업별 이상치 중 환경 데이터의 시계열 -> 시계열 데이터의 이상치 타입은 세 가지로 분류 -> 그 중 가장 많은 이상치를 차지하는 타입은 어떤 것이며 해당 타입의 연구는 어떻게 진행되고 있고 어떤 모델이 주로 사용됨. 우리는 기존 연구에서 이러이러한 문제(contribution과 연결되도록)를 해결하고자 연구를 진행함)

본 논문에서는 이상치(이상점)을 정상적으로 모인 군집에서 벗어난 관측 값으로 정상 데이터의 범주를 토대로 해당 결정 경계 밖 데이터를 이상치로 정의하였다. 하지만 환경에 따라 이상치 군집이라고 부를 수 있는 범위와 값이 달라진다. 실내 공기질은 건물의 구조와 장비에 사용되는 자재, 환기, 거주자의 행동 방식 등에 따라 실내 공기의 질이 크게 좌우된다.[13] 다시 말하면 실내 구조나 환경에 따라 농도, 수치 등 이상치라고 판명할 수 있는 범위가 달라진다. 그리하여 실내 환경이 바뀜에 따라 적용되는 센서의 선행성에 대해 검증할 필요가 있다. 이상치라고 불리우는 레이블을 따로 지정해주지 않고 축적되는 전체 데이터에서의 수치로 불리는 군집을 비지도학습을 통해 진행할 수 있다. 우리는 이전 데이터로 주요 인자들을 분석하고 이상치로 판명될 수 있는 패턴을 찾아 앞으로의 상황에서 가스를 정밀하게 측정하여 실내공기오염을 예방할 수 있다.

우리는 다변량 시계열 센서에서 이상치 탐지를 위한 LSTM-AE를 제안한다. LSTM-AE에서 생기는 인자를 추출하여 클러스터링을 진행하고, encoder와 output에서 생기는 decoder에서의 차이를 이용해 threshold를 정하여 decision rule을 정하고 이를 시계열 특성이 담긴 전체 데이터에서 훈련시켜 reconstruction 모델을 구성하고 있다. 이 후 우리의 센서에서 훈련된 데이터를 바탕으로 여러 상황들을 lab test를 통해 확인하였다.

기여점으로는(해당 연구를 통해) 실내 환경에 따라 달라질 수 있는 이상치의 범위를 정할 수 있다. 또한 비지도학습이기에 라벨링이 되어 있지 않은 상황에서의 정상 데이터와 이상 데이터를 구별할 수 있다. 그리고 실내에서의 이상치를 탐지하는데 있어 실시간으로 군집화를 이뤄 어떤 점에 위치하는지를 확인하고, 이상치가 검출되기 전의 상황을 잡아낼 수 있다. (기여점을 좀 더 풀어서 쓰면 좋을 듯)

최종적으로 우리는 시계열을 이용한 환경 다중 센서를 이용해 실내환경에서의 가스를 측정하고 이에 대한 이상치를 탐색하는 연구를 진행하였다.

1. Method

제안된 이상 탐지 방법은 그림1과 같이 크게 4단계로 (i)데이터 정리 단계, (ii) 모델링 단계, (iii) 여러 결정 규칙 단계, (iv) 자체적인 lab test 단계로 나눌 수 있다.

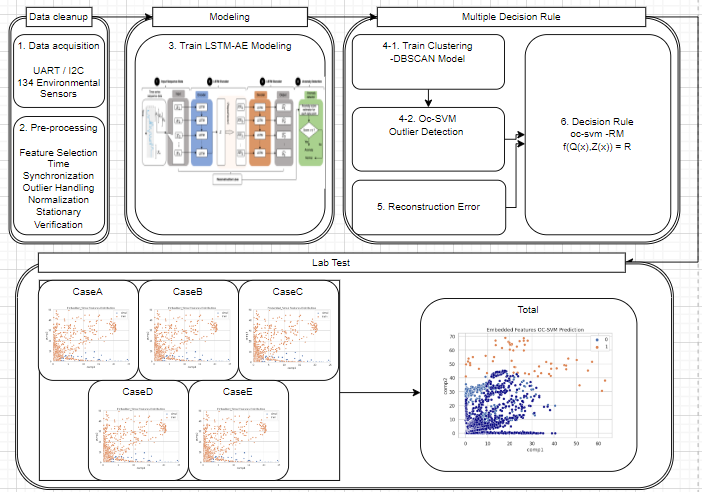
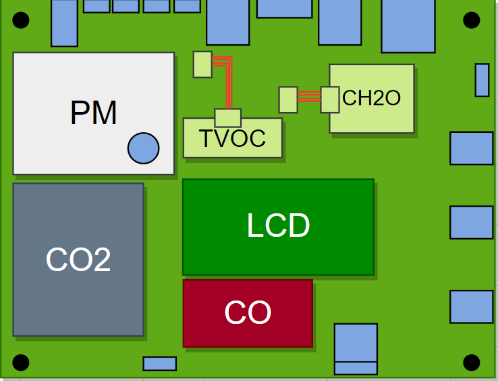


Figure . Framework of the proposed method. Flowchart for outlier detection with a total of 4 paragraphs.

먼저, 데이터 clean up에서는 센서로부터의 데이터 추출과 정제를 진행한다. 모델링 부분에는 비지도학습으로 lstm-ae 모델링과 encoder를 통한 decoder 추출에 해당한다. 여러 결정 규칙에는 클러스터링과 svm 그리고 reconstruction error를 통한 decision rule 결정 및 train의 전반적인 내용에 해당한다. 마지막으로 이를 평가하기 위한 lab test와 verification으로 나뉜다. (i)부터 (iii)까지 훈련과 검증이 시행된 후, (iv)에서 자체적인 테스트 상황과 모델의 성능을 판단할 수 있는 시스템을 통해 한 번 더 시행하여 군집화를 잘 이루어질 수 있는지 확인하는 작업을 거친다. 제안된 방식의 프레임워크는 다음 그림과 같다.

* 1. Data & Sensor

우리는 실내 공기질을 측정하기 위해 다종의 유해 물질(PM2.5,PM10,TVOC,CO2,CO,CH2O)과 영향을 줄 수 있는 물질 (온도, 습도)을 감지할 수 있는 통합 디바이스를 설계하였다. 아래 그림은 환경 센서의 회로 설계와 내구도 그리고 configuration이다.

****

(a) (b) (c)

Figure . IoT environmental sensor device : (a) actual circuit diagram, (b) sensor device structure, (c) sensor device configuration

환경 센서의 스펙 및 동작 방식을 표1에 표시하였다.

|  |  |  |
| --- | --- | --- |
| Sensor type | Detection range | Method |
| TVOC | 0 ~ 60,000 ppb | Semiconductor |
| CO2 | 400~60,000 ppm |
| PM 2.5 | 1~1000 /m3 | Optical |
| PM 10 | 1~1000 /m3 |
| CH2O | 0~5 ppm | Electrochemical |
| CO | 0 ~1000 ppm |
| Temperature | -20~80℃ | Semiconductor |
| Humidity | 0~100% (RH) |

table . Measurements of performance characteristics of the sensor device

우리가 사용한 센서의 동작 원리는 반도체식, 광학식 그리고 전기화학식을 사용하고 있으며 통신 방식으로 Wireless local area network(WLAN),Long-term evolution(LTE),Fifth-generation(5G) 환경 센서로 사용하여 다양한 환경센서용 Universal asynchronous Receiver Transmitter(UART), Inter-integrated circuit(I2C), Analog-digital converter(ADC)를 통해 정보를 수집한다. UART를 사용하여 통신하는 외부 LoRa 모뎀도 있다. 제어 장치는 원격 교정 프로토콜을 사용하여 장치 재설정 및 센서 주기 변경과 같은 기능을 수행했다.

 각각의 동작 원리나 통신 방식에 따라 발생하는 기계적 결함으로 인한 이상치가 발생할 수 있다. 하지만 우리는 이상치의 종류에 대해 분류하는 것은 어떤 상황 때문인지 작성하고 ref 삽입, 즉 이상치를 분류할 때는 비지도학습이 적합하다는 ref가 들어가야 함 해당 논문에서 제시하지 않기 때문에 센서의 특성에 따라 발생하는 이상치를 구분하지 않기로 하였다. 결론적으로 다종의 센서로 측정하여 가스의 이상치가 되는 요소를 동작 원리, 통신 방식에 상관없이 예측하고자 한다.

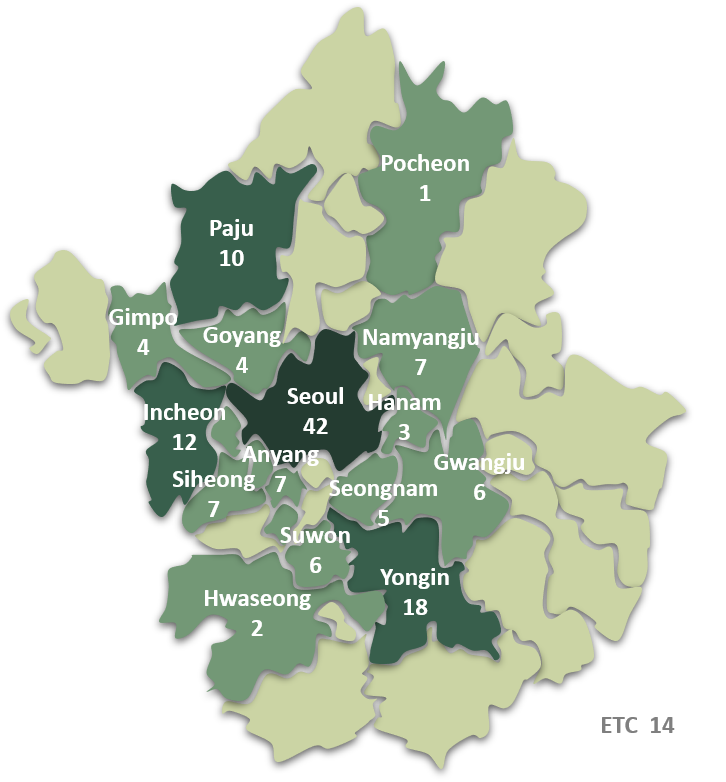


Figure . Verification of sensor stability.(a) Sensor installation at home, (b) testbed configuration

환경센서의 데이터 수집을 위해 실험 대상군과 대상지역을 설정하였다. 최소 5군집(지역, 연령, 성별, 시설, 주거환경 등) 이상 표본을 설계하고 수도권(서울, 경기) 및 중부 지역으로 설치 참가군을 모집하였다. 또한 일반가구와 산업단지지역, 스마트시티지역을 대상으로 하여 134개소에 설치를 진행하였다. 산업단지지역에서는 환기가 원활하고 특정 이상 상황이 발생하지 않는 곳에 설치하였으며, 일반 가정에는 주변의 기기로부터 자유롭고 주기적으로 환기가 원활하며 대부분의 가정에서 소유하고 있는 거실 테이블에 배치하였다. 설치장소마다 실내 환경이 다르기 때문에 최대한 동일 조건으로 측정하기 위해 환기와 높이 등을 고려하여 설치 장소를 선정하였다. 134대의 센서를 주기적으로 모니터링을 하였다. 이를 통해 센서에서 결함이나 오류가 나지 않음을 확인하였다. 또한 센서 간의 측정하는데 있어 오차가 발생하지 않는 것을 검증하여 센서의 안정성을 확인하였다.

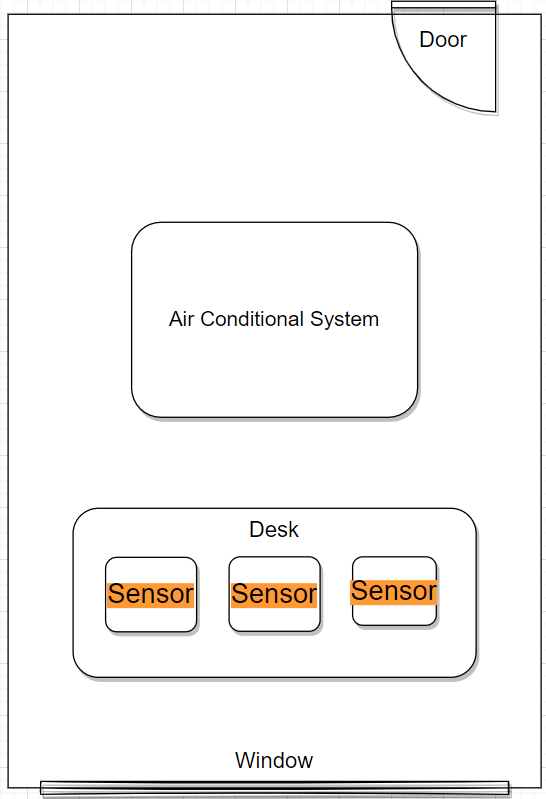
****

Figure . Verification of sensor reliability. (a) actual sensor device installation, (b) Experiment environment configuration

센서로 여러 가스를 측정할 때 너무 멀리서 측정하게 되면 측정 범위가 지나치게 넓어지는 단점이 있다[1]. 따라서 제어 가능한 범위 안에 환경 센서를 적절한 거리를 두어 배치하였다. 비이상적 센서의 신뢰도 요인을 평가하기 위해 같은 공간에 여러 센서 디바이스를 설치하였다. 위치는 서울 숭실대학교 전산관에 위치했으며 한국 환경부에서 제공하는 ‘다중이용시설 실내공기질’의 규정과 운영 매뉴얼에 따라 설치하였다. 데이터 실측 값에 대해 확신하지 못하는 부분적으로 알려진 데이터를 이를 통해 처리하였고 거리를 고려하여 여러 대의 센서의 신뢰성을 검증하였다.[2] 또한 결함이 있는 센서를 다른 센서와 겹치면 결함을 줄일 수 있다. [3] 실내 환경의 구성에는 환기와 통풍이 잘되는 환경으로 문과 창문의 거리가 어느정도 있는 곳으로 3대를 설치하여 센서의 신뢰성을 검증한 뒤, 연구를 진행하였다.

문단 문단마다의 연결을 만들 수 있는 접속어를 충분히 활용 필요

**2.2 Pre-processing**

**2.2.1  시계열 튜닝**

일변량 시계열 데이터는 연속적인 시점에서 단일 관측치의 시퀀스를 형성한다. 일반적으로 관찰 열로 간주되지만 시간은 실제로 암시적 변수이다[1]. 측정된 환경 센서의 시계열 데이터는 연속적인 특성을 가지며 순차적으로 수집되었다. 2분마다 측정되는 센서의 기계적 결함에 의해 unequally spaced series(비등간격성 시계열)의 성질을 띄기 때문에 시간을 등간격으로 맞춰주는 것이 중요하다. 그러므로 근사 및 보간법[2]을 사용하여 시간 간격을 동일한 크기를 갖게 조정해주는 작업이 필요하다.[3]

전문용어의 경우는 1차적으로 우선 영문으로 변경해둘 것. 그래야 번역할 때 해당 부분이 다른 단어로 바뀌는 문제를 방지함

**2.2.2. 정상성**

시계열 분석에서는 데이터가 안정적이고 자기상관이 없는 것이 중요하다. 환경 센서의 이상치를 감지하기 위한 클러스터링 알고리즘을 사용하기 전에 정상성 유무를 확인해야 한다. 정상성은 데이터의 평균과 표준편차가 시간에 따라 변하는 행동을 말하며, 이러한 행동을 가진 데이터는 정상적이지 않은 것으로 간주된다. 정상성을 정량적으로 검증하기 위해 Augmented Dickey fuller test를 시행한다. ADF 검정의 귀무가설의 경우 단위근이 있다는 것으로, 확인 결과 귀무가설을 기각하여 정상성을 만족하는 것을 확인하였다. ADF 검정의 경우 자유도의 손실을 우려할 수 있어, 이중적으로 필립스-페론 테스트를 통해 한 번 더 정상성을 만족하는 것을 검증하였다.[4]

**2.2.3. 결측값**

 결측값은 데이터가 없거나 데이터가 불완전함을 의미한다. 환경센서에서 결측값이 생기는 경우, 두 가지로 나뉜다. 첫째, 센서의 측정값이 급격하게 변할 때 결측값이 발생한다[1]. 둘째, 센서의 측정 범위에 따라 누락된 값이 발생할 수 있다[2]. 우리 센서에서도 약 0.92%의 결측값이 발생하였다. 해당 결측값의 경우 선형보간법을 사용했다. 선형 보간: 결측값을 추정하기 위해 양 끝점의 값을 사용하여 선형 거리에 따라 결측값을 선형으로 추정했다. 선형보간법은 유전자형 대체 및 기계 번역 분야에서 결측치 대체 성능을 개선하는 데 사용되었다.[3]

**2.2.3. 상관성 분석**

 다변량 데이터는 여러 독립 변수가 있는 데이터이다. 이 부분에서 사용된 방법은 변수 간의 종속성을 사용하는 방법이다. 상관계수는 두 시퀀스 간의 관계를 직접 반영하는 접근 방식 중 하나이다.[1] 따라서 실험 데이터에서는 먼저 Pearson 상관 계수를 사용하여 변수 간의 상관 관계를 조사했다. 우리가 제안할 모델에 있어 각 변수들의 상관성은 중요 요소로 들어가기 때문에 (좀 더 명확하게 작성할 필요가 있음) 상관성 분석을 진행하였다.

**2.3 Process Method**

**2.3.1  LSTM-AE**

본 연구에 사용한 다종 실내 대기질 데이터는 시간 도메인을 따라 특성 변화가 관측되는 시계열 데이터이다. 이와 같은 시계열 데이터 위에서 이상치 탐지를 수행할 경우, 시간에 따라 변화하는 데이터의 특성과 패턴을 매우 중요하게 고려해야한다. **[8]** 일반적으로 이러한 경우 LSTM 을 비롯한 RNN 기반의 신경망을 사용하는데, 기존에 사용되는 일반적인 형태의 RNN을 사용할 경우 기울기 소멸 / 폭발 등 이미 잘 알려진 여러 문제들로 인해 긴 시퀀스에 대해 좋은 성능을 기대하기 어렵다. **[10]** 따라서 이러한 장기 의존성(Long Dependency)의 문제를 해결하기 위해 근본적인 아키텍쳐의 변형이 필요한데, 그 해법으로 나온 신경망이 LSTM(Long Short-Term Memory) 신경망이다. LSTM layer는 비선형 게이트 유닛으로부터 메모리 셀 안팎으로 입출입하는 정보를 조절함으로써, 모델이 시퀀스의 장기적인 동적 정보를 적응적으로 학습하여 시계열 데이터 모델링에 있어 우수한 성능을 나타낸다. **[12]** 우리는 LSTM 레이어를 사용하여 실내 대기질 데이터에 숨겨진 시계열적인 특성을 자동적으로 학습하고, 이를 기반으로 비정상적인 이벤트를 감지하고 예측하는 모델을 개발하고자 한다.

 앞서 설명한 LSTM을 포함하여, 대부분의 인공신경망 프레임워크는 지도 학습 기반으로 작동하기 때문에 이상치 탐지를 비롯한 비지도학습 프레임워크에 최적화 되어있지 않다. 실증을 위해 수집되는 대다수의 데이터 또한 No Labeled Data이기 때문에 객관적인 평가기준을 잡기 매우 어렵고 이에 따른 추가적인 장치가 필요하다. **[5]** 따라서 우리는 이에 대한 해법으로 앞서 설명한 LSTM과 더불어 AutoEncoder(이후 AE) 아키텍쳐에서 아이디어를 착안한 LSTM-AE Model을 제안하고자 한다.

 AE는 Unlabelled Data로부터 데이터를 가장 잘 표현하는 잠재 표현형을 학습하는 일종의 비지도 학습 신경망이다. AE는 크게 Encoder와 Decoder로 이루어지는데, Encoder는 입력 값을 저차원의 병목 층으로 투영하여 데이터 내 상대적으로 중요치 않은 특성을 제거하고, 데이터를 가장 잘 표현하는 압축 특성만 남기는 역할을 한다. 그리고 Decoder는 위에서 병목 층으로 표현된 저차원 벡터를 다시 원 입력 값으로 복원하는 작업을 수행한다. 이렇게 복원된 출력값은 최초의 입력 값과 차이를 계산하여 해당 복원 손실을 최소화하는 방향으로 파라미터를 수정한다. 그리고 마지막으로 최종 학습된 모델으로부터 계산한 재구성 오차(Reconstruction Error)를 기준으로 임계값을 설정하여 새로운 입력에 대해 이상치를 판별한다. **[6]**

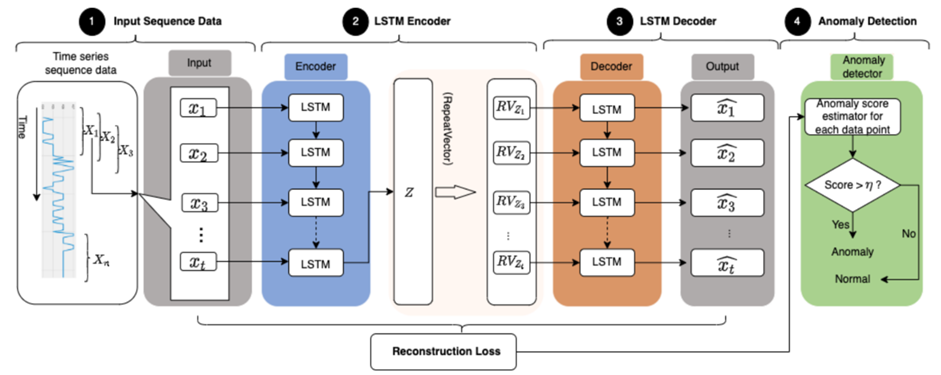
****

Figure 5. LSTM -AE 구성도 (수정바람)

**-> 사진 수정**

 우리가 본격적으로 제안하고자 하는 LSTM-AE Model은 앞서 언급한 두 신경망의 장점을 모두 가진, 시계열 데이터 기반 이상치 탐지 프레임워크이다. LSTM-AE는 데이터 내 시계열적인 특성을 스스로 잡아내어 잠재 표현형을 자동으로 학습하여 재구성 오차 기반의 이상치 탐지 결정 규칙을 제공할 뿐만 아니라, 병목 층으로부터 추출되는 잠재 표현형을 토대로 2차 서브 알고리즘을 구축할 수 있다. (ref 추가)

 LSTM-AE를 활용해 이상치 탐지를 수행하는 기존 연구들의 경우 테스트 데이터의 재구성 오차만을 가지고 사전에 계산된 임계값과 비교하는 1차원적인 결정 규칙을 가지고 있다. 하지만, 이러한 단순한 판별 방식은 학습 데이터의 품질에 따라 모델 성능이 쉽게 좌지우지되기 때문에, 노이즈가 많은 실증 데이터에 경우 적합하지 않다. 작은 노이즈에도 모델이 민감하게 반응하여 False Positive Rate (FPR) 지표가 비약적으로 증가한다면 모델 상용화 단계에서 수 많은 False Alarm을 야기할 수 있다. **[13]** 따라서 우리는 모델의 일관성과 정확성을 강화하기 위해 재구성 오차 기반의 결정 규칙과 더불어, Encoder에서 생성한 잠재 표현형을 활용한 서브 알고리즘을 정의하였다. 그리고 추가로 정의한 서브 알고리즘의 판별 결과와 기존에 정의한 재구성 오차의 판별 결과를 결합하여 보다 정확하고 일관적인 이상치 탐지 결정 규칙을 제안하고자 한다.

**2.3.2  Deep-Compact Clustering & One-Class Sub Algorithm**

 AutoEncoder 기반 이상치 탐지 모델의 경우, 훈련 데이터 대부분이 정상 데이터라는 전제 하에 학습이 진행되며, 계산된 재구성 오차의 임계값을 기준으로 이상치를 판별한다. [4] 하지만, 실증 데이터 검증 하에 데이터 내 무시할 수 없는 양의 노이즈가 포함되어 있기 때문에 해당 데이터로 학습 시 높은 FPR이 나타날 수 있음을 앞서 언급하였다. 우리는 이러한 문제점을 해결하기 위해 데이터 전처리 및 정제 작업 뿐만 아니라 알고리즘 적으로 추가적인 장치가 필요함을 주장한다. 따라서 AE의 Encoder 부분에서 출력되는 압축 표현형을 활용하여 추가적인 서브 알고리즘을 정의하였다. 또한, 해당 알고리즘과 재구성 오차의 이상치 판별 과정을 복합적으로 결합한 새로운 이상치 결정규칙을 제안하고자 한다. 본 과정을 통해 보다 복잡하고 안정된 결정 기준을 갖게 되어, 애매한 경계에 있는 포인터들에 대해 낮은 FPR과 일관적인 성능을 확보할 수 있게 된다.

 데이터 마이닝에서 널리 쓰이는 알고리즘 중 하나인 클러스터링 기법은 동일한 특징 및 분포를 띄는 데이터끼리 클러스터로 묶어주는 대표적인 비지도 알고리즘이다. **[9]** 본 프레임워크에서 클러스터링은 서브 알고리즘을 수행하기 전 사전 작업으로 정상 데이터 클러스터를 형성하여, 데이터 내 노이즈를 제거하고 2차원 공간 안에 데이터를 투영하기 위해 수행된다. 일반적으로 사용되는 클러스터링 기법의 경우 별다른 가공 없이, 압축되지 않은 순수 입력 데이터 벡터를 대상으로 그룹화를 수행한다. **[4]** 하지만, 이러한 방법은 모델이 학습한 고수준 패턴을 반영하지 않은 채로 그룹화가 수행되기 때문에 데이터의 압축률과 밀집도 면에서 설명력이 떨어지는 결과를 야기한다. **[3]** 따라서 우리는 LSTM-AE의 Encoder 부분의 병목 층 출력결과를 활용하여 해당 차원에서 그룹화를 진행하였다. 이렇게 하면 LSTM Layer가 포착한 데이터 내 시계열적인 특성과 비선형 특징들을 반영하여, 원본 데이터를 가장 잘 설명하는 함축적인 특성으로 클러스터를 생성할 수 있다.

 앞서 서브 알고리즘을 수행하기 전, 데이터를 2차원의 압축 표현형으로 투영하고 노이즈를 제거하는 방법에 대해 이야기하였다. 이제 해당 정상 데이터 군집을 기준으로 경계를 형성하여 해당 경계 밖에 있는 포인터들을 이상치로 판별하는 작업을 수행하고자 한다. 본 연구에 사용되는 실증 데이터의 경우 Unlabelled data로 이루어져 있으며, 정상 클래스만 있다는 가정 하에 이상치를 판별함으로, One-Class Classification(OCC) 기반의 접근 방법이 필요하다. **[1]** OC-SVM은 이상치 탐지를 위해 사용되는 특별한 케이스의 SVM 알고리즘으로, 대부분의 데이터가 정상 데이터라는 전제 하에 훈련 샘플을 포함하는 가장 작은 초평면을 만들어낸다. **[11]** 해당 초평면이 만들어 내는 결정경계의 안팎을 기준으로 테스트 데이터의 이상치 판별 여부를 결정한다. **[8]**

 이처럼 LSTM-AE의 Encoder 부분의 병목 층을 활용하여 데이터를 압축, 투영한 “Compact Feature Space”가 OCC 분류기의 성능을 향상시킴을 언급하였다. 이제 이렇게 만든 OC-SVM의 판별 결과와 LSTM-AE의 재구성 오차 계산 결과를 결합하여 일관적이고 정확한 이상치 탐지 결정 규칙을 제안한다.

**2.3.3  Decision Rule combining OC-SVM and Reconstruction Threshold**

일반적으로 이상치 탐지 모델을 개발할 때, 사용하는 장치마다 생성하는 시계열 데이터의 종류가 모두 다르기 떄문에 데이터의 특성에 따라 특정 이상치 탐지 기술이 다른 기술보다 좋은 성능을 보이는 경우가 일반적이다. **[7]** 또한, 이상치 탐지 모델마다 이상치를 정의하는 관점이 모두 다를 뿐만 아니라, 모델마다의 특성과 장점이 뚜렷하기 때문에 단일 모델을 통한 결정 규칙은 다소 왜곡되고 편향적인 결과를 나타내기도 한다 (ref 추가). 따라서 우리는 서로 다른 접근 방식의 두 이상치 탐지 모델을 융합하여 보다 합리적이고 일관적인 결과를 제시하는 새로운 이상치 결정 규칙을 제안하고자 한다.

최근 발표된 논문들을 보면 (어떤 논문인지 구체적으로 제시 필요), 대기질 환경 연구에서 여러 단일 모델의 응답을 앙상블 방식으로 결합하여 사용했을 때 정확성과 효율성 측면에서 우수한 성능을 보였다. **[16]** 앙상블 학습 분류기는 여러 개의 분류기로부터 출력된 응답을 하나의 결정 규칙으로 결합함으로써 보다 정확한 최종 예측을 도출하는 기법으로 기존의 앙상블 기반의 결정규칙이 다수결 방식을 채택하는 Hard Voting과 확률 기반의 Soft Voting으로 구분된다 (ref 추가). (기존 연구에서는 해당 보팅의 방식이 어떠했는지, 문제점이 무엇이고 제시하는 방식은 어떠한 장점이 있는지 (ref 추가) 그 방향에 따라) 우리는 단일 모델 별로 가중치를 조정하여 데이터에 보다 적합한 모델이 더 큰 영향력을 가지도록 하였다.

위에서 LSTM-AE의 재복원 오차를 활용한 이상치 탐지 방법에 대해 정의하였는데, 이는 모델 기반의 접근 방식이라고 볼 수 있다 (ref 추가). 모델이 정상 데이터의 특징을 포착하여, 일정 임계값을 벗어나는 포인터를 발견했을 때, 모델의 파라미터 연산 결과가 이상치를 판별하는 방식이라면, Latent Feature를 활용한 OC-SVM 알고리즘의 경우, 특정 포인트가 대다수의 정상 데이터가 그려낸 초평면 안에 포함 되는지 아닌지를 기준으로 이상치 탐지를 수행하는 컨텐츠 기반의 접근 방식이다(ref 추가). 이렇듯 서로 다른 관점으로 접근한 두 단일 모델을 결합함으로써 두 모델의 장점은 유지하면서 단점은 보완한 Generalize한 모델을 도출하고자 한다.

**2.4  Laboratory Test ( Lab test)**

모델의 안정성과 정확성을 검증하고 이상치 탐지 결과를 증명하고자 랩테스트를 적용하였다. 본 연구에서는 프로세스 검증을 위해 센서에 5가지 상황을 대입하였고, 이를 이상치가 발생한 상황이라고 정의한 뒤, 성능을 확인하는 방법으로 채택하였다. – 랩 테스트가 필요한 이유를 좀 더 부각되어야 함. 비지도 기반의 모델의 경우 라벨링이 없기 때문에 평가에 대한 한계가 있음. 이를 보완하기 위해 라벨링 되어 있는 데이터를 생성하여 해당 모델의 성능을 좀 더 정확하게 측정하였다 등으로 기존 고민했던 측정 방식에 대한 부분을 기술해야 함

프로세스 기반 모델에 자체적 랩테스트를 대입했을 때는 지도학습의 데이터로 진행한다. 이 점을 활용하여 클러스터링이 제대로 되는지를 확인하는 절차로 약 2주간 진행하였다.

실제 조건의 센서 기반 테스트에 대한 요구 사항은 다음과 같다.

1. 데이터의 신뢰성을 높이고자 동일한 3개의 센서 디바이스를 등간격으로 데이터를 수집하였다.

2. 스프레이 방식으로 하루 4번으로 제한하여 일정 간격을 두어 분사한다. 일정 간격을 두는 이유는 lstm의 특성을 고려하여 잘 일어나지 않는 이상적인 상황을 측정한 뒤 센서의 측정이 정상으로 돌아오게 하기 위함이다.

3. 방해가 될 다른 요인들을 통제하였다. 방해가 될 수 있는 요인이라고 하면, 센서의 결함을 유도하거나 이상치로 측정될 수 있는 여러 환경 요인이다. 예를 들어 흡연, 스프레이나 섬유탈취제의  사용 유무 등이다.

4. 데이터 수집 과정에서 상황을 대입하였을 때의 시간을 측정 후 안정적으로 변하는 구간을 고려하여 라벨링의 작업을 진행하였다.

1. **Results**

**3.1 pre-processing (결측치 부분이 적어서 정상성, 결측치, 상관성을 3.1 하나로 합치기)**

**3.1.1. 정상성**

표()는 ADF 검정으로 인한 p-value 값이다. 각각 가스의 p-value는 0.05보다 낮음을 알 수 있다. 이는 정상시계열이 아니라는 귀무가설을 5%의 유의수준으로 기각할 수 있으므로 단위근이 있다는 귀무가설을 기각하여 정상성이 보장된다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Type | Temperature | | Humidity | TVOC | CO | CO2 | CH2O | PM10 |
| ADF-test | | -5.336495 | -4.812659 | -5.780521 | -12.57404 | -6.217924 | -8.065061 | -5.708918 |
| Critical Value (1%) | | -3.959300 | -3.959301 | -3.959301 | -3.959301 | -3.959301 | -3.959301 | -3.959301 |
| Critical Value (5%) | | -3.410747 | -3.410748 | -3.410748 | -3.410748 | -3.410748 | -3.410748 | -3.410748 |
| Critical Value (10%) | | -3.127201 | -3.127202 | -3.127202 | -3.127202 | -3.127202 | -3.127202 | -3.127202 |
| p-value | | 0.000048 | 0.000445 | 0.000006 | 5.1291e-20 | 7.5653e-07 | 5.0535e-11 | 0.000009 |

Table . Table 2. Augmented Dicky Fuller test for time series analysis

**3.1.2. 결측치**

 외부 요인으로 인해 측정되지 않은 센서의 결측치는 약 0.92%가 나왔다. 이 결측치에 대해서 본 연구에서는 선형 보간으로 대체한다.

**3.1.3. 상관성**

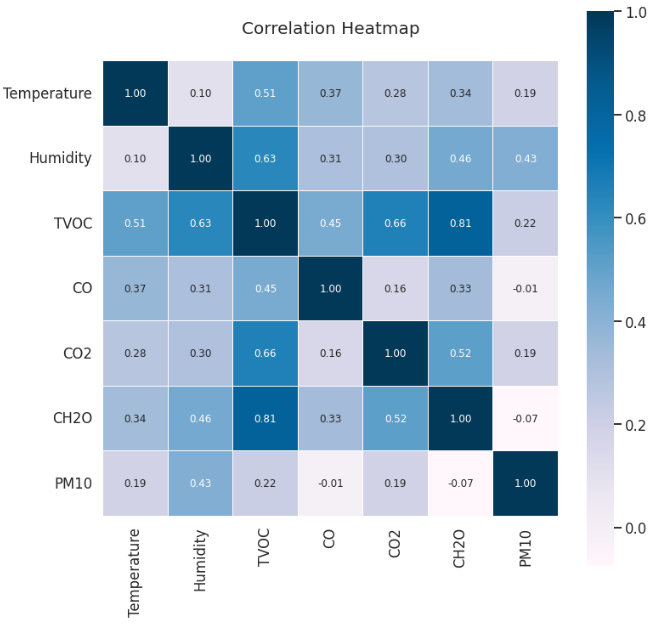


Figure . Pearson Correlation between environmental substances.

 그림과 같이 상관관계가 강한 환경물질을 식별할 수 있다. 그림과 같이 상관관계가 강한 환경물질을 식별할 수 있다. TVOC-CO2, TVOC-CH2O, TVOC-Humidity는 강한 상관성을 보인다. Temperature-TVOC가 약한 상관성을 보이는 상관 관계가 있다. 각 변수에 대해 다변량 대치를 수행하기 전에 상관 관계를 보이는 변수로 특성을 고려하여 모델링을 수행했다.다변수 layer에 상관관계가 명확한 변수를 고려하는 이유는 LSTM-AE에서 상관관계가 명확한 인자끼리가 효과(어떤 효과가 있는지 상세하게 기술하기)가 있기 때문이다 (설명요청하기).

**3.2 Modeling**

**3.2.1 LSTM-AE**

LSTM Layer 학습을 위해선 데이터를 [Samples, Timesteps, feature]와 같은 3D 형태로 변환해주어야 한다. 여기서 “Timesteps는 LSTM이 단기 기억으로 받아들일 창(Window)의 크기를 나타내는 하이퍼 파라미터로, Timestep이 클수록, 더 긴 범위의 시리즈를 단기 기억을 받아들인다. 우리는 최적의 Timestep을 찾기 위해, GridSearch 방식의 탐색 작업을 수행하였다. 총 [1, 3, 5, 7] 위 네 케이스에 대해 탐색을 수행하였으며, 그 결과 Timestep이 ‘3’일 때 재복원 오차와 클러스터 응집도 면에서 가장 우수한 성능을 보였다. 따라서 본 연구에 사용되는 데이터가 2분 주기로 측정되는 것을 고려했을 때, LSTM Layer가 6분을 단위로 창(Window)을 만들어 단기 기억을 학습한다고 할 수 있다.

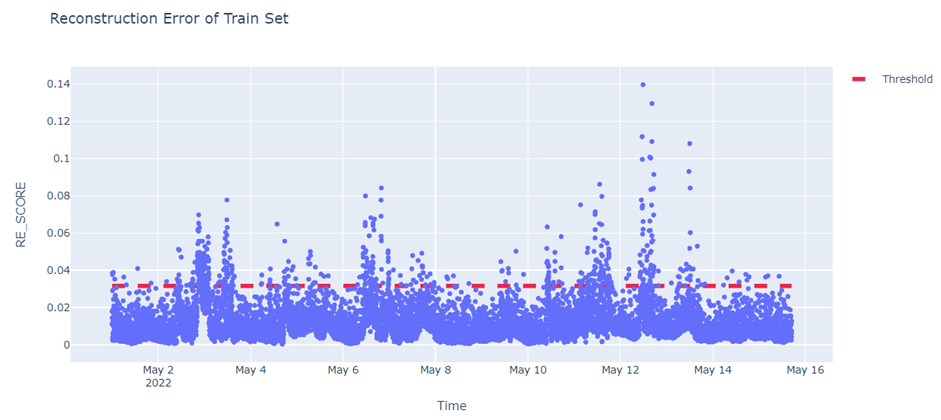


Figure . Reconstruction Error of train set

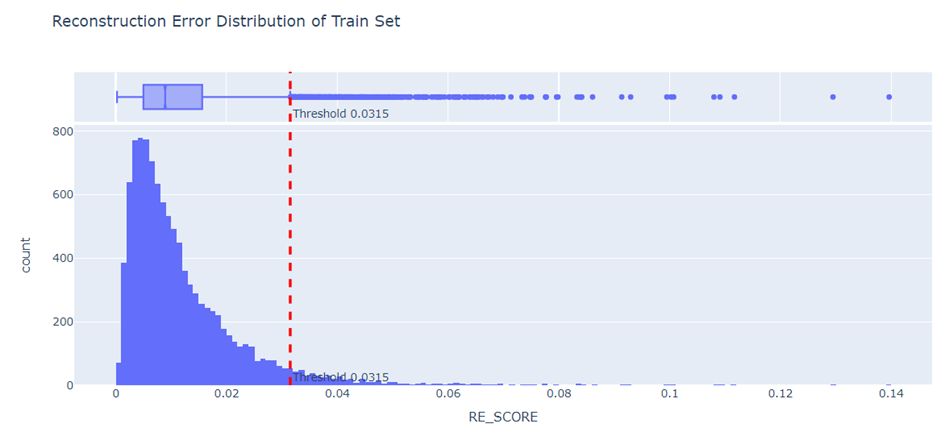


Figure . Reconstruction Error Distribution of train set

fig7과 fig8를 보면, 훈련 데이터 학습결과 IQR 범위 내에서 약 75%의 데이터가 0.03이하의 재구성 오차를 보이는 것으로 나타났다. 이는 모델이 데이터를 복원했을 때, 약 3%의 오차 범위 내에서 입력과 매우 비슷한 출력을 복원할 수 있음을 의미하며, 특징 학습이 아주 잘 이루어진 것으로 볼 수 있다. 또한, fig7의 재구성 오차 산점도를 보면, 주기적으로 관측되는 이상 포인트를 제외한 대부분의 포인트들이 강한 응집도를 보이는 것을 알 수 있는데, 이는 모델이 정상 데이터를 가장 잘 표현하는 잠재 표현형을 잘 학습한 것으로 볼 수 있으며, 추후 일괄적인 압축 표현형을 보일 것으로 기대할 수 있다.

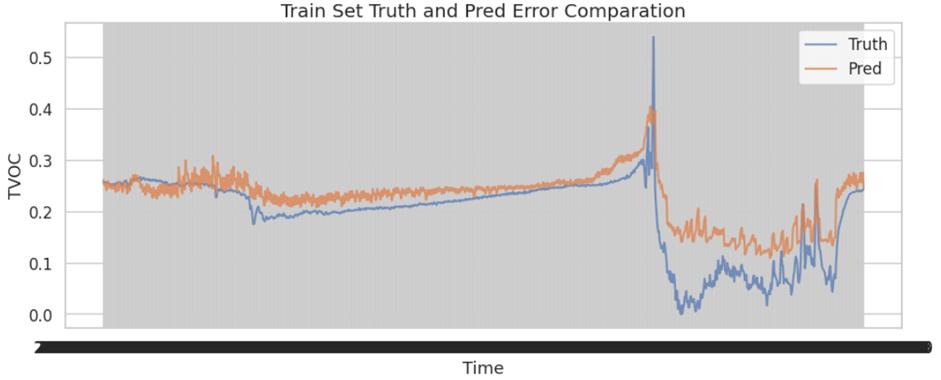


Figure . Truth and Pred error comparation of train set

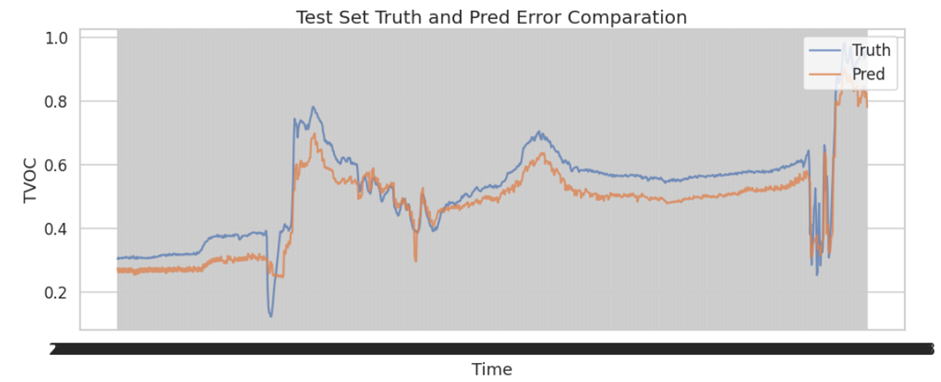


Figure . Truth and Pred error comparation of test set

  fig9은 LSTM AE로 복원한 입력값과 입력 원본을 한 캔버스(일반적인 단어??)에 표현한 그림이다. 그림을 보면, 원본 값이 변화할 때마다, 해당 변화량에 맞추어 추세를 잘 쫓아가는 것을 볼 수 있으며, 급격히 변화하는 값에 대해서는 다소 보수적인(해당 단어를 영문으로 우선 바꿔두기, 번역 때 의도치 않은 단어로 바뀔 가능성이 있음) 성향을 보이는 것을 알 수 있다. 이처럼 데이터가 기존의 추세에서 벗어나, 기존 추세와 다른 양상을 보이는 이상 구간에서는 데이터의 복원률이 낮아져, 재구성 오차가 커지게 된다. 기존의 논문들은 이와 같은 재구성 오차의 임계값 초과 유무에 따라 이상치를 판별하게 되는데, 이는 모델의 재복원 성능에 지나치게 좌지우지될 뿐만 아니라, 단순한 결정규칙으로 인해 다수의 False Alarm을 만들어낸다(ref 확인). 따라서 우리는 LSTM-AE Encoder 부분의 압축 표현형을 활용하여, 추가적인 서브 알고리즘을 정의하였다. 이렇게 정의한 알고리즘과 재구성 오차 기반의 기존 이상치 규칙을 앙상블 방법으로 결합하여 보다 복잡하고 일관적인 이상치 결정 규칙을 제안하고자 한다.

**3.2.2 Deep-Compact Clustering & OC-SVM**

LSTM AE Encoder는 LSTM Layer를 거치며 입력 데이터를 큰 특성에서부터 작은 특성으로 압축하는 역할을 수행한다. 여기서 Encoder의 Layer 중 가장 작은 수의 노드를 가지는 층인(~~을 병목 층(Bottle Neck)이라고 한다. 우리는~~) 병목 층을 4개의 노드로 구성했으며, OC-SVM 사전 작업 및 노이즈 제거를 위해 Clustering을 수행하였다.

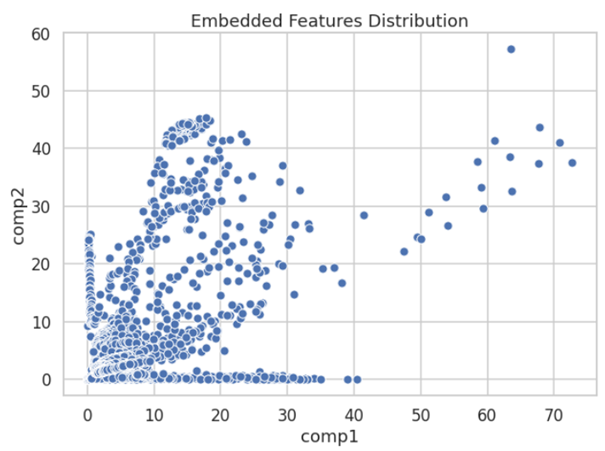


Figure 11. Embedded features distribution

 우선 데이터를 2차원의 압축 표현형을 투영하기 위해 4개의 노드 중 응집도 및 분리도가 가장 뛰어난 특성 두 가지를 선택하여 시각화였다. 왼쪽 아래에 보이는 다수의 데이터 포인트가 모여 있는 부분이 다수의 정상 클러스터이며, 오른쪽에 산개하여(해당 단어를 영문으로 우선 바꿔두기) 찍혀 있는 데이터가 학습 데이터에 포함된 노이즈로 볼 수 있다. 앞서 실증 데이터 내에 포함되어 있는 노이즈가 LSTM-AE의 정상 데이터 특성 학습에 악영향을 끼친다는 점을 언급하였다. OC-SVM 학습 또한, 훈련 데이터가 대부분 정상 데이터라는 가정 하에 진행됨으로 클러스터링 작업을 통해 이러한 노이즈를 제거하고자 하였다.

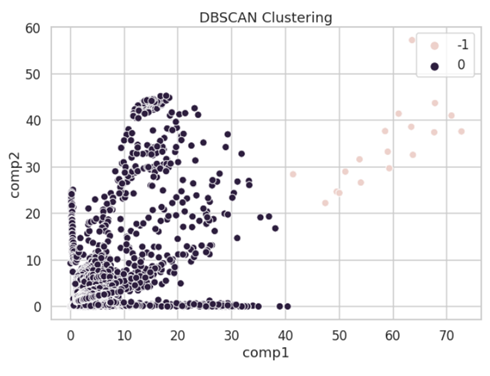


Figure 12. DBSCAN Clustering

 우리는 Encoder로 압축한 잠재 표현형의 클러스터 구축을 위해 DBSCAN 알고리즘을 사용하였다. DBSCAN은 데이터가 높은 밀도를 보이는 위치에 따라 클러스터를 구분하는 알고리즘으로 1996년에 처음 소개되었다. **[14]** DBSCAN은 사전에 클러스터의 수를 지정해주지 않아도 사용이 가능하다는 장점이 있으며, 어떤 임의의 모양의 클러스터도 모두 표현 가능하다는 장점이 있다. 특히, 비지도 학습에서 레이블이 없는 데이터셋의 경우 클러스터의 수에 대한 정보가 없기 때문에 더욱이 밀도 군집화 방식의 알고리즘이 효과적이라고 할 수 있다. 또한, 실증 데이터로 수집한 데이터는 비교적 변칙적이며, 예상 불가능한 분포를 띄는 경우가 많기 때문에 임의의 모양의 클러스터도 잘 잡아낼 수 있는 DBSCAN이 현 조건에 매우 적합하다고 할 수 있다. **[6] - 해당 부분은 method에 적합하다고 보임**

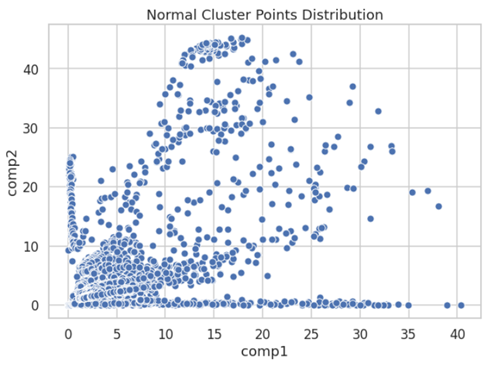
****

Figure 13. Normal Cluster points distribution

 DBSCAN은 두 개의 주요한 하이퍼 파라미터에 따라 클러스터를 생성한다. 하나는 ‘eps’로 이웃의 최대 반경을 정의하는 인자이고, 둘째는 minPts로 독자적인 클러스터를 생성할 수 있는 최소 포인트 숫자이다. 우리는 왼쪽 아래에 밀집되어 있는 포인터 집단을 정상 클러스터로 판단하고, 해당 포인트를 모두 포함할 수 있도록 하이퍼 파라미터를 조정하였다. 그 결과, fig13을 보면 알 수 있듯이 정상 데이터 분포를 표현하는 클러스터를 얻고, 불필요한 노이즈를 제거할 수 있었다.

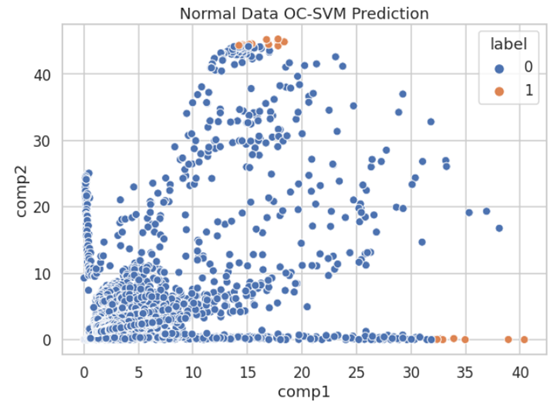


Figure . Normal data oc-svm prediction

 우리는 DBSCAN을 통해 정상 군집을 얻었다. 하지만, DBSCAN은 군집 밖에 있는 데이터를 이상치로 표현할 Scoring Method를 제공하지 않기 때문에 이상치 판별을 위한 추가적인 알고리즘을 정의하였다.

 OC-SVM은 한 개의 클래스로 구성된 서브셋을 구분하는 가장 작은 결정 경계를 계산하는 비지도 학습 알고리즘이다. **[11]**  OC-SVM은 최초 이상치 탐지 문제 해결을 위해 제안되었으며, 대부분의 정상 데이터를 포함시키는 가장 작은 초평면을 그려낸다. **[15]** 이렇게 그린 결정경계를 통해 테스트 샘플 과의 차이를 계산하여 이상치 점수로 활용이 가능하며 점수에 따른 추가적인 정량적 평가가 가능하다. **- 해당 부분은 method에 적합하다고 보임**

 fig14을 보면, 클러스터의 경계에 위치한 일부 데이터를 제외하고, 대부분의 데이터를 포함하는 결정규칙을 만든 것을 알 수 있다. 앞서 수행한 DBSCAN에서 노이즈 제거를 위해 응집도가 떨어지는 데이터 일부를 제거하였는데, 이와 같은 개념으로 데이터 경계에 위치한 애매한 데이터를 제거하였다.

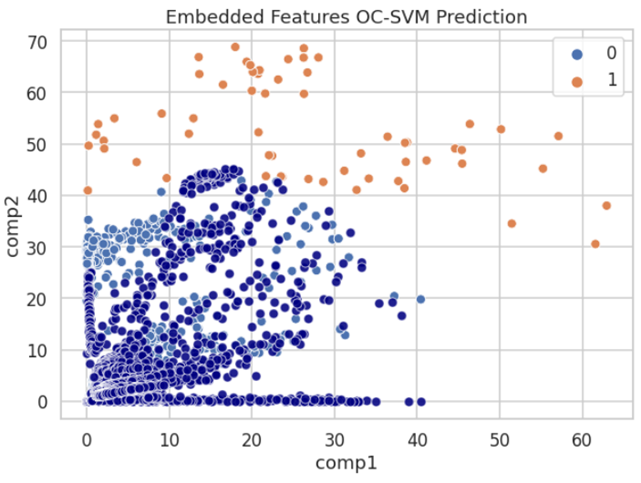


Figure . Embedded features oc-svm prediction

 fig15는 학습 데이터를 통해 형성한 클러스터와 테스트 데이터에 OC-SVM을 적용하여 얻은 이상치 탐지 결과이다. 그림을 보면, 테스트 데이터의 대다수가 정상 클러스터 내에 잘 투영되는 것을 알 수 있으며, 클러스터 바깥쪽으로 소수의 이상치 데이터가 탐지되었다. 이는 OC-SCM이 결정경계 밖으로 벗어나는 데이터에 대해 이상치로 판별한 결과이며, 이와 같은 방법을 통해 LSTM-AE Encoder의 압축 표현형을 활용한 OC-SVM 기반의 이상치 탐지 수행이 가능하다.

**3.2.3 Decision Rule combining OC-SVM and Reconstruction Threshold**

**…**

**…**

**…**

**3.3 Lab-test verification**

클러스터링 그림 추가  및 설명

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Label | comp1 mean | comp2 mean | | DCC\_score | distance | Accuracy | RMSE |
|  |  |  | |  |  |  |  |
| Case A |  |  |  | |  |  |  |
| Case B |  |  |  | |  |  |  |
| Case C |  |  |  | |  |  |  |
| Case D |  |  |  | |  |  |  |
| Case E |  |  |  | |  |  |  |
| Total |  |  |  | |  |  |  |

Table . Results of laboratory test for model validation

1. **Discussion**

비지도학습으로 이루어진 실증 데이터를 사용했다. 그렇기에 정상 데이터에 적절한 데이터로 정제하기 위해 통계적인 방법을 활용하여 진행하였다. 또한 센서로 데이터를 추출하는 동안의 환경을 제약 및 통제시켰다. (이런 방법을 통해 문제와 어떻게 극복할 수 있는지, 추가적인 장점이 어떤 것이 있었는지 작성)

앞서 LSTM-AE를 학습시키는 가정에 ~~순수~~ 정상 데이터를 통한 학습이 전제되어 있다고 언급하였다. LSTM-AE가 ~~순수~~ 정상 데이터를 학습함으로써 ~~정상 데이터의~~ 특징을 더 잘 포착할 수 있고, 해당 결과가 RE기반의 이상치 탐지 모델과, OC-SVM에도 모두 영향을 주기 때문이다. 하지만, 실증 데이터의 특성 상 최초 원본 데이터는 No-Label일 뿐만 아니라, 여러 물리적, 환경적 요소로 인한 노이즈가 포함되어 있다. 따라서 본 연구를 진행하면서 가장 중요하게 생각한 요점이 바로 실증 데이터에 대한 품질 평가와 정제이다. 정상성 검증, 상관도 측정 등의 데이터 탐색 절차를 통해 데이터 품질에 대한 객관적 합리성을 증명하였지만, ~~최초 데이터 정제에 대해선 여러 의심점이 있는 것이 사실이다~~. ~~이러한 의심점을 감안하여,~~ LSTM-AE의 재복원 오차 임계값을 선정하는데, 여러 시행착오를 거쳤으며, 클러스터링과 OC-SVM 초평면 결정과정에서 노이지를 두 번 제거해주었다. ~~이러한 데이터 정제 방식이 과연 최선이였는지에 대해선 여러 시사점이 있을 수 있지만,~~ (객관성 등을 확보하기 여러 과정을 거쳐 최대한 확보하였다 라고 확신하는 방향으로 작성) 연구를 진행함에 있어서 합리적인 선에서 수행해줄 수 있는 절차였다가 판단되어진다.

1. **Conclusion**

본 논문에서는 비지도학습에서 시간의존성에 따른 여러 환경물질과의 상관관계를 고려한 LSTM-AE 및 CLUSTERING 방법을 제안한다. TVOC, CO2, PM2.5, PM10, CH2O, CO, Temperature, Humidity로 총 8가지의 인자로 구성하여 환경 물질을 LSTM-AE로 @@@ @@@이다.

 본 연구에서는 환경센서에서 이상치가 발생할 수 있는 lab test를 진행하였다. lab test에 대해 본 기술을 적용하였고, 이에 따라 MAE와 RMSE로 보여지는 정상값과 이상값의 차이를 확인하였다. 이 과정에서 우리의 모델과 다양한 상황에서 효과적인지 판단하기 위해 5가지의 상황을 나누어 진행했다. 5가지의 상황에서의 군집이 각각 정상치와 이상치로 잘 나뉨을 확인할 수 있었고 이를 통해 모델의 검증 또한 이뤄졌다. (검증의 방법을 랩테스트를 진행하여 좀 더 명확한 성능치를 확보함이라는 명확한 contribution을 제시)

 제안된 기술과 관련하여 강조해야 할 가장 중요한 요소는 사용성 (명확하게 풀어쓰기) 이다. 시계열 데이터 중 다변수를 사용하는 모든 센서에 적용할 수 있다. 특히 라벨링이 되어있지 않은 인자들의 경우, 상관관계를 고려하여 진행하면 된다. 또한 기존 논문에서 이상치 탐지 알고리즘을 구현할 때 정상 데이터로 훈련을 시킨다는 가정이 있었지만, 정상 데이터와 이상 데이터가 혼합되어 있는 상황에서 데이터를 정제 이후 진행시켰다는 부분에서 차이가 있다. 일반적인 상황에서 환경 센서를 통해 데이터를 수집할 때, 정상치의 경우와 이상치의 경우를 확실히 구분하지 못하기 때문에 혼합 데이터에서의 해결방안이 필요하다.  이러한 이상치의 전가는 환경센서 뿐만 아니라 자율 주행, 스마트시티 등 다양한 분야에서 요구되고 있다. 환경 센서에서 비지도학습의 주어진 데이터 내에서 이상치를 탐지하여 실내공기오염을 줄이는데 있어 기여할 수 있다. (각자 맡은 부분에 대한 conclusion에 작성하기. 해당 논문의 기여점이 드러나도록)

1. **Acknowledgements**

This work was supported by a Korea Environmental Industry & Technology Institute (KEITI) grant funded by the Korea government (Ministry of Environment). Project No. RE202101551, the development of IoT-based technology for collecting and managing Big data on environmental hazards and health effects.

This section is optional. Collate acknowledgements in a separate section at the end of the article before the references and do not, therefore, include them on the title page, as a footnote to the title or otherwise. List here those individuals who provided help during the research (e.g., providing language help, writing assistance or proof reading the article, etc.).We encourage the inclusion of a brief outline of the contributions of authors in this section, for example: "Author contributions: T.L., E.K. and H.S. designed the research; T.L., E.K., W.L. and H.S. performed the research; T.L. and E.K. analyzed the data; and T.L., W.L. and H.S. wrote the paper." Other contribution categories may include model coding, programming support, research conception, etc.

1. **References**

1. Single author: the author's name (without initials, unless there is ambiguity) and the year of publication;

2. Two authors: both authors' names and the year of publication;

3. Three or more authors: first author's name followed by 'et al.' and the year of publication.

Citations may be made directly (or parenthetically). Groups of references can be listed either first alphabetically, then chronologically, or vice versa.

Examples: 'as demonstrated (Allan, 2000a, 2000b, 1999; Allan and Jones, 1999)…. Or, as demonstrated (Jones, 1999; Allan, 2000)… Kramer et al. (2010) have recently shown …'

List: References should be arranged first alphabetically and then further sorted chronologically if necessary. More than one reference from the same author(s) in the same year must be identified by the letters 'a', 'b', 'c', etc., placed after the year of publication.

Examples:

Reference to a journal publication:

Van der Geer, J., Hanraads, J.A.J., Lupton, R.A., 2010. The art of writing a scientific article. J. Sci. Commun. 163, 51–59. https://doi.org/10.1016/j.Sc.2010.00372.

Reference to a journal publication with an article number:

Van der Geer, J., Hanraads, J.A.J., Lupton, R.A., 2018. The art of writing a scientific article. Heliyon. 19, e00205. https://doi.org/10.1016/j.heliyon.2018.e00205.

Reference to a book:

Strunk Jr., W., White, E.B., 2000. The Elements of Style, fourth ed. Longman, New York.

Reference to a chapter in an edited book:

Mettam, G.R., Adams, L.B., 2009. How to prepare an electronic version of your article, in: Jones, B.S., Smith , R.Z. (Eds.), Introduction to the Electronic Age. E-Publishing Inc., New York, pp. 281–304.

Reference to a website:

Cancer Research UK, 1975. Cancer statistics reports for the UK. http://www.cancerresearchuk.org/aboutcancer/statistics/cancerstatsreport/ (accessed 13 March 2003).

Reference to a dataset:

[dataset] Oguro, M., Imahiro, S., Saito, S., Nakashizuka, T., 2015. Mortality data for Japanese oak wilt disease and surrounding forest compositions. Mendeley Data, v1. https://doi.org/10.17632/xwj98nb39r.1.

Reference to software:

Coon, E., Berndt, M., Jan, A., Svyatsky, D., Atchley, A., Kikinzon, E., Harp, D., Manzini, G., Shelef, E., Lipnikov, K., Garimella, R., Xu, C., Moulton, D., Karra, S., Painter, S., Jafarov, E., & Molins, S., 2020. Advanced Terrestrial Simulator (ATS) v0.88 (Version 0.88). Zenodo. <https://doi.org/10.5281/zenodo.3727209>.

**Reference In 2.3 , 3.2, 3.3**

**1.** **[MDPI sensors]** Deep-Compact-Clustering Based Anomaly Detection Applied to Electromechanical Industrial Systems

**2.**  **[ELSEVIER]** Minimum volume ellipsoid classification model for contamination event detection in water distribution systems + 4.3 시계열 분석 파라미터를 시행착오 방식으로 결정함. 그 이유는 비지도 분류 메서드의 경우 모델 자체로 파라미터를 정할 수 없기 때문임. 결정규칙 부분에 추가.

**3.**     **[WILEY]** Unsupervised Anomaly Detection Based on Deep Autoencoding and Clustering

**4.**     **[WACV : CVF]** Deep Unsupervised Anomaly Detection

**5.**  **[AAAI]** A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data

**6.** **[MDPI Symmetry]** Unsupervised Anomaly Detection Approach for Time-Series in Multi-Domains Using Deep Reconstruction Error

**7.** **[MDPI Sensors]** FuseAD : Unsupervised Anomaly Detection in Streaming Sensors Data by Fusing Statistical and Deep Learning Models + 1 : Ensemble 기반의 결정규칙이 이상치 탐지 분야에서 뛰어난 이유와 접근 구도

**8.**  **[IEEE]** Unsupervised Anomaly Detection With LSTM Nerual Networks

**9.** **[ELSEVIER]** Algorithm optimization and anomaly detection simulation based on extended Jarvis-Patrick clustering and outlier detection

**10.**  **[IEEE Access]** Sequential Fault Diagnosis Based on LSTM Neural Network

**11.**  **[ELSEVIER]** Patient classification as an outlier detection problem : An application of the One-Class Support Vector Machine

**12.** **[ELSEVIER]** Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management

**13.**  **[MDPI Water]** Change Point Enhanced Anomaly Detection for IoT Time Series Data

**14.**  **[AAAI]** A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise

**15.**  **[ELSEVIER]** Soft clustering using weighted one-class support vector machines

**16. [ELSEVIER]** Ensemble method based on ANN to estimate air pollution health risks