

21년도 인공지능 학습용 데이터 구축 가이드라인

< 2-17. 곤충사육데이터 >

인공지능 데이터 구축	사업 총괄	농업회사법인 푸디웜(주)
	데이터 설계	강원도농산물원종장
	데이터 수집	농업회사법인 푸디웜(주) 강원도농산물원종장
	데이터 정제	(주)아와소프트
	데이터 가공	글로벌여성ICT네트워크
	데이터 검수	서우엠에스(주)
	클라우드 소싱	글로벌여성ICT네트워크
	저작도구 개발	(주)아와소프트
	AI모델 개발 (곤충영상분석모델)	(주)아와소프트
	AI모델 개발 (곤충생육환경분석모델)	(주)케이엘큐브
가이드라인 작성	농업회사법인 푸디웜(주)	이유석
	(주)아와소프트	우영춘
가이드라인 버전	ver 1.0.0 ('22. 1. 14)	

목 차

1. 데이터 명세 정보 1

1.1 데이터 정보 요약	1
1.2 데이터 포맷	2
1.3 어노테이션 포맷	3
1.4 데이터 구성	5
1.5 데이터 통계	8
1.6 원시데이터 특성	12
1.7 기타 정보	13

2. 데이터 구축 가이드 14

2.1 데이터 구축 개요	14
2.2 문제정의	15
2.3 수집·정제	16
2.4 어노테이션/라벨링	23
2.5 검수	31
2.6 활용	39

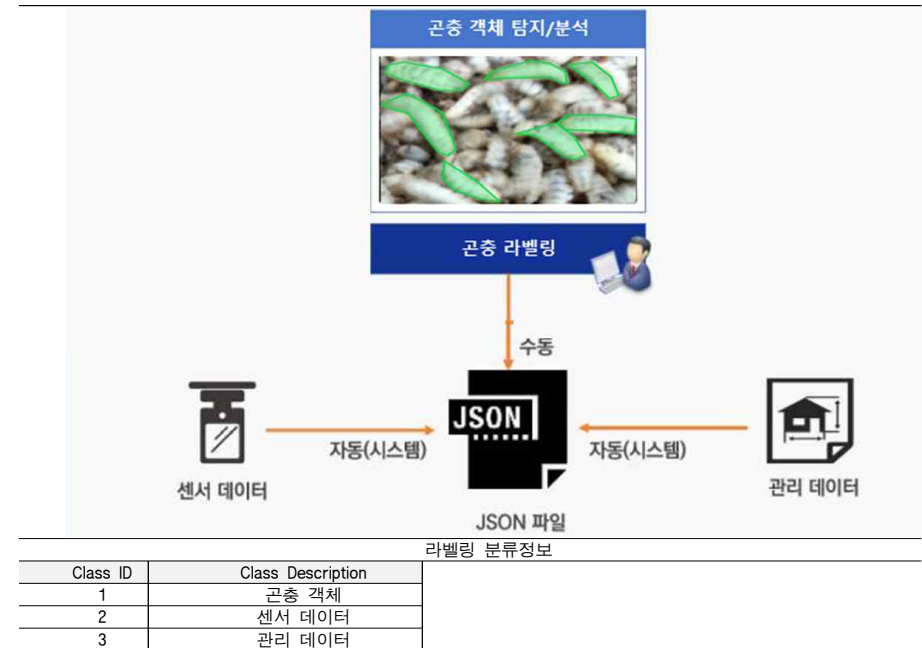
1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	지능형 곤충 사육 데이터	
활용 분야	- 곤충 영상분석 - 곤충 생육환경 분석 등	
카메라 기종	- EN-IPNC-T3	
데이터 요약	- 인공지능 학습용 곤충 사육데이터 세트 구축	
데이터셋 파일 구성	- 원본 이미지 파일 - 어노테이션 파일(Meta, Annotation, Sensor, Management 정보를 포함) - 이미지파일 : JPG파일 - 어노테이션 파일 : JSON파일	
이미지개수	18만 2000장 (RGB : 120,000장 ,열화상 : 62,000장)	
이미지 용량	약 78.33G B	
Dataset 용량	약 1.9GB	
데이터이력	배포버전	1.0.0
	개정이력	없음
	작성자/배포자	우영춘/이유석

1.2 데이터 포맷

- 원본 : 이미지데이터
- 라벨링 결과물 : image.jpg(182,000장) + annotation.json(182,000장)(이미지 한 장당 어노테이션 파일 한 장의 1:1 형식)



1.3 어노테이션 포맷

JSON
라벨 데이터(1개의 json 파일 내 아래 내용 모두 포함)
Meta_data
<pre>"META": { "Image File Name": "F01_CA01_Insect01_20210805_01.jpg", "Image File Identification": 220394, "Image Photography Date and Time": "2021.08.05. 13:00:00", "Environ. Info. MEASUREMENT Date and Time": "2021.08.05. 13:00:00", "Image File Fprmat": ".jpg", "Image File MAGNITUDE": "3000", "Equipment Information": "RGB CAM", "Camera Info.": "Canon ESO R", "Image Obtain Place Type": "Huge_farm", "Photography Area Name": "청주", "Image Photography Location Information": "2번 사육박스", "Image Copyright Information": "FOODYWORM", "Image Width MAGNITUDE": "1920", "Image Height MAGNITUDE": "1080", "Resolution": "FHD", "Bit": "24", "Image Color Group Code": "#880000", "Length AND Breadth Rate": "4:3", "Image Pixel": "1K", "Image Color Depth": "sRGB", "ISO Speed": "3200", "White Balance": "5500", "Shutter Speed": "f2.8 1/80", "F-STOP": "f2.8", "Flashlight Use ALTERNATIVE": "", "Filter Use ALTERNATIVE": "", "Focal Length": "50mm", "Feed Supply Date and Time": "", "Feed Supply Amount": "", "Feed Ingredient": "", "CCTV Name": "cam_01", "Insect Name": "누에", "Insect Farm Name": "", "Insect Breeding Room Box Number": "4", "Insect Breeding Box Size": "Huge_cage", "Scientific Name": "", "Image File Type": "", "Store Course": "", "Growth Step Name": "", "Growth Step Code": "", "MEASUREMENT Sensor INSTALLATION EXISTENCE AND NONEXISTENCE": "", "Photographer": "", "THEMA": "", "PHOTOGRAPHY YEAR": "" },</pre>
Image_Annotation_data
<pre>"ANNOTATION": { "Image File MAGNITUDE": 748979, "Image File Name": "곰뱅이1.png", "Annotation Regions": [{ "Annotation ID": 1, "Annotation Type": "polygon", "Polygon X Coordinate": [149, 139, 127, 121, 149], "Polygon Y Coordinate": [120, 124, 127, 130, 120], "Object Name": "worm", "Object Class Code": "" }] }</pre>

sensor_data
<pre>"SENSOR": { "Breeding Room Indoor Temperature": "22", "Breeding Room Indoor Humidity": "50", "Breeding Room Indoor Co2": 10, "Breeding Room Indoor Surface Moisture Quantity": 30, "Breeding Room Indoor Illumination Intensity": "300", "Breeding Room Indoor Ammonia": 50, "Breeding Room Indoor Oxygen": 50, "Breeding Room External Air Velocity": "15", "Breeding Room External Temperature": "22", "Breeding Room External Humidity": "50", "Breeding Room External Wind Direction": "N", "Breeding Room External Rain Alternative": "유", "Insect Breeding Box Temperature": "22", "Insect Breeding Box Humidity": "50", "Insect Breeding Box Co2": 30, "Insect Breeding Box Mat Temperature": "22" },</pre>
management_data
<pre>"MANAGEMENT": { "Insect Length": "5mm", "Insect Average Weight": "5", "Insect Molting Information": 10, "Feed Supply Amount": "2", "Feed Type": "밀가루", "Hydration Information": 10, "Insect infection status": "감염" }</pre>

이미지
 <알>  <성충>

1.4 데이터 구성

데이터 구성 형태			설명
경로	포맷	데이터 종류	
DATA			1) Meta
└Meta			- 메타 데이터를 포함한 리스트
└Annotation Regions	JSON	어노테이션 데이터	2) Annotation Regions
└Sensor			- 어노테이션 정보를 포함한 리스트
			3) Sensor
			- 센서 측정 정보를 포함한 리스트
└Management			4) Management
			- 관리 데이터를 포함한 리스트
└	JPG	RGB 이미지 데이터	-
└	JPG	열화상 이미지 데이터	-

구분	ID	요소명(영/한)		타입	예시	필수 여부	비고
메타정보	1	Image File Name	이미지파일명	String	FO1_CA01_INSECT01_20210805_01.JPG	Y	
	2	Image File Identification	이미지파일아이디	number	220394	Y	[0-999999]
	3	Image Photography Date and Time	이미지촬영일시	String	2021.08.05. 13:00:00	Y	YYYY.MM.DD HH:MM:mm
	4	Envrn. Info. MEASUREMENT Date and Time	환경정보 측정일시	String	2021.08.05. 13:00:00	Y	YYYY.MM.DD HH:MM:mm
	5	Image File Format	이미지파일포맷	String	JPG	Y	JPG
	6	Image File MAGNITUDE	이미지파일크기(사진파일크기)	number	3000	Y	[0-1,000,000,000]
	7	Equipment Information	장비정보	String	RGB CAM	Y	[RGB CAM, 열화상]
	8	Camera Info.	카메라정보	String	Canon ESO R	Y	카메라모델명
	9	Image Obtain Place Type	이미지획득장소유형	String	Huge_farm	Y	[Huge_farm, Middle_farm, Small_farm]
	10	Photography Area Name	촬영 지역명	String	청주	Y	[청주,제주도,강원도]
	11	Image Photography Location Information	이미지 촬영 위치 정보	String	2번 사육박스	Y	[고해상도 사육박스, 1번사육박스, 2번사육박스, 3번사육박스]
	12	Image Copyright Information	이미지저작권정보	String	FOODYWORM	Y	FOODYWORM
	13	Image Width MAGNITUDE	이미지가로크기	number	1920	Y	[0-5000]
	14	Image Height MAGNITUDE	이미지세로크기	number	1080	Y	[0-5000]
	15	Resolution	해상도	String	FHD	Y	FHD
	16	Bit	비트	number	24	Y	24
	17	Image Color Group Code	이미지색상군코드	String	# 8B0000	N	
	18	Length AND Breadth Rate	종횡비율	String	4:3	Y	[1:1, 3:2, 4:3, 16:9]
	19	Image Pixel	이미지화소	number	1K	Y	[0-1000k]
	20	Image Color Depth	이미지색상심도	String	sRGB	Y	sRGB
	21	ISO Speed	ISO 감도	number	3200	Y	[25-3200]
	22	White Balance	화이트 밸런스	number	5500	Y	[1K-10,000K]
	23	Shutter Speed	셔터 속도	String	f2.8 1/80	Y	f2.8 1/80
	24	F-STOP	조리개값	number	2.8	Y	[2.8, 4, 5.6, 8, 11,

구분	ID	요소명(영/한)	
	25	Flashlight Use	ALTERNATIVE
	26	Filter Use	ALTERNATIVE
	27	Focal Length	
	28	Feed Supply Date and Time	
	29	Feed Supply Amount	
	30	Feed Ingredient	
	31	CCTV Name	
	32	Insect Name	
	33	Insect Farm Name	
	34	Insect Breeding Room Box Number	
	35	Insect Breeding Box Size	
	36	Scientific Name	
	37	Image File Type	
	38	Store Course	
	39	Growth Step Name	
	40	Growth Step Code	
	41	MEASUREMENT Sensor INSTALLATION EXISTENCE AND NONEXISTENCE	
	42	Photographer	
	43	THEMA	
	44	PHOTOGRAPHY YEAR	
어노테이션 정보	1	Image File MAGNITUDE	
	2	Image File Name	
	3	Annotation ID	
	4	Annotation Type	
	5	Polygon X Coordinate	
	6	Polygon Y Coordinate	
	7	Object Name	

구분	ID	요소명(영/한)	객체클래스코드	타입	예시	필수 여부	비고
센서 정보	8	Object Class Code	객체클래스코드	String		N	추가
	1	Breeding Room Indoor Temperature	사육장 실내 온도	number	22	Y	0~50
	2	Breeding Room Indoor Humidity	사육장 실내 습도	number	50	Y	0~100
	3	Breeding Room Indoor Co2	사육장 실내 Co2	number	10	Y	0~100
	4	Breeding Room Indoor Surface Moisture Quantity	사육장 실내 표면 수분량	number	30	Y	0~100
	5	Breeding Room Indoor Illumination Intensity	사육장 실내 조도	number	300	Y	0 ~ 100,000
	6	Breeding Room Indoor Ammonia	사육장 실내 암모니아	number	50	Y	0~100
	7	Breeding Room Indoor Oxygen	사육장 실내 산소	number	50	Y	0~100
	8	Breeding Room External Air Velocity	사육장 외부 풍속	number	15	Y	0~100
	9	Breeding Room External Temperature	사육장 외부 온도	number	22	Y	0~50
	10	Breeding Room External Humidity	사육장 외부 습도	number	50	Y	0~100
	11	Breeding Room External Wind Direction	사육장 외부 풍향	String	N,W,E,S 중 하나	Y	[N,W,E,S]
	12	Breeding Room External Rain Alternative	사육장 외부 강우 여부	String	유/무	Y	[유/무]
	13	Insect Breeding Box Temperature	곤충 사육박스 온도	number	22	Y	0~50
	14	Insect Breeding Box Humidity	곤충 사육박스 습도	number	50	Y	0~100
	15	Insect Breeding Box Co2	곤충 사육박스 Co2	number	30	Y	0~100
	16	Insect Breeding Box Illumination Intensity	곤충 사육박스 조도	number	300	Y	0 ~ 100,000
관리 정보	1	Insect Length	곤충 길이	number	5	Y	100이하
	2	Insect Average Weight	곤충 평균 무게	number	5	Y	1000이하
	3	Insect Molting Information	곤충탈피정보	String	O	Y	O,X
	4	Feed Supply Amount	사료 공급량	number	2	Y	100이하
	5	Feed Type	사료 종류	String	밀기울	Y	x. 뽕잎, 버섯배지, 곤충젤리, 밀기울,배추,호박, 무
	6	Hydration Information	수분제공정보	String	O	Y	O,X
	7	Insect infection status	감염상태	String	O	Y	O,X

1.5 데이터 통계

1.5.1 데이터 구축 규모

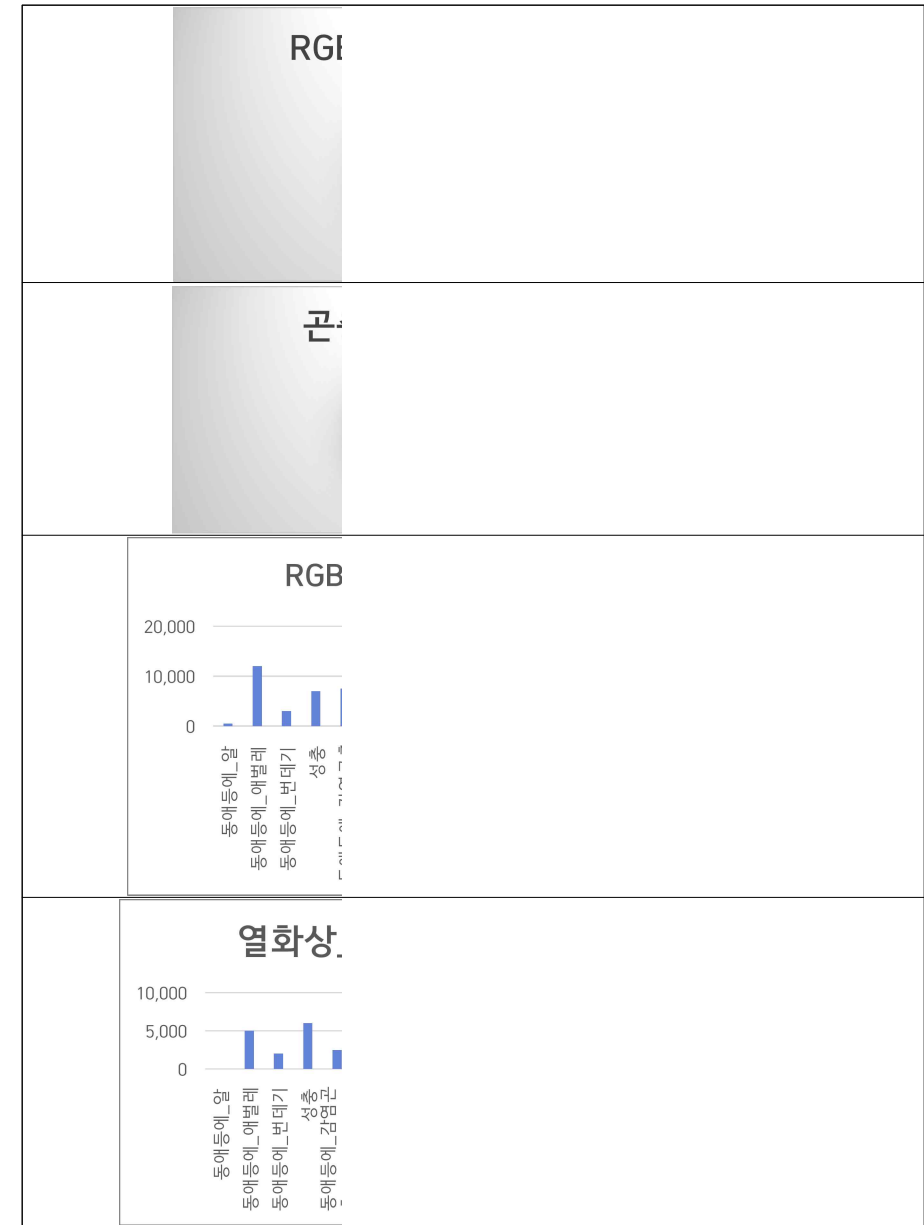
○ 데이터 규모

종류		
곤충 RGB 영상)	동애등에	
	흰점박이 꽃무지	
	갈색거저리	
	누에	
곤충 열화상 영상	동애등에	
	흰점박이 꽃무지	
	갈색거저리	
	누에	
센서데이터	RGB 영상	
관리데이터	열화상 영상	
	곤충길이 데이터(수동측정)	
	곤충평균무게데이터(수동측정)	
	곤충탈피데이터	
	사료공급량 데이터	
	사료 종류 데이터	
	수분 제공 데이터	
	감염 상태 정보 데이터	

1.5.2 데이터 분포

- 구축 데이터 종류 : 누에, 흰점박이꽃무지, 갈색거저리, 동애등에 RGB, 열화상 이미지
- 구축 데이터 클래스별 비율

종류		획득규모		비율	비고
곤충 RGB 영상	동애등에	알	500장	0.27%	
		애벌레	12,000장	6.59%	
		번데기	3,000장	1.65%	
		성충	7,000장	3.85%	
		감염곤충	7,500장	4.12%	
	흰점박이 꽃무지	알	500장	0.27%	
		애벌레	14,000장	7.69%	
		번데기	1,000장	0.55%	
		성충	7,000장	3.85%	
		감염곤충	7,500장	4.12%	
	갈색거저리	알	500장	0.27%	
		애벌레	12,000장	6.59%	
		번데기	3,000장	1.65%	
		성충	7,000장	3.85%	
		감염곤충	7,500장	4.12%	
	누에	알	500장	0.27%	
		애벌레	15,000장	8.24%	
		번데기	4,000장	2.20%	
		성충	3,000장	1.65%	
		감염곤충	7,500장	4.12%	
곤충 열화상 영상	동애등에	알	0장	0.00%	열화상 이미지로 탐지 불가
		애벌레	5,000장	2.75%	
		번데기	2,000장	1.10%	
		성충	6,000장	3.30%	
		감염곤충	2,500장	1.37%	
	흰점박이 꽃무지	알	0장	0.00%	열화상 이미지로 탐지 불가
		애벌레	7,000장	3.85%	
		번데기	0장	0.00%	
		성충	6,000장	3.30%	
		감염곤충	2,500장	1.37%	
	갈색거저리	알	0장	0.00%	열화상 이미지로 탐지 불가
		애벌레	5,000장	2.75%	
		번데기	2,000장	1.10%	
		성충	6,000장	3.30%	
		감염곤충	2,500장	1.37%	
	누에	알	0장	0.00%	열화상 이미지로 탐지 불가
		애벌레	8,000장	4.40%	
		번데기	3,000장	1.65%	
		성충	2,000장	1.10%	
		감염곤충	2,500장	1.37%	



<데이터 분포표>

1.7 기타정보

1.7.1 포괄성

- 곤충별 대표적 생육 단계를 포함하였음.
- 감염 곤충 분류를 추가 하여 이상 상태 분류함.
- 열상, RGB 영상으로 분류.

-

1.7.2 독립성

- 개인정보 보호법에 저촉되지 않는 영상 수집 필요

1.7.3 유의사항

- 고해상도, 저해상도 및 RGB, 열화상으로 구분되어 있으므로 서비스의 용도에 알맞은 사용 권고

2. 데이터 구축 가이드

2.1 데이터 구축 개요

■ 데이터 구축 프로세스의 정의

- 데이터 구축은 원시 데이터 정제 → 데이터 라벨링 → 학습 데이터 셋 검사 및 유효성 검증 과정 후, 저장소에 저장되는 절차의 프로세스로 정의



< 데이터 구축 프로세스 >

■ 데이터 포맷 정의 및 획득, 정제 방안 및 절차

데이터명	획득형태	수집장비
곤충 활동 영상 AI 데이터	사육박스 촬영	RGB/열화상 카메라 (EN-IPNC-T) 각종센서
곤충 활동 열화상 영상 AI 데이터	사육박스 촬영	
센서 시계열 데이터	IoT센서 실시간 측정	각종센서 (모델명 : 획득 도구 참조)
관리 시계열 데이터	수기를 통한 획득	수기 (시스템 입력)

2.2 문제정의

2.2.1 임무 정의

AI를 활용한 곤충 사육의 표준화 및 분석을 위하여, 곤충(동애등애, 흰점박이꽃무지, 갈색거저리, 누에) 사육장에서 CCTV를 통해 획득한 영상 데이터, 여러 가지 센서로부터 획득한 환경데이터, 사료 공급, 사육 환경 등 사육장에서 발생하는 여러 가지 이벤트에 대한 관리 데이터 구축을 목표로 함.

2.2.2 데이터 구축 유의사항

동애등애, 흰점박이꽃무지, 갈색거저리, 누에를 대상으로 데이터 수집

- 전문 농장의 사육장에 촬영장비, 센서를 설치 하여 이미지 파일 획득
- 사전 데이터 사용 동의서를 받아 진행
- 일정시간에 측정 클라우드워커 직접 측정을 통한 관리 데이터 확보

2.3 수집·정제

2.3.1 원시데이터 선정

☐ 원시 데이터 획득 시 검토사항

분류	5W1H	항목					
획득	What	- 영상 : 곤충 종별 활동영상, 곤충종별 열화상 영상 - 센서 : 사육장 / 사육 박스 환경 데이터 - 관리 : 곤충 사육관리 데이터					
	When	- 영상 : 2021.6.1. ~ 2021.12.31. 24 시간 녹화 영상 - 센서 : 2021.6.1. ~ 2021.12.31. 1분 간격 센서 데이터 - 관리 : 수시(이벤트 발생시)					
	Where	- 컨소시엄에 포함된 5개의 사육장의 28개 사육 박스 <table><tr><td>획득 장소</td></tr><tr><td>제주도 농가 1개소</td></tr><tr><td>충북 농가 2개소</td></tr><tr><td>푸디원 사육장</td></tr><tr><td>강원도 농산물 원종장 사육장</td></tr></table>	획득 장소	제주도 농가 1개소	충북 농가 2개소	푸디원 사육장	강원도 농산물 원종장 사육장
	획득 장소						
	제주도 농가 1개소						
	충북 농가 2개소						
	푸디원 사육장						
강원도 농산물 원종장 사육장							
Who	- 영상, 센서 데이터는 장비에서 자동 획득 - 관리 데이터는 직접 촬영 하여 실측한 데이터 값과 관리데이터 값을 수기로 전산입력						
How	- 영상 : 사육장 수직부에 카메라 설치 - 센서 : 사육장/사육박스/사육장외부에 설치 - 관리 : 통합라벨링 시스템에 전산입력 스마트팜이 설치된 농가는 기 구축된 스마트팜 시스템과 연계하고 데이터 수집시 메타데이터로 구분						
Why	- 지능형 곤충 사육 데이터 구축 - 지능형 곤충 사육장 구축						

2.3.2 수집·정제 절차

1) 데이터 획득 방법

□ 원시데이터 현황정보

구분	곤충명	생육단계	수집기간	주요 수집관점	획득규모
이미지 (열화상포함)	동애 등애	알	부화까지	색태, 오목현상	500장
		애벌레	용화단계까지	길이, 표면온도, 색태, 표면적	17,000장
		번데기	성충탈피까지	색태, 표면적, 표면온도	9,000장
		성충	교미/산란	색태, 형태	9,000장
	원점 박이 꽃무지	감염곤충	감염된 시점	색태, 형태, 표면 온도	10,000장
		알	부화까지	색태, 오목현상	500장
		애벌레	용화단계까지	길이, 표면온도, 색태, 표면적	21,000장
		번데기	성충탈피까지	색태, 표면적, 표면온도	1,000장
		성충	교미/산란	색태, 형태	13,000장
	누에	감염곤충	감염된 시점	색태, 형태, 표면 온도	10,000장
		알	부화까지	색태, 오목현상	500장
		애벌레	용화단계까지	길이, 표면온도, 색태, 표면적	17,000장
		번데기	성충탈피까지	색태, 표면적, 표면온도	7,000장
	갈색 거저리	누에나방	교미/산란	색태, 형태	11,000장
		감염곤충	감염된 시점	색태, 형태, 표면 온도	10,000장
		알	부화까지	색태, 오목현상	500장
		애벌레	용화단계까지	길이, 표면온도, 색태, 표면적	23,000장
		번데기	성충탈피까지	색태, 표면적, 표면온도	8,000장
		성충	교미/산란	색태, 형태	4,000장
		감염곤충	감염된 시점	색태, 형태, 표면 온도	10,000장

구분		종류		획득규모	목적	확보방안
군집 데이터	센서 데이터	온도 데이터	온도 센서	182,000건	사육장 실내 환경 정보	센서 상시제 장
		습도 데이터	습도 센서			
		Co2데이터	Co2 센서			
		표면 수분량 측정 데이터	HD 센서			
		조도데이터	조도 센서		사육장 외부 환경 정보	
		NH3데이터	NH3 센서			
		O2데이터	O2 센서			
		풍량 데이터	풍량 센서			
		온도데이터	온도 센서		사육박스 환경 정보	
		습도데이터	습도 센서			
		Co2데이터	Co2 센서			
		표면 수분량 측정 데이터	HD 센서			
		온도데이터	온도 센서			
		습도데이터	습도 센서			
Co2데이터	Co2 센서					
표면 수분량 측정 데이터	HD 센서					
조도	조도 센서					
군집 데이터	관리 데이터	곤충길이 데이터 (수동측정)	동애등애, 원점박이 꽃무지, 갈색거저리, 누에	2,500개 이상	상시 주기 입력	
		곤충평균무게데이터(수동측정)		2,500개 이상		
		곤충탈피데이터		1,000개 이상		
		사료공급량 데이터		1,000개 이상		
		사료영양성분 데이터		1,000개 이상		
		수분제공 데이터		1,000개 이상		
		감염 상태 정보 데이터		1,000개 이상		

□ 사료 공급량 데이터 정보

곤충종
동애등애
원점박이꽃무지
누에
갈색거저리

□ 개체 및 군집 데이터 측정 정보

- 개체데이터 : 곤충 개별적 수치화 가능한 데이터
- 군집데이터 : 사육상자 내 곤충그룹에 대한 전체생육상황을 나타내는 데이터
- 곤충별 개체정보 측정 후 획득기준을 제시함으로써 정확한 데이터수집방향을 확보함

구분	수집데이터
개체	곤충 길이데이터
	곤충 무게데이터
	곤충 외형 이미지변화
	곤충 표면온도
군집	사육상자내 CO2량데이터
	사육상자내 암모니아데이터
	사육상자내 활동성데이터
	사육상자내 분포데이터

□ 개체 및 군집 데이터 측정 정보

이상 종류	병징,가해증상
곰팡이	곰을, 황변(변색)
박테리아	무름, 부패
바이러스	무름, 체액분비
선충	팽창, 비대
원생동물	괴사, 무름
천적	괴사, 멜라닌반응

□ 데이터 다양성 확보

- 데이터가 촬영 시간적, 위치적으로 하나에 편중 되지 않은 다양성 있는 분포의 데이터를 구축한다.
- 즉각적인 검수를 통한 데이터의 공간적/시간적 분포를 확인하여 부족한 분포의 데이터의 즉각적 추가 촬영을 통해 원시 데이터 다양성을 유지한다.



<원시데이터 공간적/시간적 균등분포>

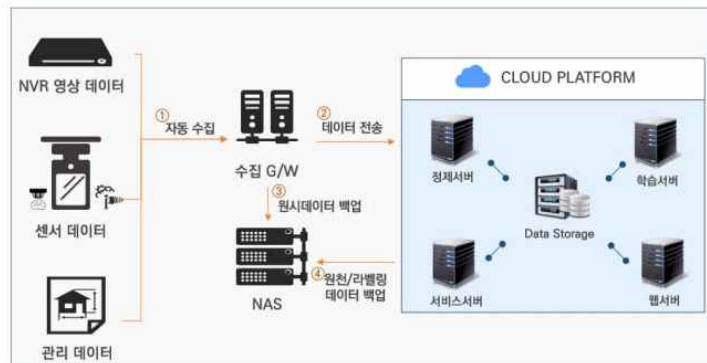
□ 획득 데이터 저장 및 관리 방안

- 획득 데이터는 영상, 열화상 데이터, 직접 촬영이미지, 센서 데이터, 관리 데이터로 나누어지며 각각의 획득 및 저장방식은 아래와 같음
- CCTV 영상/열화상 데이터 : CCTV와 함께 설치된 NVR을 통해 주기적(일별)으로 녹화 영상을 수집 G/W에서 자동으로 획득
- 곤충 생장정보 데이터 : 클라우드 위커에 의해 사육장에서 곤충 개체별 이미지 직접 촬영 후 곤충 생장정보 수집기를 통해 길이 및 무게 값을 측정하여 획득



< 곤충 생장정보 수집기 >

- 센서데이터 : 센서 데이터는 자체 개발하여 운영하는 Cloud 곤충 스마트팜을 활용하여 주기적(분당) 획득하여 자체 데이터베이스(RDB,시계열 DB)에 저장되며 수집 G/W에서 자동으로 획득
- 관리 데이터 : 곤충 사육장에 설치한 복합환경제어 장치를 통해서 자동 수집과 수동 입력하면 수집 G/W에서 자동으로 획득

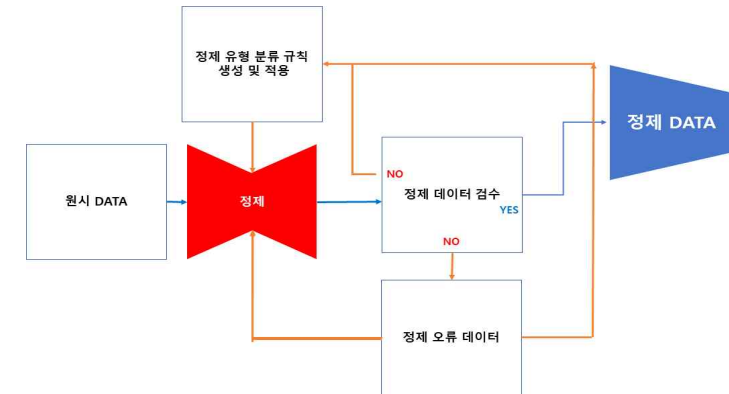


< 획득 데이터 저장 및 관리 방안 >

2) 데이터 정제 방법

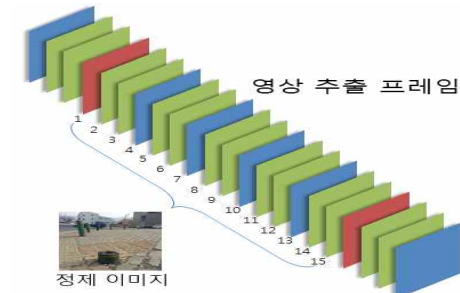
□ 원시 데이터 정제 프로세스

- 원시 데이터 정제는 크게 카메라 촬영 영상 데이터, 센서 데이터, 관리 데이터로 나누어지며 각각의 정제 방법은 데이터 특성에 따라 다르고, 영상의 경우 이미지로 변환하고 변환된 이미지를 육안검사 및 검사 알고리즘을 통해 적합하지 않은 이미지는 삭제



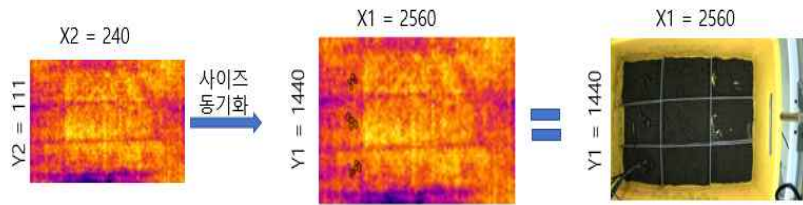
< 원시 데이터 정제 절차 >

- 원시 영상/열화상 데이터의 정제는 영상 프레임의 이미지화 정제, 해상도 개선의 정제, 개인정보 비식별을 위한 정제, 작업자의 편의를 위한 촬영 시간 및 장소에 따른 분류 정제과정을 포함
 - 영상 데이터의 초당 프레임 수 25-29개이므로 1초당 2개의 최적 프레임을 선정하여 추출
 - 추출된 영상프레임을 이미지 파일 포맷인 JPG, PNG 형태로 저장하는 정제 작업 수행
 - 정제 작업 수행은 자체 종합 데이터 정제 프로그램의 기능 중 영상 이미지 추출 기능 사용
 - 정제된 이미지 파일은 “촬영시간_촬영장소”로 생성된 폴더에 순번을 이름으로 저장

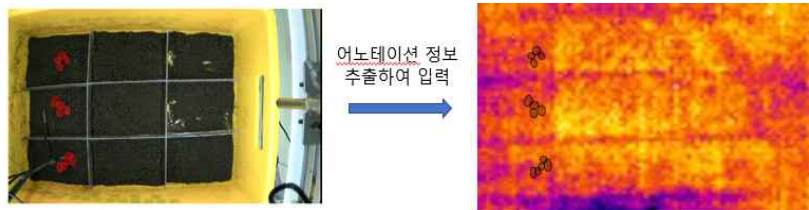


< 원시 영상/열화상 데이터에서 이미지 정제 >

- 원시 열화상 영상 데이터의 정제는 원시 영상 데이터의 정제 항목에 원시 영상 데이터와 시간 동기화 정제, 원시 영상데이터 이미지 사이즈와 동기화 정제 과정을 포함한다.
- 원시 영상데이터와 시간 동기화 정제는 초기 촬영시 시간 정보를 서로 매칭하여 동기화하는 작업으로 추후 열화상 이미지 가공을 위하여 실시한다.
- 원시 영상데이터와 이미지 사이즈 동기화 정제는 추후 가공 시 열화상 영상은 육안으로 객체를 구분이 불가 하기 때문에 앞서 시간 동기화된 원시 영상 데이터 가공을 통해 생성된 좌표 정보를 열화상 데이터에 적용하여 어노테이션 정보를 취득한다.

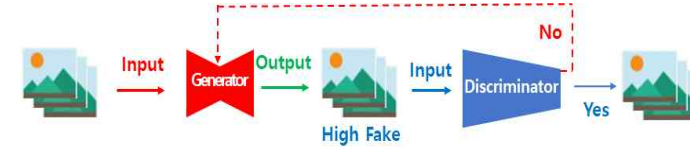


<열화상 이미지, RGB 이미지 사이즈 동기화>



<RGB 어노테이션 정보 열화상 적용 예시>

- 원시 데이터 해상도 개선의 정제
 - 이미지 정제되어 추출 된 이미지 파일 중 해상도가 떨어지거나 품질이 떨어지는 객체는 SRGAN 알고리즘을 적용하여 AI 영상 보정 수행



< GAN모델 적용한 영상 보정 프로세스 >

- 정상 해상도 및 품질의 데이터 학습하여 생성된 Generator, Discriminator 모델을 사용하여 프로세스에 맞추어 영상 파일 보정 수행
- 센서 데이터는 통합 라벨링 시스템에 내장된 Outlier 제거 알고리즘과 결측치 처리 알고리즘을 이용하여 데이터를 정제하며 마지막으로 관리 데이터의 경우 시스템에서 관리 대상별 관리주기를 확인하여 미입력된 데이터는 양식장 관리 담당자에게 재입력 요청 발송

2.3.3 수집·정제 기준

□ 원시 데이터 정제 기준

- 원시 데이터의 정제 기준은 아래와 같음
- 일반적으로 이미지를 정제할 때 이미지의 질이 낮은 경우(흐린 이미지, 객체 일부만 보이는 이미지)에도 이미지를 삭제하지만 본 과제에서는 카메라의 위치와 초점, 그리고 촬영대상이 사육장으로 고정되어 있으므로 흐린 이미지는 높은 비중을 차지하지 않음

데이터 명
곤충 객체 이미지 AI 데이터
센서 시계열 데이터
관리 시계열 데이터

2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차

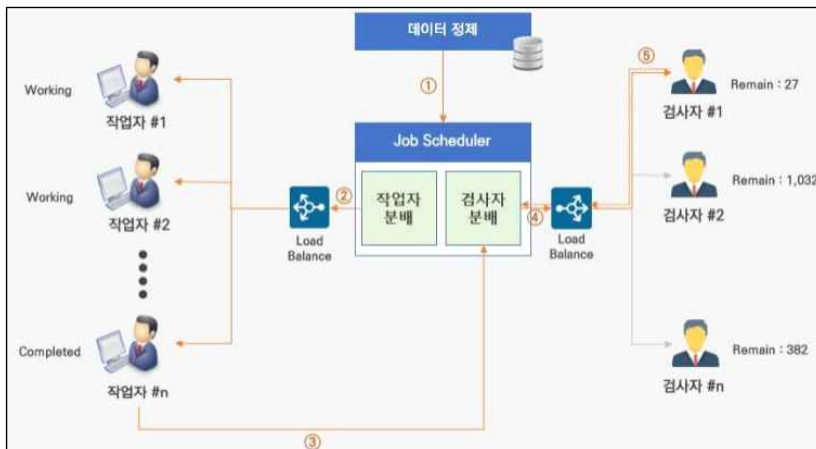
□ 라벨링 방법

- 이미지 데이터 : 사육장에서 생성된 정제 이미지를 대상으로 Polygon 라벨링 기법을 사용하여 라벨링 작업 실시
- 센서/관리 데이터 : 센서/관리 데이터는 수동 라벨링 대상이 아니며 이미지에 대한 라벨링이 완료되면 해당 이미지의 어노테이션 정보에 시스템이 자동으로 정보를 입력하며, 입력 기준은 이미지가 촬영한 시점 기준으로 가장 최근에 입력된 센서/관리 데이터가 입력



< 라벨링 시스템 개요 및 이미지 라벨링 방법 >

□ 라벨링 절차



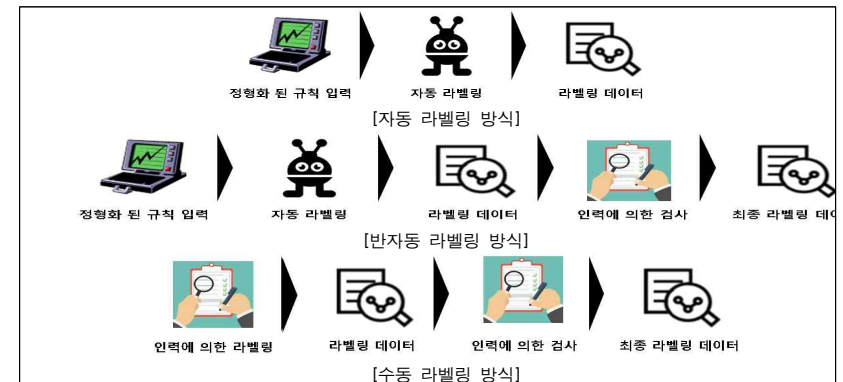
< 데이터 라벨링 절차 >

- 정제된 이미지가 데이터 스토리지에 저장되면 Job Scheduler는 자동으로 작업자 분배를 위한 작업을 진행
- 작업자의 잔여량을 파악하여 작업이 완료되었거나 가장 적은 작업량을 가진 작업자(크라우드 소싱 작업자)를 우선하여 업무 분배
- 작업 완료시 검사자에게 자동으로 분배되어 검사자는 검사조건에 맞지 않으면 해당 작업자에게 재작업 요청 실시
- 라벨링 목적
 - 이미지, 영상, 텍스트 등의 정제된 원천 데이터에 사람 또는 기계를 활용하여 인공지능이 학습할 수 있도록 다양한 정보를 목적에 맞게 입력하는 작업



< 라벨링 목적 >

- 원시 데이터 획득과 정제과정을 통해 생성된 원천데이터상에서 이미 정의한 객체 클래스의 객체를 찾아 표시하는 공정을 라벨링이라고 함
- 라벨링 방식은 특성에 따라 자동, 반자동, 수동 방식으로 분류

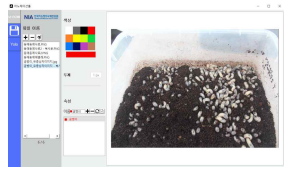
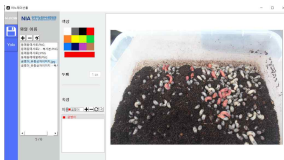
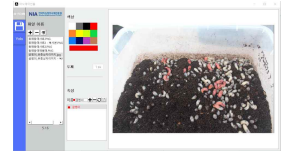


< 라벨링 방식의 비교 >

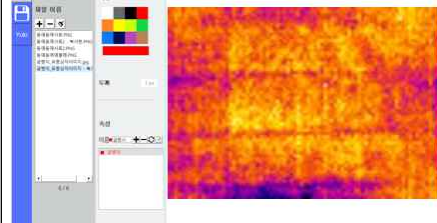
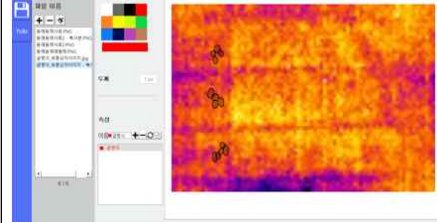
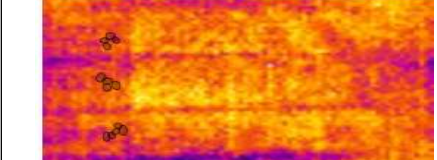
- 자동 라벨링 방식 : 원천 데이터로부터 추출하는 방식이 정형화되어 있고 자동화할 수 있는 사항인 경우, 사용하는 방식으로 원천데이터 추출 규칙을 알고리즘화하여 기계에 입력하여 실행하는 방식
- 반자동 라벨링 방식 : 자동 라벨링 방식으로 라벨링한 이후 사람이 다시 확인하여 수정하는 방식
- 수동 라벨링 방식 : 오직 사람의 판단으로 라벨링하는 방식

- 라벨링은 인공지능 학습데이터 구축 목적, 도메인, 활용 분야를 고려하여 방법 및 절차를 수립 후 실시

■ RGB이미지 라벨링 작업 절차

구분	참조 이미지	설 명
1. 원천 이미지 불러오기		작업 실시할 원천 이미지를 라벨링 프로그램에 로드
2. 객체 클래스 입력하기		- 라벨링 실시한 객체 명을 프로그램상에 입력 - 객체별 라벨링 할 색상을 선택 - 객체 리스트가 생성
3. 객체 라벨링 실시		- 육안으로 이미지 상에서 해당하 는 객체를 탐지 - 객체 탐지 시 해당 객체리스트에 서 해당 항목을 선택 - 객체의 경계선을 따라 라벨링을 실시
4. 라벨링 상태 확인		- 이미지상의 해당 객체 라벨링 완료 실시 - 객체 라벨링과 객체명(Annotation) 일치하는지 확인 - 객체 라벨링과 객체명 (Annotation) 일치하는지 확인
5. 라벨링 정보 저장		현재 이미지의 라벨링 정보를 저장하여 Annotation 파일 생성

■ 열화상 이미지 라벨링 작업 절차

구분	참조 이미지
1. 열화상 이미지 불러오기	
2. 동기화된 RGB 이미지 불러오기	
3. 동기화된 RGB 이미지 상에 라벨링 실시	
4. 열화상 이미지로	
5. 라벨링 정보 저장	

2.4.2 어노테이션/라벨링 기준

- 폴리곤 데이터 : 화면 내 해당 객체의 경계선을 따라 폴리곤으로 표현

항 목	기준 및 고려사항	어노테이션 기준
초점	대상 초점이 흔들려 경계선이 명확하지 않은 경우	육안으로 객체 구분이 가능한 경우 대략적인 경계면을 지정하여 작업 실시
흔들림/움직임	대상이 흔들리게 나와 경계선이 명확하지 않은 경우	
밝기	이미지 촬영장소 및 시간상 너무 밝거나 어두워서 형태가 뚜렷하지 않는 경우	
해상도(화질)	이미지 해상도가 낮아 대상의 경계가 뚜렷하지 않은 경우	작업하지 않고 검수자에게 제출 (이미지 해상도 개선 알고리즘 적용)
객체 잘림	이미지상에서 해당 객체가 잘려 나가 일부만 표출된 경우	객체의 형태가 60% 이상 확인 가능할 경우 작업 실시
객체 겹침	객체 기준에 없는 것과 포함된 것이 겹쳐진 경우	객체의 형태가 60%이상 확인 가능할 경우 겹침 부분을 기준으로 나누어 작업 실시

<어노테이션 시 주요 기준 및 고려사항>

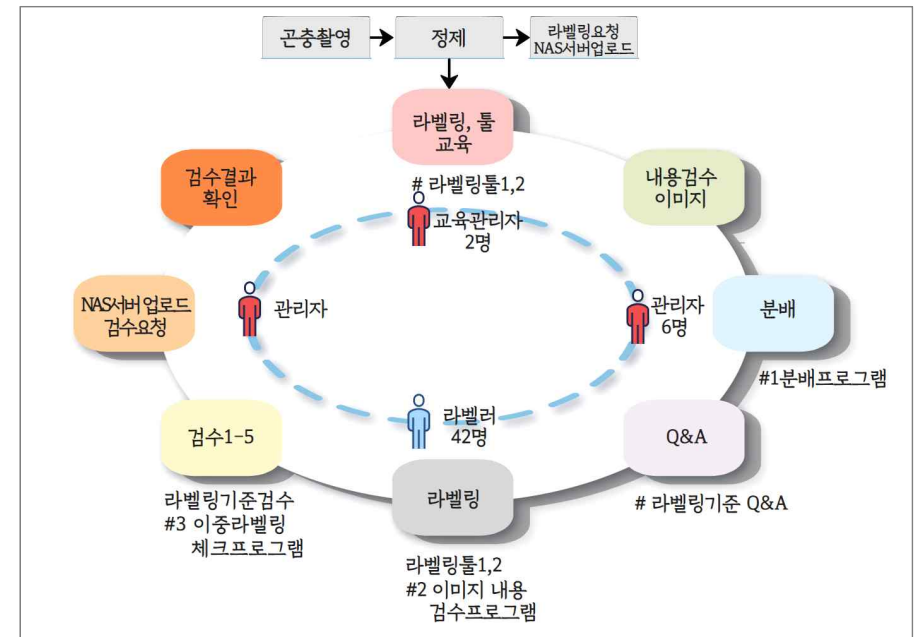
2.4.3 어노테이션/라벨링 조직

□ 라벨링 조직 및 업무수행

■ 라벨링 데이터 구축을 위해 각 역할별 업무내용

구 분	인원	업 무
교육관리자	2명	- 라벨링 툴 교육자료 제작 - 라벨러에게 라벨링 개요 및 업무수행 교육 - 라벨링 툴 교육 - 라벨링 기준 교육
관리자	6명	- 정제이미지 내용검수 - 분배(#분배프로그램 사용) - 라벨링 기준 작성 및 Q&A - 라벨링 검수(#이중라벨링 체크 프로그램) - 라벨링 데이터 검수요청 / 데이터 서버 업로드
라벨러	42명	- 라벨링 교육 수료 (공통, 입문, 기본, 심화교육) - 라벨링 툴교육 수료 - 라벨링 - 라벨링 검수후 재요청 자료 라벨링

■ 데이터 획득부터 정제, 작업자/검사자 분배, 라벨링, 검사까지의 과정 도식화



2.4.4 어노테이션/라벨링 도구

□ 라벨링 도구

- 데이터 구축 목적 달성을 위해 원천 데이터 형태, 구축 목적에 부합하는 라벨링 도구 선정
- 데이터 라벨링 도구는 본 컨소시엄에 자체 보유하고 있는 라벨링 도구를 본 과제에 맞게 변경하여 사용
- 데이터 획득부터 정제, 작업자/검사자 분배, 라벨링, 검사까지의 과정을 시스템화

구 분	설 명
원천데이터	○ 유 형 : 이미지 파일(JPG, PNG) ○ 사 이 즈 : 1920*1080(W*H)
구 축 모 델	○ Segmentation : MaskRcnn, Yolact
라벨링 형식	○ Segmentation : Segmentation(Polygon) Labelling

<데이터 구축 정보>

- 최적 모델 선정을 위하여 Seg 유형의 라벨을 수행하기 때문에 범용적으로 사용되는 오픈소스 라벨링 툴이 자체 제작 툴 사용



< 라벨링 작업에 사용될 자체 제작 툴 >

- IIDS 라벨링 툴

- 수집 정제 매니저 기능
- 작업 분배 및 라벨링 기능
- 정제 및 검수 기능
- (주)아와소프트에서 개발한 IIDS 라벨링 툴은 다양한 형태의 라벨링을 선택에 따라 적용 가능하며 한번에 Yolo, SSD, MaskRcnn, Yolact 모델 생성 가능
- 그 외에도 자체적으로 이미지 원천 데이터 로드시 적합성 유무 판별하여 체크 가능

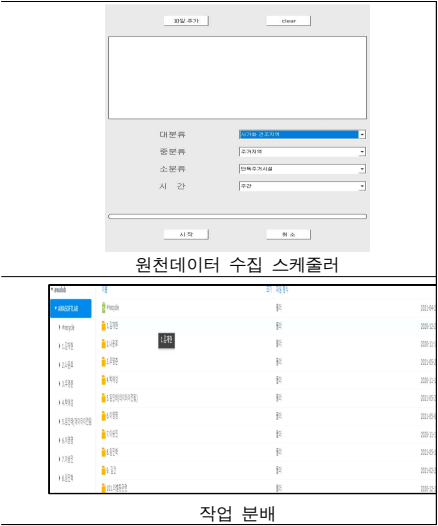


그림 2-32. IIDS 라벨링 툴의 기능

2.5 검수

2.5.1 검수 절차

1) 데이터 검사 관리 조직 운영방안

- 본 과제에서 데이터의 검사는 기존 이미지, 영상 데이터 검사 유형, 기존 방법을 참조하여 최종 구축 데이터의 사용 목적을 기준으로 사용 목적에 대한 부합성, 데이터의 유용성, 데이터 활용성 등에 대한 평가를 거쳐 유의미한 데이터를 판단 및 분류하여 향상된 품질의 데이터를 구축하는 것



< 데이터 검사 개념도 >

- 데이터의 검사는 정해진 절차로 진행
- 각각의 검사공정 중 발생하는 부적합 데이터는 검사 공정마다 취합하여 반려 조치
- 검사공정 중 수집된 오류데이터 특성 및 사례는 즉각 수집하여 공정의 기준에 포함되도록 함



< 데이터 검사 공정 절차 >

- 공정별 수행하는 검사 항목

검사 순서	검사명	항목 설명(간략)
1차	획득 데이터 동기화 검사	획득 원시 데이터 동기화
2차	데이터 정제 검사	정제 기준 명확성, 중복성 방지, 정제 작업 매뉴얼 기준 검사
3차	데이터 가공(라벨링) 검사	라벨링 가이드 작성 기준 검사
4차	전수 검사	AI 학습용 데이터 적합/부적합 판정

2) 검사절차 정의

- 데이터셋 구축 프로세스 단계마다 검사과정이 수행되며 최종적으로 전수검사를 수행, 전수검사는 과제 마지막 1달간 수행 예정이며 전체 소싱인력을 투입하여 검사 예정
- 검사를 위한 전용 툴을 이용하여 정제, 라벨링 단계마다 작업자가 작업을 완료 하면 시스템은 자동으로 검사자에게 검사를 요청하며 검사자의 승인이 떨어지면 시스템에서 자동으로 다음 단계로 진행



2.5.2 검수 기준

- ☐ 지능형 곤충 사용데이터의 상태 판단을 위하여 1차 검증된 데이터 중 30%의 데이터를 무작위 표본 추출하여 일반 작업자 그룹에서 판별이 어려운 난이도 높은 데이터를 전문 검수 그룹에서 검수하는 등 단계별 검수 수행(육안 및 자체 툴 사용, 반자동 확인)
- ☐ 검사툴을 사용하여 검사된 결과를 일반 작업자 그룹 및 검수 그룹에서도 동일한 데이터에 한해 2명 이상 참여하여 교차 검사 수행(육안으로 확인)
- ☐ 데이터의 적절성 등은 다수의 원칙을 따름
- ☐ 데이터 검수 간 오류발생 및 특정 지식이 필요한 경우 별도 수집하여 교육 커리큘럼 및 별도 교육을 통해 필요한 지식을 습득 후 재수행
- ☐ 최종적으로는 아래 표와 같은 품질을 달성하는 것을 목표로 함

＊ 각 영역별 품질목표 달성 기준

품질관리 영역	품질지표		품질목표	품질 목표 달성 기준	비고
구축 공정 품질	준비성		95% 이상	준비성 체크리스트 목록의 95% 이상 '적정'	체크리스트
	완전성		95% 이상	완정성 체크리스트 목록의 95% 이상 '적정'	체크리스트
	유용성		95% 이상	유용성 체크리스트 목록의 95% 이상 '적정'	체크리스트
구축 데이터 품질	적합성	기준적합성	95% 이상	기준적합성 체크리스트 목록의 95% 이상 '적정'	체크리스트
		기술적합성	99% 이상	준수율(%)	전수검사
		통계적다양성	99% 이상	구축 목표 대비 실적	전수검사
	정확성	의미정확성	99% 이상	정확도, 정밀도, 재현율	전수검사
		구문정확성	99.9% 이상	오류율 (검사대상건수 대비 오류건수)	전수검사

□ 수행절차 별 검사 항목

검사 절차	영상(동적/정적) 이미지 공통 항목	요구사항
1차 검사 (획득)	법·제도 준수	원시데이터 획득 시 관련 법·제도적 규정 등을 반드시 준수하여야 함
	사실적인 획득 환경 구성	원시데이터를 인위적인 환경과 조건하에 획득해야 하는 경우 사실적인 획득 환경을 구성하여야함
	데이터수집장비관리 및 수집 연속성 검토	사육장에 관리 직원이 상시 현장에서 상주 하며, 일 2회 센서의 세척 및 이물질 제거를 통한 센서 품질 관리 및 데이터 수집 여부의 상시 모니터링을 통한 안정성 확보
	데이터 동기화	다중 데이터 소스 간 정교한 동기화를 위한 절차를 마련하여야 함
	편향성 방지	데이터 편향을 방지하기 위한 절차를 마련하여야함
2차 검사 (정제)	정제 기준의 명확성	사육장 내부에 곤충 객체가 존재하는 이미지만 선별 센서 데이터의 경우 이상치 발생시 데이터 삭제 센서데이터의 경우 결측치 발생 시 정해진 룰대로 데이터 보정
	중복성 방지	데이터 정제 후 정보 비교 후 중복도 여부 검사

3차 검사 (라벨링)	정제 작업 메뉴얼	정제 작업을 위한 메뉴얼 작성 및 관리 여부 검사
	정제 도구	자체적으로 개발한 정제 도구 확보 유사성 검토를 통한 중복 데이터 제거 알고리즘 확보 결측치 보정 알고리즘 확보
	정제 작업 방식	영상을 이미지로 변환하는 작업은 정제 도구에서 자동으로 수행 자체 개발한 이미지 선별 툴에 의해 작업실시 이상치 제거는 시스템에서 제공하는 알고리즘에 의해 수행 결측치 보정은 시스템에서 제공하는 알고리즘에 의해 수행
	라벨링 가이드	목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 내용 가이드 전달
4차 검사 (전수)	어노테이션 항목	목적에 맞는 어노테이션 구성인지 여부를 검사 후 확인된 내용을 포함하도록 작업자들에게 전달 어노테이션 정확성 검사 툴 확보 클래스 의미 정확성 검사 툴 확보
	라벨링 도구	자체적으로 개발한 라벨링 툴 확보
	라벨링 검사 도구	자동화 도구를 통해 검사 후 검사자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상(동적/정적) 이미지의 누락, 번짐 및 조건 오류를 전수 검사
4차 검사 (전수)	부적합 판정 데이터 분포 확인	데이터의 오류율, 특성 분포 확인을 통한 데이터 수집, 정제, 라벨링, 부문 최적화
	외부 검사자	외부 검사자(TTA 등), 도메인 전문가, 데이터 요청자

□ 단계별 검사 방법

단계	세부 업무 내용	일정		주요 산출물
		시작일	종료일	
품질검사 준비	데이터 구조 및 구문 정확성 검증	'21.07.01	'21.07.10	수정 데이터 구조
	다양성 확보 가능성 점검	'21.07.11	'21.07.20	검토 보고서
품질검사 실시	저작 가이드 작성 및 배포	'21.06.20	'21.06.31	저작 가이드
	1차 저작 데이터 검수	'21.09.01	'21.09.15	검수 결과
	검수 결과의 교차 검수	'21.09.16	'21.09.31	교차 검수 보고서
	2차 저작 데이터 검수	'21.10.01	'21.10.15	검수 결과
	최종 검수 가이드 배포	'21.10.16	'21.10.31	저작 가이드 v2.0
품질개선 조치	1차 검수에 따른 저작 정책 개발	'21.08.16	'21.08.31	수정 가이드
	교차 검수 결과에 대한 개선	'21.08.16	'21.08.31	개선 보고서

2.5.3 검수 조직

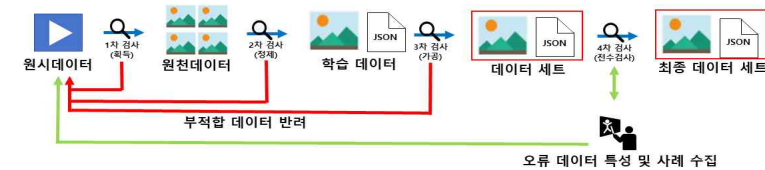
□ 검수 조직 및 업무수행

■ 데이터 검수를 위해 각 역할별 업무내용

구 분	인원	업 무
교육관리자	2명	- 검수 툴 교육자료 제작 - 검수 개요 및 업무수행 교육 - 검수 툴 교육 - 검수 기준 교육
		- 검수 기준 교육 수료
검수자	6명	- 7명 작업자 당 1명 검수자 배정 - 검수 후 오류사항 재 작업 요청
2차 검수자	3명	- 1차 검수 후 산출물 샘플링을 통한 2차 검수 진행

■ 각각의 검사과정 중 발생하는 부적합 데이터는 검사 공정마다 취합하여 반려 조치

■ 검사과정 중 수집된 오류데이터 특성 및 사례는 즉각 수집하여 공정의 기준에 포함되도록 함

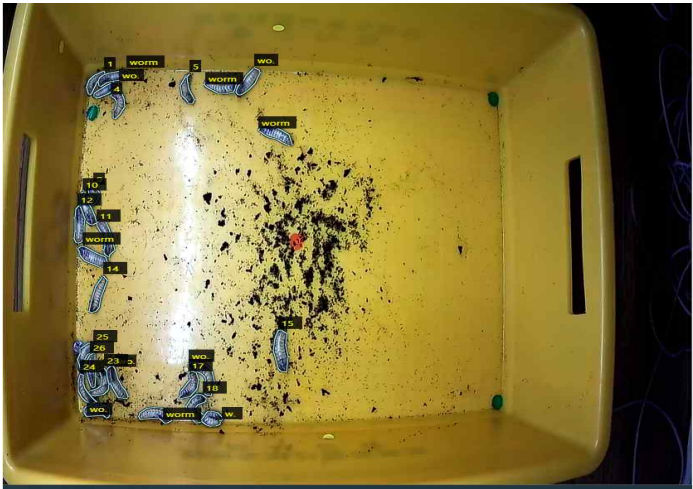


< 데이터 검사 공정 절차 >

2.5.4 검수 도구

파일로드

제어선 로드



검수 프로세스

1) 곤충 객체 instance segmentation label 검수

2) 해당 이미지에 보이는 객체의 region attribute 의 정확성 검수

3) 반려 사유가 있을시 기입하여 가공 작업자에게 일괄 반려

2.5.5 기타 품질관리 활동

외부 전문 데이터 검증 기관(TTA)과 협력 방안

■ TTA와의 협력 방안

- 한국정보통신기술협회의 협업
 - 본 과제에서 품질검증 및 멘토링을 수행하는 한국정보통신기술협회(TTA)와의 협업을 통해 과제 초기에 품질관리체계를 수립하고, 그에 따라 품질 관리를 수행
 - 품질관리체계는 품질목표, 품질관리 방법, 품질관리 일정을 포함

협력방안	상세 내용
초기 검증 진행	초기 인공지능 학습 데이터 설계 시 분류체계의 초안 공유 및 보완 의견 수렴
가이드라인 제공	월별 데이터 가이드라인을 제공하여 TTA로부터 지속적 데이터 품질 검증
체크리스트 제공	진행 상황 확인용 체크리스트를 작성하여 일정 기간별 TTA와 검증 작업 진행

<한국정보통신기술협회의와 협력 방안>

○ 가이드라인 수립 및 제공

- 과제 내 모든 사항을 안내해주는 가이드라인을 제작하여 지속적 검증 진행
- 데이터 구축에 필요한 가이드라인은 과제 초기에 작성하여 TTA의 품질 검증시 제출하여 데이터 구축에 대한 가이드라인에 대한 품질을 보증 받고 과제 내 5번의 가이드라인을 제공하여 이슈 발생 시 즉각 대응하여 문제 해결

■ 체크리스트 제공

○ 전 5차 체크리스트 수립 및 제공

- 진행상황 확인을 위한 체크리스트를 작성하여 기간 별 제공
- 데이터 구축 및 가공에 필요한 항목을 작성하여 해당 항목을 통과했을 때마다 확인할 수 있는 체크리스트를 확인하여 과제 진행 중일정 시점에서 체크리스트를 통한 과제 진척도를 확인할 수 있도록 하여 일정 점검 협의

2.6 활용

2.6.1 활용 모델

2.6.1.1 모델 학습

□ AI 곤충 영상 분석 모델

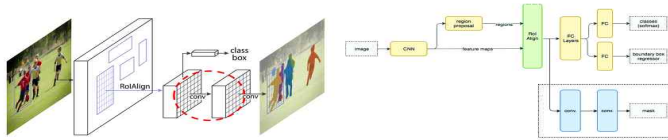
- 곤충의 영상 분석 모델은 곤충의 객체 탐지 및 인식모델과 환경 데이터 분석 모델을 합쳐 만들어지는 모델로 일반적인 Detection 방법론 중 Polygon 형태의 영역 분할 기법을 사용해야 원하는 결과 도출 가능
- 또한 Polygon 형태의 Detection 모델은 Instacne Segmentation 모델과 Semantic Segmentation 모델로 나눌 수 있는데 각 모델의 차이점은 아래와 같음

구분	Semantic Segmentation	Instacne Segmentation
공통 점	같은 종류의 객체에 대해 구분이 없음 	같은 종류의 객체라도 구분이 가능 
주요 모델	U-Net, DeepLabV3+, Semantic FPN 등	Mask R-CNN, PANet, GCNet 등
차이 점	객체를 Detection하고 객체의 영역을 분할하여 각 개체의 영역에 맞추는 기법	

- 본 과제에서는 서로 엉켜있는 곤충의 경우 서로 다른 객체로 인식해야 하기 때문에 Instatnce Segmentation 기법을 사용

■ Mask R-CNN

- Mask R-CNN은 Faster R-CNN에서 영상 내의 분할된 경계박스에서 객체 마스크 네트워크를 추가하여 성능을 향상시킨 알고리즘



<MaskRcnn 구조>

영상에서 여러 관심지역 후보를 지정하여 해당 위치의 특징점 관심 영역 Pooling 방식 대신 관심영역 Align 방식으로 추출하는 기법이며 추출된 특징점으로 객체를 분류하며 해당 객체의 위치와 크기 탐지 특히, Mask R-CNN은 사람과 자동차의 탐지에 우수한 성능을 보이고 있으며 이러한 장점은 작은 객체에 대한 마스크 레이어를 추가함으로써 처리 속도와 시스템 크기에 효율

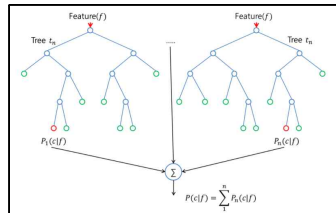
적으로 적용될 수 있음 2 stage object detection을 활용하기 때문에 데이터 정량화에 아주 적합하지만 속도는 빠른 편이 아님

□ 학습 정보

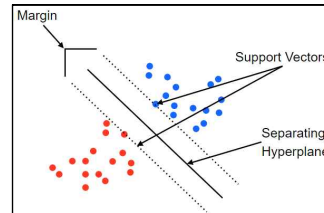
학습 최적 환경				
CPU	i9			
Memory	64GB RAM			
GPU	NVIDIA RTX 2080ti			
Storage	4TB			
OS	windows 10 professional			
학습 조건				
개발 언어	Python			
프레임워크	CUDA 10.0 tensorflow 1.15.0			
학습 알고리즘	Mask-Rcnn			
학습 조건	epoch : 3 batch : 1 optimizer : Mini-Batch Gradient Descent iteration : 30,000 learning rate : 0.01			
모델 input feature	• 학습 데이터셋: 이미지(jpg), 폴리곤 좌표(json), 클래스(json) • 평가 데이터셋: 이미지(jpg)			
모델 output feature	폴리곤 좌표(json), 클래스(json)			
학습 데이터 수량	구 분			
	<table> <tr> <th>대분류</th><th>중분류</th></tr> <tr> <td>RGB 이미지</td><td>곤충</td></tr> </table>	대분류	중분류	RGB 이미지
대분류	중분류			
RGB 이미지	곤충			

□ AI 곤충 생육환경분석 모델

- AI 곤충 생육 환경분석 모델은 곤충 생육환경에서 수집한 센서 데이터를 기본으로 사용. 추가로 곤충 이미지 파일의 폴리곤 데이터를 분석해 곤충 사이즈(길이 또는 둘레 길이, 면적)등을 계산한 후 센서 데이터와 병합 사용. 일정시간 단위로 만들어진 데이터의 라벨링 값은 곤충의 최종 생장 상태의 값을 사용(무게, 길이)
- 일반적으로 예측 알고리즘에는 선형 회귀(Linear Regression) 또는 비선형 회귀방식(Nonlinear Regression)을 많이 사용. 곤충 생산량의 경우 제공되는 온도, 습도, 조도, CO2, 암모니아 등의 입력 파라미터 중 온도, 습도 등은 비선형 속성이 강함(특정 온도 또는 습도에서 생산량이 증가하고 이 값 이상 또는 이하에서는 생산량이 떨어짐)
- 모델로는 앙상블 계열의 랜덤 포레스트, 선형 SVM(Support Vector Machine)을 사용해 서 학습



< 랜덤 포레스트 >



< 선형SVM >

■ 학습 정보

생육환경 분석모델 학습환경	
CPU	i9
Memory	64GB RAM
GPU	해당 없음
Storage	1TB
OS	Ubuntu 18.04
개발 언어	Python 3.8
프레임워크	Scikit Learn 0.24.2, Pandas 1.3.1, Matplotlib 3.4.2
학습 알고리즘	Random forest, 선형 SVM(Supported Vector Machine)
학습 조건	Random forest 하이퍼 파라미터 - max_depth : 12 - min_sample_split : 2 - max_leaf_nodes : None - min_sample_leaf : 1 - n_estimators:10 - max_samples: 0.5 - max_features : None 선형SVM : - ε(epsilon) 파라미터 :1.5 근방에서 조정 - 커널 : Linear - 입력 데이터 스케일 변경(공통) - 측정 시간 : 최초 시간 기준으로 경과시간(초)로 1차 변경 후 전 구간을 다시 0 ~ 1사이의 실수로 스케일 변경
모델 input feature	- 학습 데이터셋 - 센서데이터(json) : 사육장 실내 온도, 사육장 실내습도, 사육장 실내co2, 외 환경정보 - 관리데이터(json) : 곤충길이, 무게 외 관리정보
모델 output feature	곤충 평균 길이, 평균 무게 수치

2.6.1.2 서비스 활용 시나리오

□ 인공지능 데이터 활용 모델 개발 요약표

구분
AI 영상 분석 모델
생육환경분석 모델

□ AI 모델 입력 출력 데이터

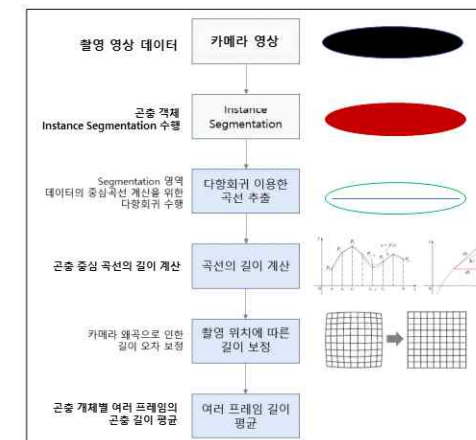
모델명	사용 알고리즘
AI 영상 분석 모델	Instance Segmentation 기반의 영상 객체 인식 알고리즘
생육환경 분석 모델	랜덤포레스트, 선형SVM

□ 인공지능 학습용 데이터 활용 응용서비스 제품 개발 방안

○ AI 영상 인식/분석 모델을 적용한 출하 연계서비스

- 서비스 구현 시나리오

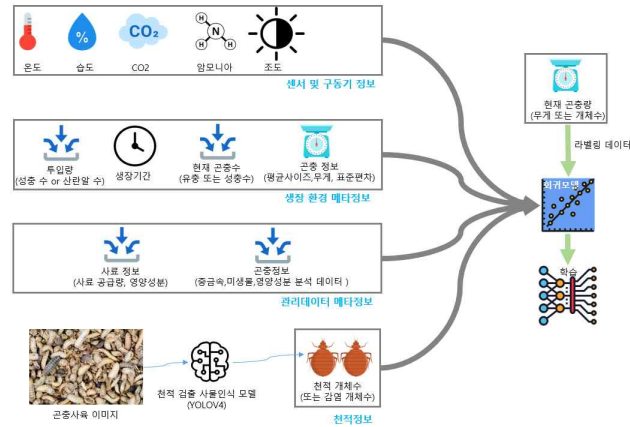
- 곤충 사육장에 설치된 카메라에서 촬영된 영상 데이터를 데이터셋을 이용하여 학습된 학습가중치 파일에 적용 영상을 통해 곤충의 종 및 생육 단계를 분류
- instance segmentation을 통하여 곤충 개별의 크기 및 길이를 산출.



< 곤충 객체 길이/크기 계산 절차 >

○ 곤충 생산량 예측 서비스

- 곤충 생산량예측 서비스는 구동기, 생육정보,환경데이터를 바탕으로 최적의 생육환경을 조성하고, 노동력과 에너지, 양분 등을 효율적으로 이용하여 최고의 곤충 생산량(출하량) 예측가능한 인공지능 곤충 서비스



< 곤충생산량예측 서비스 개념도 >

2.6.2 데이터 제공

■ 데이터셋 인프라 구축 및 운용을 통한 데이터 비즈니스 생태계 마련

- 데이터에 대한 요구사항이 지속적으로 반영되어야 하며, 이러한 요구가 수용될 수 있는 채널을 확보해야하며, 이를 통해 라벨링 오류나 검증 데이터의 오류, 그리고 더 세분화된 라벨링 요구 등 추가 데이터의 요구가 향후 데이터셋 버전에 얼마나 많이 수용되는지에 대한 평가지표를 개발하고 구축된 데이터셋 인프라의 활성화 정도를 판단함
- 구축된 데이터셋을 실제 활용할 학교, 연구기관, 기업체의 요구사항을 충분히 반영하고, 구축된 데이터셋의 초기본, 중간본 등에 대해서 품질 확인을 할 수 있는 평가지표를 개발하고, 전문기관 의 주관 하에 데이터 설계 및 구축 과정 모니터링 및 품질관리 등을 통한 데이터셋 인프라 구축

■ 데이터셋의 지속적인 개선을 위해 확장 가능한 저장도구 제공

- 공개되는 데이터셋이 구축된 이후에도 사용자의 참여에 의해 주기적으로 업데이트되는 방식으로 운용되어야 하며, 여러 사람의 피드백을 통해 발전 가능한 인프라 구축이 필요
- 본 과제를 통해 제작할 저장도구가 사용자 영상 데이터 웹 기반 업로드 기능, 개인정보 보호를 위한 영상 비식별화 엔진 연동기능, 사용자 영상 메타 정보 편집 기능, 사용자 영상 장면 어노테이션 기능, 사용자 영상의 품질 검증을 위한 학습 데이터 검증 프로세스 연동 기능, 클라우드 소싱 현황 및 통계 시각화 기능을 가지도록 하여, 클라우드 소싱을 통해 구축하고자 하는 데이터셋이 기 구축된 학습용 데이터와 동일한 형태(구조)로 저장되어 실제 서비스 분야에서도 활용이 가능하도록 함

2.6.3 데이터 유지보수

□ AI Hub의 데이터셋 저장도구를 활용한 데이터 유지보수

- 사용자 중 데이터셋의 문의 시 고객대응팀을 활용하여 문의 대응하고 필요 시 해당 데이터셋 담당자와 연결 대응
- 이슈 발생 시 데이터셋 저장도구에 검색기능을 활용하여 저장도구 내에서 이슈가 발생한 데이터 보완 및 추가 레이블링 작업 진행하여 지속적으로 고품질의 데이터셋을 유지할 수 있도록 함
- 구축한 데이터셋을 활용한 알고리즘 개발을 지속적으로 진행하여 이슈 발생시 바로 주관기관과 협의를 통해서 실시간 데이터 업데이트를 진행
- 다양한 분야의 피드백 수렴 후 필요시 데이터셋을 보완하여 추가 개선방안 적용