https://doi.org/10.7840/kics.2024.49.10.1436

CNN 기반의 유해화학물질 분류 및 판독을 위한 유해화학물질 학습데이터셋 구축 방법에 관한 연구

김 연 진◆, 최 갑 용*, 김 경 배°

A Study on the Construction Method of Artificial Intelligence Training Data to Classify and Detect for Hazardous Chemicals

Yeon-Jin Kim◆, Kap-Yong Choi*, Gyeong-Bae Kim°

요 약

최근 화학산업이 발전함에 따라 다양한 화학사고가 증가하고 있다. 증가하는 화학사고에 대응하고자 인공지능 기술을 접목한 유해화학물질 사고대응 기술에 대한 다양한 연구가 수행되고 있다. 영상 및 이미지 기반의 인공지능 유해화학물질 판독시스템은 사고 유해화학물질의 정확한 판독을 위한 충분한 학습이 가능한 양의 학습데이터가 필요하다. 그러나, 유해화학물질이 가지는 위험성으로 인해 데이터 구축에 어려움이 있어 현재 유해화학물질 판독을 위한 인공지능 연구용 학습데이터는 매우 부족하다. 따라서 본 논문에서는 물질의 상태, 시각적 특징의 유무 등 유해화학물질 및 화학사고의 특성을 반영한 인공지능 학습데이터셋 구축 방법을 제안한다. 유해화학물질 및 화학사고의 특성에 따른 학습데이터셋 구축 방법을 따라 자체 유해화학물질 실험을 통한 원시데이터 수집 및 확보, 데이터 전처리 및 가공, 학습데이터의 어노테이션 과정을 거쳐 유해화학물질 9종에 대한 학습데이터셋 약 20만 장을 구축하였다. 구축된 데이터셋은 학습과 검증을 위해 8:1:1의 비율로 나눠 학습, 검증, 테스트데이터로 활용하였다. CNN에 기반한 유해화학물질 판독의 결과로 평균 약 90%의 유해화학물질 판독 정확도가 도출되었고, 판독 결과는 실제 화학사고 현장에서 사고물질을 추정해 줌으로써, 현장대원에게 사고물질에 대해 적절하고 신속한 대응을 지원할 수 있을 것으로 기대된다.

키워드 : 유해화학물질, 재난재해, 학습데이터, 데이터셋 구축, 인공지능
Key Words : Hazardous chemicals, Disaster, Training data, Dataset construction, Artificial intelligence

ABSTRACT

Recently, as the chemical industry develops, various chemical accidents are increasing. In order to respond to the increasing number of chemical accidents, various studies are being conducted on hazardous chemical accident response technology that incorporates artificial intelligence technology to detect chemical substances. Video and image-based artificial intelligence hazardous chemical detection systems require a sufficient amount of training data to enable accurate detecting of hazardous chemicals in accidents. However, due to the risk of hazardous chemicals, it is difficult to construct data, so currently, training data for artificial intelligence

※ 본 연구는 과학기술정보통신부와 정보통신산업진흥원 주관으로 소방청 컨소시움에서 수행하는 "AI융합 유해화학물질 판독 시스템 사업" (2022~2024)의 지원을 받았음
◆ First Author : Seowon University Department of Industry Cooperation Foundation, anne6497@naver.com, 정회원
° Corresponding Author : Seowon University, Dept. of Software, gbkim@seowon.ac.kr, 종신회원
* National Fire Agency, ky6622@korea.kr, 정회원
논문번호 : 202405-092-0-SE, Received May 7, 2024; Revised June 20, 2024; Accepted July 1, 2024

research for detecting hazardous chemicals is very insufficient. Therefore, this paper proposes a method of constructing an artificial intelligence training dataset that reflects the characteristics of hazardous chemicals and chemical accidents, such as the state of the material and the presence or absence of visual features. Following the proposed training dataset construction method, about 200,000 training datasets for 9 types of hazardous chemicals were constructed through collecting and securing raw data through self-experimentation of hazardous chemical , data processing, and annotation of the training data. The constructed dataset was divided in a ratio of 8:1:1 for training and validation and used training, validation, and test data. As a result of CNN-based hazardous chemical detecting, an average hazardous chemical detecting accuracy of about 90% was obtained. The detecting results are expected to provide an estimate of the accidental hazardous chemicals at the actual chemical accidents, thereby supporting appropriate and rapid response to the accidental substances to on-site firefighter.

Ⅰ. 서 론

화학물질은 인류 문명의 발전과 의류, 석유화학 등 우리의 일상생활 또는 산업활동에 필수적으로 사용되고 있다[1]. 이처럼 화학산업이 발전하면서 화학물질의 사용량 및 유통량 또한 증가함에 따라 다양한 화학사고가 발생하기 시작하였다. 2012년 9월 구미에서 발생한 불산 누출사고는 최초 신고 당시에 소방서는 누출 물질이 불산임을 인지하였으나, 화학물질에 대한 인식과 전문성 부족으로 인한 적절한 초기대응의 부재로 큰 피해를 초래하였다[2]. 이후 2015년부터 위해관리계획제도가 시행되어 화학사고 발생건수가 일시적으로 감소하였으나, 2019년에서 2021년까지 발생한 화학사고는 58건, 75건, 92건으로 다시 매년 증가하는 추세를 보이고 있다[3]. 특히 국내 화학사고는 주로 화학산업단지에서 화학물질 생산·취급시설의 노후화 및 결함, 작업자의 부주의 등에 의해 발생하며, 화학물질의 특성상 높은 잔류성과 확산성으로 인해 환경, 인명, 재산 등에 막대한 피해를 유발하는 경우가 많다[4,5]. 소방청은 증가하는 화학사고에 대응하고자 전국 주요 화학산업단지를 중심으로 7개의 전문 화학구조센터를 운영하고 있으나, 실제 화학사고가 발생했을 때 현장관계자 및 전문가를 통한 사고물질 파악에 많은 시간이 소요되고 사고물질에 대해 비전문가인 현장대원의 전문적인 대응이 어렵다는 한계가 존재한다[6]. 화학사고가 발생하지 않도록 사전에 대비하는 것이 중요하지만, 화학사고가 발생하였다면 사고물질의 빠른 파악과 적절한 대응방안을 찾아 신속하게 대응하는 기술이 필요하다.

다양해지는 유해화학물질 사고에 대응하고자 최근 국내외에서 화학물질을 감지하고 판단하기 위한 여러 가지 연구가 진행되고 있다. 다양한 화학종 검출을 위해 개방형 적외선분석기를 이용한 실시간 가스 검출 분석 연구[7], 화학물질 오염, 화재, 폭발 등 재난상황의 사고예방 및 초기대응 강화를 위한 멀티센싱 기반의 유해화학물질 감지시스템 연구[1] 등이 수행되었다. 또한 물질 판별에 핵심이 되는 지식 쿼리를 지원하는 AllegroGraph를 활용하고 딥러닝 기반의 화학구조 정보 기반 증상예측 모델을 제안하는 SEARCH ver. 2[8], 유해화학물질의 누출/화재사고 발생 시 CNN 기반의 인공지능 기술을 활용한 유해화학물질 탐지 및 판독 방법에 대한 연구[9]가 수행되고 있다. 유해물질을 원격 탐지할 수 있는 CNN 기반의 유해물질 분류 및 식별 방법에 관한 연구[10]에서는 유해물질에 대한 1,800개의 초분광 이미지가 포함된 데이터셋을 활용하여 93%의 분류 정확도를 나타내고, 머신러닝을 활용한 유해 기름 유출 감지 연구[11]에서는 RGB와 적외선 이미지를 사용한 CNN 훈련을 통해 89%의 정확도를 달성하는 등 사고물질의 신속한 식별을 위한 인공지능 기술을 접목한 유해화학물질 사고대응 기술에 대한 연구가 수행되고 있다.

화학사고 발생 시 사고물질 식별과 특성 정보 파악 및 현장대응 지원을 위한 인공지능 기반의 유해화학물질 판독시스템을 위해서는 유해화학물질 사고영상 수집 및 학습데이터를 구축하고 판독 알고리즘을 통해 사고물질의 종류를 판독해야 한다. 유해화학물질을 정확하게 판독하기 위해서는 유해화학물질 데이터에 대한 충분한 학습이 가능한 많은 양의 학습데이터가 필요하다. 적은 양의 데이터셋을 활용하여 학습을 진행하게 되면 학습데이터의 과대적합으로 인해 유해화학물질 판독 결과의 신뢰성이 저하된다[12]. 그러나 인공지능 기술을 활용한 유해화학물질 판독 분야는 아직 미성숙한 분야이며, 유해화학물질이 가지는 위험성으로 인해 데이터 구축에 어려움이 있으므로 현재 유해화학물질 분야에서의 인공지능 연구를 위한 학습데이터셋이 매우 부족하다[8]. 현재 구축되어 있는 유해화학물질 학습데이터

셋은 대부분 화합물 및 화학구조 등을 연구하기 위함이며, 유해화학물질을 판독하기 위한 유해화학물질의 영상 및 이미지 기반의 학습데이터셋은 존재하지 않는다.

따라서 본 논문에서는 유해화학물질 판독시스템에서 활용할 수 있는 유해화학물질의 상태, 시각적 특징, 사고유형 등을 고려한 인공지능 학습데이터셋 구축 방법을 제안한다. 본 논문에서 제안한 학습데이터셋 구축 기준을 따라 구축한 데이터셋의 학습을 통해 사고가 발생한 유해화학물질의 종류를 추정하고 현장대원에게 사전정보를 제시할 수 있다. 이를 통해 실제 유해화학물질 사고가 발생하였을 때, 현장대원들의 신속하고 정확한 대응방안을 지원함으로써 유해화학물질 사고로 인한 환경 및 인적 피해를 감소시킬 수 있는 기대효과를 가진다.

본 논문의 구성은 다음과 같다. 2장에서는 인공지능 학습에 필수적 요소인 데이터셋에 대한 관련 연구를 기술하고, 3장에서는 유해화학물질 학습데이터셋 구축 방법을 제안한다. 4장에서는 유해화학물질 학습데이터셋을 활용한 유해화학물질 판독 결과를 제시하고, 5장에서는 향후 연구를 제시하며 결론을 맺는다.

Ⅱ. 관련 연구

기존의 유해화학물질에 대한 데이터셋은 화학분야의 연구를 위한 데이터셋 중심으로 구축되어 있다. Tox21 Dataset[13]은 미국 국립보건원(NIH)에서 제공하는 화학물질의 독성을 스크리닝하기 위한 데이터셋으로 화학물질의 구조와 독성 효과를 연구하는 데 사용될 수 있으며, 머신러닝 모델을 훈련에 활용할 수 있다. GHS Dataset(Globally Harmonized System of Classification and Labelling of Chemicals)[14]은 화학물질의 분류 및 라벨링에 대한 국제 기준을 제공하여 화학물질의 위험성을 판별하는 데 사용할 수 있다. 미국 환경보호청(EPA)이 제공하는 IRIS(Integrated Risk Information System)[15]는 환경에서 발견되는 화학물질의 건강 위험 평가를 위한 것으로 유해화학물질의 분석과 인식을 위한 연구에 사용할 수 있다.

유해화학물질과 관련된 연구로 화학실험 과정에서 물체 감지 향상을 위한 연구[16]에서는 화학물질 실험을 수행하는 과정이 담긴 영상에서 비커, 플라스크 등 일반적인 화학실험 도구에 대해 라벨링한 데이터셋을 활용하여 화학실험에 이미지 인식 기술을 적용한 연구가 진행되었다. 해당 연구에서 사용된 이미지 데이터셋은 유기화학 실험실에서 촬영한 5,078개의 JPG 이미지로 구성되어 있으나, 단순한 화학물질 실험 과정을 위한 데이터셋이다. 유해화학물질 사고데이터를 활용한 개체명 인식(NER) 추출 방법에 관한 연구[17]에서는 화학사고 정보를 활용한 텍스트 기반의 데이터셋을 활용하였고, 드론을 이용한 가스 누출 탐지 방법에 관한 연구[18]에서는 메탄가스 배출을 촬영한 영상 데이터인 GasVid 데이터셋을 활용하였다. 화학물질의 유사성 측정 및 독성 예측 모델 개발 연구[19]와 분자 구조와 환경 영향 사이의 관계를 학습하고 분자의 환경 영향 예측 연구[20]에서는 화학물질의 분자식, 분자량 등의 정보가 있는 PubChem 데이터를 활용하였다.

유해화학물질에 대한 데이터셋이 존재하긴 하나, 기존에 구축된 데이터셋은 표 1과 같이 주로 화학물질 독성 예측, 위험성 평가 등을 위한 목적으로 구축되었으며 유해화학물질을 판독할 수 있는 영상 및 이미지가 아닌 화학물질 연구자를 위한 분자식, 분자량 등 대부분 화학물질의 분자구조와 관련된 데이터셋이다. 이러한 데이터셋은 순수한 화학물질 자체에 대한 연구를 위해 물질의 화학구조 등과 같은 정보를 제공하고 있어 영상 기반의 인공지능 유해화학물질 판독을 위한 데이터셋으로 활용하는데 한계가 있다. 따라서 본 논문에서 구축하고자 하는 유해화학물질의 분류 및 판독을 위한 시각

표 1. 기존의 유해화학물질 데이터셋의 비교
Table 1. Comparison of existing hazardous chemical datasets

Dataset	Purpose	Characteristic	Format
Tox21	Toxicity prediction model	Chemical Structure	CSV
GHS	Risk assessment of chemicals	Chemical classification and labeling information	Excel, PDF
IRIS	Health risk assessment and analysis of hazardous chemicals	Health risk appraisal data	Database
GasVid	Gas leak detection	Methane gas emission video data	Video
PubChem	Research on measuring similarity of chemicals and predicting environmental impact	Chemical molecular formula and molecular weight	Database

적 특징 기반의 유해화학물질 인공지능 학습데이터셋 구축이 필요하다.

Ⅲ. 유해화학물질 학습데이터셋 구축

3.1 유해화학물질의 특성에 따른 분류

유해화학물질의 상태, 시각적 특징, 사고유형 등을 고려한 인공지능 학습데이터셋의 구축을 위해서는 다음과 같이 유해화학물질의 특성에 기반한 분류를 해야 한다.

첫째, 유해화학물질은 물질별 고유의 색상과 특징을 가지며, 고체・액체・기체 중 하나의 상태로 존재한다[9]. 유해화학물질은 다른 물질과 혼합되거나 공기에 노출될 경우 다른 색상이나 상태로 변화할 수 있다. 유해화학물질은 표 2와 같이 물질별로 물질의 상태, 시각적 특징의 유무, 연소 여부, 사고유형 등에 따라 세부적으로 분류할 수 있다.

표 2. 유해화학물질 분류 기준
Table 2. Classification criteria of hazardous chemicals

No.	Category	Classification criteria
1	State	Solid, Liquid, Gas
2	Visual characteristics	Presence, Absence
3	Combustion	Combustible, Noncombustible
4	Types of incidents	Fire, Explosion, Leakage

둘째, 유해화학물질은 시각적 특징의 유무에 따라 구분할 수 있고, 연소 여부에 따라 가연성, 불연성으로 구분할 수 있다. 가연성 물질에서 발생할 수 있는 사고유형으로는 화재/폭발/누출사고가 있고, 불연성 물질에서 발생할 수 있는 사고유형으로는 누출사고가 있다.

따라서 유해화학물질 특성에 따른 분류 기준을 세웠으며, 국내에서 사고발생 빈도가 높거나 취급업체 수가 많은 유해화학물질을 물질 선정 기준으로 하여 유해화학물질 17종을 선정하였다. 선정한 유해화학물질 17종을 분류 기준에 따라 표 3과 같이 분류하였다. 본 논문에서는 선정한 유해화학물질 17종 중 시각적 특징을 나타내면서 사고발생 건수나 취급업체 수가 많은 물질인 질산, 브롬, 수산화나트륨, 질산칼륨, 시너, 염화제2구리, 염소, 톨루엔, 메틸알코올 총 9종을 유해화학물질 학습데이터셋 구축 대상 물질로 선정하였다.

표 3. 유해화학물질 특성에 따른 분류 기준
Table 3. Classification criteria according to the characteristics of hazardous chemicals

Visual characteristics	Combustion	Types of incidents	State	Substance name
Presence	Combustible	Fire	Solid	Sodium hydroxide, Potassium nitrate, Copper(II) chloride
			Liquid	Thinner, Kerosene, Nitrobenzene, Toluene, Methanol
		Leakage	Gas	LPG
	Noncombustible	Leakage	Gas	Chlorine (Oxidizing)
			Liquid	Mercury
			Gas	Nitric acid, Bromine
Absence	Combustible	Fire	-	-
		Leakage	Liquid	Ammonia
	Noncombustible	Leakage	Solid	Potassium fluoride
			Liquid	Sulfuric acid
			Gas	Nitrogen

본 논문에서는 시각적 인식 기반의 학습을 통해 유해화학물질을 판독하므로 물질별로 고유의 색상, 연소 시 나타나는 불꽃색 등 시각적 특징을 나타내는 물질에 대한 데이터 확보가 필요하다. 유해화학물질 학습데이터셋은 크게 유해화학물질 9종에 대한 원시데이터 수집, 데이터 전처리 및 가공, 학습데이터의 어노테이션 과정을 거쳐 구축한다.

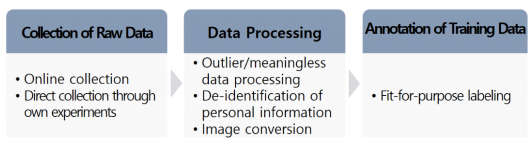


그림 1. 유해화학물질 학습데이터셋 구축 과정
Fig. 1. Construction process of hazardous chemicals training data set

3.2 유해화학물질 원시데이터 수집

유해화학물질을 판독하기 위해서는 유해화학물질이 나타내는 시각적 특징에 대한 인공지능 학습이 필요하며, 인공지능 학습을 위한 대량의 학습데이터는 필수적인 요소이다. 정확도를 높이기 위해 명확한 데이터셋을 활용하는 것은 인공지능 분야의 핵심이며, 대량의 학습

데이터를 활용한 학습은 모델 성능 및 정확도를 향상시킬 수 있으므로 매우 중요하다. 학습데이터 구축을 위해서 먼저 유해화학물질 원시데이터를 수집하는 과정이 필요하다. 본 논문에서는 유해화학물질의 불꽃이나 연기의 색상을 학습하여 물질 종류를 판독하는 것이 목적이기 때문에 유해화학물질 사고현장이나 물질 실험 등의 영상데이터를 수집하여야 한다. 영상 기반의 유해화학물질 원시데이터는 인터넷 검색을 통해 온라인으로 수집하거나, 자체 실험을 통해 직접 수집한다.

직접 수행하는 실험만으로는 물질의 특성이 뚜렷하게 나타나지 않거나, 위험성이 높은 물질에 대해서는 인터넷 검색을 통해 원시데이터를 확보한다. 그러나 온라인 수집을 통한 원시데이터는 순수한 물질이 아닌 혼합물에 대한 실험영상이 대부분이다. 또한 화학사고 영상은 화질이 현저하게 낮거나 흔들림이 심한 영상이 많으므로 학습데이터로 사용하기에 적절하지 않은 경우가 많다.

그림 2. 온라인 수집을 통한 원시데이터 및 한계
Fig. 2. Raw data and limitations of online collection

따라서 본 논문에서는 주로 자체 실험을 통해 유해화학물질 원시데이터를 직접 수집하여 학습데이터셋을 구축하였다. 본 논문에서는 색상의 특징을 나타내는 물질 중 가연성, 불연성을 나타내는 특성으로 인해 발생하는 사고유형인 화재/누출실험을 수행하였고, 잔여 가스 처리 및 집진시설이 갖춰진 장소에서 안전수칙을 준수하며 실험을 수행하였다. 유해화학물질 원시데이터로 활용할 영상을 촬영하기 위해 DSLR 카메라, 스마트폰 카메라, 열화상 카메라 등 다양한 카메라를 사용하였다.

또한 학습데이터의 다양성을 충족하기 위해 농도, 조도, 촬영거리 등 표 4와 같이 여러 가지 조건을 만족하여 실험을 수행하였다.

표 4. 유해화학물질 원시데이터의 다양성
Table 4. Diversity of raw data on hazardous chemicals

Conditions	Contents
Density	Low, General, High
luminous intensity	luminous range adjustment according to outdoor/indoor scene
Distance	1m, 2m, 3m etc.
Angle	Horizontal angle 15˚, 45˚, 90˚, 180˚ etc.
Method	Outdoor scene, Indoor scene
FPS	30fps

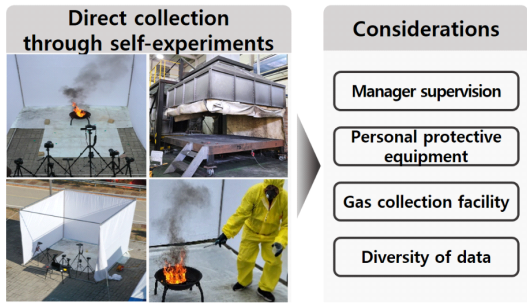


그림 3. 직접 수집을 통한 원시데이터 및 고려사항
Fig. 3. Raw data and considerations of direct collection

위와 같은 조건을 만족하여 자체 실험을 통해 유해화학물질 9종의 원시데이터를 직접 수집하였으며, 유해화학물질 목록 및 실험 유형은 표 5와 같다.

표 5. 자체 실험을 통해 직접 수집한 9종의 유해화학물질 목록
Table 5. List of 9 hazardous chemicals collected directly through self-experimentation

No.	Substance name	Combustion	Types of incidents	State	Experiment type
1	Nitric acid	Noncombustible	Leakage /Fire	Liquid	Leakage (Gas)
2	Bromine	Noncombustible	Leakage	Liquid	Leakage (Gas)
3	Sodium hydroxide	Combustible	Leakage /Fire	Solid	Fire (Solid)
4	Potassium nitrate	Combustible	Fire	Solid	Fire (Solid)
5	Thinner	Combustible	Leakage /Fire	Liquid	Fire (Liquid)
6	Copper(II) chloride	Combustible	Leakage	Solid	Fire (Solid)
7	Chlorine	Oxidizing	Leakage	Gas	Leakage (Gas)
8	Toluene	Combustible	Fire	Liquid	Fire (Liquid)
9	Methanol	Combustible	Leakage /Fire	Liquid	Fire (Liquid)

3.3 유해화학물질 데이터 전처리 및 가공

유해화학물질 학습에 적합한 형태의 데이터셋을 구축하기 위해 수집한 원시데이터에 대한 전처리 및 가공 과정이 필요하다. 원시데이터의 이상치나 불필요한 데이터를 처리하는 전처리 과정을 통해 데이터의 질을 향상시킬 수 있다. 수집한 원시데이터에는 실제 학습에 필요한 부분 외 불필요한 부분까지 모두 포함되어 있다. 본 논문에서는 불꽃이나 연기 변화가 큰 부분의 데이터 위주로 학습하기 때문에 영상 내에서 중복 및 유사한 부분의 삭제, 불꽃과 연기의 구분이 모호한 부분을 삭제하는 등 무의미한 데이터의 제거가 필요하다. 또한 영상 내에는 자동차 번호판이나 인물의 얼굴 등이 노출되는 경우가 있는데, 이처럼 개인정보에 해당하는 부분의 비식별화 과정도 필요하다.

유해화학물질은 특정 온도에서 물질의 고유한 특성이 발현되므로 물질 판독을 위한 영상 및 이미지의 취득을 위해서는 일정시간이 소요된다. 이러한 물질의 특성을 파악하기 위해서는 시계열 분석 기법을 이용하여 특정 시점이나 구간에서 발생하는 물질 고유의 색상발현을 연속적으로 분석해야 하므로 시간의 경과에 따라 변화하는 불꽃의 색상과 크기 등 초기-중기-말기까지 전 단계에서 발생하는 모든 특징이 포함된 인공지능 학습 데이터셋의 취득이 필요하다. 따라서 본 논문에서는 연속적 이미지 기반의 분석을 위해 전처리가 완료된 영상 데이터를 프레임 단위로 분할하여 각 프레임에서 시간의 흐름에 따른 변화나 패턴을 발견할 수 있도록 이미지 데이터로 변환하였다. 이미지 분할은 jupyter notebook에서 파이썬 코드를 사용해 프레임 단위로 분할하며, 유해화학물질 9종의 모든 영상에 대해 반복 수행하여 이미지로 변환한다.

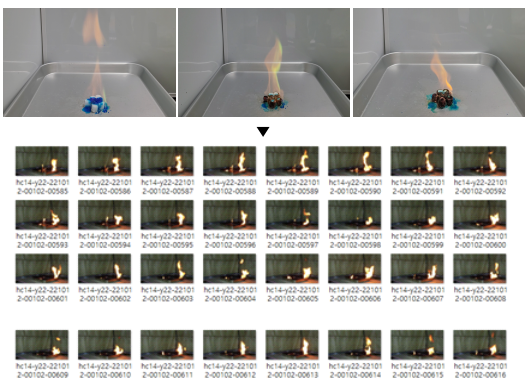


그림 4. 연속적 이미지 기반 분석을 위한 이미지 분할 과정
Fig. 4. Image segmentation process for continuous image-based analysis

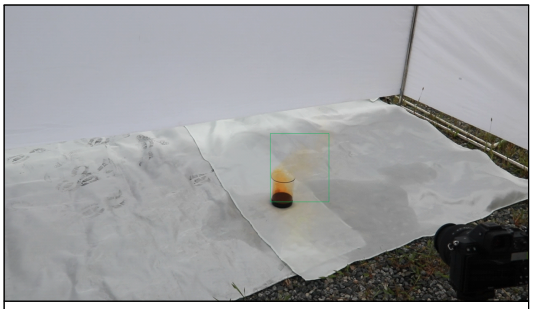
3.4 유해화학물질 학습데이터 어노테이션

원시데이터인 유해화학물질 영상을 이미지로 변환하는 과정을 거친 후, 유해화학물질 이미지 데이터에 대해 목적에 부합하는 라벨을 부착하기 위한 어노테이션 과정이 필요하다. 많은 양의 데이터도 중요하지만, 높은 품질의 데이터를 사용해 학습시키는 것도 결과에 주요한 영향을 끼친다. 어노테이션을 수행하면서 주관적 편향이 데이터셋에 반영되면, 인공지능 모델은 일관되지 않은 기준을 학습하게 되어 판독 정확도가 떨어질 수 있다. 따라서 학습 목적에 적합한 기준과 방법을 따른 데이터 어노테이션 과정은 매우 중요하며, 본 논문에서는 어노테이션 과정 및 기준에 대한 소방청 화학물질 사고대응 담당자 및 화학물질안전원의 전문가 자문을 통해서 연구가 진행되었다. 표 6과 같이 유해화학물질 이미지의 어노테이션 정보로는 사고유형, 객체정보, 기상정보, 물질정보, 위치정보 등이 있다. 사고유형 클래스는 화재/폭발/누출, 객체정보 클래스는 화염/연기/액체 등이 있다. 확산모델 등에 사용할 수 있는 기상정보 클래스는 습도/풍속/기온 등이 있고, 그 외 물질정보, 위치정보 등 유해화학물질 데이터를 학습하기 위해 필요한 다양한 클래스의 어노테이션이 필요하다.

표 6. 유해화학물질 학습데이터 어노테이션 정보
Table 6. Annotation information of Hazardous chemicals training data

Disviion	Class(example)
Types of incidents	Fire, Explosion, Leakage etc.
Information of object	Flame, Smoke, Liquid
Information of weather	Humidity(50%), Wind speed(3.4m/s), Temperatures(17°) etc.
Information of substance	Nitric acid, Bromine, Chlorine etc.
Information of location	Cheongju-si, Chungcheongbuk-do

본 논문에서는 유해화학물질 판독시스템의 성능 향상과 시각적 특징 인식 기반의 학습을 위해서 바운딩박스 기법을 사용하여 데이터 라벨링을 수행하였다. 이미지 내에서 나타나는 화염(flame), 연기(smoke), 액체(fluid)에 대해 어노테이션한 정보를 활용하였고, 어노테이션 결과는 json 파일로 도출되며, 어노테이션 이미지와 json 파일 예시는 그림 5와 같다. json 파일에는 브로민 기체 누출 이미지에서 연기(flame)를 어노테이션한 라벨 정보와 이미지에 대한 메타데이터가 포함되어있다.



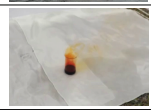

```
{
    "version": "5.3.0",
    "flags": {},
    "shapes": [
        {
            "label": "smoke",
            "points": [
                [
                    973.99968,
                    549.9997199999999
                ],
                [
                    1234.99872,
                    712.9998
                ]
            ],
            "gruop_id": null,
            "description": "",
            "shape_type": "rectangle",
            "flags": {}
        }
    ],
    "imagePath":
"..\\image\\hc04-y21-230828-001-00001.jpg",
    "imageData": null,
    "imageHeight": 1080,
    "imageWidth": 1920,
    "metadata": {
        "cam_index": "1",
        "distance": "193.0",
        "horizontal_angle": "59.0",
        "cam_location": "right",
        "origin_video_name": "MVI_6584.MP4",
        "trimmed_video_name": "hc04-y21-230828-
001.mp4",
        "fps": 29.97,
        "bit_rate": 24,
        "photographer": "서원대학교 산학협력단"
    }
}
```


그림 5. 유해화학물질 학습데이터 어노테이션 결과
Fig. 5. Annotation results of Hazardous chemicals training data

Ⅳ. 유해화학물질 학습데이터셋의 학습 및 결과

본 논문에서는 유해화학물질 원시데이터 수집과 전처리 및 가공 과정을 거쳐 유해화학물질 9종에 대해 약 20만 장의 학습데이터셋을 구축하였다. 사고유형에 따른 유해화학물질 9종의 학습데이터셋 예시는 표 7과 같다.

표 7. 사고유형에 따른 유해화학물질 학습데이터셋 예
Table 7. Example of hazardous chemical training dataset according to incident type

Experimental substance	Training data	Types of incidents	Class
Nitric acid		Leakage /Fire	Smoke
Bromine		Leakage	Smoke
Sodium hydroxide		Leakage /Fire	Flame
Potassium nitrate		Fire	Flame
Thinner		Leakage /Fire	Flame Smoke
Copper(II) chloride		Leakage	Flame
Chlorine		Leakage	Smoke
Toluene		Fire	Flame Smoke
Methanol		Leakage /Fire	Flame

본 논문에서는 구축된 유해화학물질 인공지능 학습데이터셋의 유용성을 검증하기 위해서 유해화학물질 판독을 목표로 시각 인지에 널리 사용되는 대표적인 모델인 CNN 기반의 지도학습을 통한 학습을 진행하였다.

CNN은 영상이나 이미지 데이터 분석에 특화된 인공신경망 구조로, 입력 데이터의 공간적 특성을 파악하고 패턴을 인식하는 데 뛰어난 성능을 보이므로 이미지를 기반으로 하는 유해화학물질 판독에 적합한 모델이다. 유해화학물질 학습데이터셋의 각 이미지에 라벨을 지정하여 유해화학물질의 종류를 판독할 수 있도록 학습모델을 설계하였고, 표 8과 같은 학습환경에서 유해화학물질 화재 및 누출 이미지 중 157,487장의 이미지는 학습데이터, 19,687장의 이미지는 검증데이터, 19,690장의 이미지는 테스트데이터로 학습데이터:검증데이터:테스트데이터=8:1:1의 비율로 학습과 검증을 수행하였다.

과적합을 방지하고 최적의 하이퍼파라미터를 찾기 위해 그림 6과 같은 학습모델을 구성하였으며, batch size=256, epoch=20, Activation='relu', Dropout=0.5의 값을 사용하여 학습을 진행하였다. batch size가 너무 작을 경우, 모델의 가중치가 과도하게 갱신되기 때문에 본 논문에서 활용한 데이터 수의 적합하도록 batch size를 조정하였다. 학습 횟수에 따라 훈련 손실(training loss), 검증 손실(validation loss), 훈련 정확도(training accuracy), 검증 정확도(validation accuracy)는 그림 7과 같은 그래프를 나타낸다. 학습 횟수가 증가할수록 훈련 손실과 검증 손실의 값은 감소하여 0에 근접하게 수렴하는 결과를 나타내었고, 훈련 정확도와 검증 정확도의 값은 증가하며 1에 근접하게 수렴하는 결과가 나타났다.

표 8. AI 유해화학물질 판독 시스템 환경
Table 8. AI hazardous chemical detection system environments

Disviion	Specification
OS	Ubuntu 20.04.6 LTS
CPU	Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz
Memory	94.0 GiB
GPU	NVIDIA GeForce RTX 3090

표 9. 인공지능 데이터셋의 분류
Table 9. Classification of AI datasets

Classification of dataset	Number of dataset
train set	157,487 (80%)
validation set	19,687 (10%)
test set	19,690 (10%)

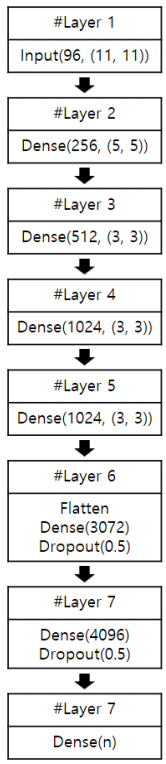


그림 6. 학습모델 구조
Fig. 6. Structure of training model

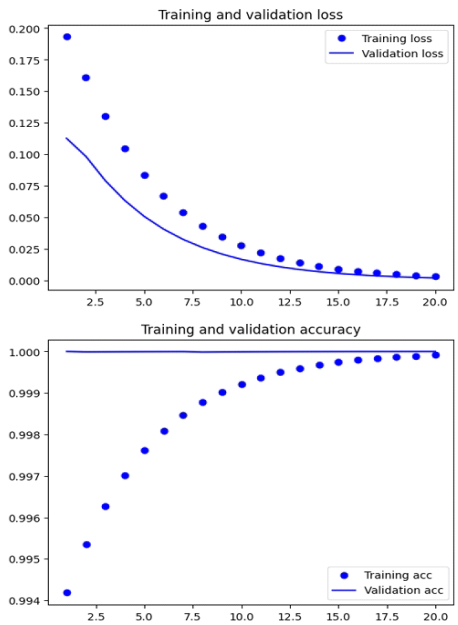


그림 7. 학습 횟수에 따른 훈련, 검증의 손실과 정확도 그래프
Fig. 7. Loss and accuracy graph of training and validation according to epochs

학습을 완료한 학습모델의 성능평가 지표로는 정확도(Accuracy)와 정밀도(Precision), 재현율(Recall) 성능을 동시에 고려한 F1-Score를 사용하였다. 학습결과로 유해화학물질 판독 정확도는 평균 약 90%이상을 나타낸다. 그러나 시너와 톨루엔과 같은 유류계열의 화학물질에서는 물질별 색상이 유사하고 서로 뚜렷하게 구분되는 고유의 특징이 보이지 않아, RGB 이미지에 대한 색상 기반의 분석에서는 판독 정확도가 낮은 문제가 발생한다.

V. 결 론

화학산업이 발전함에 따라 화학사고가 증가하고 있으며, 화학사고로 인해 초래되는 큰 피해의 방지 및 화학사고 예방을 위해 화학물질의 감지와 판단 방법과 관련된 다양한 연구가 진행되고 있다. 그러나, 유해화학물질이 지니는 위험성으로 인해 데이터 구축에 어려움이 있어 유해화학물질을 판독하기 위한 적절한 학습데이터셋은 존재하지 않으며 관련 연구 또한 미비하다. 따라서 본 논문에서는 유해화학물질 판독시스템에서 활용할 수 있는 유해화학물질 인공지능 학습데이터셋 구축 방법을 제안하고 구축하였다.

본 논문에서는 물질의 상태, 시각적 특징의 유무 등 유해화학물질 및 화학사고의 특성을 반영한 인공지능 학습데이터셋 구축 방법을 제안한다. 유해화학물질은 물질의 상태, 색상, 연소 유무, 사고유형 등에 따라 세부적으로 분류할 수 있다. 제안된 구축 방법을 통해 사고발생 건수나 취급업체 수가 많은 물질 등의 통계자료를 기반으로 유해화학물질 9종(질산, 브롬, 수산화나트륨, 질산칼륨, 시너, 염화제2구리, 염소, 톨루엔, 메틸알코올)에 대한 학습데이터셋을 구축하였다. 유해화학물질이 가지는 고유한 특징을 고려하여 온라인 수집 및 자체 실험을 통해 원시데이터를 수집하였다. 온라인 수집을 통한 원시데이터는 영상의 흔들림, 화질 저하 등으로 인해 학습데이터로 사용하기에 적합하지 않으므로 본 논문에서는 자체 실험을 통해 학습데이터의 다양성을 만족하며 화재/누출의 사고유형에 따른 원시데이터를 수집하였다.

원시데이터 영상 내에서 무의미한 데이터의 제거와 개인정보에 해당하는 부분의 비식별화 등 원시데이터의 전처리 및 가공 과정을 거친다. 연속적 이미지 기반의 분석을 위해 영상을 프레임 단위로 분할하여 이미지로 변환한다. 변환된 유해화학물질 이미지 내에서 나타나는 화염(flame), 연기(smoke), 액체(fluid)에 대한 객체정보를 바운딩박스 기법을 사용하여 어노테이션 한다.

본 논문에서는 유해화학물질 원시데이터 수집과 전처리 및 가공 과정을 거쳐 유해화학물질 9종에 대해 약 20만 장의 학습데이터셋을 구축하였으며, CNN 기반의 지도학습을 통해 학습데이터:검증데이터:테스트데이터=8:1:1의 비율로 학습과 검증을 수행하였다. 학습결과의 성능지표로써 정확도와 F1-Score를 사용하였으며, 학습결과 유해화학물질 판독 정확도는 평균 약 90% 이상의 정확도를 나타내었다. 본 논문에서 구축한 유해화학물질 학습데이터셋을 활용하였을 때 대부분 높은 정확도로 유해화학물질 종류를 판독하였다. 이를 통해 실제 화학사고 현장에서도 예상되는 사고물질을 추정해 줌으로써, 현장대원에게 사고물질에 대해 적절하고 신속한 대응을 지원할 수 있다.

향후에는 여러 가지 환경적 조건을 고려하여 불꽃, 연기의 관찰이 불가능하고, 무색/무취로 시각적 특징이 없거나 유사한 색상에 대한 유해화학물질의 판독까지 가능할 수 있도록 분광 스펙트럼 데이터 및 열화상 데이터 등을 활용하여 물질의 성분 분석 정보가 추가된 유해화학물질 학습데이터 생성 방법에 관한 연구가 지속적으로 진행되어야 한다.

References

[1] M. S. Ghil, "A study on development of hazard chemical detection system of a multi sensing type and dangerousness prediction algorithm based on artificial intelligence," Ph.D. dissertation, Dept. of Disaster Prevention Graduate School, Kangwon National University, 2021.

[2] Korea Environment Institute, *A study on the improvement of the chemical accident response system,* 2013.

[3] National Fire Agency, *Training professional firefighters to respond to chemical accidents,* https://www.nfa.go.kr, Mar. 2024.

[4] Y. J. Kim, A. R. Yoon, S. M. Ryoo, B. R. Sim, S. K. Baek, S. H. Baek, S. K. Cho, and G. B. Kim, "A study on the method of collecting training data on hazardous chemicals according to accident types," in *Proc. KICS Conf. 2024*, pp. 223-224, Pyeongchang, Korea, Jan. 2024.

[5] T. H. Lee, D. J. Lee, and C. H. Shin, "Characteristic analysis of casualty accidents

in chemical accidents," *J. Korean Inst. Fire Sci. Eng.,* vol. 31, no. 1, pp. 81-88, Feb. 2017. (https://doi.org/10.7731/KIFSE.2017.31.1.081)

[6] National Fire Agency, *Applying artificial intelligence (AI) technology to detect hazardous chemicals*, https://www.nfa.go.kr, Mar. 2024.

[7] N. W. Cho, I. G. Lee, and J. C. Lee, "A study on remote analysis of fire gas using open path FT-IR," *J. KIGAS,* vol. 17, no. 6, pp. 39-45, Nov. 2013. (http://dx.doi.org/10.7842/kigas.2013.17.6.39)

[8] S. W. Yoo, "Knowledge inference and machine learning for an interactive AI system supporting chemical accident response and handling," M.E., Dept. of Disaster and Safety Graduate School, Myongji University, 2022.

[9] S. K. Cho, S. H. Baek, B. S. Park, and G. B. Kim, "A method of detecting hazardous chemicals using artificial intelligence technology," in *Proc. KICS Conf. 2022,* pp. 264-265, Gyeongju, Korea, Nov. 2022.

[10] Y. Sun, J. Hu, D. Yuan, Y. Chen, Y. Liu, Q. Zhang, and W. Chen, "Hyperspectral classification of hazardous materials based on deep learning," *Sustainability 2023,* vol. 15, May 2023. (https://doi.org/10.3390/su15097653)

[11] T. DeKerf, J. Gladines, S. Sels, and S. Vanlanduit, "Oil spill detection using machine learning and infrared images," *Remote Sens. 2020,* vol. 12, Dec. 2020. (https://doi.org/10.3390/rs12244090)

[12] J. K. Ryu, D. K. Kwak, J. J. Kim, and J. K. Choi, "A study on fire recognition algorithm using deep learning artificial intelligence," *2018 Power Electronics Annual Conf.*, pp. 275-277, Hoengseong, Korea, Jul. 2018.

[13] G. Idakwo, S. Thangapandian, J. Luttrell, Y. Li, N. Wang, Z. Zhou, H. Hong, B. Yang, C. Zhang, and P. Gong, "Structure–activity relationship-based chemical classification of highly imbalanced Tox21 datasets," *J. Cheminformatics*, vol. 12, no. 66, Oct. 2020. (https://doi.org/10.1186/s13321-020-00468-x)

[14] L. Persson, S. K. Vinkhuyzen, A. Lai, Å. Persson, and S. Fick, "The globally harmonized system of classification and labelling of chemicals—explaining the legal implementation gap," *Sustainability 2017*, Nov. 2017. (https://doi.org/10.3390/su9122176)

[15] A. S. Persad and G. S. Cooper, "Use of epidemiologic data in Integrated Risk Information System (IRIS) assessments," *Toxicology and Appl. Pharmacology*, vol. 233, no. 1, pp. 137-145, Nov. 2008. (https://doi.org/10.1016/j.taap.2008.01.013)

[16] R. Sasaki, M. Fujinami, and H. Nakai, "Comprehensive image dataset for enhancing object detection in chemical experiments," *Data in Brief*, vol. 52, Feb. 2024. (https://doi.org/10.1016/j.dib.2024.110054)

[17] H. Dai, M. Zhu, G. Yuan, Y. Niu, H. Shi, and B. Chen, "Entity recognition for chinese hazardous chemical accident data based on rules and a pre-trained model," *Appl. Sci. 2023*, vol. 13, no. 1, Dec. 2022. (https://doi.org/10.3390/app13010375)

[18] P. Nooralishahi, F. López, and X. Maldague, "A drone-enabled approach for gas leak detection using optical flow analysis," *Appl. Sci. 2021*, vol. 11, no. 4, Feb. 2021. (https://doi.org/10.3390/app11041412)

[19] T. Luechtefeld, D. Marsh, C. Rowlands, and T. Hartung, "Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility," *Toxicological Sci.*, vol. 165, no. 1, pp. 198-212, Sep. 2018. (https://doi.org/10.1093/toxsci/kfy152)

[20] X. Zhu, C. Ho, and X. Wang, "Application of life cycle assessment and machine learning for high-throughput screening of green chemical substitutes," *ACS Publications*, 2020. (https://doi.org/10.26434/chemrxiv.12210860.v1)

김 연 진 (Yeon-Jin Kim)

2021년 2월 : 경북대학교 나노소재공학부 신소재공학전공 졸업
2023년 8월 : 서원대학교 정보통신공학 석사
2023년 4월~현재 : 서원대학교 산학협력단 연구원
<관심분야> 빅데이터, 인공지능, 재난재해

최 갑 용 (Kap-Yong Choi)

1994년 2월 : 경일대학교 경영학과 졸업
2003년 2월 : 경북대학교 컴퓨터공학과 석사
2008년 1월 : 경북대학교 공간정보학과 박사
2024년 7월~현재 : 대구소방본부 북부소방 서장
<관심분야> 재난안전 정보시스템, 빅데이터, AI, IoT

김 경 배 (Gyoung-Bae Kim)

1992년 2월 : 인하대학교 전자계산공학과 졸업
2000년 2월 : 인하대학교 컴퓨터공학 박사
2000년 3월~2004년 2월 : 한국전자통신연구원 선임연구원
2004년 3월~현재 : 서원대학교 소프트웨어학부 교수
<관심분야> 빅데이터, 인공지능, 재난재해, 클라우드 컴퓨팅