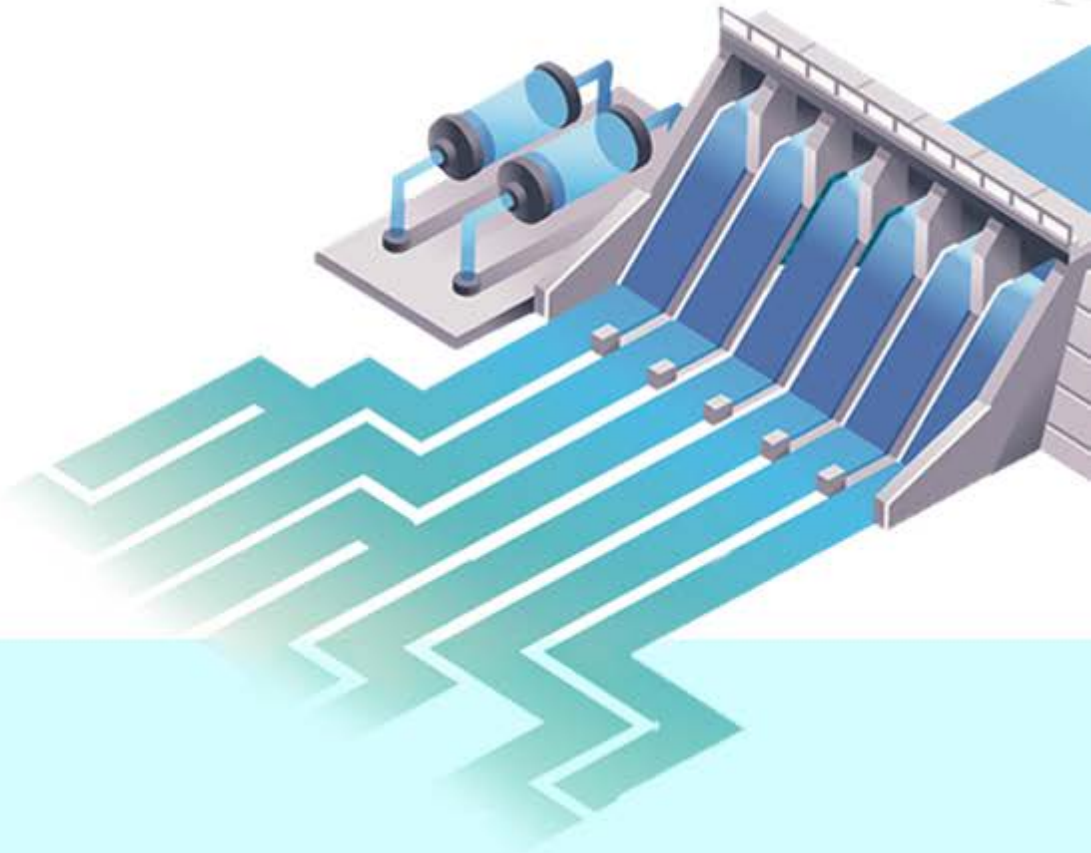
ISEC 2025
INTERNATIONAL SECURITY CONFERENCE

AI 의료 학습데이터의 현재와 미래



김성현 수석연구원
Ph.D. in Business Administration(MIS)

NIA 한국지능정보사회진흥원 NATIONAL INFORMATION SOCIETY AGENCY

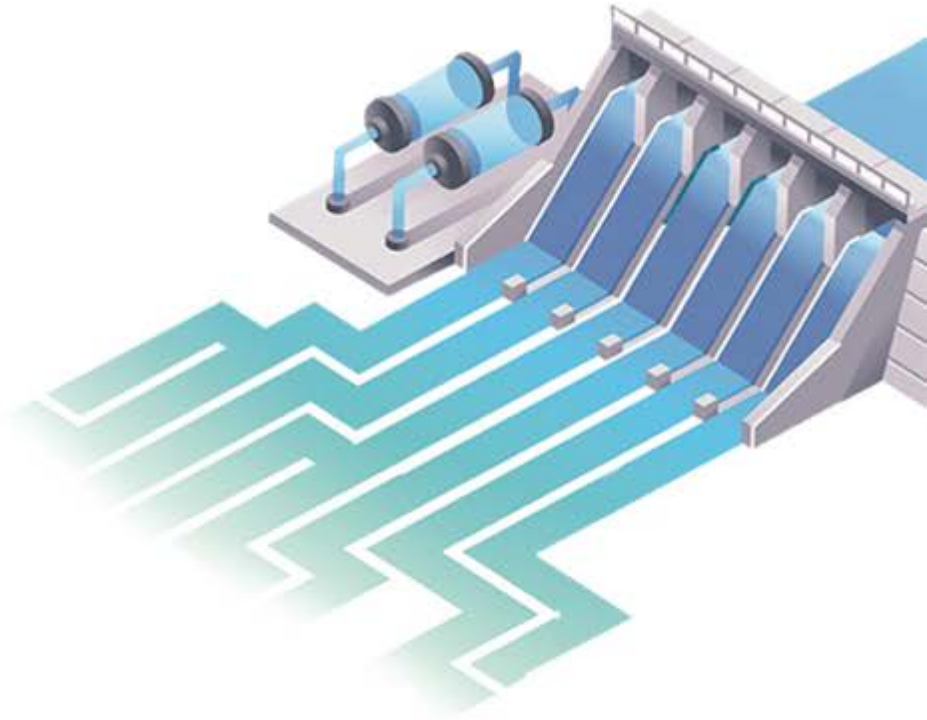
AI Hub

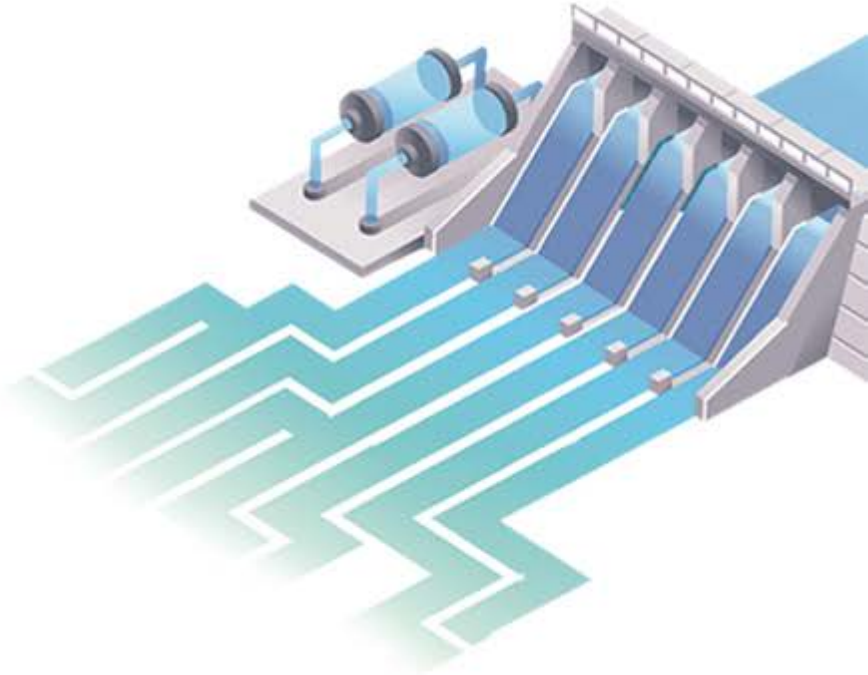
목차

Ⅰ 의료 학습데이터

Ⅱ NIA AI 학습데이터

Ⅲ 향후 발전 방향





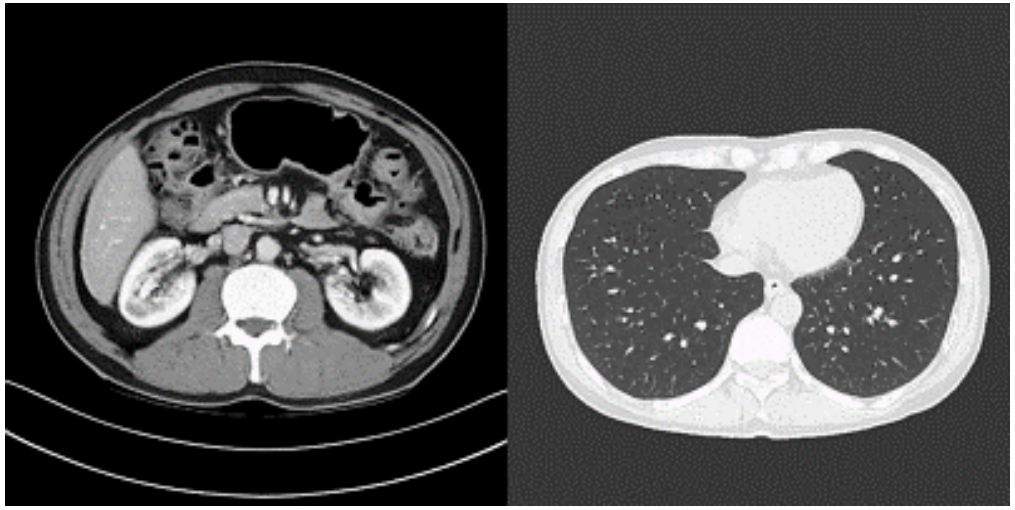
I

의료 학습 데이터

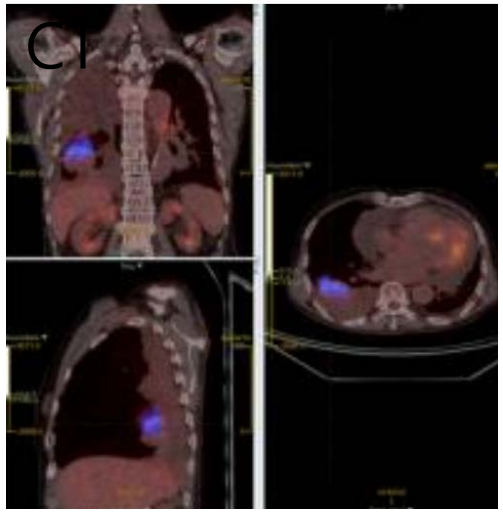
1. 이미지 데이터

X-RAY

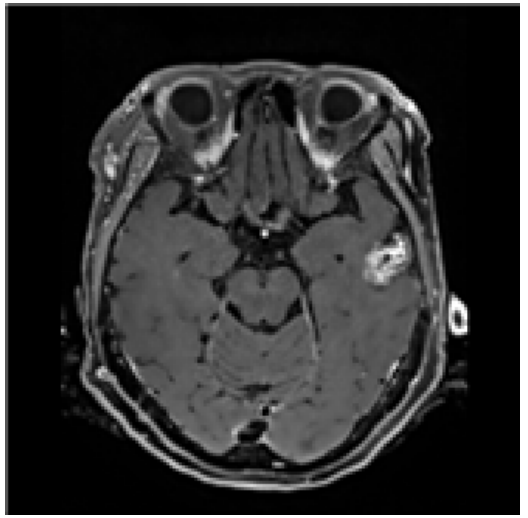
CT



PET-CT



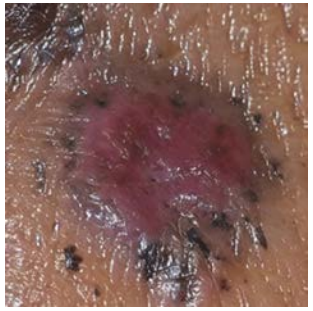
MRI



1. 이미지 데이터

RGB

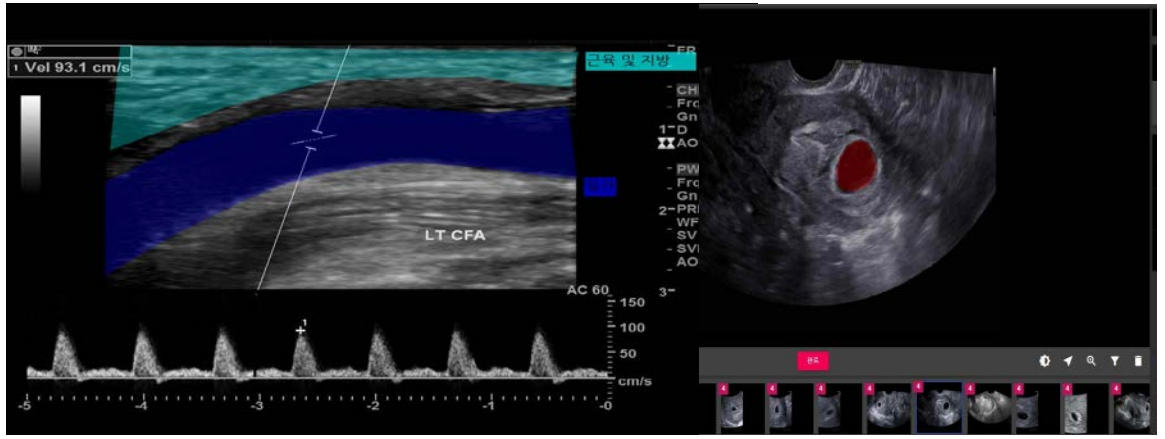
여드름



기저세포암

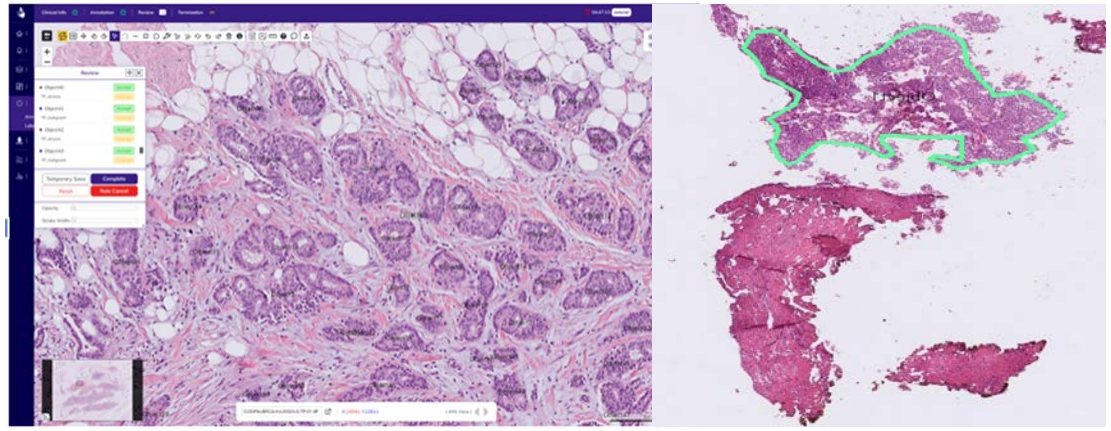
편평세포암

초음파



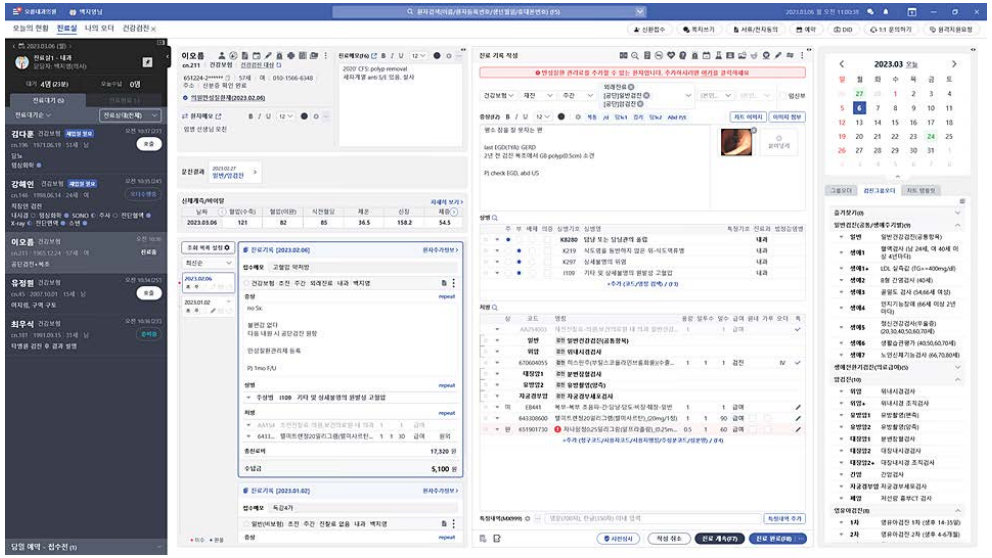
안저

병리



2. 텍스트 데이터

EMR



전문지식

최측 결장암보다 우측 결장암에서 더 많이 나오는 증상은?

가. 빈혈	나. 장폐색
다. 촉지되는 복부 종괴	라. 혈변

1) 가, 나, 다 2) 가, 다 3) 나, 라
4) 라 5) 가, 나, 다, 라

(Hint: 결장암은 대장암이라 생각하시고 풀면 됩니다.)

Q. 왼쪽 대장암보다 오른쪽 대장암에서 더 빈번한 것은?

1) 잔변감 2) 야뇻배 통증
3) 육안적 혈변 4) 철결핍빈혈
5) 배변습관 변화

(Hint: 항문으로부터 먼 쪽의 대장암의 증상을 묻는 문제입니다. 대변은 잘수록 수분이 증수됩니다.)

상담

발화자 구분	발화 텍스트
상담자, 내담자	지난 주에 출근하는데 갑자기 숨이 턱 막히고 곧 죽을 것만 같은 공포감이 들었어요. 언제 터질 지 모르니깐 계속 불안하고 가끔 가슴 두근거림이 올라와서 요즘 커피도 잘 못 마시겠는데 너무 힘들고 우울하네요. 저는 왜 이럴까요?

발화자 구분	발화 텍스트
상담자, 내담자	가끔은 제가 사라져도 괜찮을 것 같다는 생각이 들어요

해당 토큰에서 직결되는 질환별 점수

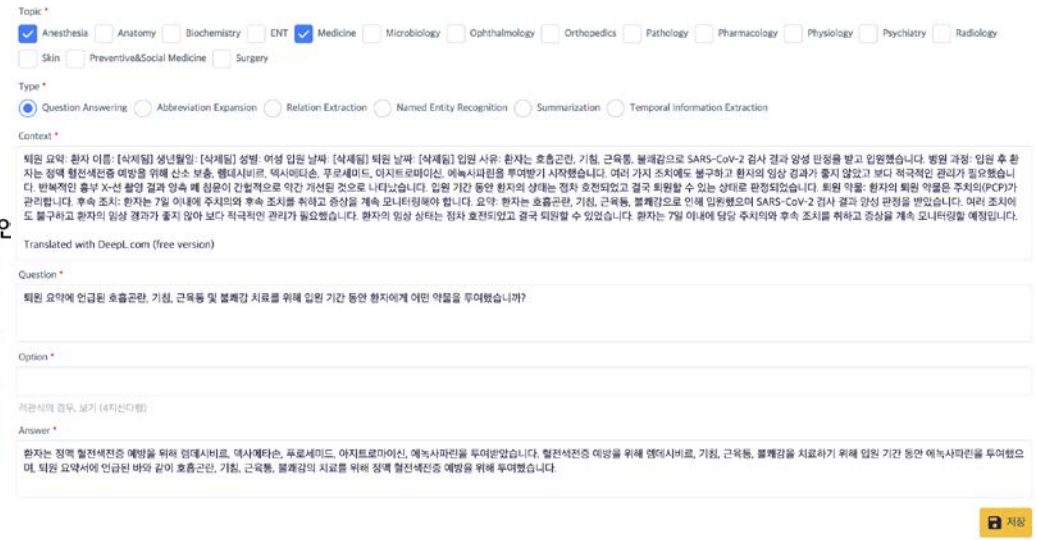
우울	불안	중독	조현병
상	없음	없음	없음

해당 토큰에서 직결되는 주요증상/위험요인

우울감	식욕변화	수면문제	자살사고	우울증상 개선
상	상	중	없음	하

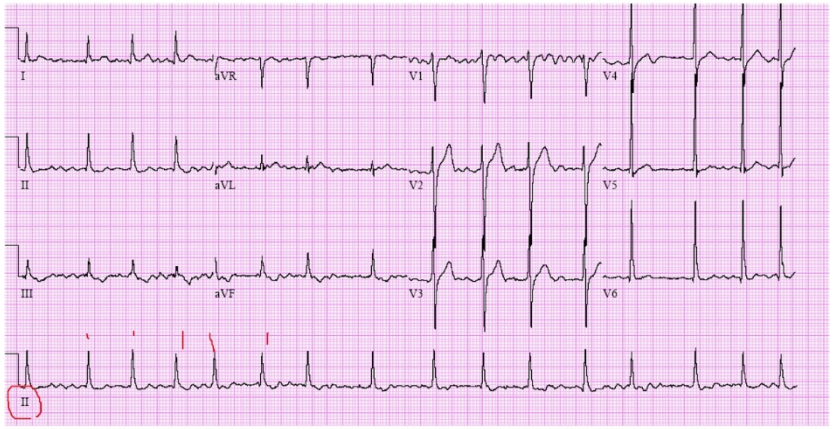
불안감	과민성	주의집중 저하	수면문제	불안증상 개선
없음	없음	중	중	없음

조절실패	현저성	문제적 결과	중독증상 개선	중독위험 요소감소
없음	없음	없음	없음	없음

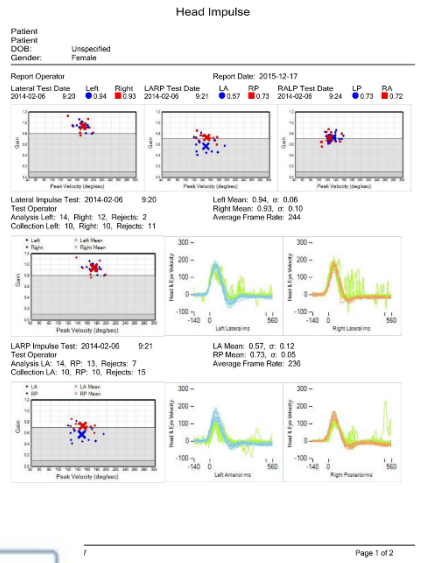


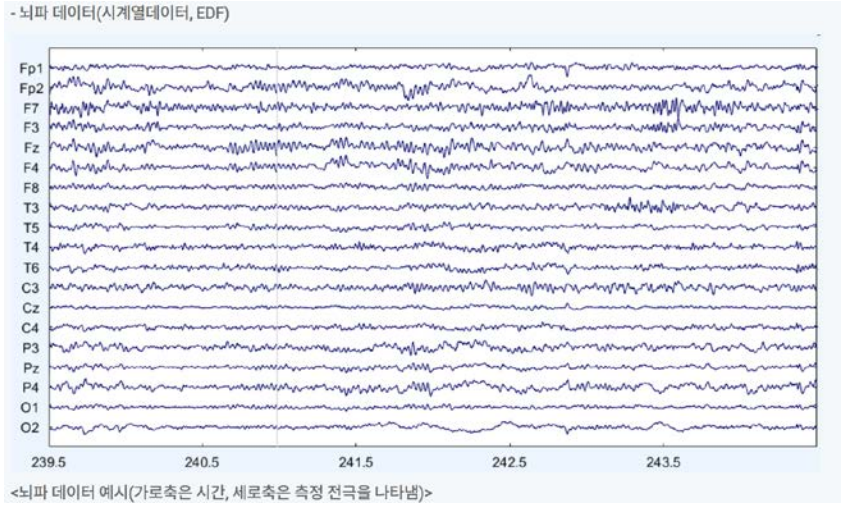
3. 센서 데이터

심전도



뇌간반응(뇌파)



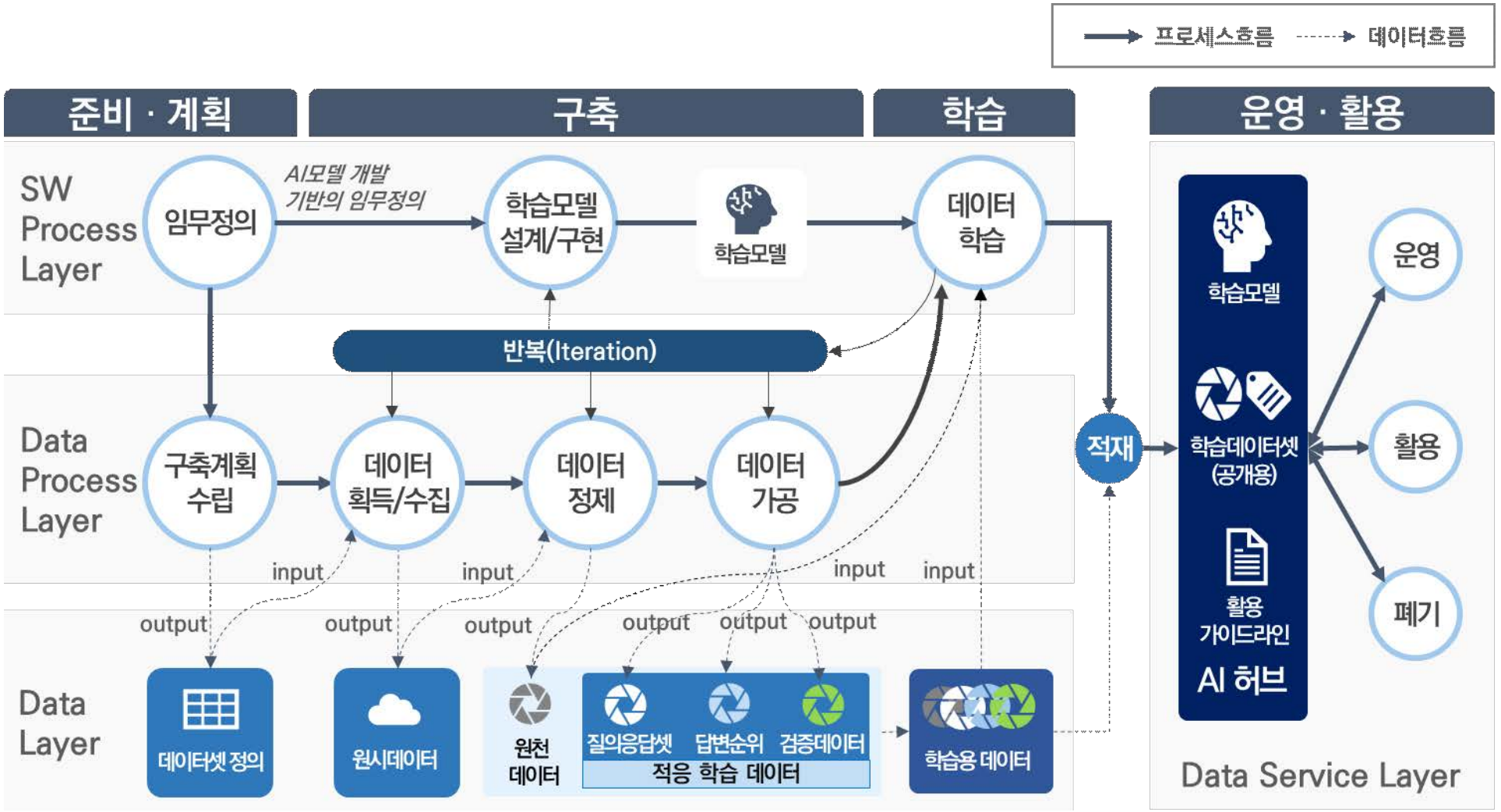


만성당뇨 추적 관찰

운동 처방(후향)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		DI_Versior	DI_Name	DI_Provide	DI_Date	PI_Patient	PI_Age	PI_Sex	MICAD_Ar	MICAD_N	MICAD_ST	MI_Conge	MIVHD_M	MIVHD_M	MIVHD_Ar	MIVHD_Ar	MIVHD_Pi	MIVHD_Pi	MIA_Atrial	MIA_Paro	MIA_Sick	MIA_Arter	MIA_Othe	MI_A
2	0	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	55	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	
3	1	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	60	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	
4	2	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	69	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	
5	3	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	58	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
6	4	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	68	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	5	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	41	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
8	6	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	53	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	
9	7	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	76	2	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
10	8	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	58	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
11	9	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	55	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
12	10	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	66	2	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
13	11	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	43	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
14	12	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	60	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
15	13	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	45	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	
16	14	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	56	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	15	0.1	심장전향	고려대학교	2022-10-31	KUMC-C-I	47	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	

4. AI 데이터 구축과 활용



5. 현 가이드라인의 성과와 한계

➢ 보건복지부, 개보위의 가이드라인은 개인정보 처리와 활용의 근거를 제공하는 효과
⇒ 보건의료 데이터 활용 가이드라인, 개인정보 처리 안내서, AI 프라이버시 리스크 관리모델

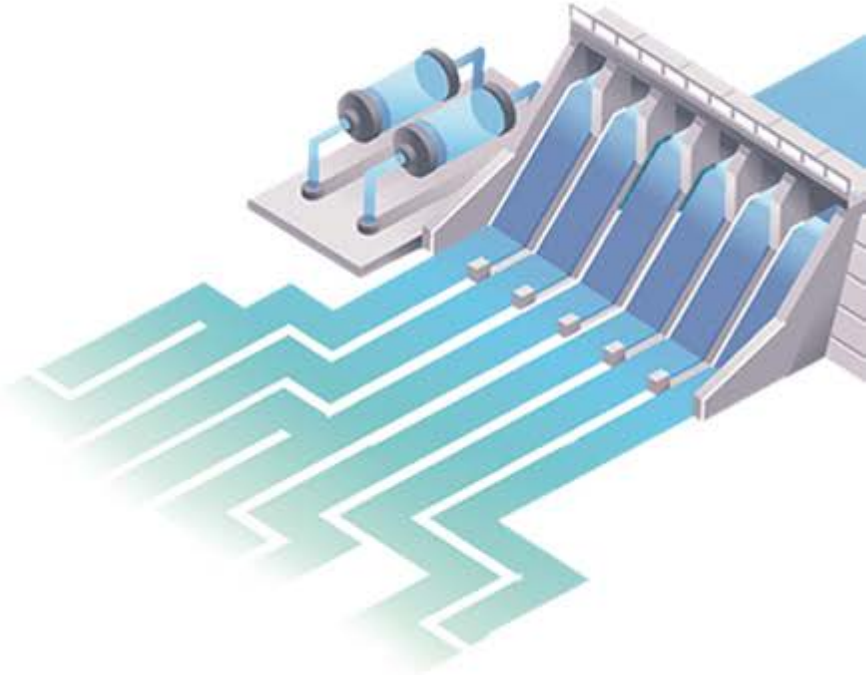
➢ 구체적인 알고리즘이나 기술 도구 제시는 미흡, 개인정보보호를 위한 원시데이터 훼손이 과도

☞ **(예시) 연구 목적에 필요 없는 영역을 블러링 처리하고, 메타데이터를 삭제하여 활용***

< 구강 촬영사진 >	**개인식별 위험성 검토**	- **구강사진 자체로는** 개인식별 위험성 거의 없음 - **충치 영역 외 부분**은 **연구에 필요 없음** - 구강사진에 대한 **메타데이터**(이름, 나이 등)는 **구강사진과 결합되어 개인식별 위험성 존재**
	데이터 처리 방안	➜ 연구에 필요한 **충치 영역은 그대로 활용**하고, 연구에 필요 없는 **그 외 영역은 블러링 처리** ➜ **메타데이터**는 연구에 필요 없어 **삭제**

(충치부분 : 그대로 활용)
(그 외 : 블러링 처리)

⇒ 충치부분을 제외하고 블러처리할 경우, 잇몸병 등 다른 목적의 AI 학습이 활용이 불가능하여 종합적인 구강질환 데이터 구축이 불가하고 활용성에 제약



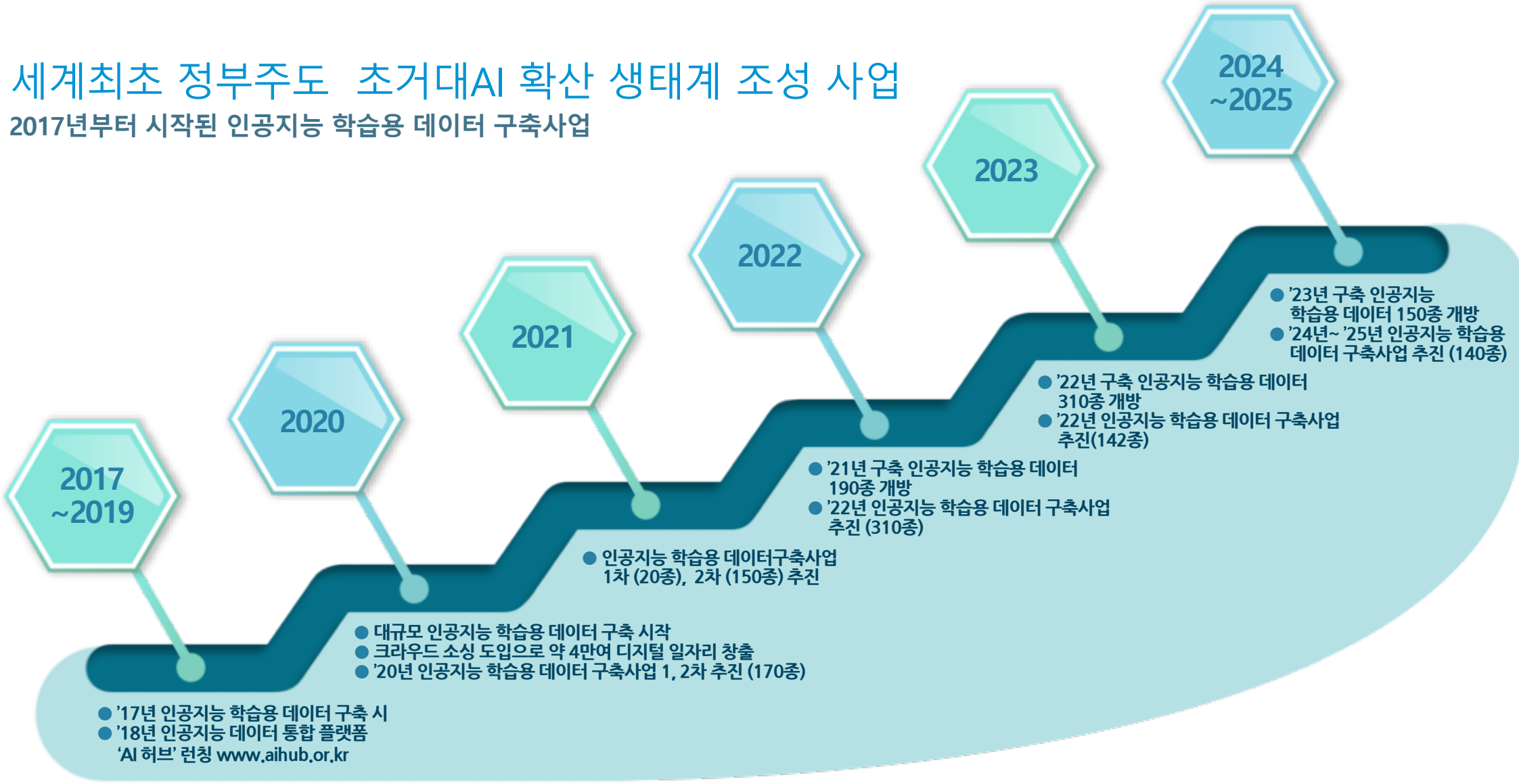
Ⅱ

NIA AI 학습 데이터

1. 초거대 AI 확산 생태계 조성사업 (AI인공지능 학습용 데이터 구축 사업)

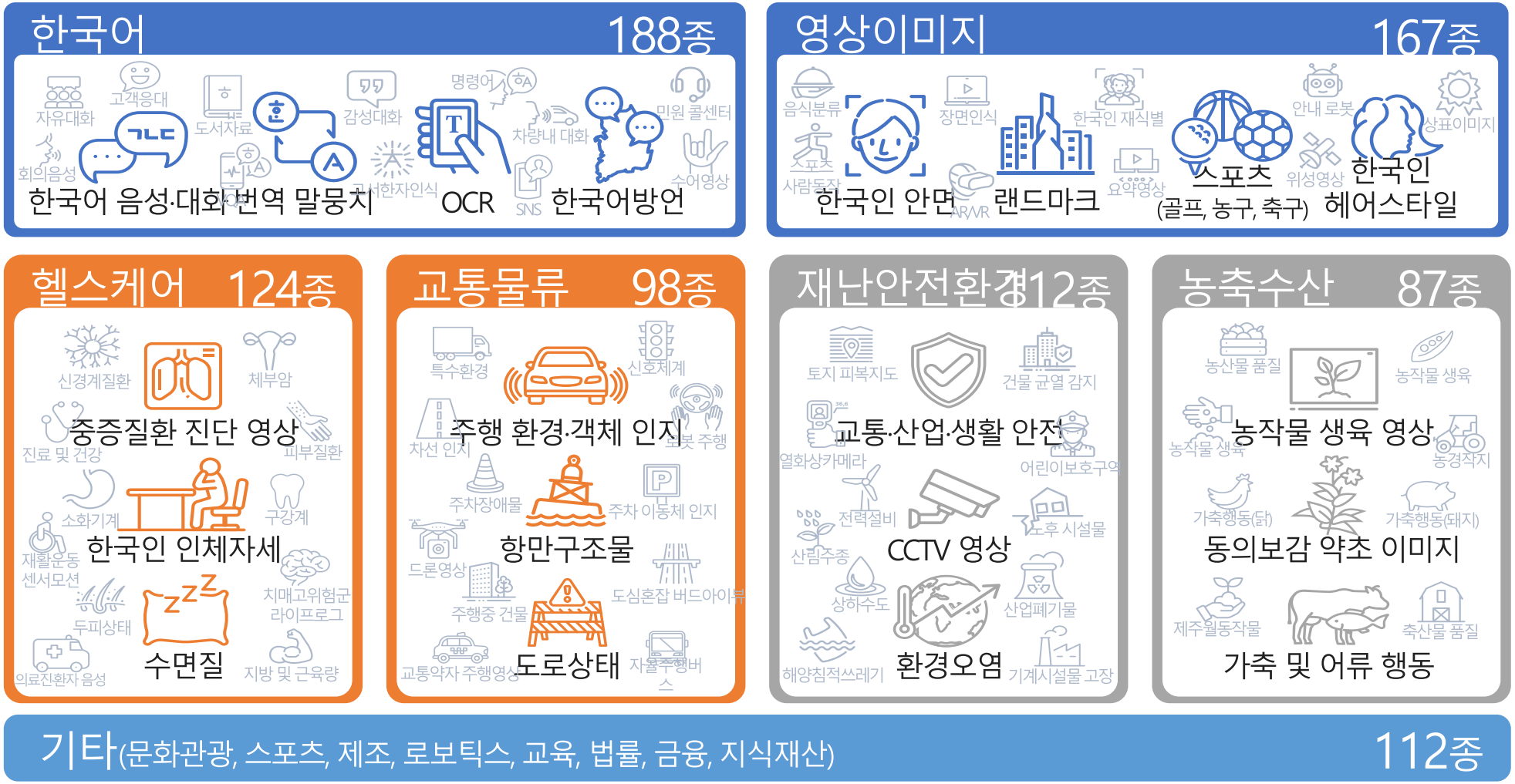
세계최초 정부주도 초거대AI 확산 생태계 조성 사업

2017년부터 시작된 인공지능 학습용 데이터 구축사업



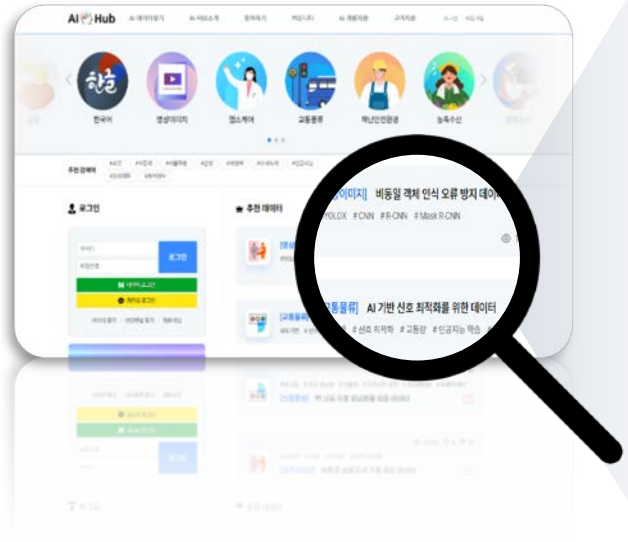
2. AI 데이터 규모

구축한 데이터는 AI 허브(aihub.or.kr)를 통해 제공, 14대 분야, 888종 데이터 (600TB 규모, '25.5.13 기준)



3. AI 데이터 제공 방식(AI허브)

www.aihub.or.kr를 통한 데이터 제공



1 **데이터 구축 단계의 모든 산출물 제공**

* 데이터 설명서, AI 모델, 소스코드, 저작도구 등

2 개발환경에서 데이터를 다운로드받을 수 있는 AI 허브 다운로드 API 제공

4. 의료 데이터의 제공

안심존을 통한 제공

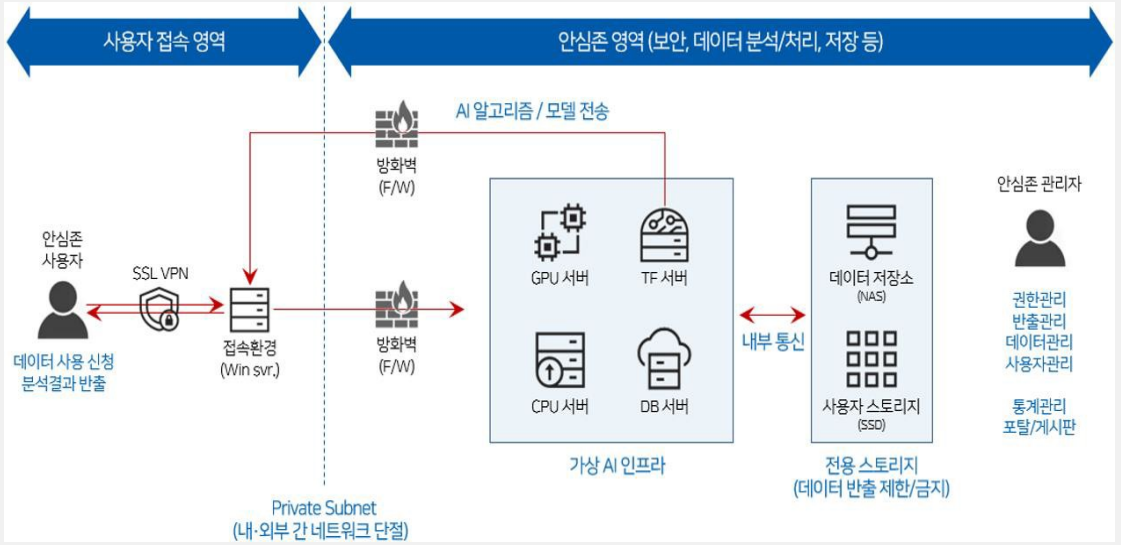
안심존 구축 운영 경과

데이터 이용을 위해
생명윤리법상 IRB심의를
받은 이용자만 접근가능한
의료데이터 등
보안 조치가 요구되는
민감정보 데이터 개방을
위해 인터넷망과 분리된
온라인 / 오프라인
'안심존' 구축('21)

서초 센터('21~'24)
원주센터('22~)
광주센터 ('22~)
대구센터('23~) 운영 중

AI Hub 안심존 구성

클라우드 기반으로 헬스케어 민감 데이터 및
고성능 GPU자원을 함께 제공하여
인공지능 기술 개발을
가능케 하는 **안전한 연구 환경제공**



13

5. ’25년 초거대 AI 데이터 구축 방향

Vertical AI로의 전환

특정 도메인에 대한 전문성을 기반으로
도메인별 심층 지식 기반의 경량화 된 형태의 AI

- 도메인의 심층적인 지식을 학습한 전문화된 초거대AI 모델 기반 AI 서비스를 통해 비즈니스 가치 창출
- 범용 AI 대비 특화된 전문 지식과 최신 AI 기술 트렌드(추론형 AI, Physical AI 등)를 반영하여 AI 서비스 경쟁력 확대

Needs

전문화된 AI 모델로의 산업 변화

전문 영역에서 요구되는 고도의 전문성과 정확성을 충족하는 AI에 대한 니즈 확대

+

Tech

기술적 발전의 가속화

모델 경량화와 최신 트렌드가 되는 알고리즘 고도화를 통해 선진화된 AI 기술로의 발전

산업 구조의 혁신을 위한 핵심 경쟁력

※ (출처) 버티컬 AI로의 변화 및 과제(AI@Data Report_2024-제3호)

Vertical AI 구현을 위한 데이터

심층 데이터

논리적 사고 능력 개선

AI 추론 성능을 강화하고 정확도 · 윤리성을 개선하는 등 AI 모델 성능을 향상·심화시키는 데이터 구축

특화 데이터

도메인 전문성 강화

국내 기업이 생성형 AI 기반의 다양한 응용서비스 시장을 선점하도록 분야별 AI 모델 품질 고도화에 필요한 특화데이터셋 구축 지원

6. ’24년 초거대 AI 데이터 구축방향

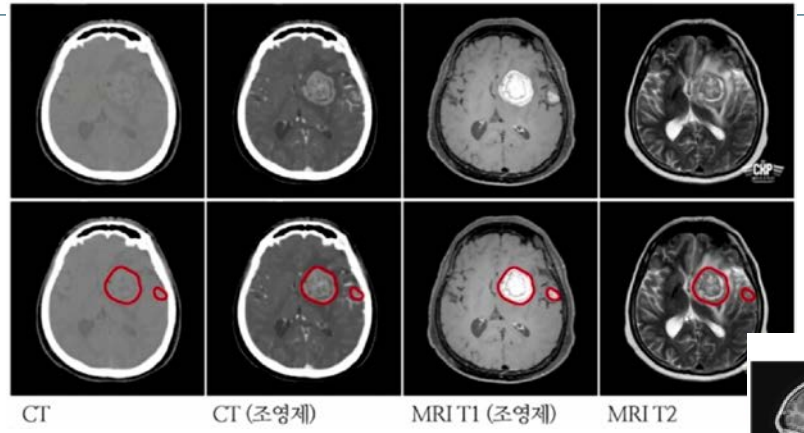
산업에 특화된 sLLM 개발용 말뭉치 데이터 구축

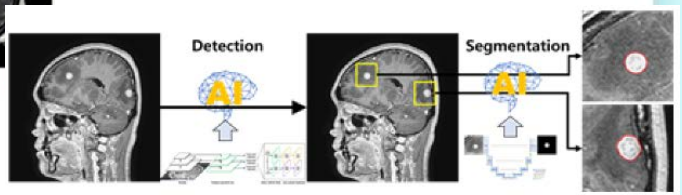
생성형AI 적용 시 필수적인 10개 영역에 양질의 원천 데이터 구축
해당 부처와 함께 기획하여 부처에서 필요로 하는 데이터 구축 (국방, 농림축수산 등)

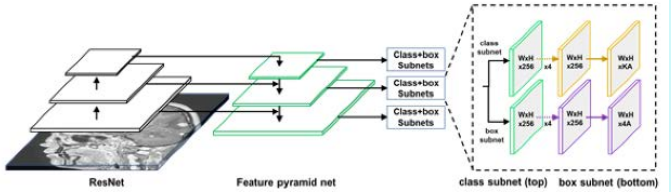
국민생활 밀착형	법률 · 의료 · 교육 · 행정사무
산업혁신	교통 · 물류 · 제조 로보틱스 · 콘텐츠
공공서비스 혁신	재난 · 안전 · 환경 · 국방 · 농림축수산

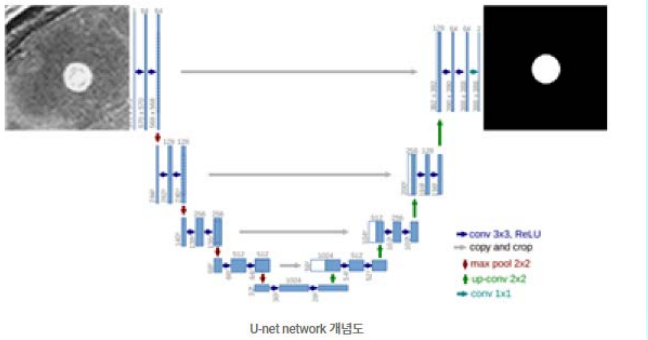
7. (2021) 뇌전이암 진단 및 치료 영상 데이터

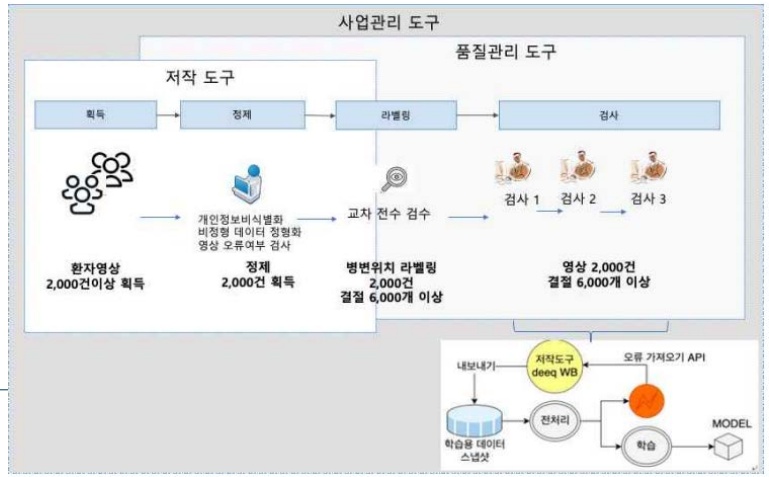
구축연도	데이터셋 명	모달리티	주요 Task
2021	뇌전이암 진단 및 치료 영상 데이터	Vision, tabular	Segmentation, Object-detection
소개	- 뇌전이암의 조영증강 뇌 MRI를 중심으로 뇌 CT 및 MRI를 이용하여 뇌전이암의 정확한 위치와 크기를 인식하기 위한 AI 시스템을 학습 시킬 수 있는 데이터셋 구축이 목표임		
원천데이터	- dcm: 627,991		
라벨링데이터	- json: 900		
하위 데이터 종류	o 치료 전 T1e(1000례) o 보조데이터(1000례) - 치료 전 MRI(T2) - 치료 후 MRI(T1e, T2) - 치료 전 CT(CT, CTE) ※ T1e MRI : 물은 흑색, 지방은 백색 ※ T2 MRI : 물은 백색, 지방은 흑색		

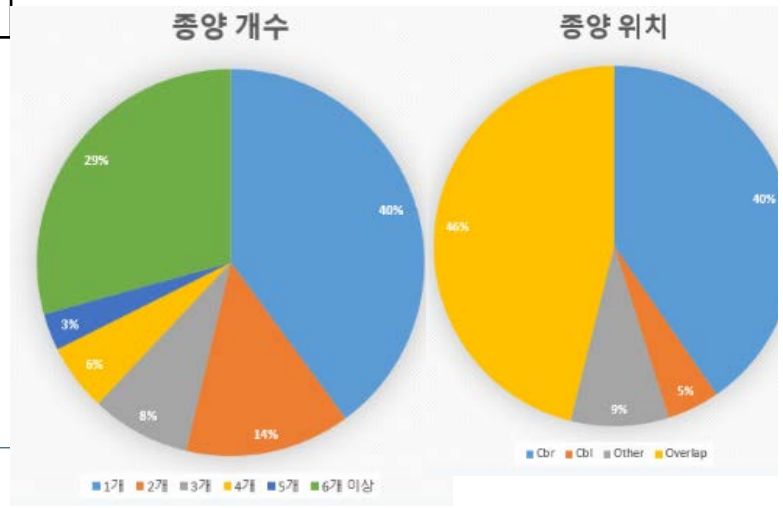








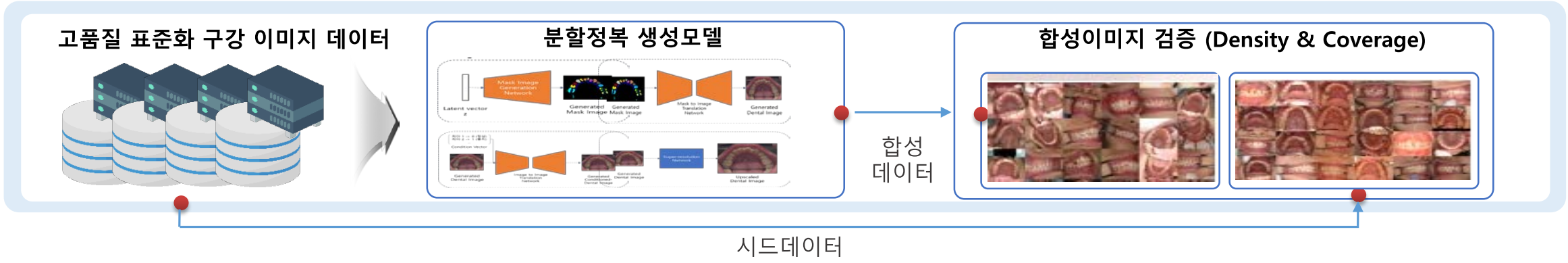




8. (2022~2023) 구강 이미지 데이터

기존 연구의 학습에 사용된 국소촬영 이미지가 아닌 전악촬영 이미지를 이용한 다양성 확보

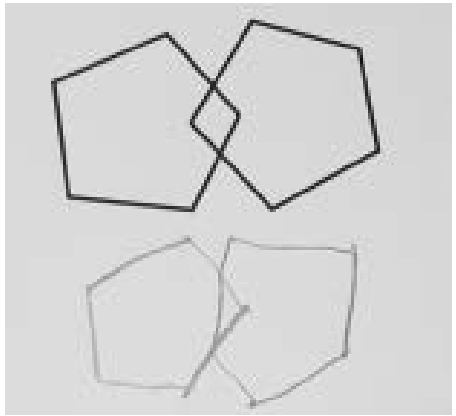
데이터 증강과 합성으로 공개 가능한 데이터셋 구축

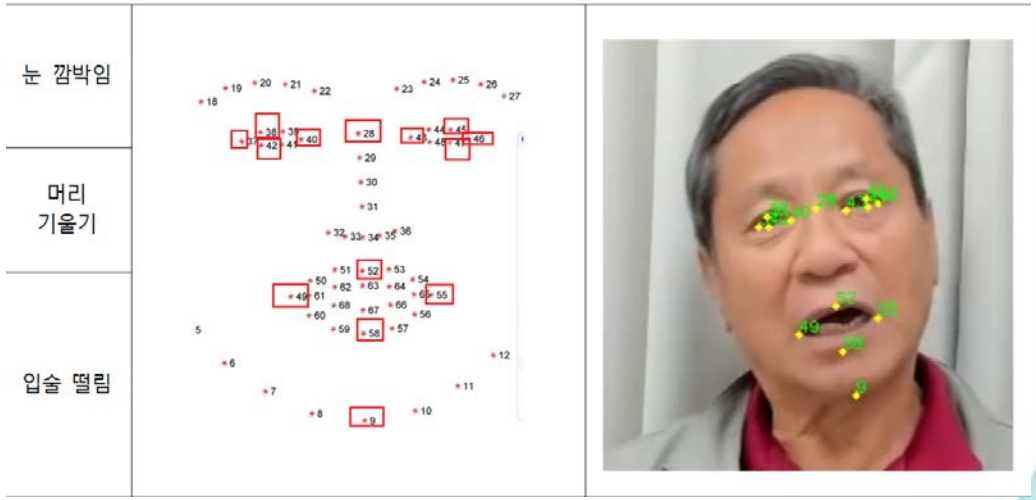


8. (2023) 치매 멀티모달 데이터

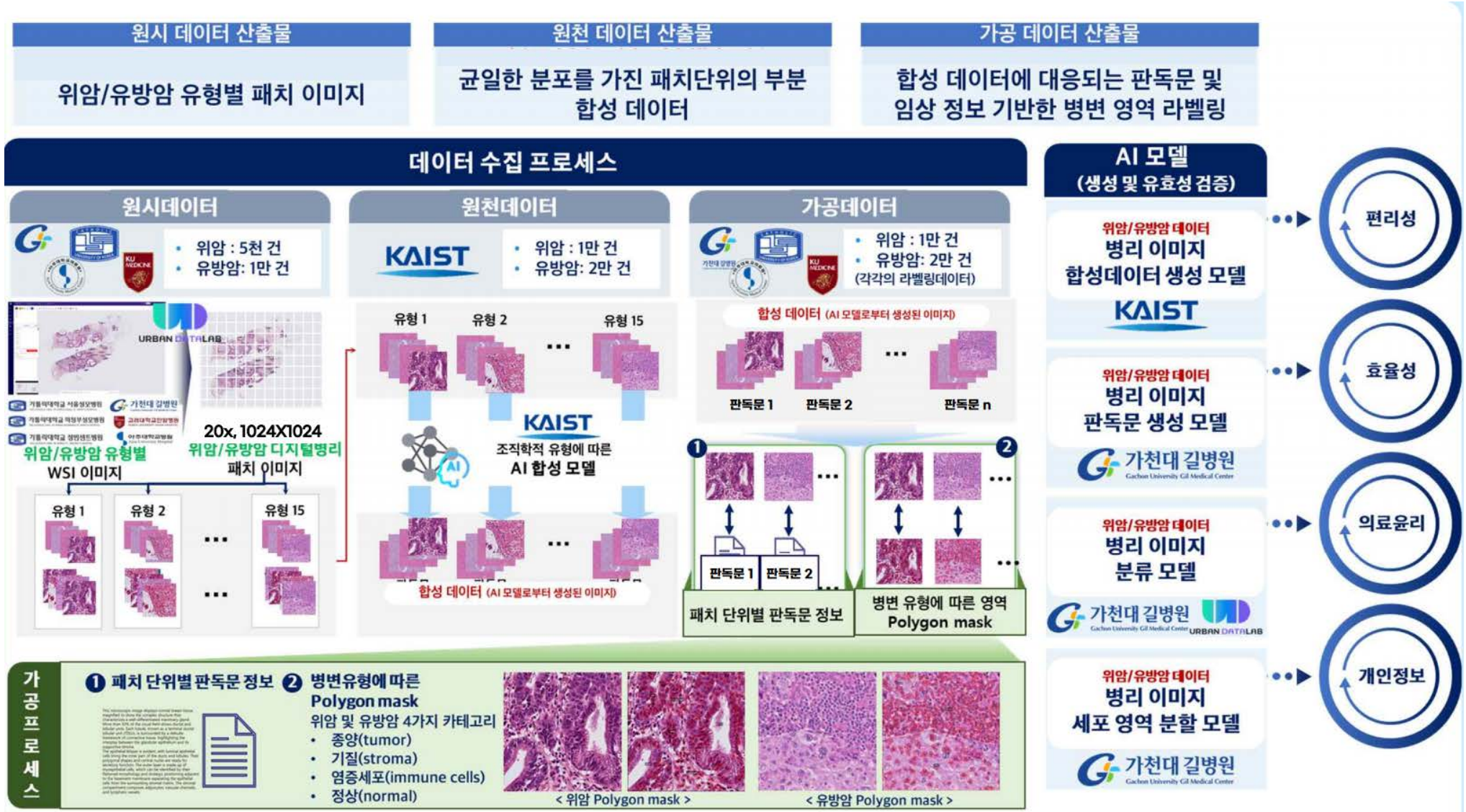
치매환자 유병률 추이

치매 판별 데이터(음성, 오각형, 발화동작)





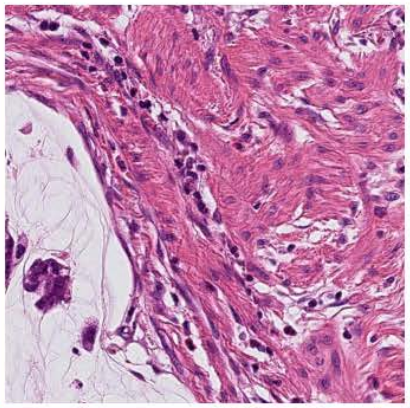
8. (2024) 병리 이미지 합성 데이터

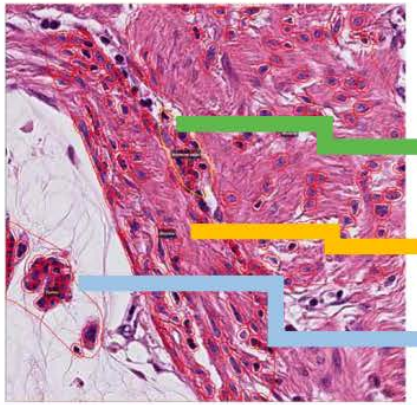


4__(2024) 병리 이미지 합성 데이터

위암 병리 이미지

라벨링 예시





class_1:
Normal(정상)

class_2:
Immune cells(염증세포)

class_3:
Stroma(기질)

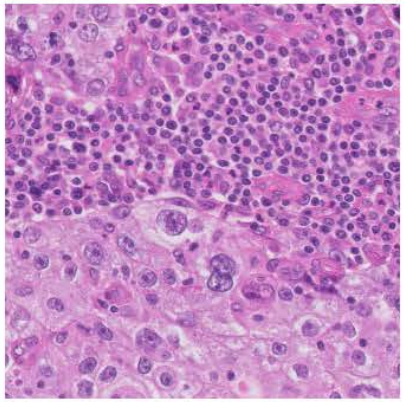
class_4:
Tumor(종양)

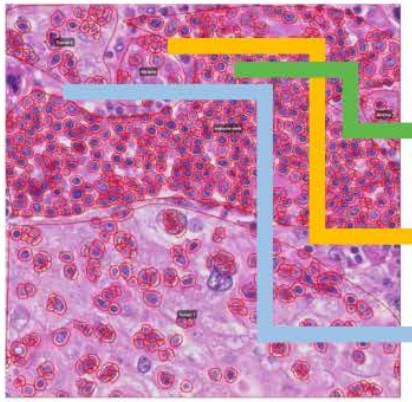
판독문 예시

Carcinoma with glandular differentiation occupies less than 50% of the tissue, and muscle cells are visible.

유방암 병리 이미지

라벨링 예시





class_1:
Normal(정상)

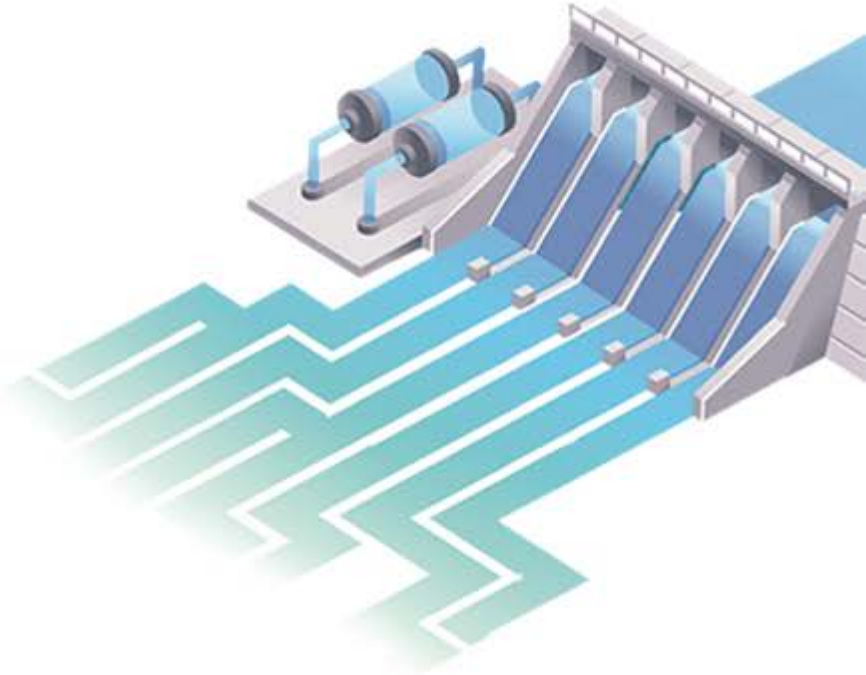
class_2:
Immune cells(염증세포)

class_3:
Stroma(기질)

class_4:
Tumor(종양)

판독문 예시

Irregular glandular structures invade the stroma, with no normal ducts visible and prominent inflammation.

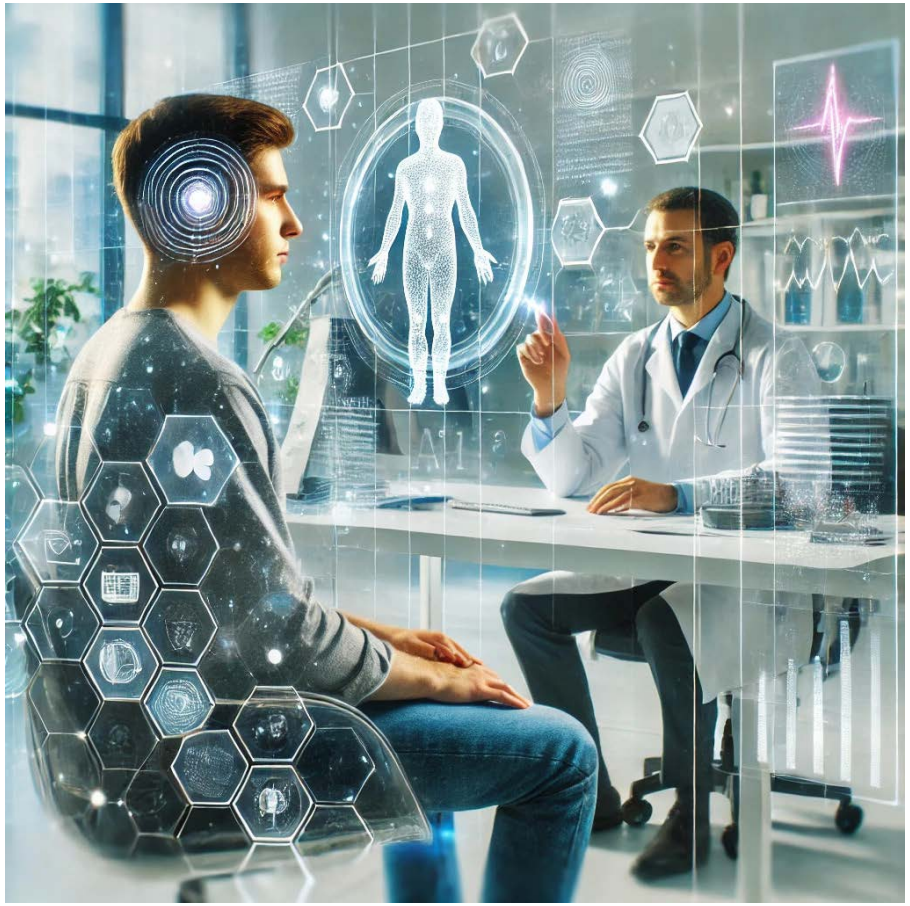


Ⅲ

향후 발전 방안

1. 보건의료에서의 AI 역할

AI의 발전은 의료 영상기기의 발전에 이어 두뇌를 탑재하는 근본적 변화를 이끌고 있음



DALL-E 2로 작성된 일러스트

➢ 국가

- 보건의료 정책 수립, 전국적인 의료 자원 효율화, 감염병 관리
- 의료 교육 및 연구 지원, 약물 개발 및 임상시험 가속화

➢ 의료기관

- 병원 운영 효율화(병상, 수술실 배정 효율화, 자동 의무기록 작성),
- 의료진 운영 최적화, 환자 관리 자동화

➢ 의료진-환자 차원

- 사전적인 질병예방 활동, 진단 및 치료에서의 의사결정 지원,
- 적극적이고 능동적인 사후관리, 개인화된 의료서비스(환자)

2. 의료 프로세스의 복잡성

IT 의료 인프라의 난제

- 데이터 다양성
- 데이터 크기
- 데이터 표준
- 데이터 비식별
- 데이터 보안

의사결정 과정에서의 어려움

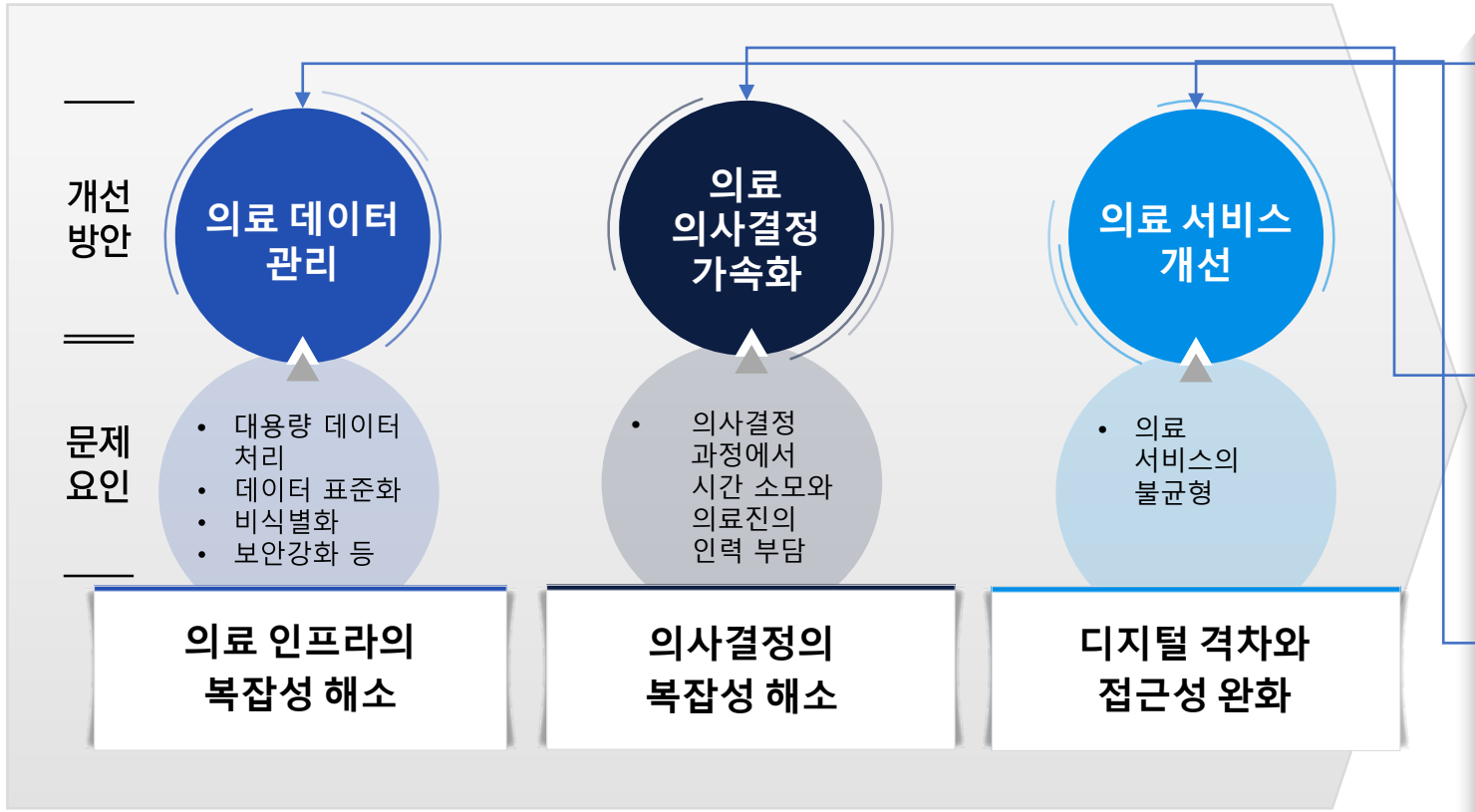
- 환자 맞춤형 의료의 증가
- 다단계 진료 과정
- 데이터 기반의 정밀의료
- 의사결정의 시간적 부담 및 인력 소모
- 전공간 Silo 문화

디지털 격차와 접근성 한계

- 기술 도입의 비용 문제
- 기술 인프라의 불균형
- 의료진의 교육 격차
- 환자 접근성의 불균형

헬스케어 산업 성장에서 인공지능(AI)은 의료 프로세스의 복잡성을 해소하는데 있어 경제와 사회의 핵심 동력으로 작용

3. AI를 통한 효율화



AI 기반 자동화된 의료 데이터 관리를 통해 의료 운영을 더욱 효율적으로 향상

- 대용량 데이터를 빠르게 처리하고, 다양한 형식의 데이터를 표준화하여 상호운용성을 높임.
- 비식별화와 보안 강화를 통해 개인정보 보호와 데이터 안전성을 확보

AI를 활용하여 의료 진단의 효율성을 개선

- AI는 데이터 통합과 자동화된 분석을 통해 진단을 신속하게 지원하고, 치료 결정을 자동으로 추천하여 의료진의 시간적 부담과 반복 업무를 줄여 의료 프로세스를 간소화 함.

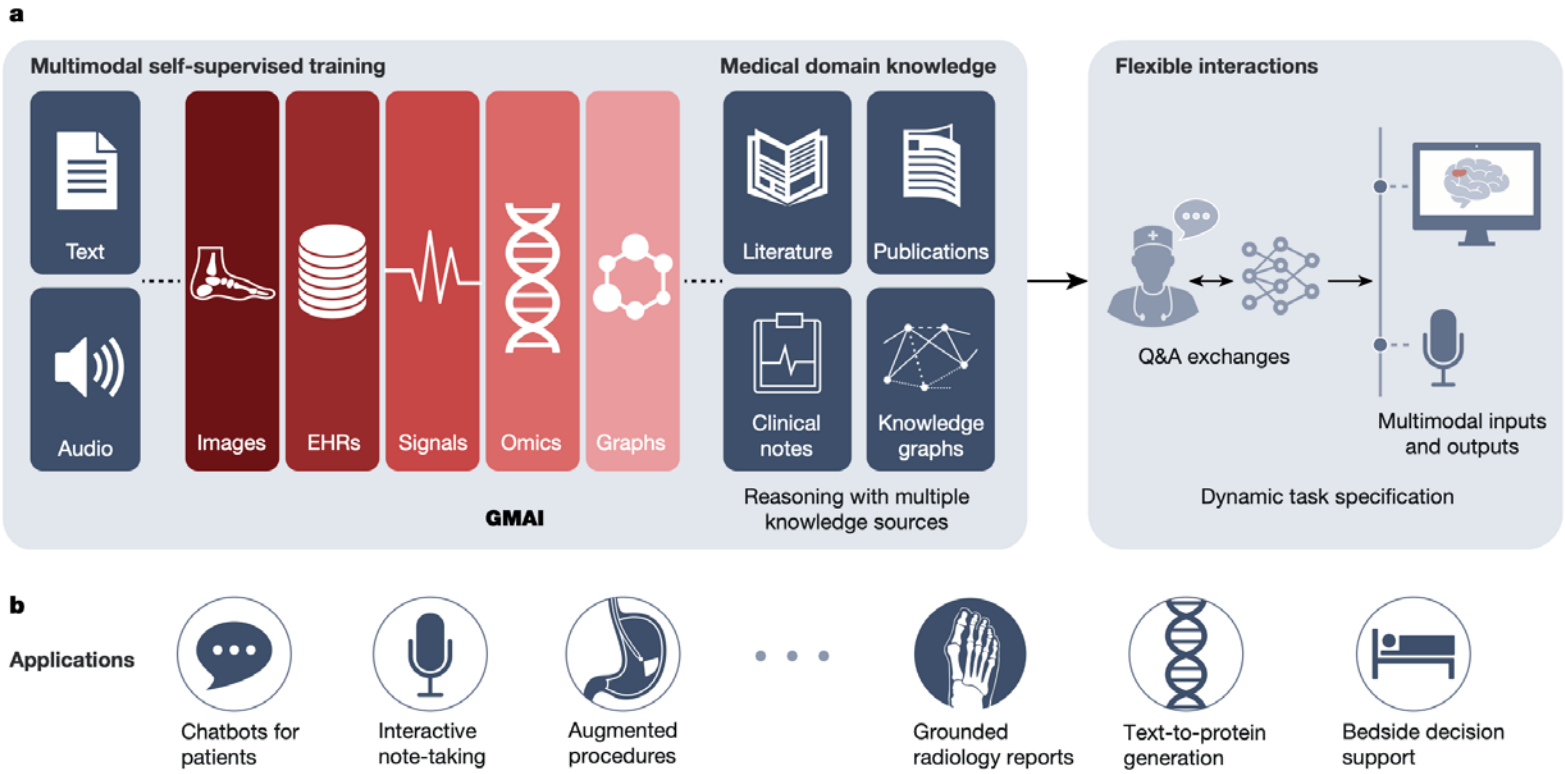
AI를 통한 건강 모니터링 및 예방치료 향상

- AI 기반 시스템은 원격의료와 자동화된 진단 시스템을 개선하고 원격의료와 AI진단을 통해 환자는 초기 단계에서 질병을 발견. 의료진은 빠르고 정확한 대응을 할 수 있어 **예방치료**의 효과가 극대화 됨.

AI는 의료 시스템의 효율성을 크게 향상시키는 중요한 도구임

4. GMAI (Generalist Medical AI)의 도래와 영향

- Generalist Medical AI 정의: 일반적인 의료 AI는 특정 질병, 진단, 또는 치료 방법에 특화된 모델들이 많지만, GMAI는 여러 의료 분야에서 폭넓은 역할을 수행할 수 있는 **범용적인 AI**를 목표로 함
- GMAI는 의료 인공지능의 최종 형태로 여겨질 수 있으며, 다양한 의료 정보를 종합적으로 분석하여 진단하고, 다양한 분야의 의료 전문가들이 통합적으로 협업할 수 있도록 돕는 역할을 하므로 **다학제 협업을 위한 통합 진료 모델의 기반**이 됨



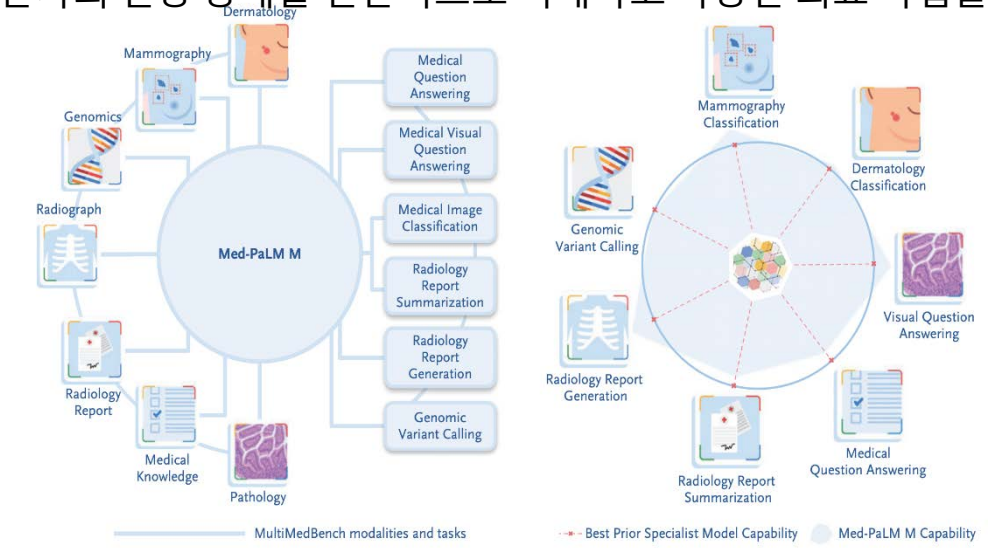
Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259-265.

4. GMAI (Generalist Medical AI)의 도래와 영향

제너럴리스트 의료 인공지능 도입 사례

Google의 Med-PaLM M

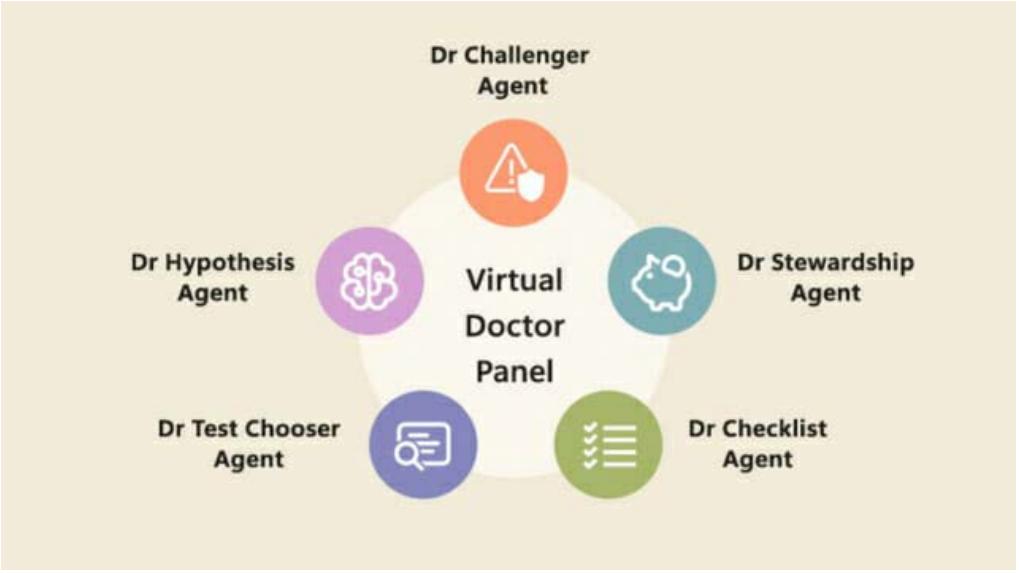
- 의료분야에서 제너럴리스트 인공지능(generalist AI)의 유의미한 성과를 보여준 첫번째 시도
- 텍스트, 이미지, 유전체 데이터 등 다양한 의료 데이터를 통합하여 하나의 아키텍처로 처리할 수 있는 멀티모달 AI 시스템. 이를 통해 환자의 건강 상태를 전반적으로 이해하고 다양한 의료 작업을 지원



Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P. C., ... & Natarajan, V. (2024). Towards generalist biomedical AI. NEJM AI, 1(3), AIoa2300138.

마이크로소프트의 MAI-DxO

- 여러 AI 에이전트들이 협력하여 질병을 진단하는 인공지능 협업 의료진단 시스템
- 의사 역할을 하는 각각의 AI 에이전트들이 오픈AI의 챗GPT, 구글의 제미나이, 앤스로픽의 클로드, 메타의 라마 등 주요 AI 모델에 질의하는 방식으로 의사보다 4배 높은 진단 정확도를 보여줌

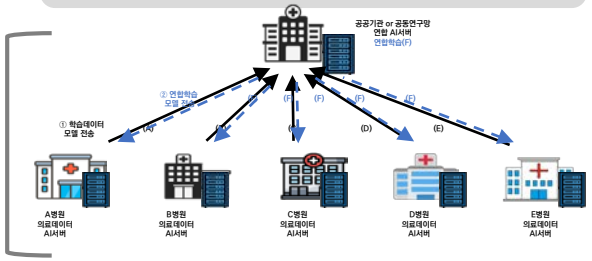


5. 미래 전망: 의료 AI의 혁신적 잠재력

As-is

의사결정지원 시스템 고도화

연합학습(Federated learning)



데이터 통합

영상
음성
표형식
자연어
유전체

합성데이터
(synthetic data)

2D
3D

적용
·
활성화

To-be

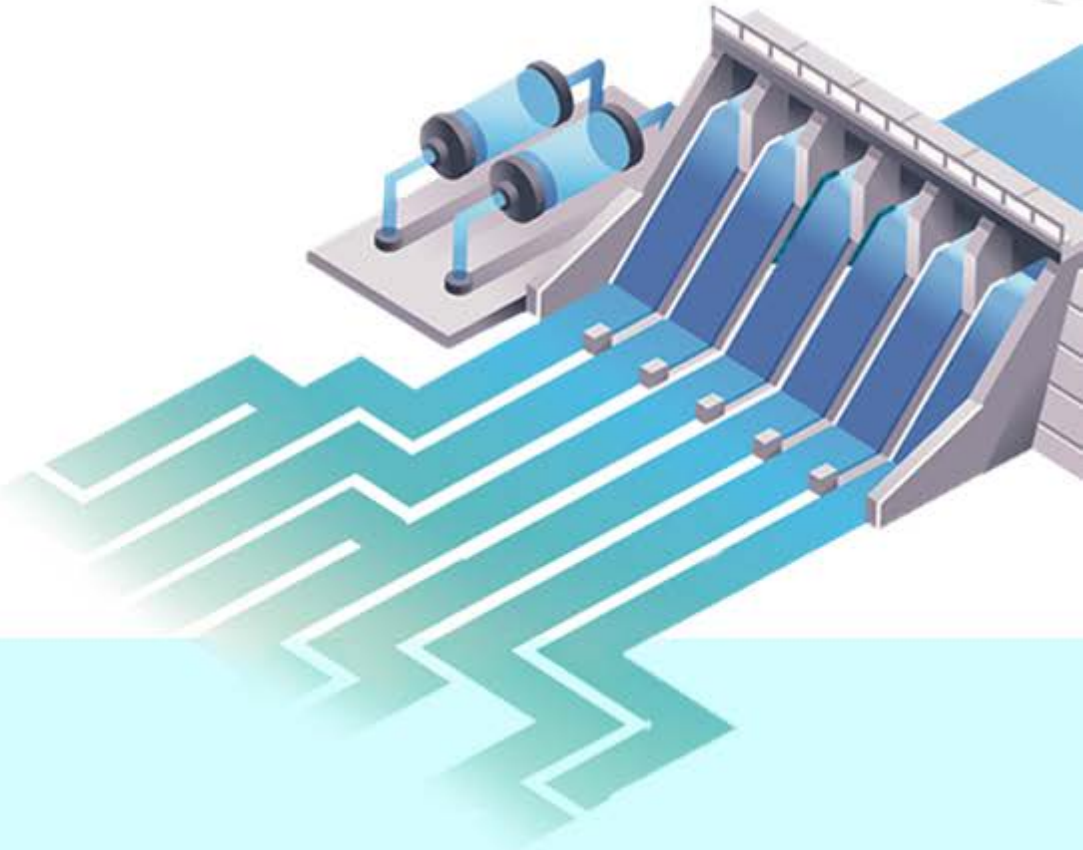
제너럴리스트 의료 인공지능 진화

고도화

다학제 협진을 위한 통합 진료 AI 모델로의 발전

AI는 다학제 진료의 미래를 변화시키는 주요 도구로, 의료진 간의 협력과 환자 진단의 정확성을 향상

감사합니다 ;)



kimcon@nia.or.kr
054-230-4286

NIA 한국지능정보사회진흥원 NATIONAL INFORMATION SOCIETY AGENCY

AI Hub