











21년도 인공지능 학습용 데이터 구축 가이드라인

< 스마트팜 통합 데이터(버섯) >

인공지능 데이터 구축	사업 총괄	 유클리드소프트
	데이터 설계	 전북대학교
	데이터 수집 및 정제	  유클리드소프트 (주)엠에이치소프트
	데이터 가공	 유클리드소프트
	데이터 검수	 유클리드소프트
	클라우드 소싱	 유클리드소프트
	저작도구 개발	 유클리드소프트
	AI모델 개발	 (주)엠에이치소프트
가이드라인 작성	 유클리드소프트	이인영
가이드라인 버전	ver 1.0 ('22. 1. 13)	

목 차

1. 데이터 명세 정보	1
1.1 데이터 정보 요약	1
1.2 데이터 포맷	1
1.3 어노테이션 포맷	2
1.4 데이터 구성	3
1.5 데이터 통계	4
1.6 원시데이터 특성	5
1.7 기타 정보	7
2. 데이터 구축 가이드	9
2.1 데이터 구축 개요	9
2.2 문제정의	10
2.3 수집·정제	13
2.4 어노테이션/라벨링	18
2.5 검수	23
2.6 활용	26

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	스마트팜 통합 데이터(버섯)	
활용 분야	- 느타리, 큰느타리, 팽이, 표고, 양송이 5종의 생육과정에 활용 - 고용인원 시 발생하는 교육과정에 활용하여 소모시간 제거	
데이터 요약	양송이, 느타리, 큰느타리, 팽이, 표고 버섯 총 5품종의 전체 생육주기와 환경요소에 따른 버섯 생육주기별 데이터를 구축하고 이를 통해 수확기 버섯 상태를 구분, 버섯 품종인식, 병충해 판단 모델 개발	
데이터 출처	전국 5개 농가 및 국립원예특작과학원 버섯과 스마트 재배사	
데이터 이력	배포버전	ver 1.0
	개정이력	신규
	작성자/ 배포자	이인영/ 유클리드소프트

1.2 데이터 포맷

가공 : 저작도구 화면

검수 : 저작도구 화면

JSON 형식

```

{
  "INFO": {
    "DATASET_NAME": "양송이 생육",
    "DATASET_DETAIL": "12.5와트형 통합데이터(버섯)",
    "VERSION": "1.0",
    "LICENSE": "-",
    "CREATE_DATE_TIME": "2021-12-06 16:25:53",
    "CONTRIBUTOR": "-",
    "URL": "https://www.klab.or.kr",
    "CATEGORY_NAME": "양송이"
  },
  "IMAGE": {
    "IMAGE_URL": "https://images.klab.or.kr/2021/11/10/5818816d8457a6d6d8172a65988b.jpg",
    "IMAGE_FILE_NAME": "양송이_생육(버섯)_1044853.jpg",
    "WIDTH": 1080,
    "HEIGHT": 720,
    "ANNOTATION_COUNT": 1
  },
  "ANNOTATION_INFO": {
    "ID": 6522626,
    "BOUNDING_BOX_X_COORDINATE": 204,
    "BOUNDING_BOX_Y_COORDINATE": 135,
    "BOUNDING_BOX_WIDTH": 295,
    "BOUNDING_BOX_HEIGHT": 142,
    "SEGMENTATION": null,
    "SEGMENTATION_AREA_TOTAL": null,
    "CROWDSOURCING_OPERATION_ALTERNATIVE": true
  },
  "META": {
    "DBYHS_SPCCHKN": "양송이(버섯)",
    "DBYHS_NORMALITY_ALTERNATIVE": false,
    "IP_CAMERA_ID": 8,
    "WIND_SPEED": 0.0,
    "AIR_VELOCITY": 0.0,
    "TEMPERATURE": 18.2,
    "HUMIDITY": 91.6,
    "ILLUMINATION_INTENSITY": 0.0,
    "CAMERA_SHOT": 1960,
    "SEQUENCE": null,
    "IMAGE_CREATE_DATE": "2021-11-30",
    "IMAGE_CREATE_TIME": "11:08:11",
    "IMAGE_CREATE_DAY_OF_WEEK": "Sunday",
    "STIR_LENGTH": null,
    "STIR_THICKNESS": null,
    "PILEUS_DIAMETER": null,
    "PILEUS_THICKNESS": null,
    "GROSS_WEIGHT": null
  }
}

```

1.3 어노테이션 포맷

분류	구분	항목명	설명
파일 정보	1-1	DATASET_NAME	데이터셋명
	1-2	DATASET_DETAIL	데이터셋상세설명
	1-3	VERSION	버전
	1-4	LICENSE	라이선스
	1-5	CREATE_DATE_TIME	파일 생성 일자
	1-6	CONTRIBUTOR	기여자
	1-7	URL	URL
	1-8	CATEGORY_NAME	카테고리 명
이미지 파일 정보	2-1	IMAGE_URL	이미지 URL
	2-2	IMAGE_FILE_NAME	이미지 파일명
	2-3	WIDTH	이미지 가로
	2-4	HEIGHT	이미지 세로
	2-5	ANNOTATION_COUNT	이미지당 라벨 개수
어노테이션 정보	3-1	ID	어노테이션 식별자
	3-2	BOUNDING_BOX_X_COORDINATE	바운딩박스 X좌표
	3-3	BOUNDING_BOX_Y_COORDINATE	바운딩박스 Y좌표
	3-4	BOUNDING_BOX_WIDTH	바운딩박스 가로
	3-5	BOUNDING_BOX_HEIGHT	바운딩박스 높이
	3-6	SEGMENTATION	세그멘테이션
	3-7	SEGMENTATION_AREA_TOTAL	세그멘테이션영역합
	3-8	CROWDSOURCING_OPERATION_ALTERNATIVE	크라우드소싱 작업여부
메타	4-1	DBYHS_SPCCHKN	병해충구분
	4-2	DBYHS_NORMALITY_ALTERNATIVE	정상여부
	4-3	IP_CAMERA_ID	IP카메라 아이디

4-4	WIND_SPEED	풍속	0~5	NUMBER	N	
4-5	AIR_VELOCITY	풍속	0~5	NUMBER	N	
4-6	TEMPERATURE	온도	0~100	NUMBER	N	
4-7	HUMIDITY	습도	0~100	NUMBER	N	
4-8	ILLUMINATION_INTENSITY	조도	-60~100	NUMBER	N	
4-9	CARBON_DIOXIDE	이산화탄소(CO2)	-20~7600	NUMBER	N	
4-10	GUIDELINE	가이드라인		array	N	NUMBER ARRAY
4-11	IMAGE_CREATE_DATE	이미지 생성 날짜	YYYY-MM-DD	String	N	
4-12	IMAGE_CREATE_TIME	이미지 생성 시간	HH:mm:ss	String	N	
4-13	IMAGE_CREATE_DAY_OF_WEEK	이미지 생성 요일	20	String	N	
4-14	STIPE_LENGTH	대길이	0	NUMBER	N	
4-15	STIPE_THICKNESS	대두께	0	NUMBER	N	
4-16	PILEUS_DIAMETER	갓직경	0	NUMBER	N	
4-17	PILEUS_THICKNESS	갓두께	0	NUMBER	N	
4-18	GROSS_WEIGHT	총중량	0	NUMBER	N	

1.4 데이터 구성

1. 폴더구조

- 1_원천데이터 / 2_라벨링데이터
 - 작물 품종(느타리, 양송이, 큰느타리, 팽이, 표고)
 - 클래스(배양, 생육, 병해)
 - 개별파일

2. 개별 파일명 규칙

작물 품종_재배실_카메라일련번호_이미지일련번호

예시: 양송이_생육실_8_12509217.json / 양송이_생육실_8_12509217.jpg



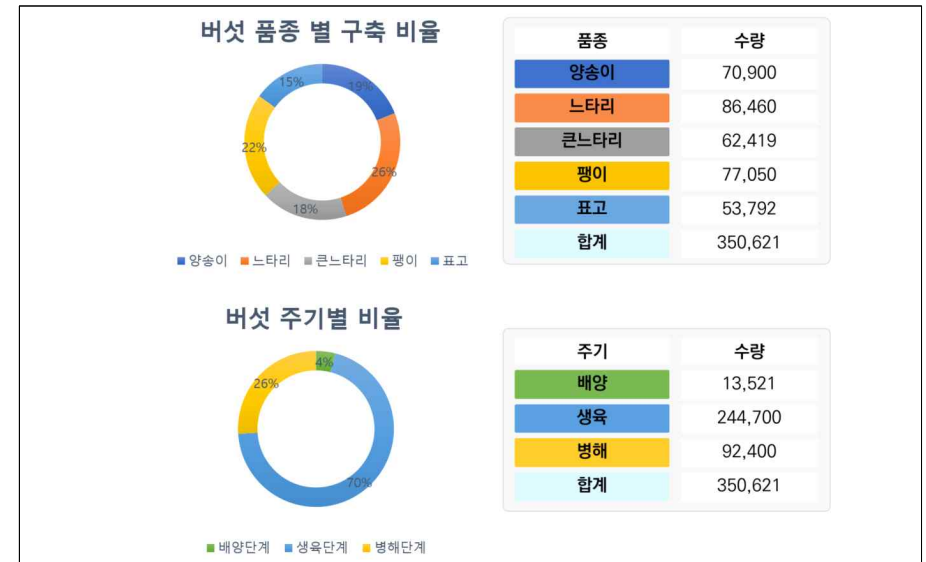
1.5 데이터 통계

1.5.1 데이터 구축 규모

- 원천데이터 35만 621건 (양송이 7만 9백 건, 느타리 8만 6천 460건, 큰느타리 6만 2천 419건, 팽이 7만 7천 50건, 표고 5만 3천 792건)을 활용하여 각각 Polygon Segmentation 5만 4백 건, Bounding box 27만 9천 721건으로 가공 하여 총 35만 621건의 데이터 도출
- 5품종의 버섯(양송이, 느타리, 큰느타리, 팽이, 표고)으로 분류하였으며, 생장별 3단계 (배양단계, 생육단계, 병해충)구분하여 데이터 구축

데이터 종류	품종	CLASS	수량	데이터 형태
버섯	양송이	생육단계	50,400	Polygon
		병해충	20,500	
		배양단계	4,960	
	느타리	생육단계	59,500	Bounding Box
		병해충	22,000	
		배양단계	2,719	
	큰느타리	생육단계	43,700	Bounding Box
		병해충	16,000	
		배양단계	4,050	
	팽이	생육단계	53,000	Bounding Box
		병해충	20,000	
		배양단계	1,792	
	표고	생육단계	38,100	Bounding Box
		병해충	13,900	
	합계		350,621 건	

1.5.2 데이터 분포



1.6 원시데이터 특성

1.6.1 대상분류



1.6.2 제약조건

□ IP카메라를 이용한 영상/이미지 데이터

- 버섯의 각 종마다 IP 카메라를 설치하여 실시간으로 영상 데이터를 획득하고 획득한 영상 데이터에서 시간 정보를 추출하여 시간 정보를 표시
- 배양단계에서는 개체별 1개의 카메라로 측면 촬영
- 생육단계에서는 개체별 2개의 카메라로 상부, 측면 동시에 2개의 각도에서 촬영
- 1종 당 30개체 이상의 버섯을 포함하고 생육 모습을 확보하기 위해 다각도로 설치하여 목적에 맞는 데이터를 확보

□ 직접 촬영을 이용한 영상/이미지 데이터 확보

- 데이터의 다양성 및 버섯 수확기를 판단하기 위해 수집장치로 획득하는 데이터 외의 추가 데이터를 확보
- 촬영 목표와 메뉴얼을 구축하여 같은 양식으로 영상 데이터를 확보하고 획득한 영상 데이터를 이미지 데이터로 변환
 - 중복된 이미지를 제거하기 위해 1초에 2장씩 이미지로 변환
 - 변환된 이미지와 적합한 환경정보 매칭

□ 시계열 데이터 확보

- 농장에 설치된 스마트 컨트롤러 기기를 통해 일정한 간격으로 환경정보를 수집
- 버섯 종류별 필수 환경요소(온도, 습도, 이산화탄소, 조도, 풍속)와 대상 요소에 맞게 데이터를 수집하고 수집된 데이터는 xlsx 형식으로 통합·취합하여 관리
- 수집 시, 시간 정보도 같이 확보하여 동 시간대 영상 정보와 관리

1.6.3 속성

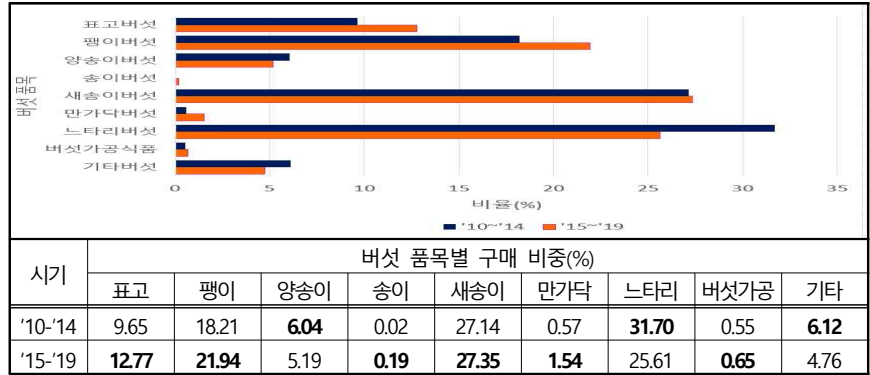
No	항목	타입
1	이미지 파일명	String
2	이미지 사이즈	NUMBER

< 데이터 속성 정의 >

1.7 기타정보

1.7.1 포괄성

- 2019년 기준 구매 비중 상위 5개 품종은 큰스타리(새송이), 느타리, 팽이, 표고, 양송이버섯으로 조사됨



주1: '19년의 경우 1월에서 11월까지 반영

주2: 버섯가공은 건표고, 버섯분말 등을 포함한 전체이며, 기타는 목이버섯 등을 포함

< 버섯 품목별 구매 비중 변화 >

□ 버섯분야 데이터셋 국내·외 현황

(국내)

- 국내 대규모 버섯 생산 농가는 대부분 공업화 시설생산에 기반하고 있으며 공장화 시설 재배 수준은 국제 선진 수준이나 품질개선&효율화 등을 위한 데이터셋은 국제 시장에서 경쟁력을 갖춘 수준이 아닌 것으로 평가
- 데이터셋을 적용한 생산개선은 프로세스 전반의 변경을 요구하기 때문에 안정적 수익구조를 보이나, 기술도입 여력이 있는 농가에서 오히려 데이터셋 구축 관심도가 낮다. 생산혁신을 추구하는 중소농가에서 수요가 높으나 데이터셋 구축 여력이 부족한 상황

(국외-기술선진국)

- 네덜란드, 미국 등은 세계 최고 수준의 기술력을 바탕으로 축적한 다양한 분야의 데이터셋 구축 경험이 버섯산업에도 영향을 미치고 있다. 데이터셋 활용을 통한 생산효율이 매우 높고, 자동화 수준 역시 데이터에 기반하고 있어 인공지능화도 시장 최고 영역에 위치

(국외-개발도상국)

- 전통적인 버섯 생산 방식에서 큰 변화가 없어 효율성 및 품질이 낮고 생산량이 높은 양극적 구조에 대한 국가적 자각을 겪은 이후, 선진기술 도입을 표방하며 데이터셋 구축을 추진하고 있으나 활용 목적성이 뚜렷하지 않아 아직 성장세는 낮으나 데이터셋 구축과 활용 잠재력이 매우 높은 것으로 평가되고 있음

1.7.2 독립성

□ 재산권 및 동의·계약서

- 작업을 내 타인의 지적재산권 활용이 필수적으로 요구되어 통상실시권을 요하거나, 기술이전 또는 지적재산권의 일부 활용이 예상되는 경우 지적재산권 보유자와 이에 관한 전용/통상실시권 관련 계약을 체결하여 법적 분쟁을 사전에 방지
- 민감정보 접근에 있어 클라우드 워커 및 과제참여인력 전원의 보안서약서와 개인정보수집제공이용 동의서를 확보하고 보안관리 전담인력 지정, 주기적 내부 보안관리교육 등을 통해 개인정보에 대한 안전 절차를 준수 작업데이터의 보안 민감성에 따라 접근 권한을 지정하고 주요 데이터 접근 이력을 누적기록
- 수집된 작업물이 지적재산 대상 데이터 여부에 포함될 가능성을 고려, 작업 완료 이후 특허 출원/신청 등 지적재산 권리확보 작업을 후속조치하고 컨소시엄의 지적재산권 및 사용&실시권리의 영역 배분은 컨소시엄 내부 협의를 거쳐 도출
- 관련 계약서는 모두 과학기술정보통신부 및 NIA에서 제공하는 표준양식을 기준으로 작성하며, 계약 당사자 간 법적 검토와 자문 등을 통해 현행법령을 위반하지 않도록 사전 근거를 확보

1.7.3 유의사항

□ 데이터 백업 방안

- 시스템에 의한 장애 발생을 사전에 차단하여 사업의 원활한 진행을 위하여 독자적인 저장 스토리지를 운영하는 것보다 주기적으로 예방 점검 등 시스템을 안정적으로 운영하는 클라우드 서버를 활용하여 학습용 데이터를 저장
- 로컬 스토리지 서버와 비교하여 보안, 성능, 관리 측면에서 우수한 클라우드 서버 활용
- 정기적으로 클라우드에 저장되어있는 데이터의 안정성을 위하여 제안사 파일 서버로 백업(2중화) 실시

□ 데이터 복구 방안

- 복구 관리 계획 수립
 - 데이터의 유실 시점과 백업 시점을 확인하여 담당자와 백업 시점 협의
 - 데이터 유실의 범위를 확인하고 담당자와 협의하여 부분복구 혹은 전체복구 진행 협의
- 데이터 속성을 고려한 데이터 복구 수행
 - 최신 복구 매뉴얼 및 절차를 참고하여 절차에 따라 복구 수행
 - 데이터 속성에 따라 데이터별 복구 검증
- 긴급 상황을 고려하여 복구 담당자를 지정하여 수행
 - 협력사 전문 엔지니어, 수행PL, 고객 담당자로 구분하여 백업복구 수행
 - 백업복구 수행 시 시스템 부하량을 고려하여 신속한 복구 검증 수행

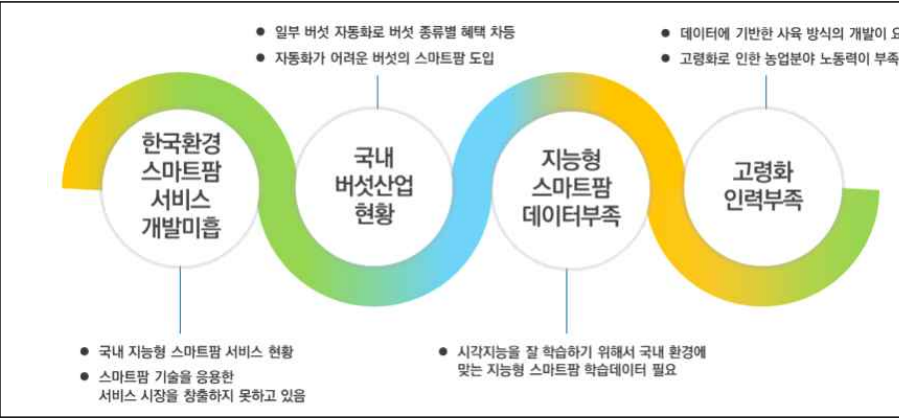
2.1 데이터 구축 개요

2.1 데이터 구축 개요

		- 재배사 안의 온도, 습도, 이산화탄소, 조도, 풍속 등 버섯의 품질에 영향을 주는 환경정보 측정
	데이터 정제	1) 중복된 이미지를 제거하기 위해 1초에 2장씩 이미지로 변환 2) 변환된 이미지와 적합한 환경정보 매칭
가공	라벨링	5품종 버섯 이미지 가공 1) Bounding Box 어노테이션 2) Polygon Segmentation 어노테이션
	저작도구	가공자가 직관적으로 작업 할 수 있는 저작도구 사용이 가능한 사이트 오픈 1) Bounding Box 저작도구 2) Polygon Segmentation 저작도구
	품질관리	-- 3단계 품질검증 및 학습데이터 정확도 향상 (교차검수 -> 제3자검수 -> 최종검수) · 유효성 검증 1) 의미정확성 및 IOU=0.8이상 2) 데이터 품질 검증 수행
활용	AI 모델링	1) Bounding Box 품종 확인, 수확기 판단, 병충해 판단에 맞춰 학습 2) Polygon Segmentation 품종 확인에 맞춰 학습 · 학습결과를 기반으로 오류 수정 및 학습

데이터 구축 개요

2.2.1 임무 정의



- **농생명분야 데이터**는 그 특성상 계절, 지역, 품종 등 다양한 요인이 있어 데이터 표준화와 수집, 관리, 분석 체계 마련을 위해서는 일정 규모 이상의 환경조성과 여기에 이어지는 광범위한 기초데이터 수집이 필요함
- **버섯 데이터셋**은 데이터 수집, 분석, 이용, 공유 전 과정을 원스톱 형태로 지원할 수 있는 장기적인 데이터

생태계 구축의 초~중반 과정을 지원하는 데이터로 활용할 수 있음

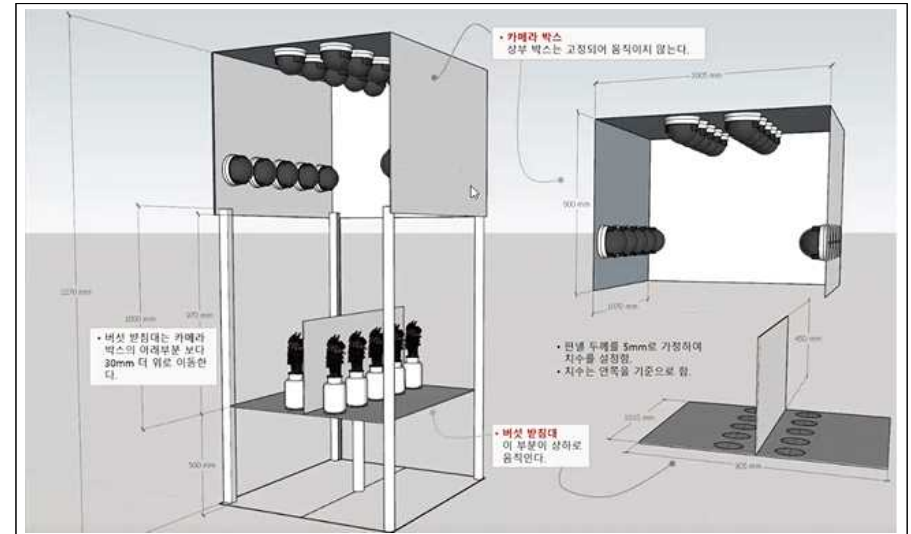
- 데이터가 전무한 버섯의 종류별 생육데이터를 확보하고 이를 활용하여 버섯 농가의 생육 환경을 개선 가능
- 버섯 품종 인식 모델, 버섯 수확기 판단 모델, 버섯 병해 판단 모델을 활용하여 버섯의 최적의 수확기를 판단하여 고품질 버섯을 균등하게 재배하는 것을 돕고, 버섯 병해를 최소화하여 버섯 생산량을 증가시킬 수 있어 버섯 산업을 활성화 함.

2.2.2 데이터 구축 유의사항

- 데이터셋을 설계할 때 가장 중요하게 고려했던 점은 현장과 동일한 실질적인 버섯 생육 데이터를 구축하는 것임.

□ 촬영환경 제어장치 및 IP카메라

- 버섯이 자라는 환경조건이 습도와 조도 때문에 이미지를 촬영하기에 부적합한 경우가 상당히 존재
- 촬영 시점에서 버섯에 영향을 주지 않고 촬영하기 위하여 환경 제어장치가 반드시 필요.
- 병 재배, 봉지 재배, 볏짚 배지 재배 등 다양한 재배방식에 따라 촬영을 위한 환경제어 장치 및 IP카메라 설치 방법 및 대수 변화
- 생육 단계의 전반적인 모습을 실시간으로 촬영하기 위한 목적으로 기기로 학습에 필요한 데이터를 얻기 위해서는 하나의 타겟 버섯 당 2개의 카메라가 필요 (상부 1개, 측면 1개)
- 배양 단계에서는 상부 촬영이 불필요하고, 중균 배양 진행이 단순하여 타겟 버섯 당 1개의 카메라만 필요
- 버섯 생육 특성상 습도가 높기에 우선적으로 카메라는 생육장소(내부/외부)와 생육시간(주/야간), 생육조건(습도)에 따라 정확하고 지속적인 데이터 수집과 추출이 가능하도록 방수가 가능한 카메라가 필요
- 설치된 IP카메라를 통해 원격지에 데이터가 저장되고 저장된 영상을 통해 이미지를 자동 추출

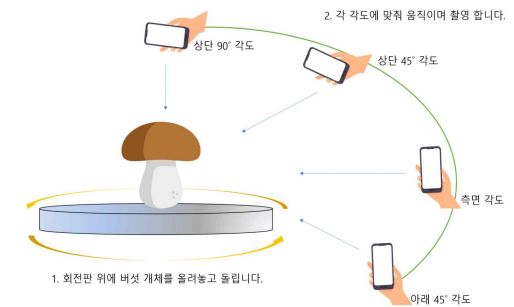


IP카메라를 이용한 영상/이미지 데이터 확보



□ 카메라 촬영기기

- 이미지 데이터 획득을 위한 기기로 200만 화소 이상의 화질을 가진 카메라 또는 장비 필요
- IP카메라를 활용하여 원격지로 영상 데이터 실시간 수집 및 처리
- 데이터의 다양성을 확보하기 위해 직접 촬영도 함께 진행
- 직접 촬영 시에는 정해진 매뉴얼을 준수하여 촬영하고 촬영시 버섯의 생육 단계를 표시하여 획득



직접 촬영을 이용한 영상/이미지 데이터 확보

2.3 수집·정제

2.3.1 원시데이터 선정

- 버섯의 생육이 이뤄지고 있으며 데이터 수집을 위한 장비 설치·제공 등 장소를 제공해 줄 버섯 농가 선정 (5개 농가 및 국립원예특작과학원 스마트배양실 4실 활용)

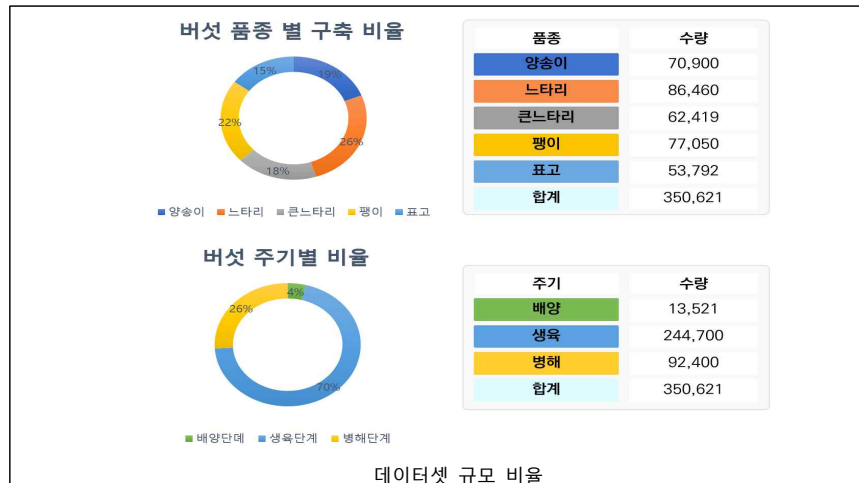
버섯 종류	농장	지역	재배방식
느타리	채인	경기 화성	병재배
큰느타리	김항익농장	전남 해남	병재배
표고	으뜸표고	충남 천안	통배재배
팽이	연우	충북 음성	병재배
양송이	정해평	충남 부여	균상재배
병해	국립원예특작과학원	충북 음성	병재배/균상재배




버섯 재배 농가 확보

- 국립원예특작과학원 버섯과와 스마트팜 사업을 진행했던 농가들로, 환경데이터 수집 경험이 있는 농가들이며, 이미지 데이터 수집을 위한 장비 설치 공간 확보
- 병해의 경우 발생 초기에 재배농가에서 다른 개체로의 확산을 방지하기 위해 즉시 제거해야 하므로, 병해 발생 후 생장 이미지 수집이 불가하여, 국립원예특작과학원 내의 스마트배양실을 이용하여 병해를 발생시켜 데이터를 수집

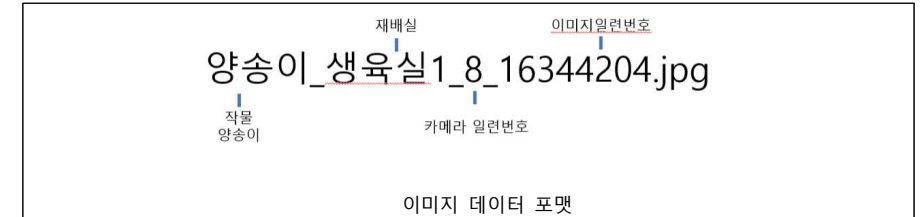
2.3.2 데이터셋 규모



- 획득하고자 하는 데이터 품종(5종)을 선정하고 설치된 기기 및 촬영 도구를 활용하여 생육 20분, 배양 120분, 병해 10분 간격으로 데이터를 수집
- 다양한 종류의 이미지 데이터를 확보하기 위해서는 버섯 생육 사이클은 최소 2회 이상 진행
- 버섯에 따라 생육주기가 상이하므로 생육주기에 맞춰 유연하게 이미지 데이터를 확보하고 환경정보 데이터는 확보된 이미지 데이터 시간과 매칭하여 수집

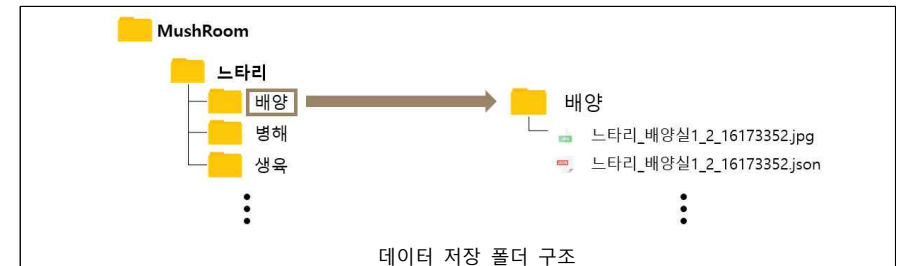
2.3.3 데이터 포맷

□ 이미지 데이터



- 학습/촬영 타겟을 지정하고, 생육장소와 주기를 촬영 각도 구분하여 구축
- 주기와 수집 날짜를 명시하고 데이터의 종류와 촬영부위, 수집 시간을 구분지어 표시하여 환경정보와 일치하도록 구성

2.3.4 저장구조 포맷



□ 저장 구조

- 영상파일(.mp4) : 농장 식별 번호 폴더(C01)에 각각의 IP카메라 번호를 부여하여 날짜별로 저장
- 이미지파일(.jpg) : 영상이미지를 분할하여 각각의 이미지에 번호 부여
- 어노테이션(.json) : 가공이 완료되면 각 농장별로 json파일 형태로 저장
- 환경데이터 : C01_CCTV1.csv 각 농장별로 얻은 데이터 저장

2.3.5 수집·정제 절차



□ 생육 단계별 데이터 수집

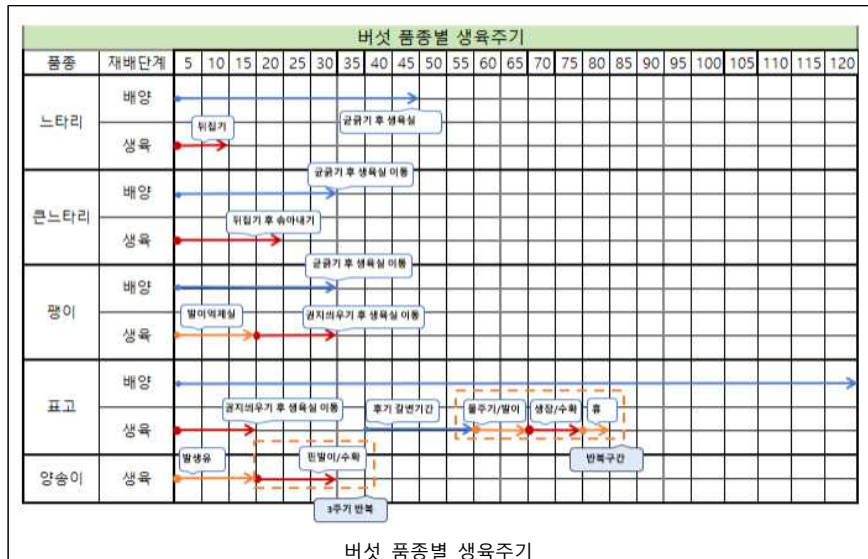
- 버섯별 의미 있는 데이터를 확보하기 위하여 측정에 대한 매뉴얼 구축
- 한 개체별 이미지데이터(촬영기기-IP카메라), 환경데이터(각종 센서), 측정데이터(수작업 측정)를 확보하기 위하여 데이터 확보 환경을 구축
- 설치된 IP 카메라를 통해 버섯 생육 단계별 영상/이미지 데이터 수집 진행
- 생장 단계의 버섯 측정 정보 수집(총무게, 갓 직경, 갓 두께, 대 길이, 대 두께, 유효경수 등)

□ 생육장소에 따른 조건

- 버섯은 품종 및 생육 단계별로 환경조건이 다양하게 설정되며, 특히 이미지 촬영 품질에 영향을 줄 수 있는 습도 및 조도 조건도 다양함
- 이에 따라 촬영 시 습도 및 조도 조건을 조절하기 위하여 별도의 제어 장치를 구성하고, 이 장치는 다른 버섯 생육에 주는 영향을 최소화할 수 있도록 제작하여 촬영을 진행
- 촬영 장치는 버섯의 종류에 따라 재배 형태가 다르므로 이에 맞게 제작하여 설치하고 개체별 이미지를 확보
- 촬영에 적합한 습도 조도를 유지하는 박스 내부에 카메라를 설치하고, 개체별로 1회 촬영 시 2장(상부, 측면) 사진 촬영
- 병충해로 인한 배지 손상, 기형버섯 등이 발생 시 해당 모습을 이미지로 수집하여 데이터 확보
- 병해 발생 시 재배사 전체로 확산됨으로 농가에서 병해 발생 이미지 데이터를 구축할 수 없어, 국립원예특작과학원 버섯과 배양실에서 별도로 병해를 발생시키고, 진행 전 과정에 대한 이미지 데이터를 구축

	데이터 획득 형태	수집장비
1	버섯 5종 이미지 데이터	IP 카메라
2	버섯 5종 환경정보 데이터	생육동 스마트컨트롤러
3	버섯 5종 이미지 데이터	스마트폰

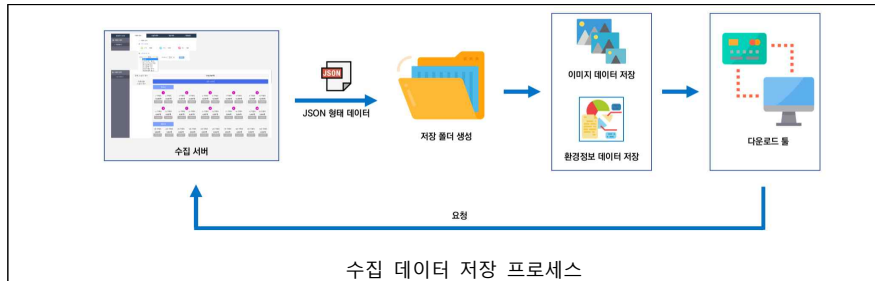
2.3.6 수집·정제 기준



□ 인공지능 학습 데이터 가공 및 제작

- IP카메라 및 스마트폰을 통해 버섯 원천 데이터(이미지)를 확보
- 확보된 데이터 정제 작업 진행
 - 버섯 종별, 날짜별, 시간별 등 분류
 - 중복 이미지 제거
 - 손상된 이미지 제거
 - 불량(카메라 오류, 타겟 손상 등) 이미지 제거
 - 환경정보 매칭
- 정제과정을 거친 데이터를 학습을 위한 가공 단계로 전환

2.3.7 수집·정제 도구

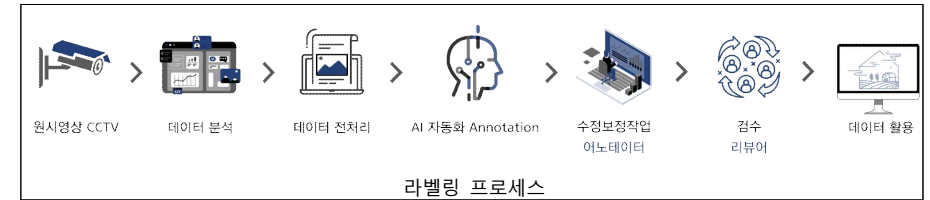


- 데이터 다운로드 도구
 - HTTP 프로토콜로 IP 카메라 촬영 서버에 요청 URL을 보내 json 형식으로 데이터 획득하도록 개발
 - 획득한 Json 데이터에서 HTTP 프로토콜 이미지 URL을 통해 이미지 폴더를 생성하고 이미지 데이터와 환경정보 데이터를 저장하도록 개발
- 데이터 중복성 제거 및 유효성 검사 도구
 - 획득한 데이터의 이름을 비교하여 같은 이름을 가진 데이터 중 1개 삭제
 - 수집 간격에 맞지 않은 데이터 삭제
Ex) 생육 데이터는 20분 간격 수집이나 10분 간격 들어올 시 해당 데이터 삭제
- 이미지 변환 도구
 - 폴더 내 모든 영상 파일을 확인하여 영상 파일의 이름으로 폴더를 생성하고 해당 폴더에 ffmpeg을 사용하여 이미지 추출 프레임 명령어 실행
 - 과한 중복성을 제거하기 위해 각 폴더 당 1초에 2장씩 추출하도록 개발
- 데이터 수집 단계 고려사항
 - 농가 버섯생육 이미지 데이터는 버섯 농가 5곳을 선정하여 선정된 농가에서 버섯 재배 환경 데이터를 수집하고, 350,000장 이상의 이미지 데이터를 획득함
 - 직접 촬영 시 정해진 매뉴얼을 준수하여 촬영하고 촬영 시 버섯의 생육 단계를 표시하여 획득함
 - 각 농장의 정보는 확약서를 통하여 개인정보 및 저작권에 대한 동의를 얻음
- 데이터 정제 시 품질 확보 고려사항
 - IP 카메라 영상 데이터에서 불필요하거나 의미 없는 데이터를 제거하여 영상 데이터가 적절한 규모로 저장 메모리를 유지하면서 분석에 이용될 수 있도록 함

- 데이터 중복으로 인한 편향 학습을 방지하기 위하여 동일한 형태의 데이터가 반복적으로 나타나는 경우 이를 제거함
- 수집한 IP카메라 영상데이터를 정제하는 전 과정은 자체 저작도구를 통하여 관리하여 정제과정의 일관성과 균질성을 보장하는 진행현황을 관리함
- 정제 과정에서 생성된 영상데이터는 데이터베이스에 저장 관리하고 가공단계에 활용되도록 함
- 추출한 Bounding box에 대한 데이터는 json 파일로 구조화함

2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차



□ 라벨링 작업 방식

- 라벨링 작업은 데이터 구축의 전체 프로세스를 진행하는데 전문화된 저작도구를 지원하는 클라우드 소싱 플랫폼을 통해 진행함
- 저작도구는 라벨링 데이터의 저장, 도구 관리자 기능, 작업자관리, Bounding Box, Polygon Segmentation 데이터입력 작업관리(작업배정, 결과제출, 반려 등)와 검수 기능을 가진 프로세스와 서버 및 DB로 구성함
- 저작도구에는 작업자의 작업 현황과 결과 검색 및 작업 이력이 관리되며, 통계자료와 작업보수의 확인이 가능하도록 기능을 구성함

□ 라벨링 프로세스

- [1 단계] 클라우드 기반 저작도구 기반 데이터 가공 업무 분배
- [2 단계] AI 학습용 데이터 저작도구를 이용한 데이터 가공
- [3 단계] 학습 데이터 저장 및 구축

절차	내 용
입력	원천데이터를 라벨링 도구를 제공하는 클라우드 소싱 플랫폼에 업로드
가이드라인 확인	라벨링 기준 및 검사 기준으로 활용되는 가이드라인 확인
라벨링 Bbox/Polygon	버섯에 대한 Bbox 라벨링, 클래스 분류 진행
1차 검수	Bbox.라벨링의 작업물에 대한 가이드 라인을 기준으로 검수를 진행 2차 검수와 교차하여 한 번 더 검수함 1. 승인 - 교차검수 진행 2. 반려 - 라벨링 재작업
2차 검수	Bbox 라벨링의 작업물에 대한 가이드 라인을 기준으로 검수를 진행 1차 검수와 교차하여 한 번 더 검수함 1. 승인 - 교차검수 진행 / 2. 반려 - 라벨링 재작업
데이터 추출	이미지+Json 한 세트의 형태로의 추출 및 제출

2.4.2 어노테이션/라벨링 기준

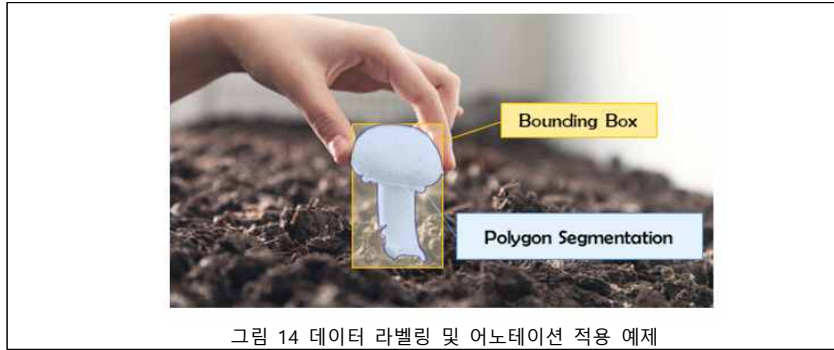


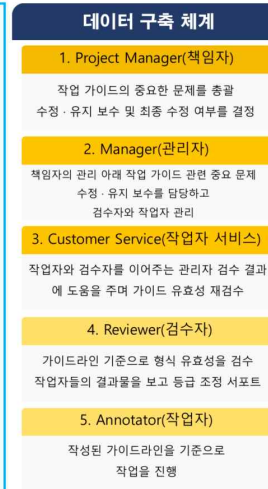
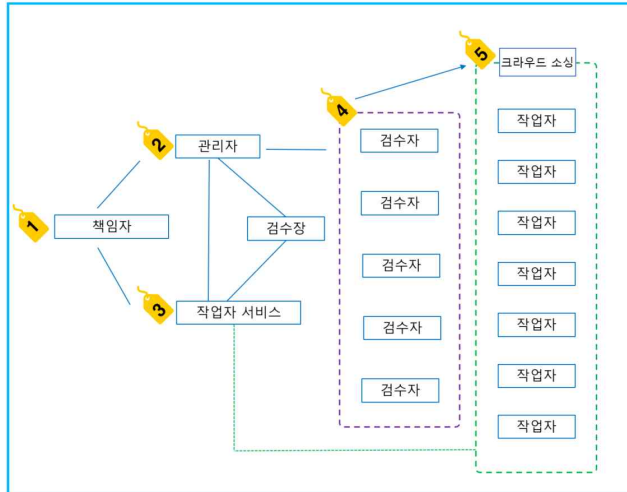
그림 14 데이터 라벨링 및 어노테이션 적용 예제

- 클라우드 소싱으로 작업자를 모집하고 온라인-클라우드 기반 작업 환경에서 작업
- 2단계 Annotation 및 보정 작업을 수행하며 학습 데이터의 정확성과 품질을 높임
- 2단계 Annotation 과정에서는 학습 데이터 가공 자동화 단계에서 탐지하지 못했거나, 라벨링 및 태깅 오류가 있는 개체들의 보정 작업을 수행

라벨링 방식	설 명	이미지 예시
Bounding Box	객체를 직사각형 모양의 박스 안에 포함되도록 그리는 라벨링 방법으로 데이터 라벨링 작업에서 가장 일반적으로 사용 - Bounding Box는 객체를 전체가 커버되도록 하며, 박스 안에 객체 이외의 여백을 최소화하도록 지정	
Polygon Segmentation	다각형 모양으로 개체의 가시 영역 외곽선을 따라 점을 찍어 그리는 라벨링 방법, 객체 이외의 포함된 빈 공간으로 인해 발생하는 오류에 대응할 수 있는 기능 - 물체를 정확하게 인식하기 위해서 사용, 사물의 테두리를 따라 그리는 것을 통해 여백 없이 정확히 물체만을 인식하기 위해 사용	

올바른 라벨링 작업 예시	설 명
	<ul style="list-style-type: none"> ● Bounding box가 버섯 개체를 모두 포함하며 box의 크기가 외곽선에서 최대한 타이트하게 라벨링 ● annotation : 표고
	<ul style="list-style-type: none"> ● 사진에서 버섯이 가려지는 것을 감안하여 라벨링을 보이는 외곽선에서 최대한 타이트 하게 라벨링 ● 보이는 않는 영역을 추정하여 라벨링 ● annotation : 표고, 표고
	<ul style="list-style-type: none"> ● 여러 개의 버섯이 겹쳐진 경우 데이터의 정확성을 위해 polygon방식을 사용하여 라벨링 ● 보이지 않는 영역을 추정하지 않고 라벨링 ● 보이는 영역에 대해서만 라벨링 ● annotation : 표고
잘못된 라벨링 작업 예시	설명
	<ul style="list-style-type: none"> ● 사진에서 특정할 수 있는 버섯의 외곽선에 비해 Bounding box가 크게 그려져 있어 잘못 라벨링된 데이터로 분류
	<ul style="list-style-type: none"> ● 사진에서 버섯의 모양을 특정할 수 있는 외곽선 부분보다 Bounding box가 작게 라벨링 되어있어 라벨링 된 데이터로 적합하지 않음

2.4.3 어노테이션/라벨링 조직



□ 작업자 대상 매뉴얼 작성

- 데이터 획득, 정제, 라벨링, 검사 단계에 참여하는 작업자들이 인공지능 학습용 데이터셋 구축 취지에 부응하여 데이터 제작이 이루어질 수 있도록 작업자들이 직접 활용하는 매뉴얼을 제작
- 작업자 대상 매뉴얼에는 구축 목적, 정의 제작 절차, 제작 도구 활용 방법과 작업 기준, 작업 결과 처리, 저장 방법 등의 내용을 포함
- 작업자 관점에서 데이터 제작 과정에서 발생할 수 있는 다양한 케이스를 포함하여 매뉴얼을 제작

□ 클라우드 소싱 인력에 대한 처우

- 플랫폼 안내서 및 작업 가이드를 통한 작업방법 안내
- 데이터 가공 방법을 알기 쉽게 설명한 작업 가이드라인 문서 제작
- 클라우드 소싱 작업 참여자들이 온라인에서 확인 가능한 문서 제공
- 문서로 작업 내용 학습이 어려운 경우 체험하기 프로젝트 제공
- 클라우드 소싱 작업 참여자들의 상시 문의 가능한 CS 창구 마련
- 각 체험에 작업자들이 참여해 사전학습을 진행하여 실제 작업시 생산성 극대화 기대
- 클라우드 소싱 인력 계약서는 기 법률 검토를 받은 계약서로 진행
- 클라우드 소싱 기간 이후 정규직 전환에 대한 검토 방안 마련



플랫폼 안내서 및 작업 가이드 예시

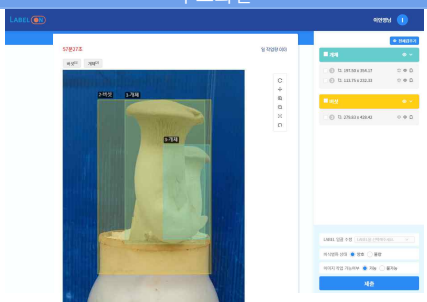
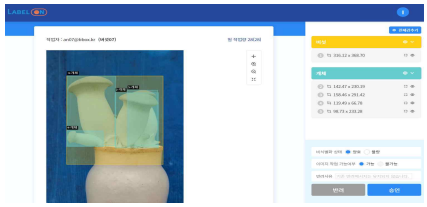
□ 클라우드 소싱 인력 업무 환경 지원

- 원격 또는 재택 작업기반의 언컨택트 클라우드 플랫폼 운영하여 작업현장의 안전성과 건전성을 유지하고, 비상사태(코로나로 인한 작업장 폐쇄 등) 발생을 최소화하고 문제 발생 시 적절하게 대응하는 비상대응체계 대비
- 클라우드 소싱 작업관리와 검수자들에게 원활한 작업이 이루어지도록 전산 환경 제공
- 클라우드 소싱 작업자들에 대한 인력 풀을 구성하여 관리하고, 향후 클라우드 소싱 데이터 구축 프로젝트에 우선 투입하는 등 지속적인 관계 유지

□ 능력별 교육 및 작업 투입

- 클라우드 소싱 작업 플랫폼에 대한 교육과 훈련이 필요하며, 데이터 구축에 대한 기본적인 이해와 저작도구 활용 교육과 훈련을 이수한 다음에 작업 투입
- 온라인 지원을 통하여 작업자를 선별하며, 능력 테스트로 작업능력을 평가하고 능력별로 3가지 등급으로 분류
- 등급이 분류된 작업자는 등급에 적합한 작업물량을 배정받게 되며, 작업결과에 대한 오류율 및 성과평가결과에 따라 등급 재조정
- 클라우드 소싱을 위해서 초기 교육 및 반복적인 피드백이 필요
- 동영상 매뉴얼을 통해 N차 교육을 진행
- 라벨링 규칙 업데이트, 엡지 케이스(애매한 사례) 발견 시 가이드를 업데이트

2.4.4 어노테이션/라벨링 도구

단계	주요화면	기능설명
데이터 가공		객체를 직사각형 모양의 박스 안에 포함되도록 그리는데 라벨링 방법으로 데이터 라벨링 작업에서 가장 일반적으로 사용 버섯 클래스 : 병목 기준으로 버섯 전체 라벨링 개체 클래스 : 날개 버섯 기준으로 버섯만 라벨링
데이터 검수		<ul style="list-style-type: none"> 작업자가 생성한 라벨링의 적절성 여부를 전수 검사 검수자는 승인 및 반려처리를 진행하고, 반려 시에는 반려사유를 입력 작업의 진행 정도를 작업 진행 대시보드를 통해 수시로 확인하고 조치하도록 안내

2.5 검수

2.5.1 검수 절차



사업 단계별 품질관리 프로세스

□ 인공지능 학습용 데이터를 구축함에 있어 수행하는 전체 공정 단계를 기획, 설계, 수집, 정제 및 가공, 검수 단계로 구분하여 각 단계별로 데이터 품질을 유지 및 개선하기 위한 활동을 수행

□ 데이터셋 정제 및 가공 단계의 품질 관리

• 수집된 데이터를 가공 과정의 프로세스, 작업 가이드, 오류 개선 활동을 중심으로 품질 관리 활동을 수행함

□ 데이터셋 검수 단계의 품질 관리

• 가공된 데이터를 검수하는 단계에서 프로세스, 도구, 조직 체계를 검토하여 일정한 품질의 데이터셋 구축 결과를 확인

□ 데이터셋 검수 후 단계의 품질 관리

• 최종 결과물의 전달 과정에 대한 확인

2.5.2 검수 기준

검사 절차	영상(동적/정적) 이미지 공통 항목
1차 검사 (획득)	법제도 준수
	사실적인 획득 환경 구성
	데이터 동기화
2차 검사 (정제)	편향성 방지
	정제 기준의 명확성
	중복성 방지
	정제 작업 매뉴얼
3차 검사 (라벨링)	정제 작업 방식
	라벨링 가이드
4차 검사 (전수)	어노테이션 항목
	라벨링 검사 도구
4차 검사 (전수)	외부 검사자

2.5.3 검수 조직

- 라벨링 완료된 데이터를 클라우드 소싱 인력을 통해 검수를 실시
- 전문 검수자(리뷰어)를 통해 2차 검수를 진행하여, 데이터 품질에 완전성을 확보
- 작업 가이드 제작 및 배포를 통해 데이터 가이드라인을 제작
- 작업자(Annotator)와 검수자(Reviewer)에게 검수 시 주의사항 및 가이드라인 제공하여 작업의 이해를 돕고, 교육을 통해 품질 확보를 함. 또한 전문 검수자와 1:1 문의 게시판 등을 통해서 지속적인 업데이트를 실시하여 구축을 진행함

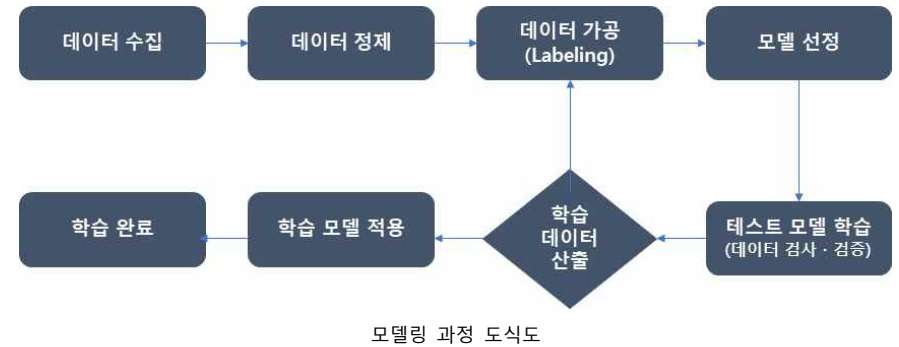


- 작업자(Annotator)가 라벨링 되어있는 이미지를 보고 1차 판단을 진행
- 검수자(Reviewer)는 작업자가 제출한 것을 2차 판단 하여 데이터의 적절성을 파악
- 검수장(Reviewer)은 검수자의 교육 및 2차 판단한 검수자의 데이터 재확인을 진행하며 전체 데이터 품질 확보

2.6 활용

2.6.1 활용 모델

2.6.1.1 모델 학습

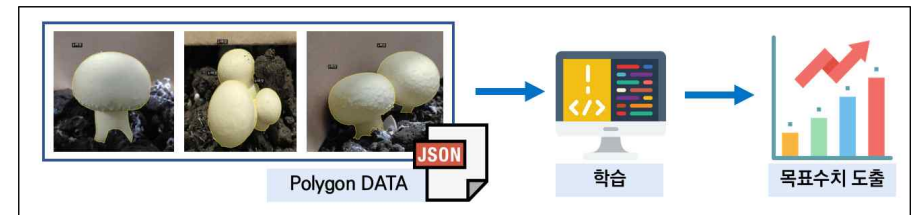


□ 모델 학습 목적

- 학습을 통해 버섯의 품종을 인식, 수확기 판단, 병충해 판단이 가능하도록 목적
- 수집된 버섯 데이터들을 정제작업을 통해 각 목적에 맞게 분류하고 이에 맞춰 가공하여 학습 데이터 변환 후 선정 모델에 학습
- 학습 결과를 통해 버섯을 인식하여 품종을 인식하고, 버섯의 수확기 및 병충해를 판별할 수 있도록 개발하여 수집된 데이터의 유효성과 가치를 입증

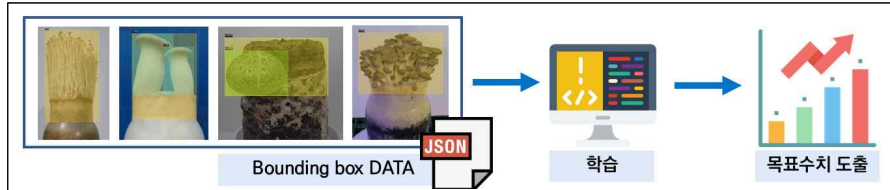
□ 목적별 학습 진행

- Polygon
 - 학습 목적에 맞게 가공된 데이터를 input 데이터로 사용하여 Mask R-CNN Inception ResNet V2 모델에 학습을 진행
 - 학습 결과를 확인하여 mAP, rate, avg loss 등이 목표수치에 근접하게 도달하는지 확인하고 이를 통해 학습 목적에 맞게 output 산출되는 것을 확인
 - 목표 값에 도달한 output이 나올 때까지 앞의 과정을 반복하여 결과를 도출



- Bounding box

- 학습 목적에 맞게 가공된 데이터를 input 데이터로 사용하여 YOLOv4 모델에 학습을 진행
- 학습 결과를 확인하여 mAP, rate, avg loss 등이 목표수치에 근접하게 도달하는지 확인하고 이를 통해 학습 목적에 맞게 output 산출되는 것을 확인
- 목표 값에 도달한 output이 나올 때까지 앞의 과정을 반복하여 결과를 도출



2.6.1.2 서비스 활용 시나리오

□ 활용방안

 <p>< 버섯 실사 모델링 및 생육 예측 ></p>	 <p>< 버섯 상태(관심필요) 인식 ></p>
 <p>< 버섯 상태(상태양호) 인식 ></p>	 <p>< 실시간 생육동 모니터링과 환경정보 ></p>

- 버섯의 상태에 따른 환경정보 제시 및 가이드 제공 서비스
- 학습 데이터를 적용한 학습 모델 구축
- 스마트폰, AR 글라스 등 각종 기기와 접목한 시스템 개발

2.6.2 데이터 제공

□ 산출물의 AI Hub 공개를 통한 자생적 데이터 확산 생태계 마련

- 본 과제를 통해 산출되는 학습용 데이터셋, 저장도구, 매뉴얼, 모델 등을 AI Hub를 통해 공개함으로써 인공지능 연구 및 서비스 개발 등에 자유롭게 활용될 수 있도록 기여
- 주관기업의 홈페이지 내 AI Hub 링크를 통해 공개 데이터셋 접근성 확대
- 인공지능을 통해 획득한 결과의 데이터와 새롭게 생성된 학습용 데이터 등을 자유롭게 거래할 수 있는 거래소를 구축하고, 데이터 거래를 통한 데이터 순환은 인공지능 모델의 고도화에 기여함으로써 정확도 향상 기대