테크니컬 리포트

2020 년 1 차

인공지능

학습용

데이터 구축

헬스케어 영역

치매진단 뇌파영상

개요: 치매진단 뇌파검사 AI 데이터셋이란?

치매는 겉으로 드러나는 증상일 뿐이고 질환의 발병 경로는 매우 다양하여 지금까지 보고된 치매의 원인질환은 약 70 여 개에 이르고 있다. 다양한 치매 원인 질환 중 알츠하이머성 치매는 절반이 넘는 비중을 차지하며, 상대적으로 치매와 연관된 전체 데이터 중에서 활용 · 분석의 안전성이 높을 것으로 예상하기에 알츠하이머성 치매 데이터를 선정하였다.

이를 통해, 인구고령화에 따라 급증하는 퇴행성 뇌질환 질병의 조기 진단 및 치유 기능의 기술 개발을 위한 딥러닝 학습용 데이터셋 구축을 통한 뇌신경 질환의 사회문제 해결에 기여 할 수 있으며, 치매의 정확한 원인을 감별하여 예측할 수 있는 장점이 있다.

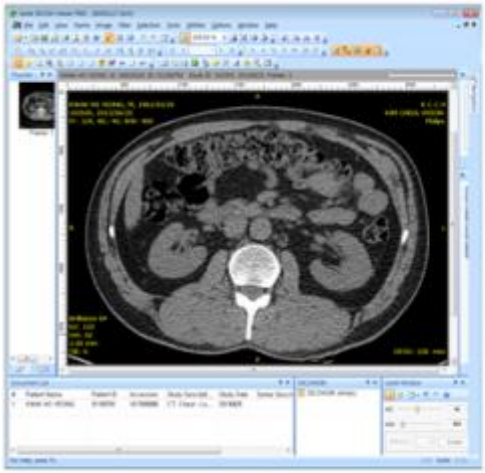
이번에 소개할 치매진단 뇌파검사 AI 데이터셋은 기계학습(딥러닝) 기반의 의료영상 진단 AI 기술의 개발·확산을 위해 치매와 경도인지장애 및 이와 관련된 질환의 영상 데이터(MRI) 및 임상전문의의 진단정보 등을 어노테이션한 학습용 데이터 셋으로 구성되어 있다

데이터셋의 구성

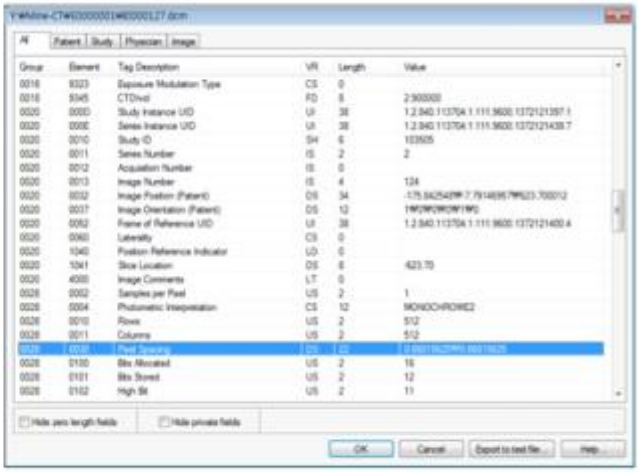
데이터셋은 크게 원천데이터와 학습데이터로 나눠진다. 학습데이터는 MRI 영상 이미지(dcm), 어노테이션 데이터셋(json), 카테고리 정의서(csv)으로 구성된다.

원천데이터인 MRI 영상 이미지(dcm) 28 만장은 약 780 명의 환자를 대상으로 한 환자당 320~360 장의 이미지 데이터(환자단 뇌의 크기가 다르기에 수집 이미지의 차이가 있음)를 수집하였다. 이 원천데이터는 Header 와 Image data 로 구성되어 있다

분류	설명
Header	• 환자 성명, 스캔 형식, 이미지차원(dimensions), sclae 등의 정보를 포함
Image Data	• Bitmap 또는 압축되지 않은 형식의 이미지 정보를 수록



[DICOM Viewer]



[DICOM 파일 뷰어로 본 이미지와 Header 정보의 예]

header information 의 비식별화를 위해 개인정보 dicom tag 를 모두 anonymization 을 처리하였다.

```
files_1(j), 1).FileModDate = '';
files_1(j), 1).ReferencedPerformedProcedureStepSequence.Item_1.InstanceCreationDate ='';
files_1(j), 1).ReferencedPerformedProcedureStepSequence.Item_1.InstanceCreationTime = '';
files_1(j), 1).ProcedureCodeSequence.Item_1.CodeValue = '';
files_1(j), 1).ProcedureCodeSequence.Item_1.CodingSchemeDesignator ='';
files_1(j), 1).ProcedureCodeSequence.Item_1.CodeMeaning = '';
files_1(j), 1).PerformedProtocolCodeSequence.Item_1.CodeMeaning ='';
files_1(j), 1).RequestAttributesSequence.Item_1.ScheduledProcedureStepDescription ='';
files_1(j), 1).Private_2001_10c8 ='';
files_1(j),1).PatientName = '';
files_1(j),1).PatientID = '';
files_1(j),1).PatientBirthDate = '';
files_1(j),1).PatientSex = '';
files_1(j),1).PatientWeight = '';
files_1(j),1).StudyDate = '';
files_1(j),1).StudyTime = '';
files_1(j),1).SeriesDate = '';
files_1(j),1).SeriesTime = '';
files_1(j),1).AcquisitionDate = '';
files_1(j),1).AcquisitionTime = '';
files_1(j),1).ContentDate = '';
files_1(j),1).ContentTime = '';
files_1(j), 1).InstanceCreationDate ='';
files_1(j), 1).InstanceCreationTime='';
files_1(j), 1).ProtocolName = 'BRAIN T1 weighted MRI';
files_1(j),1).StudyDescription = 'BRAIN T1 weighted MRI';
files_1(j),1).RequestedProcedureDescription = 'BRAIN T1 weighted MRI';
files_1(j),1).PerformedProcedureStepDescription  = 'BRAIN T1 weighted MRI';
files_1(j),1).RescaleSlope  = 1;
files_1(j),1).PerformedProcedureStepStartDate = '';
files_1(j),1).PerformedProcedureStepStartTime = '';
files_1(j),1).PerformedProcedureStepEndDate = '';
files_1(j),1).PerformedProcedureStepEndTime = '';
files_1(j),1).PerformedProcedureStepID = '';
files_1(j),1).StationName = '';
files_1(j),1).AccessionNumber = '';
files_1(j),1).DeviceSerialNumber = '';
files_1(j),1).DeviceSerialNumber = '';
```

<비식별화 처리된 Dicom header Information>

데이터셋의 설계 기준과 분포

데이터의 설계는 원천데이터 수집 단계에서부터 편향성을 최소화하도록 설계 진행하였다.

총 3class 로 나뉘어졌으며 치매군 데이터셋 10.5 만장 , 경도인지장애군 데이터셋 7 만장, 정상군 데이터셋 10.5 만장으로 총 28 만장으로 나누어져 있다.

데이터 종류	포함 내용	제공 방식
치매군 데이터셋	MRI 영상 이미지 + 어노테이션 데이터 셋	MRI 이미지 파일(dcm), JSON 포맷 파일
경도인지장애군 데이터셋		
정상군 데이터 셋		

데이터 구조

어노테이션된 데이터셋의 항목과 설명은 아래 표와 같다.

Tag	Name	VR	Value(sample)
(0008, 0005)	Specific Character Set	CS	'ISO_IR 100'
(0008, 0008)	Image Type	CS	['ORIGINAL', 'PRIMARY', 'M_FFE', 'M', 'FFE']
(0008, 0014)	Instance Creator UID	UI	1.3.46.670589.11.17015.5
(0008, 0016)	SOP Class UID	UI	MR Image Storage
(0008, 0018)	SOP Instance UID	UI	1.3.6.1.4.1.9590.100.1.2.315097729813117252127334839232492950421
(0008, 0050)	Accession Number	SH	'R203132819'
(0008, 0060)	Modality	CS	'MR'
(0008, 0102)	Coding Scheme Designator	SH	'DCM'
(0008, 1010)	Station Name	SH	'-6124000005'
(0008, 1030)	Study Description	LO	'BRAIN T1 weighted MRI'
(0008, 1032)	Procedure Code Sequence	SQ	<Sequence, length 1>
(0008, 103e)	Series Description	LO	'Sagittal 3D T1WI'
(0008, 1040)	Institutional Department Name	LO	'MRI1'

(0008, 1090)	Manufacturer's Model Name	LO	'Achieva'
(0008, 1110)	Referenced Study Sequence	SQ	<Sequence, length 1>
(0008, 1111)	Referenced Performed Procedure Step	SQ	<Sequence, length 1>
(0008, 1140)	Referenced Image Sequence	SQ	<Sequence, length 3>
(0010, 1030)	Patient's Weight	DS	None
(0010, 2160)	Ethnic Group	SH	'Asian'
(0010, 21c0)	Pregnancy Status	US	4
(0018, 0015)	Body Part Examined	CS	'BRAIN'
(0018, 0020)	Scanning Sequence	CS	'GR'
(0018, 0021)	Sequence Variant	CS	'MP'
(0018, 0022)	Scan Options	CS	'SP'
(0018, 0023)	MR Acquisition Type	CS	'3D'
(0018, 0050)	Slice Thickness	DS	"1.0"
(0018, 0080)	Repetition Time	DS	"9.91090011596679"
(0018, 0081)	Echo Time	DS	"4.606"
(0018, 0083)	Number of Averages	DS	"1.0"
(0018, 0084)	Imaging Frequency	DS	"127.749145"
(0018, 0085)	Imaged Nucleus	SH	'1H'
(0018, 0086)	Echo Number(s)	IS	"1"
(0018, 0087)	Magnetic Field Strength	DS	"3.0"
(0018, 0088)	Spacing Between Slices	DS	"0.5"
(0018, 0089)	Number of Phase Encoding Steps	IS	"240"

(0018, 0091)	Echo Train Length	IS	"240"
(0018, 0093)	Percent Sampling	DS	"100.0"
(0018, 0094)	Percent Phase Field of View	DS	"100.0"
(0018, 0095)	Pixel Bandwidth	DS	"142.0"
(0018, 1000)	Device Serial Number	LO	'17015'
(0018, 1020)	Software Versions	LO	['3.2.1', '3.2.1.1']
(0018, 1030)	Protocol Name	LO	'BRAIN T1 weighted MRI'
(0018, 1081)	Low R-R Value	IS	"0"
(0018, 1082)	High R-R Value	IS	"0"
(0018, 1083)	Intervals Acquired	IS	"0"
(0018, 1084)	Intervals Rejected	IS	"0"
(0018, 1088)	Heart Rate	IS	"0"
(0018, 1100)	Reconstruction Diameter	DS	"240.0"
(0018, 1250)	Receive Coil Name	SH	'SENSE-NV-16'
(0018, 1251)	Transmit Coil Name	SH	'S'
(0018, 1310)	Acquisition Matrix	US	[0, 240, 240, 0]
(0018, 1312)	In-plane Phase Encoding Direction	CS	'ROW'
(0018, 1314)	Flip Angle	DS	"8.0"
(0018, 1316)	SAR	DS	"0.01732053421437"
(0018, 1318)	dB/dt	DS	"59.5874379987818"
(0018, 5100)	Patient Position	CS	'HFS'
(0018, 9073)	Acquisition Duration	FD	377.4471741
(0018, 9087)	Diffusion b-value	FD	0
(0018, 9089)	Diffusion Gradient Orientation	FD	[0.0, 0.0, 0.0]

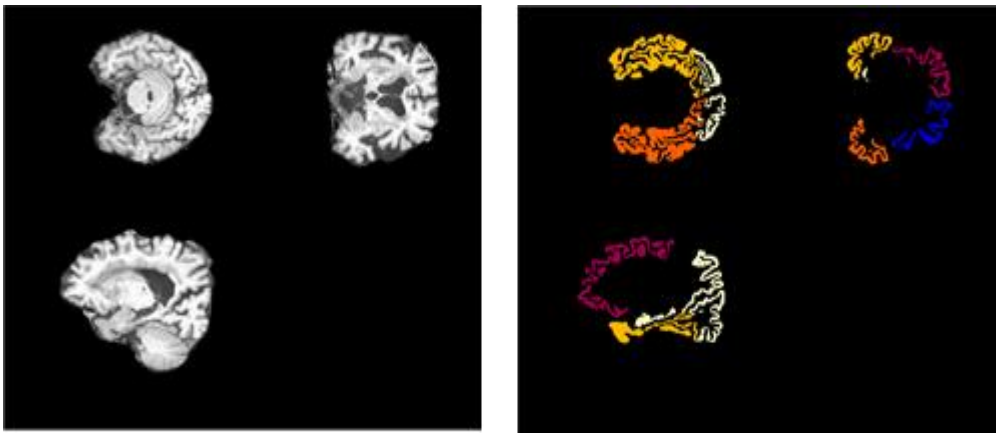
(0020, 000d)	Study Instance UID	UI	1.2.840.113619.2.182.1448922969203148.1584082306.1575035
(0020, 000e)	Series Instance UID	UI	1.2.276.0.7230010.3.1.3.702171887.59220.1591427561.4014
(0020, 0010)	Study ID	SH	'M01'
(0020, 0011)	Series Number	IS	"801"
(0020, 0012)	Acquisition Number	IS	"8"
(0020, 0013)	Instance Number	IS	"135"
(0020, 0032)	Image Position (Patient)	DS	[19.763587191701, -137.83174180984, 91.098340988159]
(0020, 0037)	Image Orientation (Patient)	DS	[0.0301144663244, 0.9995464682579, 0, 0, 0, -1]
(0020, 0052)	Frame of Reference UID	UI	1.3.46.670589.11.17015.5.0.4868.2020031316063776002
(0020, 0100)	Temporal Position Identifier	IS	"1"
(0020, 0105)	Number of Temporal Positions	IS	"1"
(0020, 1041)	Slice Location	DS	"66.9997187393155"
(0028, 0002)	Samples per Pixel	US	1
(0028, 0004)	Photometric Interpretation	CS	'MONOCHROME2'
(0028, 0010)	Rows	US	480
(0028, 0011)	Columns	US	480
(0028, 0030)	Pixel Spacing	DS	[0.5, 0.5]
(0028, 0100)	Bits Allocated	US	16
(0028, 0101)	Bits Stored	US	16
(0028, 0102)	High Bit	US	15
(0028, 0103)	Pixel Representation	US	0
(0028, 0106)	Smallest Image Pixel Value	US	0

(0028, 0107)	Largest Image Pixel Value	US	90
(0028, 1050)	Window Center	DS	"1281.0"
(0028, 1051)	Window Width	DS	"2227.0"
(0028, 1052)	Rescale Intercept	DS	"0.0"
(0028, 1053)	Rescale Slope	DS	"1.0"
(0028, 1054)	Rescale Type	LO	'normalized'
(0028, 2110)	Lossy Image Compression	CS	'00'
(0032, 1033)	Requesting Service	LO	'NR'
(0032, 1060)	Requested Procedure Description	LO	'BRAIN T1 weighted MRI'
(0040, 0241)	Performed Station AE Title	AE	'ACH17015'
(0040, 0244)	Performed Procedure Step Start Date	DA	'20200313'
(0040, 0245)	Performed Procedure Step Start Time	TM	'160735'
(0040, 0250)	Performed Procedure Step End Date	DA	'20200313'
(0040, 0251)	Performed Procedure Step End Time	TM	'160735'
(0040, 0253)	Performed Procedure Step ID	SH	'637085255'
(0040, 0254)	Performed Procedure Step Descriptio	LO	'BRAIN T1 weighted MRI'
(0040, 0260)	Performed Protocol Code Sequence	SQ	<Sequence, length 1>
(0040, 0275)	Request Attributes Sequence	SQ	<Sequence, length 1>

(0040, 1001)	Requested Procedure ID	SH	'M01'
(2050, 0020)	Presentation LUT Shape	CS	'IDENTITY'
(7fe0, 0010)	Pixel Data	OW	Array of 460800 elements

데이터 예시

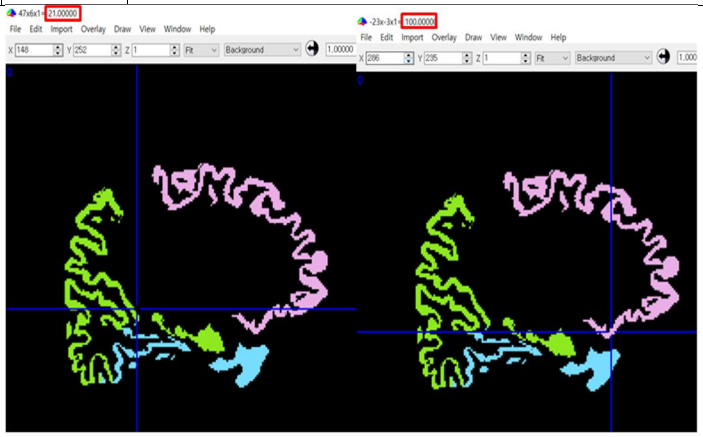
신경과 전문의가 MRI 판독 및 검수 후, regional atrophy 가 있는 유/무는 1,0 으로 구분하여 annotation 진행한다.



<두개골 제거된 뇌 MRI(좌측) 과 뇌의 관심영역 및 위축에 대한 어노테이션(우측)>

뇌 MRI 의 어노테이션 방법 (예시)

	좌	우	위축의 유/무를 정하여 x에 대입 (1/0)
Frontal cortex	1x	2x	예시) Left frontal cortex에 위축이 있다 -> 11
Temporal cortex	3x	4x	예시) Left frontal cortex에 위축이 없다 -> 10
Parietal cortex	5x	6x	
Occipital cortex	7x	8x	
Hippocampus	9x	10x	



<Right Frontal GM region atrophy 가 있는 경우 annotation:21>

<Right Hippocampus region atrophy 가 없는 경우 annotation:10>

어노테이션 데이터 예시는 아래와 같다.


```
"00400254": {
  "vr": "LO",
  "Value": [
    "BRAIN T1 weighted MRI"
  ]
},
"00400260": {
  "vr": "SQ",
  "Value": [
    {
      "00080100": {
        "vr": "SH",
        "Value": [
          "RM1027N"
        ]
      },
      "00080102": {
        "vr": "SH",
        "Value": [
          "CCG_CSTemp"
        ]
      },
      "00080104": {
        "vr": "LO"
      },
      "0008010B": {
        "vr": "CS",
        "Value": [
          "N"
        ]
      }
    }
  ]
```


데이터 구축 과정

MRI 영상 이미지 데이터 수집 process

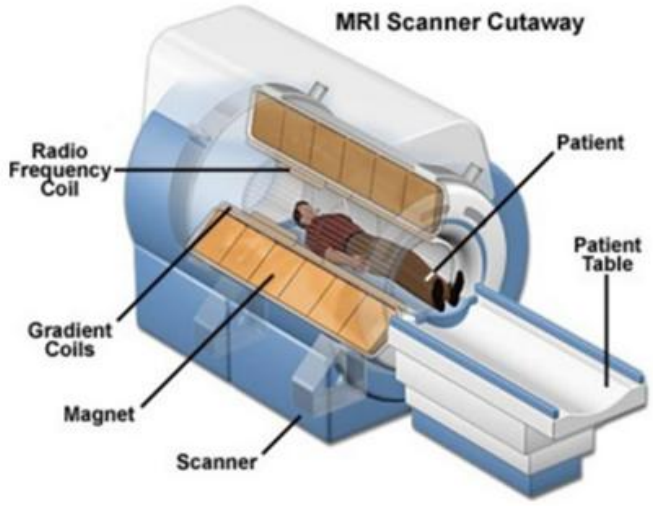
수집 절차	상세 내역	산출물
1. SNSB 실험 데이터	• SNSB 실험을 통해 치매, 경도 인지장애, 정상으로 구분	병명에 따른 데이터
2. 뇌 MRI 영상 자료수집	• 치매, 경도 인지장애, 정상 군의 뇌 이미지 영상 수집	뇌 MRI 영상 데이터
2-1. 치매 판정된 뇌 MRI 영상 분류	• 치매로 판단된 환자의 뇌 MRI 촬영 영상 데이터 수집	치매 환자의 뇌 MRI 영상 데이터
2-2. 경도 인지장애 판정된 뇌 MRI 영상 분류	• 경도 인지장애로 판단된 환자의 뇌 MRI 촬영 영상 데이터 수집	경도 인지장애 환자의 뇌 MRI 영상 데이터
2-3. 정상 판정된 뇌 MRI 영상 분류	• 정상으로 판단된 환자의 뇌 MRI 촬영 영상 데이터 수집	정상인의 뇌 MRI 영상 데이터
3. 영상 비식별화 전처리	• 영상 데이터 내에 포함된 개인정보(이름, 환자번호 등)의 비식별화 처리	비식별화된 뇌 MRI 영상 데이터

4. 비식별화 된 영상 데이터와 임상 데이터의 결합	• 레이블(치매, 경도 인지장애, 정상)한 뇌 영상 데이터(MRI)와 관련한 임상 데이터의 결합	수집 데이터가 결합된 최종 데이터

수집 환경

삼성서울병원 MRI 실에서 디지털 의료영상 장비를 사용하여 획득된 디지털 의료영상이미지를 DICOM 이라는 국제표준규약에 맞게 저장, 가공 저장하고 네트워크를 통해 병원 내·외 단말로 전송하는 환경에서 학습 데이터 수집

수집장비



MRI(Magnetic Resonance Imaging) 는 자기장을 걸고 고주파를 송신하였을 때 인체 내 수소 원자핵으로부터 발생되는 영상신호를 2 차원 혹은 3 차원 단면상으로 보여 주는 전신용 검사 장비 검사부위별로 30~60 분 정도 소요된다. 근육이나 인대와 같은 연부조직의 해상도와 대조도가 좋으며, 조영제와 같은 특별한 약물 없이도 고해상도의 혈관 영상을 찍을 수 있다. 환자의 체위를 변화시키지 않고 횡단면(axial), 시상면(sagittal), 관상면(coronal)의 영상 촬영 가능하며, 뇌신경계 영상에서 뇌경색, 뇌출혈, 뇌종양 등의 뇌질환을 비침습적 방법으로 검사할 수 있다. 뇌 MRI 검사는 부작용이 없고 통증도 없어 간단한 사전 설명만 들으면 노인들도 쉽게 받을 수 있는 안전한 검사이며, 사용상 위험성이 거의 없으며, 주변의 잡음의 영향을 거의 받지 않는 장점이 있다.

검수와 품질 확보

뇌 자기공명영상(MRI) 촬영 후 얻어지는 영상 이미지 자료에 대해 두개골 제거 및 10 개의 영역으로 분할하는 단계에서 영상 분석 연구원의 품질관리 수행하여 검수를 진행을 한다. 관심 영역 설정 후 영역별 뇌의 위축의 어노테이션을 위해 신경과 전문의가 어노테이션을 수행하며 추가 검수를 한다. 어노테이션 정보를 이용하여 관심영역 별 라벨링을 처리 후 DICOM 포맷으로 변환하는 것을 진행한다.. 두개골 제거된 뇌 MRI 또한 DICOM 포맷으로 변환 후 검수 한 후, 아래과 같이 전문의가 MRI

영상이미지 검수 및 해당 품질 평가 진행 및 검수를 최종적으로 진행한다.

검사결과	합격/불합격	2020년 11월 9일	
자료유형	DICOM	검사자	입력
관리번호			이미지
청구기호			종합
서 명			

이미지 검사						
Image File	Image 오류					
	기울기	판독불능	여백	회전	노이즈	기타
subject_010	X	X	X	X	X	X
subject_101	X	X	X	X	X	X
subject_030	X	X	X	X	X	X
subject_400	X	X	X	X	X	X
subject_643	X	X	X	X	X	X
subject_763	X	X	X	X	X	X
subject_289	X	X	X	X	X	X
subject_544	X	X	X	X	X	X
subject_398	X	X	X	X	X	X

<검수확인서>

관리번호		사업명	
작성자명		작성일	2020년 11월 9일

평가결과		평가일자	2020년 11월 9일
작업유형		평가자	김수종 박유현 (확인)
대상자료	뇌 자기 공명 영상	자료양	280,800장
NO	품질평가 중점체크 항목	에러수	비고
1	제목은 정확한가?	0	
2	파일명 명명규칙은 준수하였는가?	0	
3	해상도는 기준과 일치하는가?	0	
4	화면비율은 기준과 일치하는가?	0	
5	Annotation QC fail	0	

<품질확인서>

이렇게 만들어진 데이터셋을 전체적으로 들여다보며 데이터셋의 밸런스나 가이드라인의 적절성을 제시해주는 관리자는 삼성서울병원의 전문의로서 해당 MRI 영상이미지를 직접 분석하며 관련 분야의 경험이 풍부한 인력을 배치하였으며 이를 AI 데이터셋으로 적용가능 판단은 인공지능 데이터 분석을 2 년

이상한 디노플러스의 직원이 함께 진행함으로써 최종적인 데이터셋의 품질을 담보할 수 있었다

데이터 구축 담당자

수행기관(주관) : 디노플러스㈜ (전화: 02-548-2224), 이메일: jspark@dinnoplus.com