비정형데이터 가명처리 기준
주요 내용

2024. 2.

개인정보보호위원회

목 차

1. 추진 배경 ····· 1

2. 비정형데이터 가명처리·활용의 특수성 및 고려사항 ·· 2

3. 비정형데이터 가명처리 기본원칙 ····· 3

4. 비정형데이터 가명처리 단계별 고려사항 ····· 8

- 1 사전 준비 단계 ····· 8
- 2 위험성 검토 단계 ····· 8
- 3 가명처리 단계 ····· 11
- 4 적정성 검토 단계 ····· 12
- 5 안전한 관리 단계 ····· 12

5. 비정형데이터 개인식별 위험성 검토 체크리스트 · 14

6. 비정형데이터 위험성 검토 항목별 조치 가이드 ····· 16

[참고] 비정형데이터 가명처리 기술 및 예시 ····· 18

[부록] 비정형데이터 가명처리 시나리오 예시 ····· 28

- (시나리오 ❶) 의료 (이미지, 영상, 텍스트) 유방암·골밀도 감소 여부 진단 AI 개발 28
- (시나리오 ❷) 의료 (이미지) 구강질환 진단 AI 개발 ····· 45
- (시나리오 ❸) 의료 (이미지, 영상) 안면골 골절 진단 AI 개발 ····· 54
- (시나리오 ❹) 교통 (이미지, 영상) 자율주행차 주행 시 비정상 상황인지 AI 개발 63
- (시나리오 ❺) 교통 (이미지) 고속도로 다인승전용차로 단속 AI 개발 ····· 71
- (시나리오 ❻) 대화·검색 (텍스트) 한국어 대화가 가능한 AI 챗봇 개발 ····· 80
- (시나리오 ❼) 대화·교육 (음성) 콜센터 직원 실습용 가상상담 시나리오 생성 AI 개발· 90

1 추진 배경

□ **AI 기술 발전**과 **컴퓨팅 자원 발달**로 **데이터 활용수요**가 전통적 **정형데이터**(수치)에서 **비정형데이터**(이미지, 영상, 음성, 텍스트)**로 변화**

* 전 세계 데이터 중 이미지, 영상, 음성, 텍스트 등 비정형데이터가 최대 90%를 차지 (IDC, '23)

참고 **비정형데이터의 가명처리·활용 예시**

- **(이미지·영상) 특정 질병을 진단(보조)하는 의료 AI 연구개발**을 위해 병원이 보유한 **MRI, CT, X-ray 사진·영상**을 가명처리 후 **학습데이터로 활용**
- **(이미지·영상) 불법현수막을 탐지**하여 알려주는 **지능형 CCTV 개발**을 위해 지자체가 보유한 **공공장소 CCTV 촬영영상**을 가명처리하여 **AI 연구개발에 활용**
- **(음성·텍스트) 민원인 상담·대응**을 위한 **음성생성 AI를 개발**하기 위해 **공공기관**이 보유한 **민원상담 음성정보**와 **상담기록 정보**를 가명처리하여 학습데이터로 활용

□ 그러나, 기존 「**가이드라인**」은 정형데이터 기준으로 되어있어 **비정형데이터 고유 특징**과 **AI 시대 기술발전** 등을 반영하지 못하는 한계

참고 **정형데이터와 비정형데이터의 차이점**

▸ 정형데이터	▸ 비정형데이터
(정의) 정해진 규칙에 맞게 구조화된 형식으로 존재하는 데이터 ※ 예) DB에 열과 행으로 저장된 테이블형식의 자료 등	(정의) 일정한 규격이나 정해진 형태가 없이 구조화되지 않은 데이터 ※ 예) 사진, 비디오, 통화음성, 대화기록, 보고서, 메일 본문 등
(특징) **데이터** 연산, 분석 등 **데이터 처리 방식**, 가명처리 기술·방법이 비교적 단순	(특징) 연구목적·환경에 따라 데이터 처리방식 및 가명처리 기술·방법이 복잡·다양

➡ AI 일상화 등 디지털 심화 시대에 맞는 **비정형데이터 가명처리 기준**을 마련하여 **기업·연구자의 불확실성 해소 및 기술혁신** 지원

2 비정형데이터 가명처리 · 활용의 특수성 및 고려사항

□ (개인식별성 판단의 어려움) 개인식별 가능 정보와 그렇지 않은 정보의 구분이 상대적이며, 처리 목적·환경 등에 따라 다르게 판단될 수 있음

예시 **얼굴 CT 사진 1장**으로는 **개인식별 위험성이 낮지만**, 여러 위치·각도에서 촬영한 **얼굴 CT 사진을 여러장 결합**하면 **얼굴형상 재건**이 가능하여 **개인식별 위험성 증가**

예시 눈·코·입을 알아볼 수 없는 거리에서 찍힌 CCTV 영상은 통상 **개인식별 위험성이 낮지만**, **흉터, 문신, 머리스타일** 등 특이한 신체 특징이 있는 경우 **개인식별 위험성이 높음**

□ (가명처리 기술의 불완전성) 비정형데이터 내 개인식별 위험성이 있는 모든 항목을 완벽하게 탐지·처리할 수 있는 기술이 부재

예시 **이미지·영상 데이터**의 경우, 해상도, 조명 각도, 객체 크기 등에 따라 얼굴 등을 탐지하지 못하는 경우가 존재 → 최근 AI 기술의 **객체 탐지 정확도**는 **90~98% 수준**[1)]

예시 '신뢰역 1번출구 앞 파란건물 1층 1호가 우리집' **텍스트**를 **주소로 인식하지 않아 처리하지 않거나**, **'김신뢰 김밥'** 등 상호명을 개인정보(이름)로 인식해 **불필요하게 삭제**

□ (재식별 공격 위험) AI 및 데이터 복원기술 발달로, 다른 정보와의 연계·결합 없이도 개인 재식별 공격 위험성 증가

예시 **음성변조 규칙을 몰라도** 화자의 **원본 목소리를 복원**할 수 있는 **기술 존재**[2)]

예시 가명처리된 사진의 모자이크 패턴을 몰라도, **AI를 통해 모자이크된 사진을 원본에 가깝게 복원**해낼 수 있는 **기술 연구 중**[3)]

⇒ **비정형데이터 가명처리·활용 시 데이터 처리 맥락(context), 가명처리 기술의 한계, 재식별 공격 위험 등을 고려하여 개인식별 위험성을 낮춰야 함**

1) Kaur, J., Singh, W., "Tools, Techniques, datasets and application areas for object detection in an image: a review", Multimed Tools Appl 81 (2022)
2) Deng, Jiangyi et al, "Catch You and I Can: Revealing Source Voiceprint Against Voice Conversion", ArXivabs/2302.12434 (2023)
3) R. Dahl, M. Norouzi and J. Shlens, "Pixel Recursive Super Resolution", 2017 IEEE International Conference on Computer Vision (ICCV) (2017)

3 비정형데이터 가명처리 기본원칙

□1 **데이터 처리 목적·환경, 민감도** 등을 **종합적으로 고려**하여 **개인식별 위험성**이 있는 정보를 판단하고, **합리적인 처리 방법·수준** 설정

○ 연구목적에 맞춰 **데이터 자체 훼손을 최소화**하면서 **관리적·환경적 통제** 등 **다양한 안전성 확보 방안** 적용 가능

- **연구목적 달성에 필수적인 정보항목을 남기는 대신 그 외 정보항목에 대한 가명처리 수준**을 높이거나, **다른 정보 및 소프트웨어**(SW) **반입 제한** 등 **충분한 안전조치를 보완**하여 활용

□2 **가명처리 기술의 한계 등을 보완**하기 위해 **사전 준비단계**(연구 및 기술개발 기획)부터 **위험성을 충실히 검토**하고 **적절한 안전조치**를 수행

○ **가명처리 기술의 한계 보완**을 위해 **다음 조치를 이행**할 것을 권고

❶ 가명처리 기술의 적절성·신뢰성을 확인할 수 있는 **근거 작성·보관**
❷ 가명처리 기술 적용 이후, 처리 결과에 대해 **자체적인 검수 수행**
❸ 가명처리 **적정성 검토 과정에서 ❶, ❷를 포함하여 점검**
(외부전문가 과반 이상 참여 바람직)

○ **사전에 식별된** 개인정보 침해 위험을 **예방**하기 위해서 가명정보 활용에 참여하는 **각 기관의 내부통제 강화 노력**이 병행될 필요

○ 가명정보의 **처리목적을 달성**하면 **신속히 가명정보를 파기**하여 **사후적 위험 최소화**

□3 **데이터 복원기술 발달** 등에 대응하여, 가명처리된 비정형데이터 활용 시 **관련 시스템·SW의 접근·사용 제한** 등 **통제방안 마련**

*** 원본 복원에 활용될 수 있는 추가정보 분리보관, 복원 SW에 대한 접근권한 제한 등**

○ **AI 개발·활용** 과정에서 나타날 수 있는 **다양한 위험**을 **사전에 완벽하게 제거하는 것은 불가능**하므로, **AI 서비스 제공 과정**에서도 **개인식별 위험** 등 정보주체 권익 침해 가능성을 **지속 모니터링**

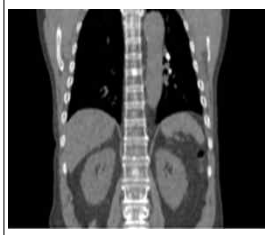
〈 주요 비정형데이터 가명처리 시나리오 예시 〉

※ 아래 가명처리 시나리오는 **실제 비정형데이터를 가명처리하여 활용했던 사례**를 관련 기업·기관 및 전문가 논의를 통해 **재구성**한 것으로 **단순 참고용**이며, 처리자 및 적정성 검토 위원회 등의 판단에 따라 **데이터 활용 분야·상황에 맞게 가명처리 방법·수준 등을 자유롭게 적용할 수 있음**

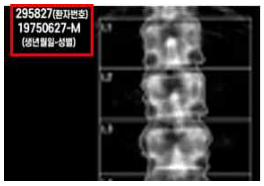
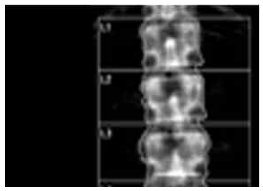
〔사례1〕 유방암 · 골밀도 감소 여부 진단 AI 개발

※ 대학병원이 보유한 **유방암 환자 CT사진(영상·이미지) 및 병리기록지(텍스트)**를 가명처리하여 **유방암 및 골밀도 감소 진단 AI 개발**을 위한 내부 연구에 활용한 사례

☞ **개인식별 위험이 발생하지 않도록 처리환경을 안전하게 통제하고 복원 SW 반입 제한 조치 등을 통해 별도의 가명처리 없이 CT 사진을 그대로 활용**

< 흉부 CT사진 >			(그대로 활용)
	개인식별 위험성 검토	- 흉부 CT사진만으로는 **개인식별 위험성 거의 없음** - **개인당 200장씩 촬영된 CT사진이 활용**되는 연구로서 3차원 재건 기술 등을 통해 신체형상의 입체적 복원이 가능하고, 복원 시 특이한 외형·흉터 등이 있는 극히 일부 환자의 경우 낮은 확률로 **개인식별 위험성 존재** - **클라우드 기반 폐쇄연구분석환경*을 이용**하고 인가되지 않은 데이터·프로그램 반입을 철저히 통제하고 있어 **3차원 재건기술 적용 불가** * **클라우드 서버에 데이터를 저장**하고 타 **외부망에서는 클라우드 서버 접속이 제한**되는 분석실에서 **인가받은 인원만** 데이터 접근 가능	
	데이터 처리 방안	➜ 3차원 재건으로 인한 개인식별 위험성이 존재하나, 환경적 통제로 인해 해당 위험의 발생 가능성이 없으므로 **별도의 가명처리 없이 그대로 활용 가능**	

☞ **이미지 내 개인식별 위험성이 있는 메타데이터를 삭제하고 활용**

< CT사진 내 환자관련정보 >			(블랙마스킹 처리)
	개인식별 위험성 검토	- 이미지 내 표시된 환자관련정보*는 타 정보와 결합되어 분석될 경우 **개인식별 위험성이 있음** * DICOM 헤더정보(환자번호, 생년월일, 성별) 표시 - 해당 정보는 **연구에 필요하지 않은 정보**임	
	데이터 처리 방안	➜ **블랙마스킹** 기법을 통해 **환자관련정보 삭제**	

☞ **비정형 텍스트데이터를 정형데이터 형태로 변환하여 활용**

< 병리기록 텍스트 >	개인식별 위험성 검토	- 암병리기록지에는 연구에 불필요한 다양한 **개인 식별가능정보가 정제되지 않은 형태**로 존재하여 개인식별 위험이 존재	(정형데이터로 변환 후 활용)
	데이터 처리 방안	➜ 자연어처리 기술을 통해 **정형데이터로 변환한 후에 활용**하고 개인식별 위험성이 있는 항목이 있으면 가명처리 수행 ➜ 자연어처리 기술 및 텍스트데이터 가명처리 기술의 정확도가 100%가 아니므로, 정형데이터 변환 후 추가 **전수검사 등을 통해 보완**	

[사례2] 구강질환 진단 AI 개발

※ 대학병원이 보유한 **구강 건강검진 촬영 사진(이미지)**을 가명처리한 뒤 기업에 제공하여, **충치·치주염 등 구강질환을 진단하는 AI 연구개발**에 활용한 사례

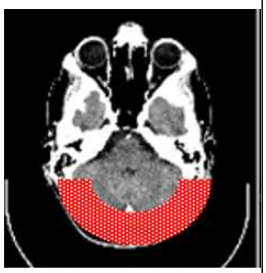
☞ **연구 목적에 필요 없는 영역을 블러링 처리하고, 메타데이터를 삭제하여 활용**

< 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**	(충치부분 : 그대로 활용) (그 외 : 블러링 처리)
	데이터 처리 방안	➜ 연구에 필요한 **충치 영역은 그대로 활용**하고, 연구에 필요 없는 **그 외 영역은 블러링 처리** ※ 블러링 수준은 현재 복원기술 발전수준 및 데이터 처리 환경(타 정보·복원기술 접근성) 등을 고려하여 설정 ➜ **메타데이터**는 연구에 필요 없어 **삭제**	

[사례3] 안면골 골절 진단 AI 개발

※ 대학병원이 보유한 **Facial CT 사진**을 가명처리하여, **안면골(얼굴뼈) 골절여부를 진단하는 AI 개발**을 위한 공동연구를 대학병원과 민간기업이 함께 수행한 사례

☞ **개인식별 위험성을 낮추기 위해 연구목적에 필요 없는 영역만 마스킹 처리하여 활용**

< Facial CT사진 >	개인식별 위험성 검토	- **CT사진 자체로는 개인식별 위험성 거의 없음** - 3차원 재건 기술을 통해 입체적 복원이 가능하며, 복원 시 **특이한 얼굴·외형, 알려진 얼굴** 등 극히 일부 환자의 경우 **낮은 확률로 개인식별 위험성 존재** - 대용량 영상·이미지를 활용하는 연구로 3차원 재건이 가능하나, **가장자리 마스킹 기법**을 활용하여 **3차원 재건 공격 위험을 낮출 수 있음** - **후두부(뇌 뒷부분) 영역은 연구에 필요 없음**	(안면부 : 그대로 활용) (후두부 : 마스킹 처리)
	데이터 처리 방안	➜ 연구에 필요한 **안면부**는 **그대로 활용**하고, 연구에 필요없는 **후두부**는 **마스킹**하여 3차원 재건위험을 낮추고 활용	

[사례4] 자율주행차 주행 시 비정상 상황인지 AI 개발

※ 연구기관이 보유한 **도로 주행상황 촬영 영상**을 가명처리한 뒤 기업에 제공하여, **자율주행자동차 운행 시 비정상 상황*을 인지하는 AI 연구개발**에 활용한 사례

* 사람이 차도에 뛰어드는 상황, 다른 차가 갑자기 앞에 끼어드는 상황, 무단횡단 등

☞ **연구 목적에 필요 없는 영역만 마스킹 처리하여 활용**

< 얼굴·차량번호판 > 	개인식별 위험성 검토	- 사람의 얼굴이 선명히 보이는 경우, 차량 번호판이 그대로 노출되어 차량탑승자 유추가 가능한 경우 등에 **개인식별 위험성 존존재** - 연구목적상 사람·차량의 전체 형상과 움직임만 파악 가능하면 되므로 **얼굴·차량번호판은 마스킹해도 무방**	(마스킹 처리)
	데이터 처리 방안	➜ 얼굴·차량번호판 영역을 사람과 컴퓨터가 식별 불가능한 수준으로 **마스킹하여 활용**	

[사례5] 고속도로 다인승전용차로 단속 AI 개발

※ 지자체가 CCTV로 촬영한 **고속도로 통행차량 이미지**를 가명처리한 뒤 기업에 제공하여, **다인승전용차로 위반*을 단속하는 AI 개발**에 활용한 사례

* 3명 이상 승차하지 않은 승용·승합자동차가 다인승전용차로를 이용한 경우

☞ **연구 목적에 필요 없는 영역만 블러링 처리하여 활용**

< 도로 통행차량 사진 > < 특이점 있는 차량 > 	개인식별 위험성 검토	- 탑승자의 얼굴이 선명하게 촬영된 경우, 차량 외부에 특이점이 있는 경우 등은 **개인식별 위험성 존재** - **연구 목적상 특정인 식별·구분이 필요없고** (1) 사람인지 아닌지 여부, (2) 차량 탑승 인원이 몇 명인지만 확인할 수 있으면 됨 - AI가 사람인지 여부는 판단할 수 있도록 하되, 탑승자가 누구인지는 판별 불가능하도록 블러링 등 처리 필요	(① 탑승자 위차범위 파악) ▼ (② 블러링 처리)
	데이터 처리 방안	➜ **특이점이 있는 차량 이미지는 삭제** ➜ 블러링 수준(1~10단계)별로 데이터를 가명처리한 후, 식별위험이 없으면서 AI 정밀도를 어느 정도 확보할 수 있는 **블러링 수준을 결정** ※ 적정성 검토 단계에서 목적 달성 가능성, 학습데이터의 개인식별 위험성 등을 검증	

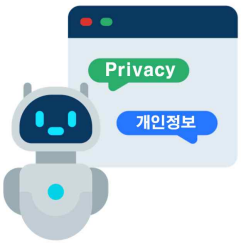
[사례6] 한국어 대화가 가능한 AI 챗봇 개발

※ 인공지능 챗봇 전문기업이 채팅앱을 통해 수집한 사용자 간 **일상대화 텍스트 데이터**를 가명처리하여 **한국어 대화가 가능한 AI 챗봇 연구개발**에 활용한 사례

☞ 개인식별 위험 항목을 엄격히 필터링하여 제거하고 메타데이터는 삭제하여 활용

< 대화텍스트 파일 >	개인식별 위험성 검토	- 일상대화 데이터(텍스트)에는 사생활 관련 정보 등 **개인식별 위험성이 높은 정보가 상당 수 포함**	(메타데이터 삭제, 개인식별정보 필터링·제거)
	데이터 처리 방안	➜ **메타데이터(대화 사용자 ID)를 삭제**하고 랜덤 ID로 대체하여 특정 개인과의 연결성을 제거 ➜ 개인식별 위험이 있는 항목들을 엄격하게 **필터링하여 가명처리(치환*, 삭제)** * 이메일 주소를 '[MAIL]'로 대체하는 등	

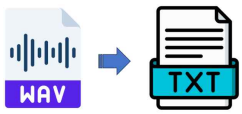
☞ 학습에 활용된 가명정보가 AI 챗봇 답변에 그대로 발화되지 않도록 조치

< 챗봇 답변 시 위험성> 	개인식별 위험성 검토	- 언어모델 학습에 활용된 **가명정보가 AI 챗봇의 답변으로 발화**될 시, 제대로 처리되지 않은 가명정보의 노출 등으로 인한 **개인식별 위험성이 높음**	(학습DB, 답변DB 분리)
	데이터 처리 방안	➜ 언어모델 학습을 위한 **'학습데이터베이스'**와 챗봇 답변을 위한 **'답변데이터베이스'**를 분리하여 학습에 활용된 문장이 그대로 노출되지 않도록 조치 **※ 답변데이터베이스에 개인 식별위험이 있는지 충분히 점검**	

[사례7] 콜센터 직원 실습용 가상상담 시나리오 생성 AI 개발

※ 기업에서 직원-고객간 **음성 상담정보**를 가명처리하여 자사 콜센터 직원들을 위한 **상담 실습교육용 AI 개발**에 활용한 사례

☞ 음성정보를 텍스트로 변환(STT, Speech To Text)한 뒤 가명처리하여 활용

< 음성 상담파일 >	개인식별 위험성 검토	- 상담음성파일에는 고객 및 상담사의 **실제 음성데이터**(목소리, 음색, 억양, 발음 등)가 포함되어 있고, 대화내용에는 다양한 **개인식별가능정보가 정제되지 않은 형태**로 존재 - 가상상담 시나리오 생성 AI 개발에는 상담목적 및 고객 특성에 따른 질의-응답과 대화의 흐름 파악이 중요하며 **실제 음성 자체는 필요하지 않음**	(① 텍스트로 변환)  ▼ (② 개인식별정보 치환삭제)
	데이터 처리 방안	➜ 음성변환(STT) 기술을 통해 **텍스트로 변환**한 뒤, 개인식별 위험성이 있는 항목들을 **가명처리(치환·삭제)하여 활용** ➜ 텍스트데이터에 대한 가명처리 기술의 정확도가 100%가 아니므로, 추가 **전수검사를 통해 식별위험성이 있는 정보를 제거**	

4 비정형데이터 가명처리 단계별 고려사항

▪ 비정형데이터의 가명처리와 관련하여서는 「가명정보 처리 가이드라인」 제2장의 **가명처리 단계별 절차를 동일하게 따르되, 비정형데이터의 특수성을 반영**한 **개인식별 위험성 검토 및 안전조치 사항**을 **추가적으로 고려하여 시행하는 것을 권고**

< 개인정보의 가명처리 단계별 절차 >

1 사전준비 단계

※ 가명정보 처리 목적을 설정·검토하고 목적에 맞는 가명처리 대상을 선정하는 단계

○ 비정형데이터 내 **개인식별 가능성이 있는 항목들을 도출**하고, **목적 달성**에 **필요한 항목의 종류와 범위**를 명확히 하여 **가명처리 대상 선정**

2 위험성 검토 단계

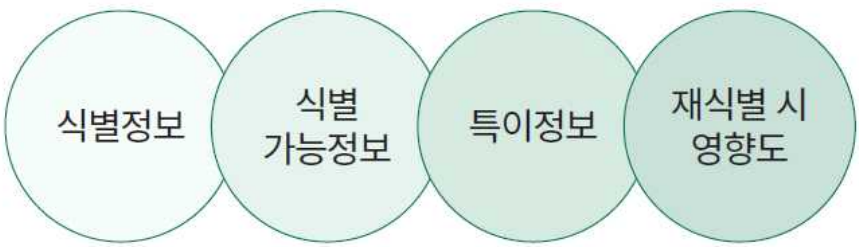
※ 가명처리 대상·처리 환경의 위험성을 검토하여 가명처리 방법·수준에 반영하기 위한 단계

○ **비정형데이터의 특성을 고려**하여 **'데이터 자체 식별 위험성'**과 **'처리 환경의 식별 위험성'**을 **종합적으로 검토**하여 가명처리 방법·수준을 결정

(1) 데이터의 식별 위험성 검토

○ 개인식별 가능성이 높은 정보(❶식별정보, ❷식별가능정보), ❸특이정보, ❹데이터 특성만으로 재식별 시 영향도가 높은 항목이 있는지 검토

< 데이터 자체 개인식별 위험성 요소 >



- **(식별성)** 비정형데이터는 정형데이터와 달리 **식별정보***와 **식별가능정보****의 **절대적인 구분이 어렵기 때문에,** 처리 목적·방법 등 **데이터 처리 맥락(context)을 고려**하여 항목별 **개인식별 가능성**을 **상대적으로 판단**하여야 함

* (식별정보) 특정 개인과 직접적으로 연결되어 다른 사람과 구분되는 정보

** (식별가능정보) 단일 항목으로는 식별 가능성이 없으나, 가명처리 대상이 되는 다른 항목과 결합하는 경우 식별 가능성이 높아지는 정보

예시 **사람 얼굴이 촬영된 CCTV 영상을 활용하고자 하는 경우**

▶ **(상황에 따라 식별 가능성이 달라지는 경우)**

- **촬영된 장소가 명확**하고 **얼굴이 선명하게 촬영**된 경우 **식별 가능성이 높음**
- **촬영된 장소가 불분명**하고 얼굴이 상당히 **작게 찍히거나 해상도가 낮아**, 특정 개인을 구분하게 어렵게 촬영된 경우 **식별 가능성이 낮음**

▶ **(처리 목적·방법에 따라 식별성이 달라지는 경우)**

- 얼굴 촬영영상을 신체특징, 걸음걸이, 이동동선 등 **영상 내 다른 항목과 결합**하여 촬영된 사람의 **성별과 행동패턴을 분석**하기 위한 목적으로 활용된다면, **식별 가능성이 높음**
- 얼굴 촬영영상이 단순히 **사람인지 사람이 아닌지를 구분하기 위한 목적**으로만 활용되거나, **다른 항목과 결합되지 않고 활용**된다면 **비교적 식별 가능성이 낮음**

예시 환자의 **MRI 사진이 메타데이터를 포함**하고 있고, 해당 메타데이터에 환자 이름, 환자번호 등 **개인식별번호가 포함**되어 있는 경우 **식별 가능성이 높음**

- **(특이정보) 개인의 특이한 신체적·외형적·행태적 특징** 또는 **개인과 연관된 객체·사물 등의 특이성**으로 인한 식별 위험성에 대해 검토

예시 **특이한 신체적·외형적 특징**으로 인해 특정 개인을 식별할 가능성이 존재하는 경우

- **(이미지·영상)** 신체적 특징, 체형, 머리스타일, 문신, 흉터 등이 특이한 경우
- **(음성)** 발음(구개 파열음 등), 음색(목소리) 등이 특이한 경우

예시 **특이한 행태적 특징**으로 인해 특정 개인을 식별할 가능성이 존재하는 경우

- **(이미지·영상)** 걸음걸이, 몸짓이나 행위 등이 특이한 경우
- **(음성)** 억양(사투리, 말투), 반복 어휘, 언어적 습관 등이 특이한 경우
- **(텍스트)** 반복 어휘, 어법, 문체, 언어적 습관 등이 특이한 경우

예시 개인과 연관된 **특이한 객체·사물 등**으로 인해, 개인 식별 가능성이 존재하는 경우

- **(이미지·영상)** 거주하는 집, 차종(희귀 슈퍼카 등), 옷차림, 반려동물 등이 특이한 경우

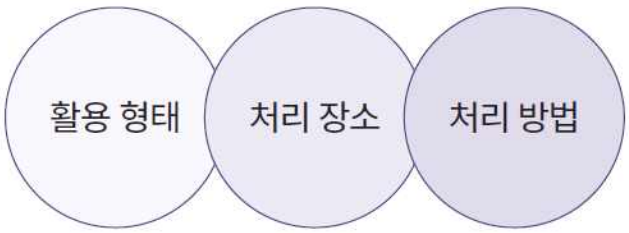
- **(재식별 시 영향도)** 가명처리된 **비정형데이터가 재식별**될 경우 특정 정보주체에게 사회적 파장 등 **영향도가 높은 항목**이 있는지 검토

> **예시** 범죄 **피해현장이 촬영**된 CCTV 영상, **민감한 사생활이 포함**된 녹취파일 등

(2) 처리 환경의 식별 위험성 검토

○ **❶가명정보 활용 형태, ❷가명정보 처리 장소, ❸가명정보 처리방법** 등 가명정보 처리 환경에 따라 발생할 수 있는 식별 위험성을 검토해야 함

< 처리 환경에 따른 개인식별 위험성 요소 >



- **(활용 형태)** 처리자(또는 취급자)가 **보유하고 있는 정보** 또는 **접근·입수 가능한 정보**와 **이용 범위·유형을 고려**하여 식별가능한 항목이 있는지 검토

> **예시** 의사가 분석하고자 하는 **환자 X-ray 사진 데이터셋**에 **본인이 직접 진료한 환자의 데이터가 포함**된 경우, **해당 의사가 보유**하고 있는 **진료데이터**와 **진료경험**을 토대로 특정 환자를 추정할 수 있는 **위험성이 높아짐**
>
> **예시** CT 영상에 포함된 **메타데이터를 삭제**할 경우, 데이터셋에 포함된 **특정 환자에 대한 배경지식을 알고 있는 자**가 해당 환자를 추정할 수 있는 **위험성이 낮아짐**
>
> **예시** AI 특성에 따라 텍스트정보를 **학습하여 그대로 암기**하고, **개인식별 위험이 있는 정보**들을 **그대로 생성·출력**하는 형태로 구현될 경우, **추론 공격** 등에 노출될 **위험성이 높아짐**
>
> **예시** AI가 CCTV에 촬영된 **사람의 형상만 파악·집계**하여 **통계값만 추출**하는 경우, **개개인에 대한 특징을 분석하지 않기** 때문에 특정 개인이 식별될 **위험성이 낮아짐**

- **(처리 장소)** 가명정보가 해당 가명정보 외에 **다른 정보의 접근·입수, 재식별 기술의 접근**이 **제한된 환경·장소에서 처리**되는지 검토

> **예시** 가명처리(모자이크)된 사진을 **복원시킬 수 있는 기술이 존재**하는 경우, 해당 사진에 대해 **복원기술을 적용할 수 있는 환경에서 처리**된다면 **식별 위험성이 높아짐**
>
> **예시** 개인식별이 어렵도록 **음성변조된 음성정보를 처리**할 시, 해당 음성정보와 비교·대조할 수 없는 **다른 정보를 입수할 수 없는 환경**에서 처리된다면 **식별 위험성이 낮아짐**

- **(처리 방법)** 가명정보가 **다른 정보와 연계·결합**되거나 **반복 제공** 등이 예정된 경우, 식별가능성이 높아지는 항목이 있는지 검토

예시 각 환자를 다양하게 촬영한 **100장의 두경부 CT사진** 활용 시, **영상 재건 기술 등을** 활용하여 **얼굴 형상을 3차원으로 복원**해낼 수 있는 **위험성이 존재**

예시 **환자당 1장의 CT사진만 활용**하여 연구를 수행하는 경우, **영상 재건 기술을 활용**하여 **얼굴 형상을 복원해낼 수 없으므로 개인식별 위험성이 낮음**

3 가명처리 단계

※ 위험성 검토 결과 및 항목별 가명처리 계획을 기반으로 실제 가명처리를 수행하는 단계

○ **가명처리가 필수적이지 않은 항목**과 **가명처리가 필요한 항목을 구분**하고, 가명처리가 필요한 항목은 **합리적인 가명처리 방법과 수준을 결정**

➜ 비정형데이터 항목 중 **처리목적 달성을 위해 반드시 필요**하지만, **개인식별 위험성이 낮은 정보**는 **가명처리하지 않고 그대로 사용** 가능

➜ 비정형데이터 항목 중 **가명처리해도 처리목적 달성이 가능**하고, **개인식별 위험성이 높은 정보**는 **가명처리하여 활용**하여야 함

예시 **CT 사진을 활용하여 얼굴뼈 골절 진단을 위한 AI 연구개발 시**

▶ **얼굴 앞부분(안면부)**
: **얼굴뼈 골절 진단에 반드시 필요하므로, 그대로 활용**

▶ **뇌 뒷부분(후두부)**
: **얼굴뼈 골절 진단에 필요하지 않으므로, 마스킹처리하여 활용**

예시 **구강사진을 활용하여 충치를 분석·진단하는 AI 개발 시**

▶ **충치 의심영역으로 라벨링된 치아부분**
: **충치 분석·진단에 반드시 필요하므로, 그대로 활용**

▶ **충치 의심영역 외 일반치아 및 잇몸 부분**
: **충치 분석·진단에 필요하지 않고, 구강 내 특이구조나 치료흔적 등을 통한 개인식별 위험성이 존재하므로, 블러링 처리하여 활용**

○ 비정형데이터 가명처리 기술 적용시, 해당 **기술의 적절성·신뢰성을 평가**하고 **관련 근거***를 **작성하여 보관**할 것을 권고

* 예) CT사진의 가장자리를 마스킹 솔루션을 적용하여 가명처리한 경우, 해당 솔루션의 관련 가명처리 기능, 솔루션의 객체 인식률·처리 정확도(오류율)에 대한 증빙자료 등

○ 비정형데이터 가명처리의 **기술적 한계를 보완**하고 잔존 위험(Residual Risk)을 **낮추기** 위해, 처리결과에 대한 **자체적인 추가검수** 등이 필요

- **가명처리 목적, 데이터의 성격, 적용한 기술의 특성, 처리환경 통제수준** 등을 고려하여 **위험도에 비례한 적절한 검수방법**을 적용하되, **검수 과정**에서 **발견된 위험**을 낮추기 위한 **조치사항을 기록·보관**하고 이에 대한 **적정성 검토*를 받을 것을 권고**

* 예) 사람에 의한 육안 전수검사, 통계적으로 신뢰할 수 있는 샘플링 검사 등을 통해 합리적으로 예측가능한 위험을 낮추기 위해 충분한 노력이 시행되었는지 검토

4 적정성 검토 단계

※ 외부전문가를 포함한 적정성 평가 위원회 등을 구성하여 처리 목적의 적합성, 위험성 검토 결과의 적정성, 가명처리 결과의 적정성, 목적 달성 가능성 등을 검토하는 단계

○ **비정형데이터의 특성**과 **처리 목적·환경 등을 고려**하여 **합리적인 방법·수준**으로 **가명처리를 수행하였는지 검토**

○ 비정형데이터 가명처리에 활용한 **기술의 적절성·신뢰성을 검토**하고, 해당 기술의 한계 등으로 인한 **잔존 위험(Residual Risk)을 충분히 낮추기** 위한 **추가검수**를 하였는지 검토

○ 비정형데이터는 가명처리 시 **데이터의 특성**과 관련 **기술 발전 수준, 재식별 위험** 등을 **종합 고려**하여야 하고, 이를 위한 **전문성이 필요**하므로,

- 적정성 검토 시 **외부전문가를 과반수 이상으로 구성**하여 **객관성·전문성 있는 검증을 받을 것을 권고**

5 안전한 관리 단계

※ 적정성 검토 이후 가명정보 활용 과정에서 재식별 가능성 등을 모니터링·관리하는 단계

○ 가명정보 특례를 활용하여 **AI 학습·연구개발**을 하고자 하는 자는 **AI 기술·서비스의 특성을 고려**하여 **사전·사후적으로 발생**할 수 있는 다양한 **위험을 낮추기 위한 충분한 조치**를 취해야 함

○ 다만, **AI 맥락에서 나타날 수 있는 다양한 위험**을 **사전에 완벽히 제거**하는 것은 현재 기술상 불가능하므로,

- **잔존 위험(Residual Risk)을 최소화시키기 위한 노력 정도**에 따라 **사후관리의 이행 수준을 판단**하여야 함

○ **AI 서비스 운영 과정**에서도 **개인식별 위험성** 및 **프라이버시 침해 가능성**이 높아지지 않는지 **지속적으로 모니터링**하여야 함

- 또한, 개인식별 위험 증가 및 프라이버시 침해 등 관련 **문제 발견 즉시**, 해당 **가명정보 처리를 중단**하고 **관련 위험을 제거**하여야 함

5 비정형데이터 개인식별 위험성 검토 체크리스트

※ 해당 체크리스트는 가명처리 계획을 세우기 전에 개인식별 위험성을 검토하기 위한 것으로, 검토 결과가 "예"에 해당하더라도, 해당 위험을 낮추기 위한 적절한 가명처리 방안 적용 후 활용 가능

구분		개인식별 위험성 검토 사항	
데이터	식별성	개인 식별이 가능한 항목 여부	
		검토 항목	검토 결과
		① 그 자체로 특정 개인을 식별할 가능성이 높은 정보가 있는가 * (예시) △ 얼굴 전체가 선명히 보이는 경우(안면, 옆면, 성형수술 전·후 얼굴, 거울·유리 등에 비치거나 반사된 사람의 얼굴 등) △ 명찰(이름표) △ 이미지나 또는 영상에 포함된 메타데이터 상에 사람 이름이나 환자 번호 등 그 자체로 개인의 식별성이 높은 정보들이 그대로 포함된 경우 등	□ 예 □ 아니오
		② 두 개 이상의 정보를 조합하여 식별가능성이 높아지는 정보가 있는가 * (예시) △ 휴대폰 사진에 일시, 장소, 촬영자 정보 등이 포함된 경우 △ 흉부 CT 영상 이미지에 환자정보를 나타내는 정보가 포함되어 있거나 환자의 특이사항에 대한 의사의 소견이 텍스트 등으로 적혀있는 경우	□ 예 □ 아니오
	특이정보	개인의 특이한 특징으로 인하여 식별 가능성이 있는 이용 항목 여부	
		검토 항목	검토 결과
		③ 개인의 특이한 신체적·외형적 특징으로 인해 특정 개인을 식별할 가능성이 있는가 * (예시: **이미지·영상정보**) 신체적 특징, 체형, 머리스타일, (특정 위치의) 문신, 흉터 등이 특이한 경우 * (예시: **음성정보**) 발음(구개 파열음 등), 음색(목소리) 등이 특이한 경우	□ 예 □ 아니오
		④ 개인의 특이한 행태적 특징으로 인해 특정 개인을 식별할 가능성이 있는가 * (예시: **이미지·영상정보**) 걸음걸이, 몸짓이나 행위 등이 특이한 경우 * (예시: **음성정보**) 억양(사투리, 말투), 반복 어휘, 언어적 습관 등이 특이한 경우 * (예시: **텍스트정보**) 반복 어휘, 어법, 문체, 언어적 습관 등이 특이한 경우	□ 예 □ 아니오
		검토 항목	검토 결과
		⑤ 개인과 연관된 객체·사물 등의 특이성으로 인하여 식별가능성이 있는 이용 항목 여부	□ 예 □ 아니오

<table>
<tr><td rowspan="12">처리 환경</td><td></td><td>* (예시: 이미지·영상정보) 거주하는 집, 차종(희귀 슈퍼카 등), 옷차림, 반려동물 등이 특이한 경우</td><td></td></tr>
<tr><td rowspan="3">재식별 시 영향도</td><td colspan="2">재식별 시 정보주체에게 심각한 피해 또는 불이익을 초래할 수 있는 이용 항목 여부</td></tr>
<tr><td>검토 항목</td><td>검토 결과</td></tr>
<tr><td>⑥ 사회통념상 차별 등으로 인해 정보주체가 피해 또는 불이익을 받을 수 있는 정보가 있는가</td><td>□ 예
□ 아니오</td></tr>
<tr><td rowspan="2">⑦ 재식별로 인하여 받는 피해 또는 불이익의 정도와 규모가 상당히 클 수 있는 정보주체에 관한 정보가 있는가
* (예시: 음성·텍스트) 민감한 사생활 또는 질병관련 내용이 포함된 병원 진료·상담 녹취파일, 상담보고서</td><td>□ 예
□ 아니오</td></tr>
<tr><td rowspan="4">이용 및 제공</td><td colspan="2">가명정보 활용 형태 및 이용 기관의 개인정보 보호 수준 등을 고려하여 식별가능성이 있는 항목 여부</td></tr>
<tr><td>검토 항목</td><td>검토 결과</td></tr>
<tr><td>⑧ 처리주체가 보유하고 있는 정보 또는 접근·입수 가능한 정보와 이용 범위 및 유형을 고려하여 식별가능한 항목이 있는가
* (예시: 이미지·영상정보) △ 의사가 분석하고자 하는 환자 X-ray 사진 데이터셋에 본인이 직접 진료했던 환자의 데이터가 포함된 경우, 해당 의사가 보유한 진료데이터·경험을 통해 특정 환자 추정 가능 △ CT 영상에 환자에 대한 메타데이터가 포함되어, 환자에 대한 배경지식을 알고 있는 자가 해당 환자 추정이 가능한 경우 등</td><td>□ 예
□ 아니오</td></tr>
<tr><td>⑨ 추가정보를 삭제하지 않고 보관하는가</td><td>□ 예
□ 아니오</td></tr>
<tr><td rowspan="4">처리 장소</td><td>⑩ 가명정보 제공 시 제공받는 자의 개인정보 보호 수준 및 신뢰할 수 있는 인증을 받았는가 (ISMS, ISMS-P, ISO 27001 등)</td><td>□ 예
□ 아니오</td></tr>
<tr><td colspan="2">가명정보가 관리적·기술적·물리적으로 안전한 장소에서 처리되는지 여부</td></tr>
<tr><td>검토 항목</td><td>검토 결과</td></tr>
<tr><td>⑪ 가명정보 처리 시 다른 정보를 접근·입수할 수 있는 장소인가
* (예시) △ 누구나 접근 가능한 개방형 형태의 장소 및 네트워크인지 △ 허가된 인원만 출입할 수 있는 장소 및 네트워크인지 △ 다른 정보, 데이터 복원 기술 등에 대한 제한이 가능한 환경인지 등</td><td>□ 예
□ 아니오</td></tr>
<tr><td rowspan="3">다른 정보와 결합</td><td colspan="2">가명정보를 다른 정보와의 결합하여 활용 시 식별가능성이 있는 항목 여부</td></tr>
<tr><td>검토 항목</td><td>검토 결과</td></tr>
<tr><td>⑫ 다른 정보와의 연계 분석이 예정되어 있는가</td><td>□ 예
□ 아니오</td></tr>
<tr><td></td><td>⑬ 처리주체가 보유하거나 접근·입수 가능한 정보 등 다른 정보와 연계 또는 결합하여 식별가능한 항목이 있는가</td><td>□ 예
□ 아니오</td></tr>
</table>

6 개인식별 위험성 검토 항목별 조치 가이드

구분		조치 가이드
데이터	식별성	o **(이미지·영상정보)** - 지문, 홍채 등 개인을 직접적으로 알아볼 수 있는 생체인식정보 등이 고해상도로 촬영된 경우 원칙적으로 삭제하여야 함 - 사람얼굴, 신체부위, 차량번호판 등의 개인식별가능정보가 처리 목적 상 반드시 필요한 경우 안전한 방식으로 가명처리하여 활용 가능 * 가명처리 방법으로는 이미지 필터링(블러링, 픽셀화(모자이크), 마스킹(블랙박스), 이미지 암호화, 얼굴 합성(k-same 모델 등), 인페인팅 기법 등 활용 가능 - 특히, 블러링이나 픽셀화 기술은 가명처리 수준에 따라 기술적으로 원본 복원이 가능하므로, 추가정보(이미지 필터링 알고리즘, 이미지 필터링 강도 등)는 삭제하는 것이 원칙 - 체형, 옷차림, 머리스타일, 문신 등이 특이한 경우 개인을 알아볼 수 있는 가능성이 높으므로, 이용목적 상 반드시 필요하지 않다면 삭제하고 필요 시 이미지 필터링 등 가명처리 필요 o **(음성정보)** - 발화자나 혹은 대화 상대자 등의 음성 내용 중 개인을 식별할 수 있는 정보(이름, 주소, 전화번호 등)나 대화 맥락상 개인이 식별될 수 있는 정보 등은 삭제하거나 대체할 수 있는 정보를 생성하여 대체하여야 함 ***** 특히 음성 내용 중 민감한 담화주제(예. 성적취향, 종교, 질병)가 포함된 경우 개인이 식별되지 않도록 각별한 주의 필요 - 음성정보는 음성내용 외에도 발화자의 음성 자체를 통해서도 개인식별이 가능하므로, 규칙 기반 개인정보 단순 삭제나 혹은 대체, 음성변형 원리에 기반한 음성 변형(transformation) 및 변환(conversion), STT(Speech To Text) 등을 통해 가명처리 필요 - 특히, 음성을 변조한 경우, 변조한 규칙을 알 경우 이를 역으로 실행하면 음성복원이 가능하므로 추가정보(음성변조 알고리즘 등)는 삭제하는 것이 원칙 o **(텍스트정보)** - 자유 형식의 텍스트, STT, 영상의 자막 등 텍스트 내용에 포함된 개인을 식별할 수 있는 정보(이름, 주소, 전화번호 등)는 삭제하거나 대체할 수 있는 정보를 생성하여 대체하여야 함 * 가명처리 방법으로는 규칙 기반 개인정보 단순 삭제나 혹은 대체, 스크러빙(scrubbing), 정규표현식, 주석달기(annotation) 등 활용 가능

	특이정보	o 특이정보는 그 정보만으로 개인을 식별할 수 있는 정보는 아니더라도 고유(희소)한 특성 때문에 개인을 알아볼 수 있는 가능성이 높으므로, 이용목적 상 반드시 필요하지 않다면 삭제하고 처리 목적 달성에 반드시 필요한 경우 안전하게 가명처리하여 활용
	재식별 시 영향도	o 사회통념상 차별, 기본권 침해 등 파급영향이 클 수 있는 정보는 재식별 시 다른 일반정보와 다르게 개인의 피해와 더불어 사회적 파장이 있을 수 있으므로, 꼭 필요한 항목 이외에는 삭제 등 조치
처리 환경	**이용 및 제공**	o 이용 및 제공의 위험성이 있는 경우 이용자와 제공자가 서로 위험성을 낮추기 위한 처리 환경에 대한 안전성 입증 관련 협의가 필요 - 예를 들어, 제공자 입장에서 내부 이용에 비하여 외부 제3자 제공의 경우가 처리환경으로 인한 식별 위험성이 높아 데이터의 가명처리 수준이 높게 책정될 수 있음. 이 경우, 가명정보 이용자가 해당 가명처리 수준으로 처리 목적 달성이 어렵다고 판단한다면, 제공자와의 협의를 통하여 처리 환경에 대한 통제 강화를 전제로 데이터의 가명처리 수준을 낮출 수 있는지 협의 가능 - 가명처리한 이후에도 계속 원본정보를 보유할 경우, 원본정보와의 대조 및 연결을 통하여 식별 위험이 높아질 수 있으며, 원본정보 내에 개인을 식별할 수 있는 메타데이터가 포함되어 있거나 그 밖에 원본정보에 대한 배경지식 등으로 인하여 식별 위험이 높아질 수 있어 각별한 주의 필요
	처리 장소	o 데이터 자체의 가명처리 수준을 낮춰서 활용해야 할 경우, 물리적·관리적·기술적 조치 등 처리 장소의 안전성을 강화하여 종합적인 개인식별 위험성을 낮출 수 있음 - 가명처리 기술의 적절성·신뢰성을 담보하기 어렵거나 안전한 환경적 통제방안을 갖추기 어려운 경우, 데이터 자체에 대한 통제 강화를 통해 종합적인 식별위험성을 낮출 필요 o 데이터 복원기술 등에 취약한 경우, 다른 정보 접근 및 복구 기술(SW)에 대한 접근·사용을 제한할 수 있는 환경을 갖출 필요가 있음
	다른 정보와의 결합	o 다른 정보와 연계·결합 예정에 있는 경우 연계·결합되는 정보와 결합하여 식별가능성이 높아지는 항목이 있는지 추가 검토 필요 o 처리주체가 보유하거나 접근·입수 가능한 정보를 통해 식별가능한 항목 있는지 검토 - 가명정보를 제공받아 활용하게될 자가 가진 과거 유사 정보에 대한 수행 경험이나 지식 등은 가명정보를 제공하려는 자가 자체적으로 판단하기 어렵기 때문에 가명정보를 제공받아 활용하게 될 자에게 사전에 확인 및 검토 필요. 사전 검토가 어려운 경우, 가명처리의 수준을 높이는 방법 등으로 위험성을 낮춰야 함

참 고 비정형데이터 가명처리 기술 및 예시

※ 아래 가명·익명처리 기술은 비정형데이터의 가명처리 시 처리자의 단순참고를 위해 국내·외 자료와 전문가 논의를 통해 구성된 예시자료로, 해당 가이드라인에서 제시되지 않은 기술도 처리자의 판단에 따라 데이터 활용 분야상황에 맞게 자유롭게 활용할 수 있음

1. 영상정보

① 이미지 필터링 기술

○ 필터링 기술(❶ 블러링, ❷ 픽셀화, ❸ 마스킹)로 처리된 영상 이미지

* 얼굴 이외에도 사람 형상, 신체, 옷차림 등, 사람과 관련된 사물/동물 등도 필터링 대상이 될 수 있음

※ 출처 : Tao Li and Lei Lin, Natural Face De-identification with Measurable Privacy(CV-COPS, 2019)

○ 필터링 기술 : 블러링(Blurring), 픽셀화(Pixelization), 마스킹(Masking)

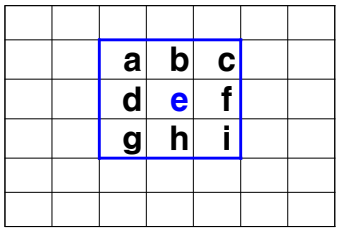
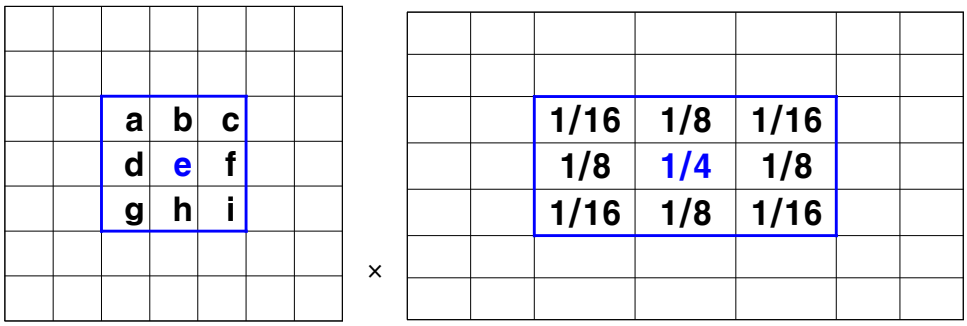
❶ 이미지 블러링(Blurring)

◦ 필터의 격자 사이즈 : 3*3 또는 5*5 등

◦ 필터링의 종류 : 평균, 가우시안, 중간값, 바이레터럴(Bilateral)등

< 필터링의 종류 >

종류	설 명
평균	- 입력 이미지의 현재 위치에서 예를 들어 3×3 격자 범위의 주변 픽셀값의 평균을 구하여 원 픽셀값을 결과 이미지의 픽셀값으로 대체 **=> e = 1/9(a+b+c+d+e+f+g+h+i))**

<table>
<tr><td></td><td> 
</td></tr>
<tr><td>가우시안</td><td>- 입력 이미지의 현재 위치에서 예를 들어 3×3 격자 범위의 주변 픽셀값에 가중치를 부여(중심으로 갈수록 높은 가중치를 부여)하여 원 픽셀값을 결과 이미지의 픽셀값으로 대체
※ 입력값으로 커널크기, 가우시안 표준편차, 필터의 데이터 타입이 사용됨
=> e = 1/16*a+1/8*b+1/16*c+1/8*d+1/4*e+1/8*f+1/16*g+1/8*h+1/16*i
 

원본값 필터</td></tr>
<tr><td>중앙값</td><td>- 입력 이미지의 현재 위치에서 예를 들어 3×3 범위의 주변 픽셀값의 중앙값을 구하여 원 픽셀값을 결과 이미지의 픽셀값으로 대체</td></tr>
<tr><td>바이레터럴
(bilateral)</td><td>- 원본 이미지로부터 최대한 노이즈는 제거하고 에지는 보존하기 위한 것으로 공간과 밀도를 함께 고려하여 원 픽셀값을 결과 이미지의 픽셀값으로 대체</td></tr>
</table>

○ 추가정보 : 원본 정보, 필터 사이즈, 가우시안 표준편차, 필터링의 종류 등

※ 블러링 처리된 이미지를 원본 이미지로 직접 복원은 불가하나 최근에는 자연 영상의 통계적 특성(natural image statistics)에 기반한 모션 블러 추정 및 제거 등 디블러링 방법들이 제안되고 있음

❷ 이미지 픽셀화(Pixelization) (=모자이크(Mosaic))

○ 이미지 블러링의 평균 필터와 유사하나 계산한 평균값을 해당 픽셀뿐만 아니라 적용한 주변 모든 픽셀(예를 들어 3×3 범위의)에 대체한다는 점이 다름

○ 추가정보 : 원본 정보, 필터 사이즈 등

< 블러링과 픽셀화 처리 시 참고 사항 >

- 표준 QCIF(Quarter Common Intermediate Format) 화상회의 이미지 크기(176×144 픽셀, 픽셀당 24비트)에서 초당 24프레임의 프레임 속도로 변환하고 PC에서 재생하기에 적합한 AVI 파일로 저장한 5개 이미지 스틸컷에 대하여 설문을 통한 실험 결과, 식별가능 임계값을

블러링시 최소 레벨 5이상을 픽셀화시 최소 레벨 6이상을 임계값으로 제시

※ 출처 : Michael Boyle 외 2인, the effects of filtered video on awareness and privacy(2000)

- 또한 현재 인터넷 상에 이미지 블러링 및 픽셀화 처리가 가능한 OpenCV 등 라이브러리나 소스코드 등이 공개되어 있어 쉽게 처리가 가능

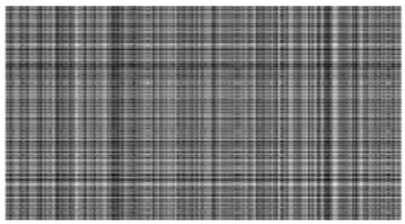
❸ 이미지 마스킹(Masking) (=블랙박스(Blackbox))

2 이미지 암호화

○ 원본 이미지의 일부를 암호화하여 데이터 주체를 알아볼 수 없도록 하는 기법

○ 이미지 암호화로 처리된 영상 이미지

원본 이미지


암호화된 이미지

복호화된 이미지

※ 출처 : Lihua Gong 외 3인, Image compression-encryption algorithms by combining hyper-chaotic system with discrete fractional random transform(Optics and Laser Technology, 2018)

○ 이미지 암호화 기술 : **이산코사인변환 기반 암호화, 픽셀 위치 기반 암호화**

❶ 이산코사인변환 기반 암호화(DCT, Discrete Cosine Transform)

◦ 영상을 주파수 영역으로 바꾸어 특정 부분만 암호화

❷ 픽셀 위치 기반 암호화(Pixel location)

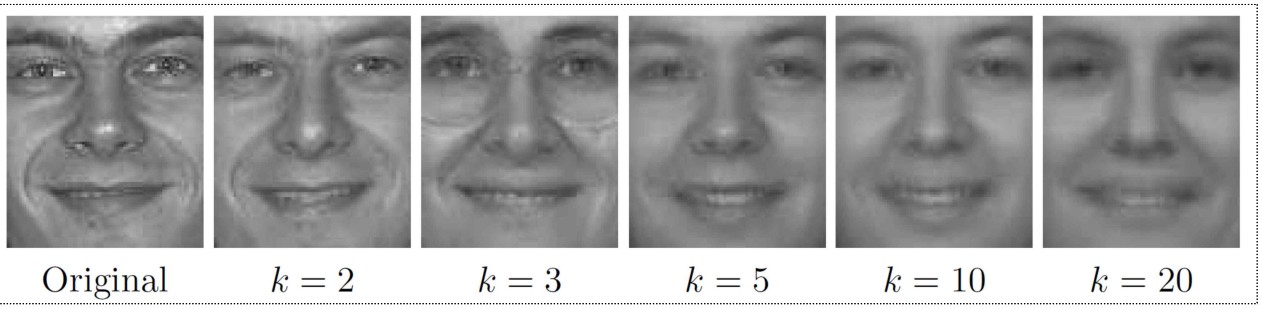
◦ 픽셀의 위치를 일정한 규칙에 따라 바꾸는 방식으로 구현하는 암호화

3 얼굴 합성(프라이버시 보존형 데이터 마이닝)

○ K-익명성 프라이버시보호 모델을 확장하여 K명의 얼굴을 합성한 기술로 **K-same 모델**로도 부름

※ 출처 : E.M.Newton 외 2인, preserving privacy by de-identifying face images(2005)

○ K-same 기법으로 처리된 영상 이미지



※ 출처 : R. Gross 외 2인, Integrating utility into face de-identification(2005)

○ K-same 모델의 개선

- K-same 모델을 보다 개선하여 개인정보 보호와 유용성의 균형*을 맞추기 위한 **K-Same-Select 모델** 등장

* 공개 위험(즉, 이미지 난독화 수준)과 분류 정확도 간의 균형

※ 출처 : R. Gross 외 2인, Integrating utility into face de-identification(2005)

- **K-same 모델의 정확도를 유지하면서** 데이터 셋의 민감한 정보를 숨기기 위해 기존 샘플에 **무작위 노이즈를 추가하거나 새 샘플을 생성하는 난독화(obfuscate) 기능을 추가**로 설계

* 안전성은 높으나 성별이 바뀌는 등 이미지 왜곡이 심함

※ 출처 : T. Zhang, Privacy-preserving machine learning through data obfuscation(2018)

4 인페인팅(Inpainting)

○ 영상 내 개인 식별 영역을 제거한 후 다른 물체 또는 배경으로 대체하여 신원을 보호하는 기술

○ 인페이팅 기법으로 처리된 영상 이미지

※ 출처 : Takahiro Ogawa 외 1인, Image inpainting based on sparse representations with a perceptual metric(2013)

○ 인페인팅 기술 : **패치 기반 인페인팅, 객체 기반 인페인팅**

❶ 패치 기반 인페인팅 기술

○ 영상 프레임 내 공백과 가장 비슷한 영역을 찾아 채우는 방식

❷ 객체 기반 인페인팅 기술

○ 영상을 배경과 객체로 구분해 객체를 제거 후 남은 부분은 배경으로 채우는 방식

5 AI 이용 영상정보 가명처리

○ **얼굴 보존형 가명·익명처리 기술**(De-identification without losing faces)

- 원본 얼굴의 요소를 변경하거나 얼굴을 완전히 합성하는 대신, 훈련된 얼굴 속성 전달 모델을 사용하여 동의한 대상의 소수(보통 2~3명)인 기증자의 얼굴에 비 신원 관련 얼굴 속성을 매핑

원본 이미지 K-same 이미지 제안 이미지

※ 출처 : Yuezun Li, De-identification Without Losing Faces(2019)

○ **AnonymousNet 프레임워크**

- 얼굴 보존형 가명·익명처리 기술의 프레임과 비슷해 보이지만 다음의 4단계 절차를 통하여 아이덴티티가 완전히 다른 이미지를 생성

※ (1단계) 얼굴 특징 추출, (2단계) 의미 기반 속성 난독화, (3단계) 익명화된 얼굴 생성, (4단계) 적대적 교란 절차

원본 이미지 AnonymousNet

※ 출처 : Tao Li 외 1인, AnonymousNet: Natural Face De-Identification with Measurable Privacy(CV-COPS 2019)

○ AI 딥러닝 기반의 알고리즘을 활용

- 얼굴 및 차량 번호판 등을 추출한 후 각종 SW 라이브러리를 활용하여 블러링 처리

※ 미검출이나 오검출 얼굴 등에 대해서는 수작업을 통하여 추가 보정

○ AI 적대적 생성 신경망 기반 모델(GAN, Generative Adversarial Networks)

- 이미지에서 보존해야할 영역을 수치화하여 연속형 데이터로 처리한 후 해당 개별 데이터에 노이즈를 추가하거나 재현 처리

2. 음성정보

① 음성정보 자체에 대한 가명처리

■ 대부분 텍스트로 변환하여 처리하고 있으며 그 외 발화자의 개인정보(비속어, 성적, 정치적 발언 등 포함)에 대한 규칙을 정하여 해당 구간을 단순히 삭제하거나 혹은 음성 변형의 원리에 기반하여 음조를 변형하여 처리

○ 규칙기반 개인정보 단순 삭제

- 음성 데이터 상에서 규칙*을 정하여 개인식별가능 정보를 단순히 삭제

* (예) 인명, 지명 등 개인식별가능 정보가 포함된 카드번호, 주민등록번호, 전화번호 등 삭제하고, 비속어, 성적및 정치적 발언의 경우도 삭제하거나 다른 언어로 대체하는 등

※ 이 방법은 완전 자동으로 수행이 어려운 측면이 있으므로 추가적인 검수가 필요할 수 있음

- 음성 변형의 원리에 기반하여 음성 데이터 내용에 영향을 주지 않으면서 주어진 발화의 비언어적 특징들을 수정

❶ 음성 변형(VT, Voice Transformation)

○ 원본 정보 소스, 필터, 소스와 필터의 조합 등을 통하여 음성을 변형하는 기법

※ 소스로는 시간량(time-sclae), 음조(pitch) 또는 음량(energy)을 변조할 수 있으며, 필터로는 음성 트랙 시스템에 기반하여 음폭을 변형 가능

종류	설 명
시간량 (time-scale) 수정	- 원래 음성의 지각적(perceptual) 품질에 영향을 주지 않으면서 조음 속도를 변경
음조 (pitch, 음높이, 고저) 수정	- 짧은 시간 스펙트럼 엔벨로프(포먼트의 위치 및 대역폭)와 시간을 보존하면서 스펙트럼의 하모닉 구성 요소 사이의 간격을 압축하거나 확장하기 위해 기본 주파수를 변경

음량(energy) 변형	- 입력 음성의 인지된 크기를 수정
필터 수정	- 성도(vocal tract) 시스템의 주파수 응답의 크기 스펙트럼 수정(예 : 여성의 목소리를 수정하여 아이처럼 들리게 함)
소스 수정과 필터 수정의 조합	- 일반적으로 사용되는 기법으로 예를 들어 화자의 음성을 다른 화자의 음성처럼 들리도록 수정하려면 운율(prosody)과 성도(vocal tract) 수정을 결합해야 함 → 이때 특정 대상 화자가 제공되는 경우 이를 음성 변환(voice conversion)이라 부름

❷ 음성 변환(VC, Voice Conversion)

○ 음성 변형(VT)의 특별한 형태로 발화자의 음성 특성을 특정(대상) 화자의 음성 특성으로 매핑하는 것으로 원래의 목소리를 다른 사람의 목소리로 변환하는 기법

○ 음성 변환(VC)은 '텍스트 종속적'이거나 '텍스트 독립적'일 수 있음

종류	설 명
텍스트 종속적	- AI 학습 단계에서 병렬 말뭉치(동일한 텍스트를 발화하는 소스 및 대상 화자의 훈련 자료)가 필요 ※ 실제 응용 프로그램에서 음성 비식별화에 이러한 방식을 사용하는 것은 주요 제약 사항임
텍스트 독립적	- 학습 단계에서 병렬 말뭉치가 필요하지 않으며 발화자의 개인정보 보호 차원에서 보다 효과적이고 현실적임

❸ GMM(Gaussian Mixture Model) 매핑기반의 음성 변환 방식

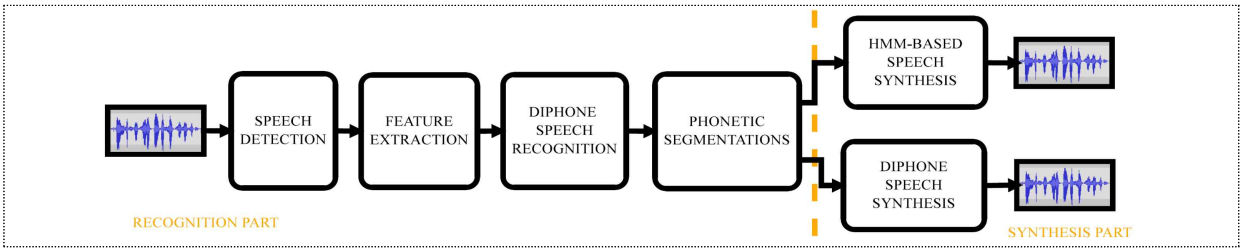
○ 영상정보에서의 K-same 방식과 유사한 텍스트 독립적인 GMM 매핑기반의 음성 변환 기법

○ 화자 인식 시스템이 비식별화된 음성에서 우연보다 더 나은 성능으로 실제 화자의 신원을 인식할 수 없는 반면, 비식별화된 음성은 대부분의 경우 이해할 수 있음

○ 모든 음성이 동일한 대상 화자에 매핑되기 때문에 서로 다른 화자를 구별할 수 있는 (비식별화된) 음성을 생성하는 기능은 부족

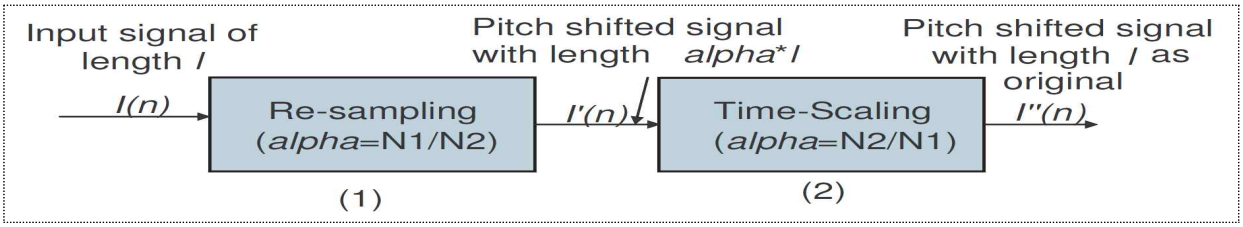
❹ HMM(Hidden Markov Model) 기반과 TD-PSOLA(diphone Time-Domain Pitch Synchronous Overlap and Add) 기술을 기반으로 한 음성 변형 방식(DROPSY)

○ HMM과 TD-PSOLA 기술 기반 DROPSY 알고리즘



※ 출처 : T. Justin 외 5인, Speaker de- identification using diphone recognition and speech synthesis(2015)

○ PitchScale SOLA(Synchronous Overlap and Add) 알고리즘(2단계)



※ 출처 : Y. Stylianou, Voice transformation : a survey(2009)

② 음성을 텍스트로 변환(STT, Speech To Text) 후 가명처리

○ 개인식별정보가 포함된 음성을 텍스트로 변환 후 변환한 텍스트에서 개인식별정보를 가명처리하고 다시 음성으로 변환하는 방식

3. 텍스트정보

① 규칙기반 개인정보 단순 삭제 혹은 마스킹

○ 사전에 텍스트 내 개인식별(가능)정보들을 정의하고 정의된 형태(포맷)에 기반하여 해당 정보를 삭제하거나 마스킹, 대체 처리 등의 방법으로 제거

② 스크러빙(Scrubbing)

○ 원 텍스트의 내용과 구조를 보존하면서 즉석해서 파싱을 통하여 혹은 파싱 이후 개인식별(가능)정보만을 제거(마스킹 혹은 대체)하는 것으로 이 경우 다수의 정보 주체와 해당 속성들 사이의 명확한 연관성이 없어질 수도 있음

- 즉, 특정 정보 주체가 어느 속성을 지칭하는지 알 수가 없게 될 수도 있음
- 단순 삭제 혹은 마스킹 방법과 유사하며 수작업이 아닌 자동화 SW를 이용한다는 점이 다름

3 정규표현식(Regular Expression)

○ 문자나 혹은 문자열의 일정한 패턴을 표현하는 일종의 형식 언어

	항목	텍스트 예시	정규표현식
1	전화번호	010-1111-2222	^d{3}₩-d{4}₩-d{4}$
2	이메일주소	aa@bb.net	^[0-9a-zA-Z]([-_.]?[0-9a-zA-Z])*@[0-9a-zA-Z]([-_.]?[0-9a-zA-Z])*.[a-zA-Z]{2,3}$/i;
3	IP주소	111.111.111.111	^₩d{1,3}₩.₩d{1,3}₩.₩d{1,3}₩.₩d{1,3}$
4	주민등록번호	220101-1111111	^₩d{6}₩-[1-4]₩d{6}$

※ 사용시 오타나 혹은 문자 숫자 혼합 등에 따라 누락되는 경우 존재하므로 추가검수가 필요할 수 있음

4 주석 달기(Annotation)

○ 주어진 텍스트를 논리적으로 분할한 후 분할된 단어(들)에 주석을 첨가하는 기법

종류	설 명
규칙(Rule) 기반	- 구문 문법의 규칙에 따라 텍스트를 토큰(예 : 사전 정의된 수의 단어)으로 나누는 것으로 고급 규칙의 경우 정규표현식을 사용하여 정의
사전(Dictionary) 기반	- 개인식별(가능)정보들을 미리 사전으로 정의한 후 개체명 인식(NER, Named Entity Recognition) 기술을 이용하여 주어진 텍스트(이름, 주소, 전화번호 등)와 일치시킴 * 사전기반의 경우 AI 기술을 활용하여 각 단어들을 자동으로 인식하지만 모든 단어(들)을 완벽하게 인식하는 것은 어려우며, 따라서 맥락(Context)에 의한 개인식별 가능성에 대한 조치는 어려울 수 있음

5 AI 기반 텍스트정보 가명처리

○ 규칙, 정규표현식 등을 통한 개인정보 검출 및 마스킹은 정확도 측면에서 한계가 있을 수 있으며, 이를 보완하기 위해 딥러닝 기술 등을 적용한 자연어 처리 언어 모델을 통해 사전에 정의되지 않은 패턴의 개인정보를 검출하고 마스킹할 수 있음

○ 학습방법에 따라 다양한 형태의 인공지능 기반 개인정보 검출 기법 존재 (HMM, MEM, CRFs, structural SVM, Deep-Learning)

○ 규칙기반의 유연성 부족을 해결하기 위해 패턴이나 규칙을 수동이나 반자동으로 작성하고 인공지능을 통해 사전(dictionary)을 확장하여 규칙에 적용되는 개인식별가능정보를 새롭게 검출할 수 있음(단계별 규칙을 순차적으로 적용하여 가중치에 따라 인명, 지명, 조직명 등으로 범주를 결정하는 등)

6 텍스트를 테이블 형식으로 변환

○ 주어진 텍스트를 구문 문법의 규칙에 따라 파싱(parsing, 정해진 규칙에 따라 문장의 구문을 분할)한 다음 분할된 각 세그먼트들을 열과 행이 있는 테이블 형태로 정렬한 후 나머지 데이터들은 삭제

- 변환된 테이블은 기존 정형 데이터에 대한 가명처리를 적용
- 실제로는 텍스트 자체의 복잡성으로 인하여 구조화된 테이블로의 변환이 불가능할 수도 있음

부 록 비정형데이터 가명처리 시나리오 예시

※ 아래 가명처리 시나리오는 **실제 비정형데이터를 가명처리하여 활용했던 사례**를 관련 기업·기관 및 전문가 논의를 통해 **재구성**한 것으로 **단순 참고용**이며, 처리자 및 적정성 검토위원회 등의 판단에 따라 **데이터 활용 분야·상황에 맞게 가명처리 방법·수준 등을 자유롭게 적용할 수 있음**

시나리오 ❶ : 유방암 · 골밀도 감소 여부 진단 AI 개발 사례

의료 분야 (이미지, 영상, 텍스트)

- 한국대학교병원은 병원 내부 연구자에게 과학적 연구 수행을 위한 데이터 제공을 요청받아 유방암 환자의 이미지・영상 데이터(병리조직, 흉부CT, 골밀도검사(DEXA) 기록 등)과 임상 데이터(외과병리 기록지(텍스트), 정형데이터)를 제공하려 한다.

☑ 데이터의 이용 목적

○ 유방암 진단 자동화 및 골밀도 감소 여부 확인을 위한 AI 개발 연구

☑ 데이터 특징

○ (이미지·영상 데이터) 암 진단 및 치료 관련 병리조직이미지, 흉부CT, 골밀도 검사(DEXA) 기록

○ (임상 데이터) 암 진단 시 발생한 암 관련 외과병리보고서(텍스트), 암등록정보 등 정형데이터

☑ 데이터의 이용 환경

○ (폐쇄연구분석환경 활용) 한국대학교병원에서 제공하는 클라우드 기반의 폐쇄연구분석환경이 갖춰진 분석실에서 데이터 활용, 승인된 사용자 외에는 접근 불가

○ (자료 반입) 자료 반입시 한국대학교병원 관리자에게 요청(관리자가 자료 확인 후 반입)

○ (자료 반출) 분석결과 반출 시 한국대학교병원 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	SG-20230110001	접수일	2023년 1월 10일

신청자	조직/부서명	A대학교병원 외과		
	담당자 직위	교수	담당자 성명	한의료
	전화번호	03-1234-5678	이메일주소	research_han@corea.ac.kr

처리목적		☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존 유방암 진단 자동화 및 골밀도 감소 여부 확인을 위한 AI 개발 연구
활용 형태		☑ 내부이용 ☐ 제3자 제공 ☐ 결합전문기관을 통한 결합
이용 주체		☑ 동일 개인정보처리자 ☐ 제3자(제공받는 자) 제공
처리 환경		☑ 내부 ☐ 외부
처리 장소		☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부		☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법		☑ 온라인* ☐ 오프라인 * 폐쇄형연구분석환경에서만 접속 가능한 클라우드 서버에 업로드
제공 받는 자		(내부) ☐ 동일부서 ☑ 타 부서 (외부) ☐ 제3자, 결합전문기관 제공
신청 명세	가명처리 대상 명칭	임상데이터, 병리 및 CT, DEXA 영상 이미지
	데이터 내역	2012년부터 2019년까지 유방암 병리 및 CT, DEXA 검사를 받은 유방암 환자 (500명)의 임상데이터, 병리 및 CT, DEXA 영상 이미지
	가명정보 이용기간	2023년 5월 1일 ~ 2025년 4월 30일(2년)

위와 같이 가명정보 이용·제공을 신청합니다.

2023년 3월 2일

신청인(부서장명) 한의료 (서명 또는 인)

첨부서류
1. 연구계획서
2. 개인정보 유형 분류표
3. 가명정보에 대한 안전조치 이행 확약서
4. 보안 서약서 등

② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

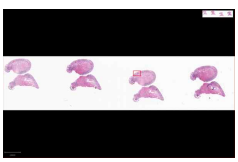
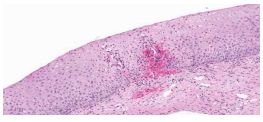
과학적 연구 계획서

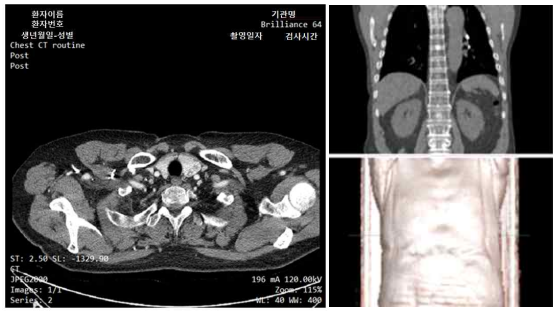
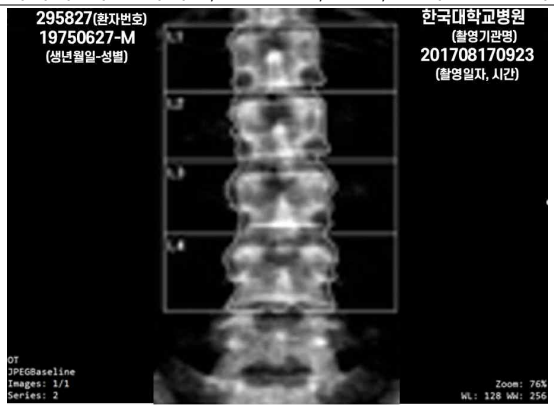
연구명	유방암 진단 자동화 및 골밀도 감소 여부 확인을 위한 AI 개발 연구
연구진	한국대학교병원 외과 한의료
연구 배경 및 목적	○ 유방암은 2020년 기준 우리나라 여성에서 가장 많이 발생하는 암으로 정확한 진단을 위해 조직병리검사를 수행하고 있으며, 암 수술 후 전이, 재발 등을 검진하기 위해 정기적으로 CT 등을, 치료 중 골밀도 감소 여부 확인을 위해 DEXA 검사를 수행하고 있음 ○ 유방암 병리조직이미지를 이용한 AI 분석을 통해 유방암 조직학적진단의 민감도를 높이고자 함 ○ 또한 유방암 치료 중 골밀도 감소증 발생이 흔하여 주기적인 추적 검사가 필요한 바, CT 및 DEXA 이미지 학습을 통해 CT 검사를 통한 골밀도 감소 사전 예측을 수행하고자 함
예상 연구 기간	2023. 5. 1. ~ 2025. 4. 30.(2년)
연구 대상자 수	한국대학병원에서 유방암으로 진단받고 수술한 여성 환자 500명 - 연구대상자 선정기간 (2012.1.1. ~ 2019.12.31.)
연구 방법	유방암 병리 및 CT, DEXA 검사 등 임상데이터에서 유방암 발생 여부, 골밀도 감소 여부 등을 판별 및 분류하는 AI 모듈 개발을 위한 학습에 활용
연구 내용	○ 병리 이미지와 조직 병리 검사결과(기록지)를 이용한 AI 학습을 통해 병리 이미지를 이용한 진단 결과 도출 -450명은 학습용 데이터셋으로, 50명은 테스트 데이터셋으로 활용 ○ 흉부CT 이미지와 DEXA 검사결과를 이용한 AI 학습을 통해 CT에서 골밀도 감소 여부 결과 도출 -450명은 학습용 데이터셋으로, 50명은 테스트 데이터셋으로 활용
기대 효과 및 활용 방안	○ 조직 병리 진단 자동화를 통해 유방암 진단의 정확도를 높이고 예후와의 관련성을 확인 ○ 유방암 환자의 재발 등을 확인하기 위해 검사한 CT 이미지를 이용하여 골밀도 감소 진단을 수행함으로써 유방암 환자들에게 필요한 DEXA 검사 등의 부담을 감소시킬 수 있음

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

개인정보 유형 분류표 (요약)				
연번	항목명	데이터 유형	데이터 규모	비고
1	병리조직 데이터	이미지	2,000장 (500명*4장)	비정형데이터
2	흉부CT 데이터	영상·이미지 (DICOM*)	100,000장 (500명*100장*2회 촬영)	
3	골밀도 검사 (DEXA) 데이터	영상·이미지 (DICOM*)	1,000장 (500명*2장*1회 촬영)	
4	외과병리 보고서	텍스트(비정형, 관찰입력정보)	500건	
4-1	Organ(장기명)	텍스트(정형)	500건	비정형데이터 ⇒ **정형데이터** 변환
4-2	Location(유방 수술의 위치)	텍스트(정형)	500건	
~	~	~	500건	
4-17	Intraductal_comp(상피내암 유무)	텍스트(정형)	500건	
5	gender(성별)	텍스트(정형)	500건	**정형데이터 (환자 임상데이터)**
6	death(사망일자)	숫자(정형)	500건	
~	~	~	500건	
32	MIEX_YMD(영상검사 일자)	숫자(정형)	500건	

* DICOM(Digital Imaging and Communications in Medicine) : 의료용 디지털 영상 및 통신 표준

개인정보 유형 분류표 (상세 : 비정형데이터)					
연번	항목명	데이터 유형	데이터 규모	구분	예시
1	병리조직 데이터	이미지	2,000장 (500명* 4장)	① 비율슬라이드 사진	
				② 병리슬라이드 번호 사진	
				③ 검체 슬라이드 사진	
				④ 검체 확대 사진	

2	흉부CT 데이터	영상·이미지 (DICOM)	100,000장 (500명 * 100장 * 2회 촬영)	 * 이미지 내 환자이름, 생년월일, 성별, 환자번호 존재
3	골밀도 검사 (DEXA) 데이터	영상·이미지 (DICOM)	1000장 (500명 * 2장 * 1회 촬영)	 * 이미지 내 환자 생년월일, 성별, 환자번호, 촬영기관명, 촬영일자·시간 존재
4	외과병리 보고서	텍스트 (관찰입력정보, 자유입력텍스트)	500명 (11.5MB)	

➜ 4. 외과병리보고서(비정형 텍스트데이터)는 자연어 처리기술 및 수기작업을 통해 정형데이터로 변경하여 활용하고자 함

*** (예시) 비정형 외과병리보고서의 정형데이터 변환**

외과병리보고서는 암의 치료, 예후 예측을 위해 매우 중요한 정보를 포함하나, 반정형의 기록지 형태로 되어 있어 정형자료와 비정형자료의 애매한 경계선에 있는 자료로 그대로는 활용이 쉽지 않고, 병리보고서 내에 연구에 필요하지 않은 다양한 개인식별가능정보들이 정제되지 않은 형태로 활용될 위험이 있으므로, 비정형 기록지 활용을 위해서는 정형화된 형태로의 작업을 통해 활용이 가능하도록 처리가 필요할 수 있음

기존에는 병리과 수작업 등의 방법으로 추출 및 활용하였으나, 최근에는 자연어처리 기술의 발달로 학습에 의한 정형화된 정보로 추출 및 관리가 가능해짐

< 비정형 외과병리보고서의 정형화 데이터 변환 과정 예시 >

비정형 외과병리 검사결과

INVASIVE DUCTAL CARCINOMA
1) Site of tumour: 1.5cm
2) Histologic grade: 3/3
(tubule formation: 3/3, nuclear pleomorphism: 3/3, mitotic court: 3/3)
3) Intradutal component:present, intratuoural/extratumoral (50%)
.
.
.
10) Microcalcification present, tumoral/non=tumoral
11) Pathological TN category (AJCC 2017) : pT1c(Paget)No(sn)

자연어 처리기술

정형화 외과병리 검사결과

path_no	EFEATURE	EVALUE	ENTART	EEND
0	ORGAN	Stomach	859	906
1	OP_NAME	laparoscopy assisted distal gastrectomy	859	907
2	GROSS_TYPE	type 0-I+IIc	892	944
3	LOC	lower third, center at antrum and anterior wall	946	980
4	HISTO_TYPE	Adenocarcinoma, tubular, well differentiated (...	1000	1066
5	T_SIZE	8.0x4.3x0.2cm	1081	1094
6	DEPTH_INV	lamina propria	1122	1137
7	S_MARGIN	negative	1170	1178
8	LYMPH_NODE	no metastasis in 17 lymph nodes	1232	1264
9	LYMPH_INV	not identified	1452	1466
10	VEIN_INV	not identified	1482	1508
11	Perineur_INV	not identified	1538	1552
12	T_BORDER	pushing	1573	1580
13	P_STAGE	pT1aN0	1657	1665
14	ORGAN	Stomach	714	721
15	OP_NAME	laparoscopy assisted distal gastrectomy	725	782
16	GROSS_TYPE	type 0-IIa	787	797
17	LOC	lower third, center at antrum and posterior wall	799	847
18	HISTO_TYPE	Signet ring cell carcinoma with tubular, poo...	858	957
19	T_SIZE	2.4x2.0x0.35cm	982	996

개인정보 유형 분류표 (상세: 비정형데이터 ⇒ 정형데이터 변환)

연번	항목명(영문)	항목명(한글)	예시	비고
<외과병리 정보>				
4-1	Organ	장기명	Breast	
4-2	Location	유방 수술의 위치	Right, left	
4-3	Tumor_size	종양의 크기	2.2cm	
4-4	Histologic_type	조직학적 유형	Invasive ductal carcinoma	
4-5	Histologic_grade	조직학적 등급	2/3	
4-6	Surgical_margins	수술절제면	deep margin	
~	~	~	~	~
4-17	Intraductal_comp	상피내암 유무	present	

개인정보 유형 분류표 (상세: 정형데이터)

연번	항목명(영문)	항목명(한글)	예시	비고
<기본정보>				
5	gender	성별	F	
6	death	사망일자	19910525	
<암 등록정보>				
7	fdx	초진연월일	20200505	
8	age	암진단당시 나이	50	
~	~	~	~	
<수술 정보>				
20	sur_Date	수술일자	20200605	
21	sur_Date1	처방일자	20200605	
~	~	~	~	
<외과병리 정보>				
25	hist_date1	병리검사일자	20200605	
26	hist_date2	처방일자	20200605	
<DEXA 검사 결과>				
27	MIEX_YMD	검사일자	2017-04-21	
28	MIEX_CD	검사코드	RX211	
29	MIEX_NM	검사명	Dexa Bone Densitometry	
30	RGN_NM	검사부위명	L1-L4	
<CT영상 검사 결과>				
31	ORD_YMD	영상검사 처방일자	2020-08-24	
32	MIEX_YMD	영상검사 일자	2020-09-27	

활용데이터 요구 수준표

※ 가명정보 처리 목적 달성에 필요한 데이터 항목과 가명처리 요구수준을 검토

< 비정형데이터 >

연번	항목명	요구 수준	비고
1	병리조직데이터 (이미지)	1) 비율슬라이드 사진 : 연구목적 달성을 위해 그대로 사용 필요 2) 병리슬라이드 번호 사진 : 분석에 필요 없어 삭제·대체 가능 3) 검체 슬라이드 사진 : 연구목적 달성을 위해 그대로 사용 필요 4) 검체 확대 사진 : 연구목적 달성을 위해 그대로 사용 필요	
2	흉부CT 데이터 (DICOM)	1) 흉부 촬영부분: 연구목적 달성을 위해 그대로 사용 필요 2) 이미지 내 환자관련정보 2-1) 환자이름 : 분석에 필요 없으므로 삭제·대체 가능 2-2) 생년월일 : 연별 분석을 위해 연도 자료 필요(일자는 삭제·대체 가능) 2-3) 환자성별 : 여성환자로 한정하여 분석하므로, 남성환자 정보는 모두 삭제 삭제·대체해도 연구에 영향 없음 2-4) 환자번호 : 환자구분만 되면 되므로 일련번호로 대체 가능	
3	골밀도 검사 데이터 (DICOM)	1) 골밀도 촬영부분: 연구목적 달성을 위해 그대로 사용 필요 2) 이미지 내 환자관련정보 2-1) 생년월일 : 분석에 필요 없으므로 삭제·대체 가능 2-2) 환자성별 : 분석에 필요 없으므로 삭제·대체 가능 2-3) 환자번호 : 분석에 필요 없으므로 삭제·대체 가능 2-4) 촬영기관명 : 분석에 필요 없으므로 삭제·대체 가능 2-5) 촬영일자·시간 : 분석에 필요 없으므로 삭제·대체 가능	
4	외과병리 보고서	연구수행을 위해 정형화된 검사결과만 필요하므로, 정형데이터로 변환 후 삭제 가능	

< 비정형데이터 → 정형데이터 >

연번	항목명(영문)	항목명(한글)	요구 수준	비고
<외과병리 정보>				
4-1	Organ	장기명	연구목적 달성을 위해 그대로 사용 필요	
4-2	Location	유방 수술의 위치	연구목적 달성을 위해 그대로 사용 필요	
4-3	Tumor_size	종양의 크기	연구목적 달성을 위해 그대로 사용 필요	
4-4	Histologic_type	조직학적 유형	연구목적 달성을 위해 그대로 사용 필요	
4-5	Histologic_grade	조직학적 등급	연구목적 달성을 위해 그대로 사용 필요	
4-6	Surgical_margins	수술절제면	연구목적 달성을 위해 그대로 사용 필요	
~	~	~	~	~
4-17	Intraductal_comp	상피내암 유무	연구목적 달성을 위해 그대로 사용 필요	

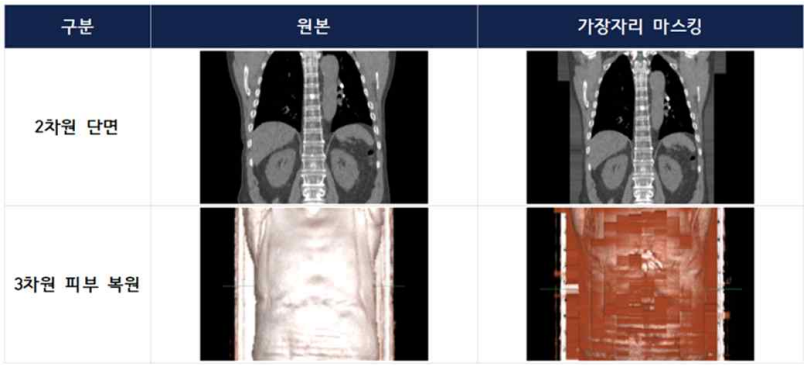
< 정형데이터 >

연번	항목명(영문)	항목명(한글)	요구 수준	비고
<기본정보>				
5	gender	성별	여성환자 대상 분석으로, 여성환자 정보만 필요 남성환자 정보는 삭제	
6	death	사망일자	연별 분석 수행 예정으로, 연도 자료만 필요	
<암 등록정보>				
7	fdx	초진연월일	월단위 분석 수행 예정으로, 연월만 필요	
8	age	암진단당시 나이	통계 분석에 무리가 없는 선에서 일정 수준의 범주화 가능(협의 필요)	
~	~	~	~	
<수술 정보>				
20	sur_Date	수술일자	월단위 분석 수행 예정으로, 연월만 필요	
21	sur_Date1	처방일자	월단위 분석 수행 예정으로, 연월만 필요	
~	~	~	~	
<외과병리 정보>				
25	hist_date1	병리검사일자	월단위 분석 수행 예정으로, 연월만 필요	
26	hist_date2	처방일자	월단위 분석 수행 예정으로, 연월만 필요	
<DEXA 검사 결과>				
27	MIEX_YMD	검사일자	월단위 분석 수행 예정으로, 연월만 필요	
28	MIEX_CD	검사코드	연구목적 달성을 위해 그대로 사용 필요	
29	MIEX_NM	검사명	연구목적 달성을 위해 그대로 사용 필요	
30	RGN_NM	검사부위명	연구목적 달성을 위해 그대로 사용 필요	
<CT영상 검사 결과>				
31	ORD_YMD	영상검사 처방일자	월단위 분석 수행 예정으로, 연월만 필요	
32	MIEX_YMD	영상검사 일자	월단위 분석 수행 예정으로, 연월만 필요	

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

<table>
<tr><td>가명정보
활용목적</td><td colspan="2">○ 유방암 진단 자동화 및 골밀도 감소 여부 확인을 위한 AI 개발 연구</td></tr>
<tr><td>가명처리
대상 데이터
항목</td><td colspan="2">○ 병리조직데이터(이미지) <연번 1>
○ 흉부CT데이터(영상·이미지) <연번 2>
○ 골밀도 검사(DEXA) 데이터(영상·이미지) <연번 3>
○ 환자 임상데이터(외과병리보고서, 텍스트) <연번 4 (4-1~4-17)>
○ 환자 임상데이터(환자 기본정보, 암 등록정보, 수술정보, 외과병리정보, DEXA검사결과, CT영상 검사결과) <연번 5~32></td></tr>
<tr><td rowspan="2">데이터
위험성</td><td>식별성
유무</td><td><비정형데이터>
○(병리이미지) 병리슬라이드 번호사진 외에는 개인식별 가능성 거의 없음
○(CT 영상·이미지) 영상·이미지 자체로는 개인식별 가능성 거의 없음
- DICOM 영상·이미지에 포함된 환자관련정보(환자이름, 생년월일, 환자성별, 환자번호)는 개인식별 가능성이 있어 가명처리 필요
○(DEXA 영상·이미지) 영상·이미지 자체로는 개인식별 가능성 거의 없음
- DICOM 영상·이미지에 포함된 환자관련정보(환자이름, 생년월일, 환자성별, 환자번호 등)는 개인식별 가능성이 있어 가명처리 필요
○(외과병리 보고서) 광학문자인식(Optical Character Recognition, OCR) 기법을 통해 이미지내 데이터를 정형화데이터로 변환하여 활용할 예정이며, 병리보고서 내 외과병리정보만으로는 개인식별 가능성 거의 없음
<정형데이터>
○ 한국대학교병원 암관련 환자의 사망일자, 초진연월일, 암진단 연령, 병리검사일자, 처방일자, CT영상검사 처방일자, CT영상검사 일자 등은 조합되었을 때 개인의 식별이 가능한 개인식별가능정보임</td></tr>
<tr><td>특이정보
유무</td><td><비정형데이터>
○(CT 영상·이미지) 개인당 200장의 이미지가 촬영되었기 때문에 3차원 재건 등의 기술을 활용하면 신체 이미지를 입체적으로 복원가능하며, 복원시 특이한 외형·흉터 등이 있는 환자의 경우 낮은 확률로 식별위험이 생길 수 있음
* 가장자리 마스킹 기법을 활용하여 3차원 재건 공격 위험을 막을 수 있음
<정형데이터>
○‘암진단 연령’의 경우 상당히 적거나 높은 경우 특이치로 인한 개인식별성이 발생할 수 있음</td></tr>
</table>

<table>
<tr><td rowspan="2"></td><td></td><td> 
</td></tr>
<tr><td>재식별시 영향도</td><td>○개인에 대한 진료·진단 정보로, 재식별시 영향도는 높은 편</td></tr>
<tr><td rowspan="3">처리 환경 위험성</td><td>이용 및 제공 형태</td><td>○기관내 연구자에게 제공</td></tr>
<tr><td>처리 장소</td><td>○한국대학교병원에서 제공하는 클라우드 기반의 폐쇄연구분석환경이 갖춰진 분석실
○한국대학교병원은 개인정보(가명정보)처리시스템에 대한 ISMS-P인증 취득</td></tr>
<tr><td>다른 정보 결합 가능성</td><td>○폐쇄환경분석실 관리자 승인하에 제한된 데이터·프로그램(프로그램 패키지, 라이브러리, 코드설명서 등)만 반입 가능
○분석대상 가명정보와 결합가능성 있는 데이터는 반입 제한</td></tr>
<tr><td>최종 검토의견</td><td colspan="2">▪ 해당 가명정보는 기관내 연구자에게 제공될 뿐만 아니라, 클라우드 기반의 폐쇄연구 분석환경(외부 인터넷 이용, 다른 데이터의 반입 및 결합, 데이터 외부반출 등이 제한)에서만 접속하여 분석 가능하므로 식별 가능성이 낮은 편이며 안전한 처리 환경을 고려할 때 다음과 같은 조치가 필요함
- 병리슬라이드 번호 사진, 개인 신상정보 중 ‘성별’ 등 연구 수행에 필요없는 정보는 삭제
- 병리조직·CT·DEXA 영상·이미지는 연구목적에 필요한 범위 내에서 그대로 활용하여도 식별 가능성이 거의 없으나 영상·이미지에 포함된 환자관련정보(환자이름, 생년월일, 환자성별, 환자번호 등)은 삭제·대체 등 필요
- 개인당 200장의 이미지가 촬영된 CT의 경우 3차원 재건 등을 통한 식별위험이 존재하나, 연구자가 기관 내 폐쇄연구 분석환경에서 연구를 수행하고 외부 데이터·프로그램 활용이 제한되기 때문에 관련 식별위험이 없을 것으로 판단되므로, 가장자리 마스킹 등 별도 가명처리 없이 그대로 활용 가능
- ‘사망일자’, ‘초진연월일’, 연령’, ‘병리검사일자’, ‘처방일자’, ‘CT영상검사 처방일자’, ‘CT영상검사 일자’, ‘DEXA영상검사 일자’는 그대로 사용하는 경우 조합에 의한 개인식별 가능성이 있으며, 연구목적상 연·월단위 분석까지만 수행하므로 필요에 맞게 사망일자는 연단위까지, 나머지 데이터는 월단위까지만 활용 필요
- ‘암진단 연령’ 또한 ‘39세 이하’, ‘40 ~ 84세: 1세단위’, ‘85세 이상’으로 범주화하여 연구 필요
- 자유입력데이터인 외과병리보고서는 그대로 활용이 쉽지 않으므로 자연어 처리 기술 등을 통해 정형데이터로 변환 후 활용하는 것이 적절</td></tr>
</table>

④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

< 비정형데이터 >

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
1	병리조직 데이터	1) 비율슬라이드 사진	☑ 그대로 사용	별도 처리하지 않음
		2) 병리슬라이드 번호 사진	☑ 삭제	분석에 필요 없고, 식별 가능성 있으므로 삭제
		3) 검체 슬라이드 사진	☑ 그대로 사용	별도 처리하지 않음
		4) 검체 확대 사진	☑ 그대로 사용	별도 처리하지 않음
2	흉부CT 데이터	1) 흉부 촬영부분	개인당 200장의 사진이 촬영되어, 3차원 재건 등의 기술을 활용하면 신체 이미지를 복구할 수 있고 복원시 특이한 외형·흉터 등이 있는 환자의 경우 낮은 확률로 식별가능성 존재	
			□ 표면 가장자리 삭제	3차원 재건으로 인한 개인식별 가능성을 더욱 낮출 수는 있으나 이 경우, 데이터 손실로 인해 연구목적 달성이 어려움
			☑ 그대로 사용	이미지를 그대로 사용하되, 3차원 재건 기술을 적용할 수 없도록 처리 환경을 통제 * 클라우드 기반 폐쇄연구 분석환경에서만 활용, 철저한 외부 데이터·프로그램 반입 관리 수행
		2) 이미지 내 환자관련정보 * DICOM 형식 그대로 연구에 활용이 필요하기 때문에 블랙마스킹 처리가 불가하며, 이미지 내 표시된 DICOM 메타데이터를 손쉽게 일괄 변경할 수 있도록 자체개발한 도구를 활용하여 가명처리 수행		
		2-1) 환자이름	☑ 대체	연구에 필요 없는 정보이므로, 자체개발한 DICOM 데이터 변경도구를 활용하여 Anonymized 값으로 대체 (블랙마스킹으로 삭제해도 무방)
		2-2) 생년월일	☑ 대체(일부)	"일자" 정보는 연구에 필요 없으므로 자체개발한 DICOM 데이터 변경도구를 활용하여, "00"으로 대체 (블랙마스킹으로 삭제해도 무방)
		2-3) 환자성별	☑ 삭제·대체	여성환자만 분석에 사용하므로 남성 환자가 있다면 데이터를 삭제하고 이후 성별값은 구분에 의미가 없으므로, 자체개발한 DICOM 데이터 변경도구를 활용하여, "F"(Female) 값을 "S"(Sex)로 대체 (블랙마스킹으로 삭제해도 무방)
		2-2) 환자번호	☑ 대체	환자구분만 되면 되므로 자체개발한 DICOM 데이터 변경도구를 활용하여, 단순일련번호로 대체
3	골밀도검사 (DEXA) 데이터	1) 골밀도 촬영부분	개인당 2장의 사진만 촬영되었을 뿐만 아니라, 골밀도 촬영 사진의 특성상 3차원 재건 등의 기술을 활용한 신체 이미지 복구는 어려움	
			☑ 그대로 사용	3차원 재건 등을 통한 식별 위험성이 없으므로, 별도 처리하지 않음
		2) 이미지 내 환자관련정보 * PNG 포맷으로 연구에 활용할 계획으로, 파일 포맷 변환(DICOM→PNG) 후, 블랙마스킹 기법을 통해 이미지 내 환자관련정보 삭제		
		2-1) 생년월일	☑ 마스킹	분석에 필요 없으므로, 블랙마스킹 기법으로 삭제 처리
		2-2) 환자성별	☑ 마스킹	분석에 필요 없으므로, 블랙마스킹 기법으로 삭제 처리
		2-3) 환자번호	☑ 마스킹	분석에 필요 없으므로, 블랙마스킹 기법으로 삭제 처리
		2-4) 촬영기관명	☑ 마스킹	분석에 필요 없으므로, 블랙마스킹 기법으로 삭제 처리
		2-5) 촬영일자·시간	☑ 마스킹	분석에 필요 없으므로, 블랙마스킹 기법으로 삭제 처리

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
4	외과병리 보고서	검사결과 내용	☑ 삭제	정형데이터로 변환시켜 분석에 필요한 주요 검사결과만 추출 (아래 4-1~17 참조)한 뒤 삭제

< 비정형데이터 → 정형데이터 >

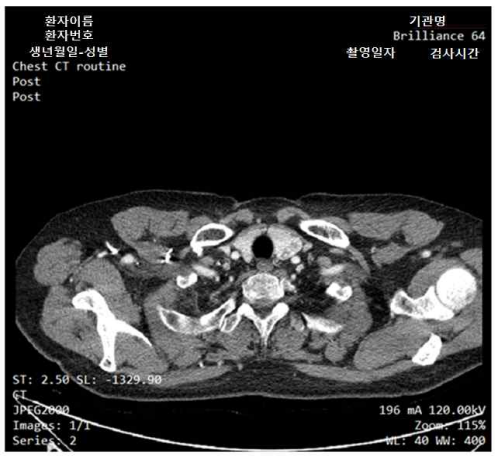
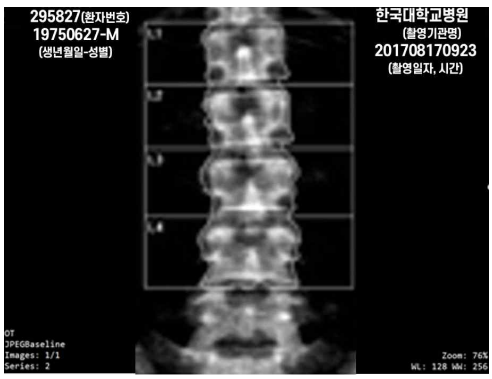
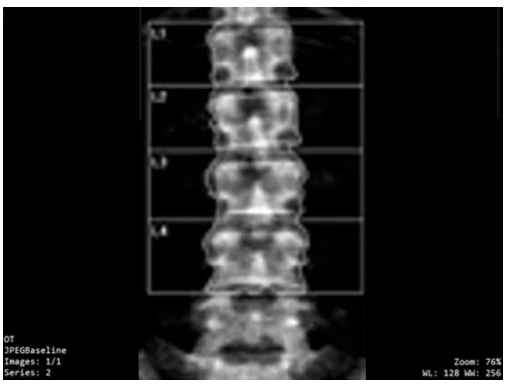
연번	항목명(영문)	항목명(한글)	처리방법	세부방법 및 처리수준
<외과병리 보고서>				
4-1	Organ	장기명	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음
4-2	Location	유방 수술의 위치	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음
4-3	Tumor_size	종양의 크기	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음
4-4	Histologic_type	조직학적 유형	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음
4-5	Histologic_grade	조직학적 등급	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음
4-6	Surgical_margins	수술절제면	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음
~	~	~	~	~
4-17	Intraductal_comp	상피내암 유무	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음

< 정형데이터 >

연번	항목명(영문)	항목명(한글)	처리방법	세부방법 및 처리수준	가명처리 전	가명처리 후
<개인 신상정보>						
5	gender	성별	☑ 삭제	분석에 필요 없으므로 삭제 처리		
6	death	사망일자	☑ 부분삭제	연도만 필요하므로 월일은 삭제 처리	20230404	2023
<암 등록정보>						
7	fdx	초진연월일	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리		
8	age	암진단당시 나이	☑ 범주화	39세 이하, 40-84세(1세단위), 85세 이상으로 구분	29세 50세 93세	39세 이하 50세 85세 이상
~	~	~	~	~		
<수술 정보>						
20	sur_Date	수술일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리		
21	sur_Date1	처방일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리		
~	~	~	~	~		
<외과병리 정보>						
25	hist_date1	병리검사일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리		
26	hist_date2	처방일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리		
<DEXA 검사 결과>						
27	MIEX_YMD	검사일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리		
28	MIEX_CD	검사코드	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음		
29	MIEX_NM	검사명	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음		
30	RGN_NM	검사부위명	☑ 그대로 사용	개인 식별가능성이 낮아 별도 처리하지 않음		
<CT영상 검사 결과>						
31	ORD_YMD	영상검사 처방일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리	20200610	202006
32	MIEX_YMD	영상검사 일자	☑ 부분삭제	연월만 필요하므로 일자는 삭제 처리	20200927	202009

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

주요 비정형데이터 가명처리 수행 결과 (예시)

연번	항목명	가명처리 전	가명처리 후
2	흉부CT 이미지	 1) 흉부촬영부분 2) DICOM 이미지 내 환자관련정보 2-1) 환자이름 2-2) 생년월일 2-3) 환자성별 2-4) 환자번호	 1) 흉부촬영부분 : 그대로 유지 2) DICOM 이미지 내 환자관련정보 2-1) 환자이름 : Anonymized로 대체 2-2) 생년월일 : 연도정보로만 변경 2-3) 환자성별 : 성별값을 "S"로 대체 2-4) 환자번호 : 환자 구별을 위한 단순 일련번호로 대체 * 자체개발한 DICOM 데이터 변경도구 활용
3	골밀도 검사 DEXA 이미지	 1) 골밀도 촬영부분 2) 이미지 내 환자관련정보 2-1) 생년월일 2-2) 환자성별 2-3) 환자번호 2-4) 촬영기관명 2-5) 촬영일자·시간	 1) 골밀도 촬영부분 : 그대로 유지 2) 이미지 내 환자관련정보 2-1) 생년월일 : 블랙마스킹 2-2) 환자성별 : 블랙마스킹 2-3) 환자번호 : 블랙마스킹 2-4) 촬영기관명 : 블랙마스킹 2-5) 촬영일자·시간 : 블랙마스킹 * PNG 포맷으로 연구에 활용할 계획으로, 파일 포맷 변환(DICOM→PNG) 후, 이미지 블랙마스킹 기법을 통해 이미지 내 환자관련정보 삭제

가명처리 결과 자체검증

○ 비정형 데이터 가명처리 기술의 적절성・신뢰성 관련 근거(예시)

<table>
<tr><th>연번</th><th>대상
항목</th><th>데이터
유형</th><th>처리 기술명</th><th>처리 기술의 적절성·신뢰성 관련
근거 또는 배경</th><th>비고</th></tr>
<tr><td>2</td><td>흉부CT
데이터</td><td>이미지</td><td>- DICOM 데이터 변경 도구를 사용한 환자 정보 변경·대체</td><td>▪ 한국대학교병원에서 자체 개발한 DICOM 데이터 변경 도구를 활용하여 이미지 내 환자관련정보(환자이름, 생년월일, 환자성별, 환자번호) 대체
- 자체 개발한 DICOM 데이터 변경도구는 CT사진에 대해 DICOM 형식 그대로 연구에 활용이 필요할 경우(블랙마스킹 처리불가) 등을 위해 이미지 내에 표시된 DICOM 메타데이터를 손쉽게 일괄 변경할 수 있도록 개발한 솔루션
- 개발용역사의 테스트 결과, 객체 인식률 99.2%, 처리 정확도 100%로 측정
※ 객체 인식률, 처리 정확도(오류율) 증빙자료 별첨
- 환자관련정보가 특이한 경우(환자이름이 너무 길어 도구 인식가능 영역을 넘어가는 경우 등) 인식이 안되는 경우가 존재하므로, 내부정책상 DICOM 데이터 변경도구를 이용한 가명처리 후 추가적인 자체 전수검사 수행</td><td></td></tr>
<tr><td>3</td><td>골밀도
검사
(DEXA)
데이터</td><td>이미지</td><td>- 영상이미지 블랙 마스킹</td><td>▪ 파일 포맷 변환(DICOM→PNG) 후, 이미지 블랙마스킹 프로그램을 통해 이미지 내 환자관련정보 삭제
- 파일 포맷 변환(DICOM→PNG) 후에도 이미지 내 환자관련 정보가 동일한 위치에 존재하므로 환자관련정보가 위치한 영역을 지정하여 블랙마스킹 프로그램 적용
- 프로그램 개발사 테스트 결과, 객체 인식률 98%, 처리 정확도 98.4%로 측정
※ 객체 인식률, 처리 정확도(오류율) 증빙자료 별첨
- 환자관련정보가 특이한 경우(환자이름이 너무 길어 도구 지정 영역을 넘어가는 경우 등), 인식·처리가 정확히 안되는 경우가 존재하므로, 내부정책상 블랙마스킹 프로그램을 이용한 가명처리 후 추가적인 자체 전수검사 수행</td><td></td></tr>
<tr><td>4</td><td>외과병리
보고서</td><td>텍스트</td><td>- 자연어 처리 기술(NLP)을 통해 정형 데이터로 변환

* NLP(Natural Language Processing)는 인간의 언어를 해석, 조작 및 이해하는 능력을 컴퓨터에 부여하는 기계학습 기술로서, 과거 규칙기반 처리에서 최근 머신러닝 알고리즘 기반으로 발전 중</td><td>- 한국대병원은 EMR 내 암종별 병리기록지 현황을 파악하여 암종별 특성에 따라 자연어 처리가 가능한 학습모델을 자체개발
- 사전에 검토한 자연어처리 모델 중 병리기록지 처리에 가장 적합하다고 판단한 BioBERT 모델을 선정하고, 의료정보관리사가 암종별로 주석처리한 병리기록지를 바탕으로 학습모델을 개발하여 검증
- 한국대병원의 병리기록지 중 4개 암종(간암, 대장암, 위암, 유방암)을 대상으로 검증을 진행하였으며, 유방암은 95.8%의 처리 정확도를 보임
<table>
<tr><th>암종명</th><th>건수</th><th>변수 개수</th><th>암종별
평균 처리 정확도</th></tr>
<tr><td>간암</td><td>2,312</td><td>22</td><td>97.8%</td></tr>
<tr><td>대장암</td><td>7,869</td><td>24</td><td>94%</td></tr>
<tr><td>위암</td><td>5,772</td><td>15</td><td>95.3%</td></tr>
<tr><td>유방암</td><td>18,492</td><td>24</td><td>95.8%</td></tr>
</table>
※ 관련 증빙자료 별첨
- 검증된 모델을 통해 실제 암 병리 기록지의 텍스트 정보를 추출하여 정형데이터로 변환해 활용
- 다만, 처리 정확도가 100%가 아니므로 변환 이후 추가적인 자체 전수검사를 수행하여 보완작업 진행</td><td></td></tr>
</table>

비정형데이터 가명처리 결과에 대한 자체 검증 결과서 (1)

<table>
<tr><td rowspan="9">검증 대상
데이터 명세</td><td colspan="3">개요</td></tr>
<tr><td colspan="3">한국대학교병원에서 수집한 흉부CT 데이터에 대해 영상이미지 내 환자관련정보(환자이름, 생년월일, 환자성별, 환자번호)를 DICOM 데이터 변경 도구를 활용하여 대체처리</td></tr>
<tr><td>데이터 유형</td><td colspan="2">영상·이미지</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td colspan="2">DICOM</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td colspan="2">DICOM</td></tr>
<tr><td>데이터 규모</td><td colspan="2">100,000장 (500명 * 100장 * 2회)</td></tr>
<tr><td>데이터 크기(용량)</td><td colspan="2">80GB</td></tr>
<tr><td>대상 데이터 항목명</td><td colspan="2">흉부CT데이터 <연번 2></td></tr>
<tr><td>가명처리 적용 기술</td><td colspan="2">- 한국대학교병원에서 자체 개발한 DICOM 데이터 변경 도구를 활용하여 이미지 내 환자관련정보(환자이름, 생년월일, 환자성별, 환자번호) 대체처리</td></tr>
<tr><td>자체 검증
기간</td><td colspan="3">2023년 4월 1일 ~ 2023년 4월 7일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="3">회의실 (원격으로 분석실의 가상컴퓨터 접속)</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="3">(검증방법)
- 검증은 한국대학교병원 내부 가관리 개인정보보호부장의 주도하에 의료정보관리실장, 개인정보보호부 직원 4인이 함께 진행
- 실장 및 직원이 데이터를 나누어 환자관련정보 가명처리 정상 수행 여부를 육안으로 검수하고, 특이사항이 발생한 표본만 선별하여 전체인원이 추가 합동검수·처리 수행
(검증 시 확인사항)
1. DICOM 이미지 내 환자관련정보가 다음과 같이 처리되었는지 전수 확인
① 환자이름 : Anonymized 로 대체
② 생년월일 : 연도정보만 표시되고, 일자는 ‘00’으로 대체
③ 환자성별 : 환자성별값(‘M’, ‘F’)을 ‘S’로 대체
④ 환자번호 : 환자구별을 위한 단순일련번호로 대체
2. 위 4가지 사항 중 하나라도 만족하지 않았을 경우 DICOM 데이터 변경도구 또는 수작업을 통해 다시 한 번 추가처리를 수행
3. 처리가 제대로 되지 않는 표본은 최종 연구 표본에서 삭제처리</td></tr>
<tr><td>자체 검증
결과</td><td colspan="3">확인 결과 이상 없음
(검수 결과, 724장이 인식되지 않았으나, 수작업을 통해 최종 가명처리 완료)</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>한국대학교 개인정보보호부장</td><td>가관리</td><td></td></tr>
<tr><td>한국대학교 의료정보관리실장</td><td>나정보</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>다보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>라보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>마보호</td><td></td></tr>
</table>

비정형데이터 가명처리 결과에 대한 자체 검증 결과서 (2)

<table>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="3">개요</td></tr>
<tr><td colspan="3">한국대학교병원에서 수집한 골밀도 검사(DEXA) 데이터를 DICOM 포맷에서 PNG 포맷으로 변환 후, 영상이미지 내 환자관련정보(생년월일, 환자성별, 환자번호, 촬영기관명, 촬영일자·시간)을 블랙마스킹 프로그램을 활용하여 마스킹처리</td></tr>
<tr><td>데이터 유형</td><td colspan="2">영상·이미지</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td colspan="2">DICOM</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td colspan="2">PNG</td></tr>
<tr><td>데이터 규모</td><td colspan="2">1,000장 (500명 * 2장 * 1회)</td></tr>
<tr><td>데이터 크기(용량)</td><td colspan="2">1GB</td></tr>
<tr><td>대상 데이터 항목명</td><td colspan="2">골밀도 검사(DEXA) 데이터 <연번 3></td></tr>
<tr><td>가명처리 적용 기술</td><td colspan="2">- DICOM 데이터 삭제
- 파일 포맷 변환(DICOM -> PNG)을 통한 메타데이터 제거
- 영상이미지 내 환자관련정보 블랙마스킹</td></tr>
<tr><td>자체 검증
기간</td><td colspan="3">2023년 4월 10일 ~ 2023년 4월 11일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="3">회의실 (원격으로 분석실의 가상컴퓨터 접속)</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="3">(검증방법)
- 검증은 한국대학교병원 내부 가관리 개인정보보호부장의 주도하에 의료정보관리실장, 개인정보보호부 직원 4인이 함께 진행
- 실장 및 직원이 데이터를 나누어 환자관련정보 가명처리 정상 수행 여부를 육안으로 검수하고, 특이사항이 발생한 표본만 선별하여 전체인원이 추가 합동검수·처리 수행
(검증 시 확인사항)
1. PNG 이미지 내 환자관련정보가 문제 없이 블랙마스킹 처리되었는지 전수 확인
- 생년월일, 환자성별, 환자번호, 촬영기관명, 촬영일자·시간
2. 처리가 완벽히 수행되지 않았을 경우 수작업으로 블랙마스킹 처리 수행</td></tr>
<tr><td>자체 검증
결과</td><td colspan="3">확인 결과 이상 없음
(검수 결과, 22장이 제대로 처리되지 않아, 수작업을 통해 최종 가명처리 완료)</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>한국대학교 개인정보보호부장</td><td>가관리</td><td></td></tr>
<tr><td>한국대학교 의료정보관리실장</td><td>나정보</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>다보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>라보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>마보호</td><td></td></tr>
</table>

<table>
<tr><th colspan="4">비정형데이터 가명처리 결과에 대한 자체 검증 결과서 (3)</th></tr>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="3">개요</td></tr>
<tr><td colspan="3">암 조직 관련 외과병리보고서를 자연어 처리 기술을 적용하여 정형데이터로 변환하여 활용</td></tr>
<tr><td>데이터 유형</td><td colspan="2">텍스트</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td colspan="2">DB</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td colspan="2">DB</td></tr>
<tr><td>데이터 규모</td><td colspan="2">500건</td></tr>
<tr><td>데이터 크기(용량)</td><td colspan="2">11.5MB</td></tr>
<tr><td>대상 데이터 항목명</td><td colspan="2">외과병리보고서 내 아래 데이터 항목을 추출하여 정형데이터로 변환
- Organ, Location, Tumor_size, Histologic_type, Histologic_grade, Surgical_margins, Lymph_node, P_stage, Necrosis, Nuclear grade, Microcalcification, Skin/nipple invasion, Arteriovenous_invasion, Lymphovascular invasion, Tumor Border, Architectural pattern, Intraductal_comp</td></tr>
<tr><td>가명처리 적용 기술</td><td colspan="2">자연어처리기술을 통해
텍스트정보를 정형데이터로 변환</td></tr>
<tr><td>자체 검증 기간</td><td colspan="3">2023년 4월 13일 ~ 2023년 4월 14일</td></tr>
<tr><td>자체 검증 장소</td><td colspan="3">회의실 (원격으로 분석실의 가상컴퓨터 접속)</td></tr>
<tr><td>자체 검증 과정
및 방법</td><td colspan="3">(검증방법)
- 검증은 한국대학교병원 내부 가관리 개인정보보호부장의 주도하에 개인정보보호부 직원 2인이 함께 진행
- 직원이 데이터를 나눠 정형데이터로의 정상 변환 여부를 육안으로 검수하고, 특이사항이 발생한 표본만 선별하여 전체인원이 추가 검수·처리 수행
(검증 시 확인사항)
1. 17가지의 데이터 항목이 문제없이 인식되어 정형화되었는지 확인
2. 처리가 완벽히 수행되지 않을 경우 수작업으로 DB에 데이터 입력</td></tr>
<tr><td>자체 검증 결과</td><td colspan="3">확인 결과 이상 없음
(검수 결과, 57건의 표본이 제대로 처리되지 않아,
수작업을 통해 최종 가명처리 완료)</td></tr>
<tr><td rowspan="4">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>한국대학교 개인정보보호부장</td><td>가관리</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>다보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>라보호</td><td></td></tr>
</table>

시나리오 ❷ : 구강질환 진단 AI 개발 사례

의료 분야 (이미지)

- 한국대학교병원 임상시험센터는 헬스케어 분야 AI개발 스타트업인 ㈜헬스케어소프트로부터 과학적 연구 수행을 위한 데이터 제공을 요청받아 구강 건강검진 촬영 이미지 데이터를 제공하려 한다.

☑ 데이터의 이용 목적

○ 구강 내 충치 및 치주염 관련 질환을 진단하는 AI 개발 연구

☑ 데이터 특징

○ (이미지 데이터) 구강 건강검진 촬영 사진

- 5,000명의 구강 건강검진 촬영영상에서 각 10장씩 이미지를 추출하여 저장한 데이터 (JPEG포맷)
- AI 모델 학습을 위해 각 구강 이미지의 구개, 혀, 위 아래 잇몸, 상악치, 하악치, 절치, 구치, 충치영역, 치주염 영역을 각각 라벨링처리하였음
- 각 이미지에 촬영된 자에 대한 메타데이터(성별, 이름, 나이, 촬영 날짜) 포함

☑ 데이터의 이용 환경

○ (폐쇄연구분석환경 활용) 한국대학교병원에서 제공하는 클라우드 기반의 폐쇄연구분석환경이 갖춰진 분석실에서 데이터 활용, 승인된 사용자 외에는 접근 불가

○ (자료 반입) 자료 반입시 한국대학교병원 관리자에게 요청(관리자가 자료 확인 후 반입)

○ (자료 반출) 분석결과 반출 시 한국대학교병원 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	SG-20230110002	접수일	2023년 1월 5일

신청자	조직/부서명	㈜ 헬스케어소프트 데이터연구팀		
	담당자 직위	팀장	담당자 성명	홍길동
	전화번호	03-1234-5678	이메일주소	kildong@healthcaresoft.com

처리목적	☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존
	구강 내 충치와 치주염 관련 질환을 진단하는 AI 모듈 개발
활용 형태	☐ 내부이용 ☑ 제3자 제공 ☐ 결합전문기관을 통한 결합
이용 주체	☐ 동일 개인정보처리자 ☑ 제3자(제공받는 자) 제공
처리 환경	☑ 내부 ☐ 외부
처리 장소	☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부	☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법	☑ 온라인* ☐ 오프라인 * 폐쇄형연구분석환경에서만 접속 가능한 클라우드 서버에 업로드
제공 받는 자	(내부) ☐ 동일부서 ☐ 타부서 (외부) ☑ 제3자, 결합전문기관 제공

신청 명세	가명처리 대상 명칭	구강 건강검진 촬영 영상 및 이미지 데이터
	데이터 내역	2022년 구강 건강검진 및 질환 치료를 위해 촬영한 영상 데이터 5만건을 대상으로 학습 데이터로 필요한 단층 이미지로 변환해둔 데이터를 가명처리・제공하여 구강 질환 진단 AI 모듈을 개발하는데 활용
	가명정보 이용기간	2023년 3월 16일 ~ 2024년 3월 15일(1년)

위와 같이 가명정보 이용·제공을 신청합니다.

2023년 1월 5일

신청인(부서장명) 홍길동 (서명 또는 인)

첨부서류	1. 연구계획서 2. 개인정보 유형 분류표 3. 가명정보에 대한 안전조치 이행 확약서 4. 보안 서약서 등

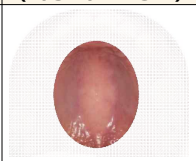
② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

과학적 연구 계획서

연구명	구강 내 충치와 치주염 관련 질환을 진단하는 AI 모듈 개발 연구
연구진	㈜헬스케어소프트 데이터연구팀 (책임자: 홍길동 팀장)
연구 배경 및 목적	○ 구강검사 시 단순한 기초 검사에도 대기와 진행시간이 소요되고 질환 치료의 의사결정까지 오래 걸림 ○ 대형병원의 충치 및 치주염 등 기초검사 진행시간 단축 및 정확도 향상 필요 ○ 다양한 구강질환의 진단, 초기 발견 등 치료에 필요한 구강검진 기록입력 자동화 등을 지원 가능 ○ 국민의 Self 구강검진을 통한 조기검진 및 예방관리 지원 기술 개발 필요
예상 연구 기간	2023년 3월 16일 ~ 2024년 3월 15일(1년)
연구 대상자 수	한국대학병원에서 구강 건강검진을 받은 5,000명의 내원자 - 연구대상자 선정기간 (2012.1.1. ~ 2022.12.31.)
연구 방법	○ (데이터 전처리) 데이터를 전처리하여, 이미지 크기 조정, 노이즈 제거, 이미지 보정 등의 과정 수행 ○ (데이터 라벨링) 데이터에 대해 라벨링 작업 수행, 라벨링 작업은 전문가들이 수행하며, 충치와 치주염 여부, 질환의 위치, 크기 등을 라벨링 ○ (모델 개발) 데이터를 바탕으로 구강 질환 판별 모델 개발 ○ (모델 평가) 개발한 모델을 평가(정확도, 재현율, F1-score 등의 지표 검증)
연구 내용	구강 건강검진 및 질환 치료를 위해 촬영한 영상 데이터 5,000건(단층 이미지 50,000장)을 학습 데이터로 학습하여 구강질환 진단 모듈 개발
기대효과 및 활용방안	구강진단 AI 모듈 개발로 진료 정확성 및 효과성 향상
붙임. 상세 연구계획서 등	

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

개인정보 유형 분류표 (요약)					
연번	항목명	데이터 유형	데이터 규모	예시	비고
1	구강 촬영데이터	이미지 (JPEG 포맷)	이미지 50,000장 (5,000명*10장) * 개인 영상 당 단층 촬영물 10개의 이미지 파일 추출 * 개인당 평균 약 6MB, 총 300GB		**비정형데이터 (메타데이터 포함)**

개인정보 유형 분류표 (상세)					
연번	항목명	구분	설명	예시 (해당 부분 강조)	비고
1	구강 촬영데이터	① 구개	입천장		이미지 (비정형데이터)
		② 혀	구강의 바닥에 위치한 근육 조직		
		③ 윗잇몸	위잇몸(상악점막)		
		④ 아래잇몸	아랫잇몸(하악점막)		
		⑤ 상악치	윗니		
		⑥ 하악치	아랫니		
		⑦ 충치 영역	절치와 구치 중 일부		
		⑧ 치주염 영역	상악점막과 하악점막 중 일부		
	1-1	성별	'남성', '여성'으로 구분	여성	구강 촬영데이터의 메타데이터
	1-2	이름	내원자의 이름	홍길동	
	1-3	나이	12세~89세까지 분포	27세	
	1-4	촬영날짜	2012.1.1.~2022.12.31.까지 분포	2021.8.28.	

활용데이터 요구 수준표

연번	항목명	구분	요구 수준	비고
1	구강 촬영데이터	① 구개	▪ 충치 및 치주염 관련 진단에 직접적으로 활용되지 않으므로, 블러링 기법 등을 통해 가명처리해도 상관 없음 ▪ 다만, 다른 영역과 구분하여 인식하여야 하므로 삭제하거나 마스킹 등으로 완전히 삭제하거나 대체하면 안 됨	이미지 (비정형데이터)
		② 혀		
		③ 윗잇몸		
		④ 아래잇몸		
		⑤ 상악치		
		⑥ 하악치		
		⑦ 충치 영역	▪ 연구목적 달성을 위해 그대로 사용 필요(학습에 직접적으로 활용)	
		⑧ 치주염 영역		
1-1	성별		▪ 분석에 필요 없으므로 삭제·대체 가능	메타데이터
1-2	이름			
1-3	나이			
1-4	촬영날짜			

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

가명정보 활용목적	**구강 내 충치와 치주염 관련 질환을 진단하는 AI 모듈 개발 연구**	
가명처리 대상 데이터 항목	○ 구강촬영데이터(이미지) *<연번 1>* ○ 구강촬영데이터의 메타데이터(성별,이름,나이,촬영날짜) *<연번 1-1~1-4>*	
데이터 위험성	**식별성 유무**	**<구강촬영이미지>** ○**(구개, 혀, 윗잇몸, 아래잇몸)** 해당 부위는 개인식별 가능성 거의 없으며, 타 촬영부위와 연계되어도 식별위험은 높아지지 않을 것으로 판단됨 ○**(상악치, 하악치)** 구강구조 등을 통한 개인식별 가능성은 상당히 낮으나, 메타데이터와 결합되어 분석될 시 개인식별 가능성 존재 ○**(충치 영역, 치주염 영역)** 충치·치주염 이미지를 통한 개인식별 가능성은 상당히 낮으나, 메타데이터와 결합되어 분석될 시 개인식별 가능성 존재 **<메타데이터>** ○ 한국대학교병원 진료환자의 성별, 이름, 나이, 촬영날짜 등은 조합되었을 때 개인식별이 가능한 개인식별가능정보이며, 구강촬영이미지와 연계되면 개인식별 가능성이 더욱 높아질 수 있음
	특이정보 유무	**<구강촬영이미지>** ○**(구개, 혀, 윗잇몸, 아래잇몸)** 특이사항이 발생하기 어려운 부위임 ○**(상악치, 하악치)** 구강구조, 치열 등이 특이한 경우 개인식별 가능성 존재 ○**(충치 영역, 치주염 영역)** 심각한 구강질환이 다수 존재하는 경우, 특

		이한 시술 등이 이뤄진 경우 등엔 개인식별 가능성 존재 <메타데이터> ○'나이'의 경우 상당히 적거나 높은 경우 특이치로 인한 개인식별성이 발생할 수 있음
	재식별 시 영향도	○개인 구강에 대한 촬영정보는 재식별시 영향도는 낮은 편
처리 환경 위험성	**이용 및 제공 형태**	○기관외부 기업에게 제공 ○환자를 진료한 자와 관련 없는 제3자가 데이터를 활용하므로, 처리자가 보유한 경험·데이터로 인한 특정 환자 추정 위험성은 낮음
	처리 장소	○한국대학교병원에서 제공하는 클라우드 기반의 폐쇄연구분석환경이 갖춰진 분석실 ○한국대학교병원은 개인정보(가명정보)처리시스템에 대한 ISMS-P인증 취득
	다른 정보 결합 가능성	○폐쇄환경분석실 관리자 승인하에 제한된 데이터·프로그램(프로그램 패키지, 라이브러리, 코드설명서 등)만 반입 가능 ○분석대상 가명정보와 결합가능성 있는 데이터는 반입 제한
최종 검토의견		▪ 해당 연구는 그 자체로는 개인식별 위험성이 낮고, 재식별시 영향도가 미미한 구강촬영데이터를 다루는 연구로 전반적인 위험성이 낮은 편임 ▪ 구강구조, 특이한 시술 등으로 인한 특이정보가 발생할 수는 있으나, 해당 연구는 타 데이터의 반입 및 결합이 불가능한 클라우드 기반의 폐쇄연구 분석환경에서 연구될 뿐만 아니라, 데이터셋에 포함된 환자와 전혀 관련이 없는 제3자에 의해 연구되므로 환자에 대한 재식별·추정 가능성은 상당히 낮은 편임 ▪ 다만, 연구에는 충치·치주염 영역만 학습데이터로 활용되고, 촬영된 영역이 이미 라벨링을 통해 구분되어 있어 가명처리도 용이한 편이므로, 다음과 같이 최소한의 가명처리 후 제공 필요 - 구개, 혀 윗잇몸, 아랫잇몸 부분은 개인식별 가능성이 거의 없으므로 그대로 활용 가능 - 상악치, 하악치 부분은 연구에 필요 없고 구강구조, 치열 등이 특이한 경우 개인식별 가능성이 존재하므로 충치 영역을 제외한 부분만 블러링 처리하여 활용 - 충치 영역, 치주염 영역은 연구목적 달성을 위해 그대로 활용 가능 - 이미지의 메타데이터는 연구에 전혀 필요 없으며, 이미지와 연계되어 개인식별 위험성이 높아지므로 전체 삭제하여 활용

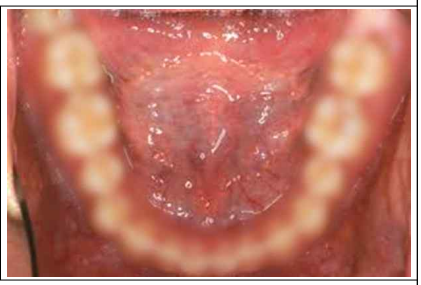
④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
1	구강 촬영데이터	① 구개	☑ 그대로 사용	별도 처리하지 않음
		② 혀	☑ 그대로 사용	별도 처리하지 않음
		③ 윗잇몸	☑ 그대로 사용	별도 처리하지 않음
		④ 아래잇몸	☑ 그대로 사용	별도 처리하지 않음
		⑤ 상악치	☑ 블러링 처리	충치 영역 외 부분 블러링 처리
		⑥ 하악치	☑ 블러링 처리	충치 영역 외 부분 블러링 처리
		⑦ 충치 영역	☑ 그대로 사용	별도 처리하지 않음
		⑧ 치주염 영역	☑ 그대로 사용	별도 처리하지 않음
1-1	성별		☑ 삭제	연구에 필요 없으므로 삭제
1-2	이름		☑ 삭제	연구에 필요 없으므로 삭제
1-3	나이		☑ 삭제	연구에 필요 없으므로 삭제
1-4	촬영날짜		☑ 삭제	연구에 필요 없으므로 삭제

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

주요 비정형데이터 가명처리 수행 결과 (예시)

<table>
<tr><th>연번</th><th>항목명</th><th>세부 항목</th><th>가명처리 전</th><th>가명처리 후</th></tr>
<tr><td rowspan="3">1</td><td rowspan="3">구강
촬영데이터</td><td>⑤ 상악치</td><td> </td><td> </td></tr>
<tr><td colspan="2">○ 충치 영역 외 부분 블러링 처리
- A사의 '이미지 가명처리 솔루션' 활용 (블러링 수준 3단계 적용)</td></tr>
<tr><td>⑥ 하악치</td><td> </td><td>  </td></tr>
<tr><td colspan="3">○ 충치 영역 외 부분 블러링 처리
- A사의 '이미지 가명처리 솔루션' 활용 (블러링 수준 3단계 적용)</td></tr>
</table>

○ 비정형 데이터 가명처리 기술의 적절성·신뢰성 관련 근거(예시)

<table>
<tr><th>연번</th><th>항목명</th><th>세부 항목</th><th>처리 기술명</th><th>처리 기술의 적절성·신뢰성 관련
근거 또는 배경</th></tr>
<tr><td rowspan="2">1</td><td rowspan="2">구강
촬영데이터</td><td>⑤ 상악치</td><td>블러링(부분)</td><td rowspan="2">▪ 한국대학교병원은 A사에서 구매한 이미지 가명처리 솔루션을 활용하여 상악치·하악치에서 충치 영역으로 라벨링되어 있는 영역 외의 부분을 부분 블러링 처리
- A사 이미지 가명처리 솔루션은 설정한 객체를 인식하여 블러링 처리 수준을 1~5단계로 설정 가능함
* A사의 솔루션 내부 테스트 결과, 객체 인식률 92%, 처리 정확도 97%로 측정
※ 객체 인식률, 처리 정확도(오류율) 증빙자료 별첨
- 현재 복원기술의 발전 수준 및 데이터 처리 환경(타 정보·복원기술 활용 불가) 등을 고려하여 블러링 수준을 3단계로 설정
* 블러링 수준의 적정성은 외부전문가를 과반 이상 포함한 적적성 평가 위원회를 구성하여 검토 완료
※ 적정성 검토위원회 결과보고서 및 회의록 별첨
- A사 가명처리 솔루션이 상악치·하악치를 인식하지 못하는 경우, 블러링 처리에 오류가 있는 경우 등이 존재하므로, 가명처리 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사 수행</td></tr>
<tr><td>⑥ 하악치</td><td>블러링(부분)</td></tr>
</table>

비정형데이터 가명처리 결과에 대한 자체 검증 결과서

<table>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="2">개요</td></tr>
<tr><td colspan="2">한국대학교병원에서 수집한 구강검진 촬영데이터 중 상악치·하악치 부분(충치영역 이외)을 A사의 '이미지 가명처리 솔루션'을 활용하여 블러링 처리</td></tr>
<tr><td>데이터 유형</td><td>이미지</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td>JEPG</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td>JEPG</td></tr>
<tr><td>데이터 규모</td><td>50,000장 (5,000명 * 10장)</td></tr>
<tr><td>데이터 크기(용량)</td><td>300GB</td></tr>
<tr><td>대상 데이터 항목명</td><td>구강검진 촬영데이터 <연번 1></td></tr>
<tr><td>가명처리 적용 기술</td><td>- 한국대학교병원에서 구매한 A사의 '이미지 가명처리 솔루션'을 활용하여 각 이미지의 상악치·하악치로 라벨링된 부분을 블러링 처리(3단계 수준)</td></tr>
<tr><td>자체 검증
기간</td><td colspan="2">2023년 1월 20일 ~ 2023년 1월 27일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="2">병원 내 임상시험센터 분석PC</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="2">(검증방법)
- 검증은 한국대학교병원 내부 가관리 개인정보보호부장의 주도하에 의료정보관리실장, 개인정보보호부 직원 4인이 함께 진행
- 실장 및 직원이 데이터를 나누어 블러링 처리 정상 수행 여부를 육안으로 검수하고, 특이사항이 발생한 표본만 선별하여 전체인원이 추가 합동검수·처리 수행

(검증 시 확인사항)
1. 구강검진 이미지 내 상악치·하악치가 다음과 같이 처리되었는지 전수 확인
① 충치영역으로 라벨링된 부분을 제외한 상악치·하악치 모든 영역이 블러링처리 되었는지?
② 충치영역으로 라벨링된 부분까지 블러링처리되지는 않았는지?
2. 위 사항 중 하나라도 만족하지 않았을 경우 수작업으로 영역을 지정한 뒤, 솔루션을 활용하여 블러링 처리를 수행
3. 처리가 제대로 되지 않는 표본은 최종 연구 표본에서 삭제처리</td></tr>
<tr><td>자체 검증
결과</td><td colspan="2">확인 결과 이상 없음
(검수 결과, 최초 1,241장이 제대로 처리되지 않았으나,
수작업을 통해 최종 가명처리 완료)</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>한국대학교 개인정보보호부장</td><td>가관리</td><td></td></tr>
<tr><td>한국대학교 의료정보관리실장</td><td>나정보</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>다보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>라보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>마보호</td><td></td></tr>
</table>

시나리오 ❸ : 안면골 골절 진단 AI 개발 사례

의료 분야 (이미지, 영상)

- 한국대학교병원은 임상시험센터는 한국대학병원 내 영상의학과 연구자와 헬스케어 분야 AI개발 스타트업인 ㈜퓨처비전데이터(공동연구 컨소시움)로부터 과학적 연구 수행을 위한 데이터 제공을 요청받아 안면골(얼굴뼈) 골절 관련 Facial CT 이미지·영상 등의 데이터를 제공하려 한다.

☑ 데이터의 이용 목적

○ 안면골 골절을 진단하는 AI 솔루션 개발 연구

☑ 데이터 특징

○ (영상·이미지 데이터) 안면골 골절 증상으로 CT검사를 받은 80,000명(2007년~2020년)의 환자에 대한 Facial CT 영상·이미지 파일 (DICOM포맷)

- DICOM 헤더정보 포함 (환자번호, 환자이름, 성별, 생년월일 등)

○ (증례기록지 데이터[정형데이터]) 안면골 골절 증상으로 Facial CT를 촬영한 환자에 대한 기본정보(차트번호, 검사일자, 성별, 나이)와 진단명(진단코드)

☑ 데이터의 이용 환경

○ (폐쇄연구분석환경 활용) 한국대학교병원에서 제공하는 클라우드 기반의 폐쇄연구분석환경이 갖춰진 분석실에서 데이터 활용, 승인된 사용자 외에는 접근 불가

○ (자료 반입) 자료 반입시 한국대학교병원 관리자에게 요청(관리자가 자료 확인 후 반입)

○ (자료 반출) 분석결과 반출 시 한국대학교병원 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	SG-20230110003	접수일	2023년 2월 1일

신청자	조직/부서명	한국대학교병원 영상의학과		
	담당자 직위	연구원	담당자 성명	강진단
	전화번호	03-1234-5678	이메일주소	kjd@hankook.ac.kr

처리목적		☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존
		안면골 골절에 대한 진단 AI 및 영상분석 솔루션 개발
활용 형태		☑ 내부이용* ☐ 제3자 제공 ☐ 결합전문기관을 통한 결합 * 한국대병원 영상의학과-(주)퓨처비전데이터(처리위탁) 컨소시움을 통한 공동연구
이용 주체		☑ 동일 개인정보처리자 ☐ 제3자(제공받는 자) 제공
처리 환경		☑ 내부 ☐ 외부
처리 장소		☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부		☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법		☑ 온라인* ☐ 오프라인 * 폐쇄형연구분석환경에서만 접속가능한 클라우드 서버에 업로드
제공 받는 자		(내부) ☐ 동일부서 ☑ 타 부서 (외부) ☐ 제3자, 결합전문기관 제공
신청 명세	가명처리 대상 명칭	Facial CT 영상・이미지(비정형) 및 증례기록지(정형)
	데이터 내역	2007년부터 2020년까지 안면골절의 진단을 위해 Facial CT 검사를 받은 사람 중 안면골절 환자 40,000명과 정상인자 40,000명의 CT 영상・이미지 및 증례기록
	가명정보 이용기간	2023년 3월 1일 ~ 2025년 3월 1일 (3년)

위와 같이 가명정보 이용·제공을 신청합니다.

2023년 2월 1일

신청인(부서장명) 유관순 (서명 또는 인)

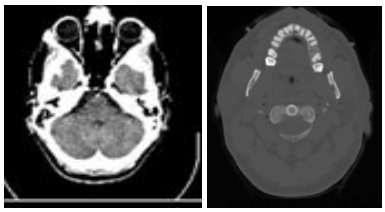
첨부서류
1. 연구계획서
2. 개인정보 유형 분류표
3. 가명정보 처리 위탁 계약서 등
4. 가명정보에 대한 안전조치 이행 확약서
5. 보안 서약서 등

② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

과학적 연구 계획서

연구명	**안면골 골절을 진단하는 AI 솔루션 개발 연구**
연구진	한국대학교병원 영상의학과 강진단 연구원(연구책임자) ㈜ 퓨처비전데이터 데이터연구팀(공동연구 참여, 처리위탁 : AI모듈 개발)
연구 배경 및 목적	○ 2020년 심사평가원에서 제공한 보건의료 통계에 따르면 연간 70만명에 이르는 안면골 골절 환자가 발생하고 있음 ○ 안면골 골절은 사고나 상해 등의 이유로 매우 빈번하게 일어나므로 CT 촬영이 가능한 모든 의료기관에서의 진단이 필요하나 많은 병원에서 실시간 판독이 가능한 전문의가 없는 경우가 많음 ○ 안면골 골절 진단이 가능한 영상의학과, 성형외과, 구강외과 등의 전문의가 없더라도 당직 의사에게 빠른 의사결정을 위한 안면골 골절을 진단할 수 있는 AI솔루션 개발 필요
예상 연구 기간	2023년 3월 1일 ~ 2025년 3월 1일 (3년)
연구 대상자 수	2007년부터 2020년까지 한국대학교병원에서 안면골 골절 의심 증상으로 Facial CT검사를 받은 환자의 CT촬영 영상 80,000건
연구 방법	○ 본 연구는 후향적 연구로, 안면골 골절에 대한 AI를 이용한 판독을 지원하는 AI솔루션을 개발하는 것을 목표로 함 ○ AI 기계학습과 대조군을 이용한 검증 수행
연구 내용	○ 안면골 골절 증상으로 CT검사를 받은 80,000건의 환자 데이터 활용 - 안면골 골절진단을 받은 40,000건을 학습데이터 사용하고 대조군의 40,000건을 대상으로 검증 시행 예정 ○ 기계학습을 통해 S02의 다양한 진단에 대해 학습하고 검증분석할 예정임
기대효과 및 활용방안	○ Facial CT영상 분석을 통해 성형외과, 정형외과, 영상의학과 전문의가 없는 CT를 보유하고 있는 병원에서 안면골 골절에 대한 진단에 도움을 받을 수 있음 ○ 안면골 골절의 정확한 영상 판독을 통한 적절한 치료로 후유증 발생률 최소화 가능
붙임. 상세 연구계획서 등	

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

개인정보 유형 분류표 (요약)					
연번	항목명	데이터 유형	데이터 규모	예시	비고
1	Facial CT 데이터	영상·이미지 (DICOM*)	80,000건 (97TB) - 안면골 골절 판정 데이터 (40,000건) - 정상 판정(안면골 골절X) 데이터 (40,000건)		**비정형데이터 (DICOM 헤더정보 포함)**
2	증례기록지	텍스트·숫자 등 (정형)	80,000건 (52MB)	차트번호, 검사일자, 환자성별, 환자나이, 진단코드 포함	**정형데이터**

* DICOM(Digital Imaging and Communications in Medicine) : 의료용 디지털 영상 및 통신 표준

개인정보 유형 분류표 (상세)				
연번	항목명	구분	설명	예시
1	Facial CT 데이터	① 안면부	▪ 안면골 골절 여부 판단에 필요한 안면부 촬영부분	
		② 후두부 (뇌 뒷부분)	▪ 안면골 골절 여부 판단에 필요없는 후두부(뇌 뒷부분) 촬영부분	
1-1	DICOM 헤더정보	환자 번호	▪ 환자 구분을 위해 고유하게 부여되는 번호	P158687
1-2		환자 이름	▪ 환자의 이름	김철수
1-3		환자 성별	▪ 1: 남자, 2: 여자	1
1-4		환자 생년월일	▪ 환자의 생년월일 정보	1968.5.26.
2-1	증례기록지	① 차트번호	▪ 환자진료기록 차트에 고유하게 부여되는 번호로, 각 환자의 Facial CT 사진에 매핑하여 활용	01768062
2-2		② 검사일자	▪ 안면골 골절에 대한 Facial CT 검사일자	2019-06-23
2-3		③ 환자 성별	▪ 1: 남자, 2: 여자	2
2-4		④ 환자 나이	▪ 1세 단위, 12세~98세까지 분포	58
2-5		⑤ 진단명 (진단코드)	▪ S02 두개골 및 안면골의 골절 ▪ S02.0 두개원개의 골절 ▪ S02.1 두개저의 골절 ▪ S02.2 비골의 골절 ▪ S02.3 안와바닥의 골절 ▪ S02.4 광대뼈 및 상악골의 골절 ▪ S02.5 치아의 파절 ▪ S02.6 하악골의 골절 ▪ S02.7 두개골 및 안면골을 침범한 다발골절 ▪ S02.8 기타 두개골 및 안면골의 골절 ▪ S02.9 두개골 및 안면골의 상세불명 부분의 골절	S02

활용데이터 요구 수준표

※ 가명정보 처리 목적 달성에 필요한 데이터 항목과 가명처리 요구수준을 검토

연번	항목명	구분	요구 수준	비고
1	Facial CT 데이터	① 안면부	▪ 연구목적 달성에 필수적인 정보로 그대로 활용 필요	이미지 (비정형데이터)
		② 후두부 (뇌뒷부분)	▪ 연구목적 달성에 필요하지 않으므로, 마스킹 기법 등을 통해 가명처리해도 상관 없음	
1-1	DICOM 헤더정보	환자 번호	▪ 연구에 필요 없는 정보로 삭제·대체 가능	메타데이터
1-2		환자 이름	▪ 연구에 필요 없는 정보로 삭제·대체 가능	
1-3		환자 성별	▪ 연구에 필요한 정보이나, 증례기록지를 통해 확인 가능한 정보이므로 삭제·대체 가능	
1-4		환자 생년월일	▪ 연구에 필요한 정보이나, 증례기록지를 통해 확인 가능한 정보이므로 삭제·대체 가능	
2-1	증례기록지	① 차트번호	▪ 환자 구분을 위해 필요한 정보로, 단순일련번호로 대체 가능	정형데이터
2-2		② 검사일자	▪ 연구에 반드시 필요한 정보이며, 연 단위까지 범주화 가능	
2-3		③ 환자 성별	▪ 연구에 반드시 필요한 정보이며, 그대로 사용 필요	
2-4		④ 환자 나이	▪ 연구에 반드시 필요한 정보이며, 10세 단위까지 범주화 가능	
2-5		⑤ 진단명 (진단코드)	▪ 연구에 반드시 필요한 정보이며, 그대로 사용 필요	

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

<table>
<tr><td>가명정보
활용목적</td><td colspan="2">안면골 골절에 대한 진단 AI 및 영상분석 솔루션 개발</td></tr>
<tr><td>가명처리
대상 데이터
항목</td><td colspan="2">○ Facial CT 영상·이미지(비정형데이터) <연번 1>
○ Facial CT 영상·이미지의 메타데이터(DICOM헤더) <연번 1-1~1-4>
○ 증례기록지(정형데이터) <연번 2></td></tr>
<tr><td rowspan="4">데이터
위험성</td><td rowspan="2">식별성
유무</td><td><Facial CT 영상·이미지>
○ 데이터 그 자체로는 개인식별 위험성이 상당히 낮은 편이나, 개인에 대한 대용량의 CT 영상·이미지에 대해 3차원 재건 등의 기술을 활용하면 얼굴 외형을 입체적으로 복원가능하며, 복원시 특이한 얼굴·외형 등이 있는 경우, 연예인 등 유명인인 경우 등엔 낮은 확률로 식별 위험이 생길 수 있음
* 가장자리 마스킹 기법을 활용하여 3차원 재건 공격 위험을 막을 수 있음
<Facial CT DICOM 헤더정보>
○ 환자번호, 환자이름, 생년월일 정보는 다른 항목, 다른 정보와 결합될 시 개인식별 가능성이 있어 가명처리 필요
<증례기록지>
○ 차트번호, 환자이름은 환자 개인식별정보로, 삭제 또는 대체 필요
○ 환자 성별, 검사일자, 환자 나이 정보는 다른 정보와 결합되어 개인을 식별할 수 있는 가능성이 존재</td></tr>
<tr><td>특이정보
유무</td><td>○ Facial CT 영상·이미지에 특이한 안면골 골절 사항 등이 존재할 수는 있으나, 데이터셋에 포함된 환자와 관련 없는 제3자에 의해 연구될 경우 식별가능성이 거의 없음</td></tr>
<tr><td>재식별시
영향도</td><td>○ 안면골 CT 촬영정보는 재식별시 영향도는 낮은 편</td></tr>
<tr><td rowspan="3">처리 환경
위험성</td><td>이용 및
제공
형태</td><td>○ 한국대학교병원 영상의학과 강진단 교수 연구팀 및 ㈜퓨처비전데이터가 함께 분석할 예정으로 자체 활용과 제공(위탁)의 형태를 모두 가짐
○ 다만 사업주체가 컨소시움형태로 데이터의 제3자 제공은 해당하지 않으며, 공동의 목적 달성을 위해 한국대학교병원 내 폐쇄공간 내에서만 데이터가 활용되므로 비교적 관리가 용이한 편
○ 클라우드 분석환경에서의 데이터 접근 권한은 연구 컨소시움 참여자 6명에게만 부여되며, 데이터 수정 권한은 강진단 교수에게만 부여</td></tr>
<tr><td>처리 장소</td><td>○한국대학교병원에서 제공하는 클라우드 기반의 폐쇄연구분석 환경이 갖춰진 분석실
○한국대학교병원은 개인정보(가명정보)처리시스템에 대한 ISMS-P인증 취득</td></tr>
<tr><td>다른 정보
결합
가능성</td><td>○ 다른 정보와의 연계 분석이나 결합은 예정되어 있지 않음
○ 폐쇄환경분석실 관리자 승인하에 제한된 데이터·프로그램(프로그램 패키지, 라이브러리, 코드설명서 등)만 반입 가능
○ 분석대상 가명정보와 결합가능성 있는 데이터는 반입 제한</td></tr>
</table>

최종 검토의견	▪ 해당 연구는 그 자체로는 개인식별 위험성이 낮고, 재식별시 영향도가 미미한 Facial CT 촬영데이터를 다루는 연구로 전반적인 위험성이 낮은 편임 ▪ Facial CT 영상·이미지를 대용량으로 활용하므로 3차원 재건 시 복원 위험이 일부 있을 수는 있으나, 타 데이터의 반입 및 결합이 불가능한 클라우드 기반의 폐쇄연구 분석환경에서 연구될 뿐만 아니라, 데이터셋에 포함된 환자와 전혀 관련이 없는 타부서 연구자, 제3자((주)퓨처비전데이터)에 의해 연구되므로 환자에 대한 재식별·추정 가능성은 상당히 낮은 편임 - 다만, 연구에는 '안면부 부분'만 활용되므로 필요없는 '후두부 부분'은 가장자리 마스킹 처리하여 활용하는 것이 바람직 ▪ Facial CT DICOM 헤더정보는 연구에 필요 없는 정보('환자번호', '환자이름')이거나, 증례기록지를 통해 확인이 가능한 정보('환자 성별', '생년월일')이므로 삭제 ▪ 증례기록지의 '차트번호'는 환자구분에 필요하므로 단순일련번호로 변환하여 사용하고, 검사일자, 환자 나이 등은 연구 목적이 달성한 수준으로 범주화하여 활용 ▪ 연구목적 달성에 반드시 그대로 사용이 필요한 '환자 성별'과 '진단코드'는 다른 정보 항목들이 충분히 가명처리된 경우, 그대로 활용해도 식별 위험이 크지 않음

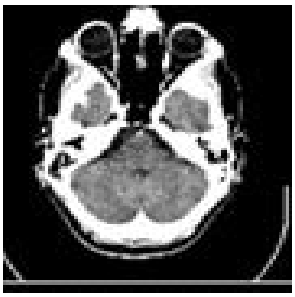
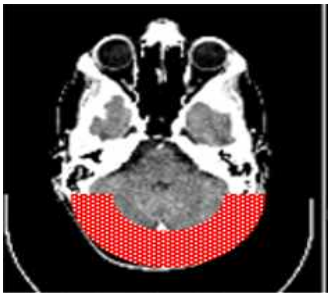
④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
1	Facial CT 데이터	① 안면부	☑ 그대로 사용	별도 처리하지 않음
		② 후두부 (뇌뒷부분)	☑ 마스킹	연구에 필요 없으므로, 마스킹 처리하여 안전하게 활용
1-1	DICOM 헤더정보	환자 번호	☑ 삭제(포맷변환)	연구에 필요 없으므로 삭제 - DICOM 포맷을 TIFF 포맷으로 변경하여 저장(헤더 삭제)
1-2		환자 이름	☑ 삭제(포맷변환)	
1-3		환자 성별	☑ 삭제(포맷변환)	증례기록지를 통해 확인 가능한 정보이므로 삭제 - DICOM 포맷을 TIFF 포맷으로 변경하여 저장(헤더 삭제)
1-4		환자 생년월일	☑ 삭제(포맷변환)	
2-1	증례기록지	① 차트번호	☑ 대체	환자 구분을 위해 단순일련번호로 대체
2-2		② 검사일자	☑ 범주화	연 단위로 범주화
2-3		③ 환자 성별	☑ 그대로 사용	별도 처리하지 않음
2-4		④ 환자 나이	☑ 범주화 (10세 단위)	10세 단위 범주화 적용, 90세 이상은 90대로 상단코딩 적용
2-5		⑤ 진단명 (진단코드)	☑ 그대로 사용	별도 처리하지 않음

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

주요 비정형데이터 가명처리 수행 결과 (예시)

연번	항목명	가명처리 전	가명처리 후
1	Facial CT 데이터		

가명처리 결과 자체검증

○ 비정형 데이터 가명처리 기술의 적절성・신뢰성 관련 근거(예시)

연번	대상 항목	처리 기술명	처리 기술의 적절성·신뢰성 관련 근거 또는 배경	비고
1	Facial CT 데이터	표면 가장자리 마스킹	▪ 한국대학교병원은 B사에서 구매한 이미지 가명처리 솔루션을 활용하여 후두부 영역만 부분 마스킹 처리하였음 - B사 이미지 마스킹 솔루션은 설정한 객체를 인식하여 설정한 영역에 대해 부분적으로 마스킹 처리 가능 * B사의 솔루션 내부 테스트 결과, 객체 인식률 92%, 처리 정확도 98%로 측정 *※ 객체 인식률, 처리 정확도(오류율) 증빙자료 별첨* - B사 마스킹 솔루션이 후두부 부위를 제대로 인식하지 못하는 경우, 마스킹 처리에 오류가 있는 경우 등이 존재할 수 있으므로 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사 수행	

비정형데이터 가명처리 결과에 대한 자체 검증 결과서

<table>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="2">개요</td></tr>
<tr><td colspan="2">한국대학교병원에서 수집한 Facial CT 데이터(DICOM 포맷을 TIFF 포맷으로 변환)에 대해 연구에 필요하지 않은 후두부(뇌 뒷부분) 영역을 B사의 '이미지 마스킹 솔루션'을 사용하여 마스킹 처리</td></tr>
<tr><td>데이터 유형</td><td>영상·이미지</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td>DICOM</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td>TIFF</td></tr>
<tr><td>데이터 규모</td><td>80,000건</td></tr>
<tr><td>데이터 크기(용량)</td><td>97TB</td></tr>
<tr><td>대상 데이터 항목명</td><td>Ficial CT 데이터 <연번 1></td></tr>
<tr><td>가명처리 적용 기술</td><td>- 한국대학교병원에서 구매한 B사의 '이미지 마스킹 솔루션'을 활용하여 각 환자의 후두부 영역만 마스킹 처리</td></tr>
<tr><td>자체 검증
기간</td><td colspan="2">2023년 2월 20일 ~ 2023년 2월 27일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="2">회의실 (원격으로 분석실의 가상컴퓨터 접속)</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="2">(검증방법)
- 검증은 한국대학교병원 내부 가관리 개인정보보호부장의 주도하에 의료정보관리실장, 개인정보보호부 직원 4인이 함께 진행
- 실장 및 직원이 데이터를 나누어 환자관련정보 가명처리 정상 수행 여부를 육안으로 검수하고, 특이사항이 발생한 표본만 선별하여 전체인원이 추가 합동검수·처리 수행
(검증 시 확인사항)
1. Facial CT 후두부 영역이 다음과 같이 처리되었는지 전수 확인
① 후두부 영역 전체가 확인 불가능하도록 마스킹처리 되었는지?
② 안면부 영역까지 마스킹처리되지는 않았는지?
2. 위 사항 중 하나라도 만족하지 않았을 경우 수작업으로 영역을 지정한 뒤, 솔루션을 활용하여 마스킹 처리 수행
3. 처리가 제대로 되지 않는 표본은 최종 연구 표본에서 삭제처리</td></tr>
<tr><td>자체 검증
결과</td><td colspan="2">확인 결과 이상 없음
(검수 결과, 4,235장이 제대로 처리되지 않았으나,
수작업을 통해 최종 가명처리 완료)</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>한국대학교 개인정보보호부장</td><td>가관리</td><td></td></tr>
<tr><td>한국대학교 의료정보관리실장</td><td>나정보</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>다보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>라보호</td><td></td></tr>
<tr><td>한국대학교 개인정보보호부</td><td>마보호</td><td></td></tr>
</table>

시나리오 ❹ : 자율주행차 주행 시 비정상 상황인지 AI 개발 사례

교통 분야 (이미지, 영상)

- 한국영상연구원은 국가 R&D 사업의 일환으로 자율주행차를 통해 한국대학교 교내 도로 주행영상을 촬영하여 보유하고 있다(근거법률:「개인정보 보호법」 제25조의2 제2호). 한국영상연구원은 자율주행차의 주행환경에 대한 비정상 상황인지 AI 기술개발을 수행하고자 하는 ㈜한국자율테크에 해당 영상을 제공하고자 한다.

☑ 데이터의 이용 목적

○ 자율주행자동차의 주행환경에 대한 비정상 상황 인지 AI 기술 연구

○ 과학적 연구 목적으로 가명처리 후 AI 기술개발을 위한 인공지능 신경망 학습용 데이터로 활용

☑ 데이터 특징

○ (이미지·영상 데이터) '22년 1월~2월 동안 자율주행차가 한국대학교 교내 도로 주행 상황을 촬영한 영상

- 차량 탑재 카메라로부터 MP4데이터를 취득하고 이를 PNG파일로 생성 (327GB, 5만장)
- 이미지·영상 내 객체는 정적 객체(신호등, 도로 등)와 동적 객체(사람, 차량, 현수막 등)으로 구성
- 이미지·영상과 관련된 메타데이터는 별도로 생성·저장되지 않음

☑ 데이터의 이용 환경

○ (한국영상연구원 내부 분석실 이용) USB보안 기능이 설치된 PC에서 관리하며, 이용자는 외부망이 차단된 PC에서 이용 및 활용

* 물리적 망분리, 가상 단말기(PC) 보안 준수

○ (자료 반입) 자료 반입(오프라인으로 전달)시 한국영상연구원 관리자에게 요청(관리자가 자료 확인 후 반입), 사전 특정된 연구원 외 데이터 접근 불가

○ (자료 반출) 분석결과 반출 시 한국영상연구원 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	IMAGEINSTI_001_002	접수일	2023년 01월 03일

신청자	조직/부서명	한국영상연구원 / 자율주행연구실		
	담당자 직위	책임	담당자 성명	이테크
	전화번호	02-123-4567	이메일주소	lee_korea@imageinti.re.kr

처리목적	☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존 자율주행차의 비정상 주행 상황인지 AI 개발
활용 형태	☐ 내부이용 ☑ 제3자 제공 ☐ 결합전문기관을 통한 결합
이용 주체	☐ 동일 개인정보처리자 ☑ 제3자(제공받는 자) 제공
처리 환경	☑ 내부 ☑ 외부
처리 장소	☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부	☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법	☐ 온라인 ☑ 오프라인
제공 받는 자	(내부) ☐ 동일부서 ☐ 타부서 (외부) ☑ 제3자, 결합전문기관 제공

신청 명세	가명처리 대상 명칭	도로 주행 영상 데이터
	데이터 내역	한국대학교 교내에서 주간과 야간, 새벽 시간대에 취득한 도로 주행 영상 데이터
	가명정보 이용기간	2023년 02월 01일 ~ 2023년 12월 31일(11개월)

위와 같이 가명정보 이용·제공을 신청합니다.

2023년 01월 03일

신청인(부서장명) 김 부 장 (서명 또는 인)

첨부서류	1. 연구계획서 2. 개인정보 유형 분류표 3. 가명정보에 대한 안전조치 이행 확약서 4. 보안 서약서 등

② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

과학적 연구계획서

연구명	자율주행차의 주행환경에 대한 비정상 상황 인지 기술개발	
연구진	소속	㈜한국자율테크
	연구책임자	이 테 크
연구 배경 및 목적	o ㈜한국자율테크에서는 2023년 1월부터 산업부 R&D 사업의 일환으로 클라우드 기반 자율주행차 인지 기술 개발을 진행하고 있음 o 상기 과제에서 본 기관은 자율주행차의 주행 경로상에서 발생할 수 있는 비정상 상황을 인지하여 자율주행차가 원활한 주행을 할 수 있도록 상황 인지 SW 개발을 추진하고 있음 o 본 데이터를 자율주행차의 주행 경로상에서 발생할 수 있는 비정상적인 상황을 인지할 수 있는 인공지능 SW 개발의 학습데이터로 활용하고자 함	
예상 연구 기간	2022년 02월 01일 ~ 2023년 12월 31일 (11개월)	
연구 대상자 수	2022년 1월부터 2월까지 한국대학교 교내 캠퍼스에서 새벽, 주간, 야간에 수집한 자율주행 영상(20시간)·이미지(50,000장) 데이터	
연구 방법	o 데이터의 학습에 필요한 대상 객체를 가공(라벨링)하여 활용 o 상황 인지를 위한 딥러닝 인공지능 알고리즘의 학습용 데이터를 활용하며, 일부 데이터는 테스트용으로 활용함	
연구내용	o 자율주행차 주행환경 비정상 상황인지 기술개발 ① 주행 경로 동적, 정적 객체인지 - 자율주행차의 OEDR와 OEDR에 명시된 상황을 통해 예외 상황을 정의 - 예외 상황을 구성하는 객체(사람, 자동차, 쓰레기 더미 등)를 분류하고, 가명처리 데이터에서 학습용으로 활용할 수 있는 객체들을 구분함 - 상기 분류된 객체들을 가명처리 데이터에서 라벨링하고, 인공지능 학습 데이터로 활용 ② 비정상 상항인지 기술 개발 - 상기에서 인지한 객체 정보를 활용하고, 객체 간의 관계와 시간, 장소, 날씨 등의 연관 관계를 통해 OEDR에 명시된 예외 상황 인지	
기대효과 및 활용방안	o 자율주행차의 상용화를 위한 인지 학습데이터로 활용함으로써 향후, 자율주행 모빌리티 산업의 활성화에 기여할 수 있음 o 가명처리 데이터는 비정상 상황인지와 더불어 자율주행차의 동적 객체 인지에 활용될 수 있음	

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

<table>
<tr><th colspan="4">개인정보 유형 분류표</th></tr>
<tr><th colspan="2">항목명</th><th>데이터 유형</th><th>데이터 규모</th></tr>
<tr><td colspan="2">자율주행차 촬영
영상 및 이미지</td><td>영상(MP4),
이미지(PNG)</td><td>영상: 20시간 분량
이미지 : 50,000장(327GB)</td></tr>
<tr><th>연번</th><th>항목명</th><th>설명</th><th>예시</th></tr>
<tr><td>1</td><td>사람 전체형상</td><td>사람의
신체·형상 전체</td><td> </td></tr>
<tr><td>2</td><td>사람 얼굴</td><td>사람의
정면과 측면 얼굴</td><td>
※ 이해를 돕기 위해 해당 부분을 확대하여 표시하였음</td></tr>
<tr><td>3</td><td>차량 전체형상</td><td>차량의 형상 전체</td><td> </td></tr>
<tr><td>4</td><td>차량 번호판</td><td>차량의 번호판 영역</td><td>
※ 이해를 돕기 위해 해당 부분을 확대하여 표시하였음</td></tr>
</table>

활용데이터 요구 수준표

※ 가명정보 처리 목적 달성에 필요한 데이터 항목과 가명처리 요구수준을 검토

연번	항목명	요구 수준	비고
1	사람 전체형상	▪ 연구목적 달성을 위해 사람의 전체형상은 최대한 보존되어야 함 ▪ 해당 영역 분석을 통해 사람인지 아닌지 여부를 판단할 수 있어야 하며, 사람이 움직이는 방향·거리·속도 정보를 추출할 수 있어야 함	비정형 데이터
2	사람 얼굴	▪ 비정상적인 주행 상황 인지와 직접적인 관련이 없는 정보로, 일부 마스킹 가능 ▪ 다만, 얼굴 영역이 지나치게 과도하게 가명처리되어 AI가 사람인지 아닌지 여부를 판단할 수 없어질 경우 학습데이터로서의 가치가 무의미해질 수 있음	
3	차량 전체형상	▪ 차량의 차종(승용, SUV, 버스 등)을 구분할 수 있어야 하며, 차량의 움직이는 방향·거리·속도 정보를 추출할 수 있어야 함	
4	차량 번호판	▪ 비정상적인 주행 상황 인지와 관련 없는 정보로, 삭제되어도 무방함	

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

가명정보 활용목적	자율주행차 비정상 주행 상황 인지 AI 개발	
가명처리 대상 데이터 항목	한국대학교 교내 주행 촬영영상 (새벽, 주간, 야간) *<연번 1~4>*	
데이터 위험성	**식별성 유무**	○ 영상 내 사람 얼굴, 차량번호판 정보는 개인식별 위험성이 존재
	특이정보 유무	○ (사람) 개인의 특이한 신체적·외형적 특징으로 인한 개인 식별 가능성 존재 ○ (차량) 차량의 경우 특이한 차종(희귀 슈퍼카), 특이한 색상 등으로 인한 개인식별 가능성 존재
	재식별시 영향도	○ 교내의 일반적인 주행상황에 대해 촬영된 영상으로, 재식별에 대한 영향도가 높지 않을 것으로 판단됨
처리 환경 위험성	**이용 및 제공 형태**	○ 한국영상연구원에서 보안 USB(암호 처리)를 통해 가명처리된 영상 데이터를 제공
	처리 장소	○ 한국영상연구원 연구실의 외부망이 차단된 PC에서 연구를 수행하며, 철저한 접근권한 통제 및 자료 반입·반출 수행
	다른 정보 결합 가능성	○ 관리자 승인하에 제한된 데이터·프로그램(프로그램 패키지, 라이브러리, 코드설명서 등)만 반입 가능 ○ 분석대상 가명정보와 결합가능성 있는 데이터는 반입 제한
최종 검토의견	○ 영상 내 사람 얼굴, 차량번호판만 마스킹 처리하여 활용 ○ 한정된 공간(교내)의 일반적인 주행상황을 촬영한 정보로 재식별시 영향도가 크지 않으며, 다른 정보와 결합이 불가능하도록 처리환경을 통제하고 있는 점 등을 고려하였을 때 사람얼굴, 차량번호판 블러링 처리 외에 특이정보 삭제 등 별도의 추가 가명처리는 불필요할 것으로 판단됨	

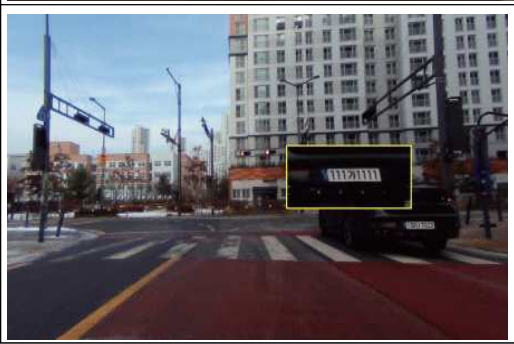
④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

연번	항목명	처리방법	세부방법 및 처리수준
1	사람 전체형상	☑ 그대로 사용	▪ 연구목적 달성을 위해 반드시 필요하며, 처리환경의 안전성을 고려하였을 때 개인식별 위험이 높지 않으므로 그대로 사용
2	사람 얼굴	☑ 마스킹	▪ 얼굴 영역을 사람과 컴퓨터가 식별 불가능한 수준으로 마스킹 처리
3	차량 전체형상	☑ 그대로 사용	▪ 연구목적 달성을 위해 반드시 필요하며, 처리환경의 안전성을 고려하였을 때 개인식별 위험이 높지 않으므로 그대로 사용
4	차량 번호판	☑ 마스킹	▪ 차량 번호판 영역을 사람과 컴퓨터가 식별 불가능한 수준으로 마스킹 처리

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

주요 비정형데이터 가명처리 수행 결과 (예시)

연번	항목명	가명처리 전	가명처리 후
2	사람 얼굴		
4	차량 번호판		

※ 이해를 돕기 위해 해당 부분을 확대하여 표시하였음

가명처리 결과 자체검증

○ 비정형 데이터 가명처리 기술의 적절성・신뢰성 관련 근거(예시)

<table>
<tr><th>연번</th><th>항목명</th><th>처리 기술명</th><th>처리 기술의 적절성·신뢰성 관련
근거 또는 배경</th></tr>
<tr><td>2</td><td>사람 얼굴</td><td rowspan="2">블러링(부분)</td><td rowspan="2">▪ 한국영상연구원은 자체보유하고 있는 '임베디드 장치 기반 영상데이터 개인정보 비식별화 시스템'을 활용하여 사람얼굴 영역과 차량 번호판 영역에 대해 마스킹을 수행
※ 해당 시스템은 아래와 같이 10가지 객체 종류 인식 가능
<table>
<tr><th>구분</th><th>연번</th><th>종류</th></tr>
<tr><td rowspan="4">사람얼굴</td><td>1</td><td>앞면(마스크 미착용)</td></tr>
<tr><td>2</td><td>앞면(마스크 착용)</td></tr>
<tr><td>3</td><td>측면(마스크 미착용)</td></tr>
<tr><td>4</td><td>측면(마스크 착용)</td></tr>
<tr><td rowspan="6">차량번호판</td><td>5</td><td>최신 일반(흰색)</td></tr>
<tr><td>6</td><td>이전 일반(흰색) 1:2사이즈</td></tr>
<tr><td>7</td><td>이전 일반(녹색) 1:2사이즈</td></tr>
<tr><td>8</td><td>특수 - 영업(노란색)</td></tr>
<tr><td>9</td><td>특수 - 이륜차(흰색)</td></tr>
<tr><td>10</td><td>특수 - 친환경차(파란색)</td></tr>
</table>
- 현재 복원기술의 발전 수준 및 데이터 처리 환경(타 정보·복원기술 활용 불가) 등을 고려하여 사람과 컴퓨터가 식별 및 복원이 불가능하도록 마스킹 처리
* 마스킹 방법 및 수준의 적정성은 외부전문가를 과반 이상 포함한 적적성 평가 위원회를 구성하여 검토 완료
※ 적정성 검토위원회 결과보고서 및 회의록 별첨
- 600개 파일을 통한 테스트 결과, 시스템의 객체 인식률은 99.75%, 마스킹 정확도 99.87%로 측정
※ 객체 인식률, 처리 정확도(오류율) 관련 한국기계전기전자시험연구원의 공인인증시험 성적서 별첨
- 사람 얼굴 및 차량 번호판을 인식하지 못하는 경우나 마스킹 처리가 제대로 되지 않는 경우가 존재하므로, 시스템 적용 후 처리결과에 대해 추가적인 자체 전수검사 수행</td></tr>
<tr><td>4</td><td>차량 번호판</td></tr>
</table>

비정형데이터 가명처리 결과에 대한 자체 검증 결과서

<table>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="2">개요</td></tr>
<tr><td colspan="2">과학적연구 목적으로 2022년 1월부터 2월까지 한국대학교 교내에서 촬영한 자율주행 영상데이터 중 사람얼굴 영역과 차량 번호판 영역 마스킹 처리</td></tr>
<tr><td>데이터 유형</td><td>영상·이미지</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td>MP4, PNG</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td>MP4, PNG</td></tr>
<tr><td>데이터 규모</td><td>영상: 20시간 분량
이미지 : 50,000장</td></tr>
<tr><td>데이터 크기(용량)</td><td>327GB</td></tr>
<tr><td>대상 데이터 항목명</td><td>도로주행 영상데이터</td></tr>
<tr><td>가명처리 적용 기술</td><td>‘임베디드 장치 기반 영상데이터 개인정보 비식별화 시스템’을 활용하여 영상의 사람 얼굴과 차량번호판을 인식하고마스킹 기법을 적용하여 가명처리</td></tr>
<tr><td>자체 검증
기간</td><td colspan="2">2023년 01월 10일 ~ 2023년 01월 20일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="2">한국영상연구원 내(외부 망 차단) 보안이 적용된 PC(보안 USB)</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="2">1. 전수조사 및 조사 보고서 작성
- 총 5명의 검수자가 10일간 SW 적용 후 육안 전수조사
➜ 객체가 인식되지 않아 마스킹 처리가 안된 경우 검토
➜ 마스킹 처리가 제대로 되지 않아 얼굴·차량번호판이 식별 가능한 경우 검토
- 전수조사 이후 마스킹 처리되지 않은 파일은 마스킹 SW로 재처리하였으며, 재처리되지 않는 파일은 수작업을 통해 마스킹 처리함
- 보고서 작성
2. 조사 내용 검토
- 조사 보고서에 대해 책임자가 검토 후 조사 완료</td></tr>
<tr><td>자체 검증
결과</td><td colspan="2">항목별 가명처리 계획에 맞게 처리되었음을 확인함</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>자율주행연구실 책임</td><td>가관리</td><td></td></tr>
<tr><td>자율주행연구실 선임</td><td>나정보</td><td></td></tr>
<tr><td>자율주행연구실 선임</td><td>다보호</td><td></td></tr>
<tr><td>자율주행연구실 주임</td><td>라보호</td><td></td></tr>
<tr><td>자율주행연구실 주임</td><td>마보호</td><td></td></tr>
</table>

시나리오 ❺ : 고속도로 다인승전용차로 단속 AI 개발 사례

교통 분야 (이미지)

- A지자체의 교통정보센터는 교통단속 및 교통정보의 수집·분석 등을 위하여 고속도로에 CCTV를 설치하고 고속도로 통행차량 영상·이미지를 촬영하여 보관하고 있다. (근거법률: 「개인정보 보호법」 제25조 제1항 제4호, 제5호). A지자체는 AI솔루션 개발 전문기업 COREA-AI사에 과학적 연구 수행을 위한 데이터 제공을 요청받아 해당 고속도로 통행차량 영상·이미지를 제공하려 한다.

☑ 데이터의 이용 목적

○ 재차인원 검지기반 다인승전용차로 단속 AI 개발 (학습데이터로 활용)

○ (이미지 데이터) 고속도로 통행 차량을 촬영한 이미지 (120,000장)

- 재차인원 검지만을 위해 별도 제작된 카메라를 활용해 촬영된 사진으로, 차량번호판이 포함된 차량 앞부분은 촬영되지 않음

- 탑승된 인원의 얼굴이 함께 촬영된 경우가 많음
  * 얼굴이 선명하게 촬영된 경우, 흐릿하게 촬영된 경우, 그림자·장애물 등으로 절반만 보이는 경우 등 다양

- 각 이미지마다 메타데이터 존재(촬영일시, 촬영장소)

☑ 데이터의 이용 환경

○ (망분리 폐쇄형 개발 환경 활용) COREA-AI 개발본부에서 구축한 Private 클라우드 기반의 망분리된 폐쇄형 AI 개발 환경에서 데이터를 이용 및 활용

- 취급자에게만 접근권한 부여, 접근통제 솔루션을 통해 비인가자의 접근 통제

○ (자료 반입) COREA-AI 관리자의 검토 및 승인 절차를 수립하여 운영 (관리자가 자료 확인 후 반입)

○ (자료 반출) 분석결과 반출 시 COREA-AI 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	SG-20230110004	접수일	2023년 5월 8일

신청자	조직/부서명	COREA-AI 개발본부 인공지능개발팀		
	담당자 직위	수석	담당자 성명	박채주
	전화번호	03-1234-5678	이메일주소	chaejoo@corea-ai.kr

처리목적	☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존
	재차인원 검지기반 다인승전용차로 단속에 활용될 AI 기술개발
활용 형태	☐ 내부이용 ☑ 제3자 제공 ☐ 결합전문기관을 통한 결합
이용 주체	☐ 동일 개인정보처리자 ☑ 제3자(제공받는 자) 제공
처리 환경	☐ 내부 ☑ 외부
처리 장소	☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부	☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법	☐ 온라인 ☑ 오프라인
제공 받는 자	(내부) ☐ 동일부서 ☐ 타부서 (외부) ☑ 제3자, 결합전문기관 제공

신청 명세	가명처리 대상 명칭	고속도로 통행 차량 사진
	데이터 내역	2022년 4월부터 2023년 4월까지 1년간 A지자체 OO고속도로 통행 차량에 대한 스냅사진 약 12만장
	가명정보 이용기간	2023년 6월 1일 ~ 2025년 6월 1일(2년)

위와 같이 가명정보 이용· 제공을 신청합니다.

2023년 5월 8일

신청인(부서장명) 박채주 (서명 또는 인)

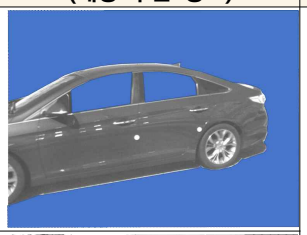
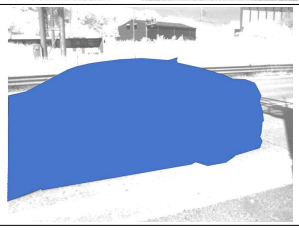
첨부서류	1. 연구계획서 2. 개인정보 유형 분류표 3. 가명정보에 대한 안전조치 이행 확약서 4. 보안 서약서 등

② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

과학적 연구 계획서

연구명	재차인원 검지기반 다인승전용차로 단속 AI 개발
연구진	COREA-AI 개발본부 인공지능개발팀 (책임: 박채주 수석)
연구 배경 및 목적	o 고속도로의 수용 효율을 증대하기 위하여 다인승전용차로 제도 시행 중(1995년 2월~) - 9인승이상 승용자동차 및 11,12인 승합자동차의 경우, 6인 이상이 실제 승차하였을 때만 다인승전용차로 이용 가능 o 그러나 다인승전용차로 이용자의 상당 수가 불법으로 이용 중인 것으로 파악됨 o 불법이용을 단속하고 교통흐름을 개선하기 위해 차량통행사진을 기반으로 재차인원을 파악하고 불법 이용자를 단속할 수 있는 시스템 개발 필요
예상 연구 기간	2023년 6월 1일 ~ 2025년 6월 1일(2년)
연구 대상자 수	o 2022년 4월부터 2023년 4월까지, A지자체 OO고속도로 통행 차량을 촬영한 이미지(12만장)
연구 방법	o A지자체가 고속도로 통행 차량 촬영 사진을 가명처리하여 COREA-AI에 제공 o COREA-AI 개발본부 내 담당부서에서 라벨링 진행 o 라벨링 결과를 받아서 AI 학습 수행 o 수행 결과를 바탕으로 인원계수 알고리즘 개발
연구 내용	o 다인승전용차로 시행효과 극대화를 위해 재차인원 검지기술 개발 및 실증 수행 o 다인승 전용차로 단속기능을 구현하기 위해 1) 영상촬영 기술 개발 2) 학습데이터 수집, 라벨링 및 학습 3) 차량 내 인원계수 알고리즘 개발 o 다인승 전용차로 단속시스템 관련 AI 모듈 개발
기대 효과 및 활용 방안	o 고속도로 버스전용차로 운영 효율화 및 단속에 따른 불필요한 지체를 감소 o 재차인원 검지기반 기술을 통해 재해재난 시 조난 대상자 파악 등 사회적 현안 해결에 활용 가능

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

개인정보 유형 분류표					
연번	항목명	구분	설명	예시 (해당 부분 강조)	비고
1	고속도로 통행차량 촬영 데이터	① 차량 외형	특수차, 차종 등을 판단하기 위해 차량 외형이 촬영된 부분		이미지 (비정형데이터)
		② 차량 탑승자	차량 내에 탑승한 사람이 촬영된 부분		
		③ 주변 환경	차량, 탑승자 영역 외 도로 주변환경이 촬영된 부분		
1-1	촬영 일시		사진이 촬영된 날짜 및 시각	2023.1.2.14:00	각 사진에 대한 메타데이터
1-2	촬영 장소		사진이 촬영된 주소 정보	OO시 OO동 OO로 123-4	

활용데이터 요구 수준표

※ 가명정보 처리 목적 달성에 필요한 데이터 항목과 가명처리 요구수준을 검토

연번	항목명	구분	요구 수준	비고
1	고속도로 통행차량 촬영 데이터	① 차량 외형	▪ 차종에 따라 불법성 여부 판단, 재차인원 파악을 위해 감지해야 하는 영역 등이 달라지므로 연구에 필요한 정보 ▪ 다만, 회사로고가 크게 적혀있는 경우, 소방차·구급차 등 특수 차량 등 일반적이지 않은 소수 차량의 경우는 표본에서 제외하여도 연구목적 달성에 문제 없음	이미지 (비정형데이터)
		② 차량 탑승자	▪ 재차인원 파악, 사람인지 아닌지 여부 등만 인식하면 되므로, 블러링처리하여 활용해도 상관 없음	
		③ 주변 환경	▪ 차량부분과 주변환경을 구분하기 위해 필요하며, 개인식별 가능성이 없다면 그대로 활용 필요	
1-1	촬영 일시		▪ 분석에 필요 없으므로 삭제·대체 가능	각 사진에 대한 메타데이터
1-2	촬영 장소		▪ 분석에 필요 없으므로 삭제·대체 가능	

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

<table>
<tr><td>가명정보
활용목적</td><td colspan="2">재차인원 검지기반 다인승전용차로 단속 AI 개발</td></tr>
<tr><td>가명처리
대상 데이터
항목</td><td colspan="2">○ 고속도로 통행 차량을 촬영사진(이미지) <연번 1>
○ 고속도로 통행 차량 촬영사진의 메타데이터(촬영일시, 촬영장소) <연번 1-1~1-2></td></tr>
<tr><td rowspan="3">데이터
위험성</td><td>식별성
유무</td><td><고속도로 통행차량 촬영사진>
○(차량 외형, 주변환경 부분) 이미지 그 자체만으로는 개인식별 위험성이 거의 없으나, 촬영일시·장소 등 메타데이터와 결합될 시 식별성이 생길 수 있음
○(차량 탑승자) 얼굴이 선명하게 촬영된 경우, 동승자 파악이 가능한 경우 등 개인식별 가능성이 높으며, 촬영일시·장소 등 메타데이터와 결합될 시 이동동선까지 파악 가능
- 다만, 개발하려는 AI는 특정인을 분간하지 않으며 사람인지 아닌지 여부, 재차 여부를 파악하기 위한 목적으로만 활용되므로 특정 개인식별 위험성은 낮은 편
<메타데이터>
○(촬영 일시, 촬영 장소) 고속도로 통행차량 촬영사진과 결합될 시 개인식별 위험이 존재함</td></tr>
<tr><td>특이정
보 유무</td><td>○‘차량 외형’ 촬영부분의 경우 회사로고가 크게 적혀있는 경우, 소방차·구급차 등 특수차량의 경우 등 특이한 차량외형으로 인한 개인식별 위험이 존재</td></tr>
<tr><td>재식별
시
영향도</td><td>○차량통행 촬영사진은 재식별시 영향도는 낮은 편</td></tr>
<tr><td rowspan="3">처리 환경
위험성</td><td>이용 및
제공
형태</td><td>○외부 기업(제3자)에 제공</td></tr>
<tr><td>처리
장소</td><td>○개발본부 내 Private 클라우드를 구축하여 망분리된 폐쇄형 개발환경에서 처리
○가명정보를 제공받는 기업은 개인정보(가명정보)처리시스템에 대한 ISMS-P인증을 취득하고 있음</td></tr>
<tr><td>다른
정보
결합
가능성</td><td>○ 폐쇄환경분석실 관리자 승인하에 제한된 데이터·프로그램(프로그램 패키지, 라이브러리, 코드설명서 등)만 반입 가능
○ 분석대상 가명정보와 결합가능성 있는 데이터는 반입 제한</td></tr>
<tr><td>최종
검토의견</td><td colspan="2">▪ 해당 연구는 차량 내 재차 인원을 파악하기 위한 목적으로 고속도로 통행차량 사진을 활용하므로, 활용 목적 및 처리 환경을 고려하였을 때 전반적인 개인식별 위험성은 낮은 편임
▪ 차량탑승자 촬영부분은 개인을 특정하여 활용할 필요가 없고, 탑승 여부만 확인하면 되므로 블러링 처리하여 활용하는 것이 적절
▪ 차량 외형의 경우 특이정보로 판단될 수 있는 차종 외형을 가명처리하여 활용할 필요가 있으며 연구 목적 달성에 심각한 영향을 주지 않는다면, 삭제 필요</td></tr>
</table>

	▪ 메타데이터(차량일시, 차량장소)는 연구목적과 관계 없으므로 삭제하여 활용함으로써 데이터 간 연계·결합으로 인한 개인식별 위험성을 제거할 필요가 있음 - 이미지의 메타데이터는 연구목적 달성에 필요하지 않으며, 이미지와 연계되어 개인식별 위험성이 높아지므로 전체 삭제하고 이미지 정보만 활용 필요

④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
1	고속도로 통행차량 촬영 데이터	① 차량 외형	☑ 특이정보 파일 표본 제외	▪ 차량 외형이 일반적이지 않고, 특이하여 탑승자의 개인식별 위험성이 비교적 높은 경우(회사로고가 크게 적혀있는 경우, 소방차·구급차 등 특수차량 등)해당 파일을 표본에서 제외
		② 차량 탑승자	☑ 블러링 처리	▪ 탑승자의 위치 및 범위를 파악하고 얼굴·상체 등 차량 탑승자 영역 전체를 블러링 처리 (블러링 수준 5단계 적용)
		③ 주변 환경	☑ 그대로 사용	▪ 연구목적 달성에 필요한 정보이며 차량 외형(특이한 경우 연구표본에서 삭제), 차량 탑승자(블러링 처리), 메타데이터(삭제)를 가명처리하므로 타 정보와의 결합을 통한 개인식별 가능성이 거의 없어 별도 처리하지 않고 그대로 사용
1-1	촬영 일시		☑ 삭제	▪ 연구에 필요 없으므로 삭제
1-2	촬영 장소		☑ 삭제	▪ 연구에 필요 없으므로 삭제

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

주요 비정형데이터 가명처리 수행 결과 (예시)

연번	항목명	세부 항목	가명처리 전	가명처리 후
1	고속도로 통행차량 촬영 데이터	① 차량 외형	 	연구 표본에서 제외 (연구에 활용하지 않음) ※ 특수차량(응급차, 경찰차, 회사차 등), 특이한 스티커·로고·광고 등이 포함된 차량 등
		② 차량 탑승자		 (탑승자 위치·범위 파악) ▼ (1~10단계 중 5단계 수준 블러링 처리)

○ 비정형 데이터 가명처리 기술의 적절성・신뢰성 관련 근거(예시)

연번	항목명	세부 항목	처리 기술명	처리 기술의 적절성·신뢰성 관련 근거 또는 배경
1	고속도로 통행차량 촬영 데이터	② 차량 탑승자	블러링(부분)	▪ A지자체는 R&D사업을 통해 자체 개발한 '이미지 블러링 프로그램'을 활용하여 차량 탑승자 영역에 대한 블러링을 수행 ※ 해당 프로그램은 이미지 인식 알고리즘을 이용하여 차량 내 탑승자의 위치 및 범위를 판단할 수 있으며, 블러링 설정 값을 1단계에서 10단계로 구분하여 설정 가능 - 현재 복원기술의 발전 수준 및 데이터 처리 환경(타 정보·복원기술 활용 불가) 등을 고려하여 블러링 수준을 5단계로 설정 * 블러링 수준의 적정성은 외부전문가를 과반 이상 포함한 적적성 평가 위원회를 구성하여 검토 완료 *※ 적정성 검토위원회 결과보고서 및 회의록 별첨* - 자체 개발 이미지 블러링 프로그램은 내부 테스트 결과 객체 인식률 94%, 처리 정확도 99%로 측정 *※ 객체 인식률, 처리 정확도(오류율) 증빙자료 별첨* - 탑승자를 인식하지 못하는 경우, 블러링 처리에 오류가 있는 경우 등이 존재하므로, 가명처리 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사 수행

비정형데이터 가명처리 결과에 대한 자체 검증 결과서

<table>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="3">개요</td></tr>
<tr><td colspan="3">A지자체에서 수집한 고속도로 통행차량 촬영데이터 중 차량 탑승자 영역을 자체개발한 ‘이미지 블러링 프로그램’을 활용하여 블러링 처리, 차량특이치는 삭제</td></tr>
<tr><td>데이터 유형</td><td colspan="2">이미지</td></tr>
<tr><td>원본 데이터 형식
(파일 포맷)</td><td colspan="2">JEPG</td></tr>
<tr><td>처리 결과 데이터 형식
(파일 포맷)</td><td colspan="2">JEPG</td></tr>
<tr><td>데이터 규모</td><td colspan="2">120,000장</td></tr>
<tr><td>데이터 크기(용량)</td><td colspan="2">18.31GB</td></tr>
<tr><td>대상 데이터 항목명</td><td colspan="2">고속도로 통행차량 촬영데이터 <연번 1></td></tr>
<tr><td>가명처리 적용 기술</td><td colspan="2">- A지자체에서 자체개발한 ‘이미지 블러링 프로그램’을 활용하여 각 이미지 내 차량탑승자 영역을 블러링 처리(5단계 수준)
- 차량특이치는 직접 검수하여 삭제</td></tr>
<tr><td>자체 검증
기간</td><td colspan="3">2023년 5월 12일 ~ 2023년 5월 30일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="3">COREA-AI 개발본부 AI품질팀 사무실</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="3">(검증방법)
- 개인정보보호팀 담당 과장 주도하에 총 2명의 검수자가 Workday 10일간 전수조사 (1인당 1일 평균 약 6,000개 검수)
- 전수조사 이후 파일대상 진단 툴 조사 및 보고서 작성
(처리된 이미지의 영역정보를 기준으로 원본 이미지와 가명처리된 이미지의 영역을 비교하여 진단)

(검증 시 확인사항)
① 차량의 색상, 특수차량, 회사로고 등 특이한 차량인가?
➜ 해당 시 표본 삭제 (연구 대상에서 제외)
② 탑승자를 식별할 수 없도록 블러링 처리가 되었는가?
➜ 제대로 처리되지 않았을 시 표본 삭제 (연구 대상에서 제외)</td></tr>
<tr><td>자체 검증
결과</td><td colspan="3">차량외형의 특이치가 존재하여 표본을 삭제한 경우 : 1,871장
탑승자 블러링처리가 제대로 안되어 표본을 삭제한 경우 : 142장
나머지 이미지는 정상 처리되어 연구 활용 가능</td></tr>
<tr><td rowspan="5">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>개인정보보호 팀장</td><td>박지원</td><td></td></tr>
<tr><td>개인정보보호팀 과장</td><td>최우선</td><td></td></tr>
<tr><td>AI 품질팀 인턴</td><td>김용두</td><td></td></tr>
<tr><td>AI 품질팀 인턴</td><td>이용미</td><td></td></tr>
</table>

시나리오 ⑥ : 한국어 대화가 가능한 AI 챗봇 개발

대화·검색 분야 (텍스트)

- 인공지능 챗봇 전문기업 ㈜한국데이터테크는 그간 채팅앱을 통해 수집한 고객 간의 다양한 대화 텍스트 데이터를 가명처리하여 한국어 대화가 가능한 AI 챗봇 개발에 활용하고자 한다.

☑ 데이터의 이용 목적

○ 한국어 대화가 가능한 AI 챗봇 언어모델 개발

☑ 데이터 특징

○ 채팅앱을 통해 수집된 고객들 간 일상 대화 데이터(텍스트)

- 고객 1,500명이 대화한 대화 데이터셋 총 20,000개 파일
- 대화 당사자 간의 일상 대화 내용이 포함되어 있으며 데이터에 정해진 형식과 법칙성이 없고 대화 주제가 다양함

* 격식을 갖추지 않은 일상의 구어체 대화가 대부분이며 한국어의 문법, 어법, 철자 등을 철저하게 준수하기보다는 편리함에 초점을 둔 대화 데이터임

- 대화 당사자의 이름, 연락처, 계좌번호, 사생활의 영역까지 개인식별위험이 있는 다양한 정보가 포함되어 있음
- 대화파일마다 대화에 참여한 사용자 계정정보가 메타데이터로 포함

☑ 데이터의 이용 환경

○ (폐쇄형 내부 개발환경 활용) ㈜한국데이터테크 내부 공간에 마련한 철저한 폐쇄망 환경에서만 데이터 활용

- 취급자에게만 접근권한 부여, 접근통제 솔루션으로 비인가자 접근 통제

○ (자료 반입) 내부 보안팀의 검토 및 승인 절차에 따라 운영(관리자가 자료 확인 후 반입)

○ (자료 반출) 분석결과 반출 시 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	SG-20230110083	접수일	2023년 3월 2일

신청자	조직/부서명	㈜한국데이터테크 AI 데이터 구축팀		
	담당자 직위	책임	담당자 성명	이신뢰
	전화번호	03-1234-5678	이메일주소	security@kor_datatech.kr

처리목적		☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존
		한국어 대화가 가능한 AI 챗봇 언어모델 개발
활용 형태		☑ 내부이용 ☐ 제3자 제공 ☐ 결합전문기관을 통한 결합
이용 주체		☑ 동일 개인정보처리자 ☐ 제3자(제공받는 자) 제공
처리 환경		☑ 내부 ☐ 외부
처리 장소		☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부		☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법		☑ 온라인 ☐ 오프라인
제공 받는 자		(내부) ☑ 동일부서 ☐ 타부서 (외부) ☐ 제3자, 결합전문기관 제공
신청 명세	가명처리 대상 명칭	채팅앱 내 고객 간 일상대화 데이터
	데이터 내역	2018~2020년동안 자사 채팅앱 내에서 1,500명의 고객들 간 이루어진 대화 데이터셋 20,000개 파일
	가명정보 이용기간	2023년 4월 1일 ~ 2026년 4월 1일(3년)

위와 같이 가명정보 이용·제공을 신청합니다.

2023년 3월 2일

신청인(부서장명) 이 신 뢰 (서명 또는 인)

첨부서류
1. 연구계획서
2. 개인정보 유형 분류표

② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

과학적 연구 계획서	
연구명	한국어 대화가 가능한 AI 챗봇 언어모델 개발
연구진	㈜한국데이터테크 AI 데이터 구축팀 (연구책임: 이신뢰 책임)
연구 배경 및 목적	o ㈜한국데이터테크는 사람과 한국어로 자유롭게 일상적인 대화를 할 수 있는 AI 챗봇 서비스를 제공하고자 하며, 이를 위해 한국어 대화의 맥락과 행간을 깊이 있게 이해하고 실제 사람과 대화하는 것 같은 대화 경험을 제공하기 위한 한국어 AI 언어모델을 개발하고자 함
예상 연구 기간	2023년 4월 1일 ~ 2026년 4월 1일 (3년)
연구 대상자 수	2018~2020년동안 자사 채팅앱 내에서 1,500명의 고객들 간 이루어진 20,000개의 대화 텍스트 데이터셋 파일
연구 방법	o 대화 텍스트 데이터셋을 가명처리하여 학습용 데이터베이스를 구축하고, 이를 컴퓨터가 이해할 수 있는 벡터 형태로 변환한 뒤, 대화 문맥 내 단어들 간 등장 확률을 학습
연구 내용	o 학습용 데이터베이스를 통해 상대방의 발화를 이해하고 적절한 답변을 선택하기 위한 모델, 문맥을 파악하고 부적절한 답변을 걸러내는 모델을 개발 o AI 챗봇의 대답은 외부에 공개되기 때문에 **가명정보를 통해 학습된 대화가 그대로 발화되지 않도록, 답변에 활용되는 데이터베이스는 별도로 구축**하여 **가명정보 노출 가능성을 근본적으로 차단** - 답변 데이터베이스는 ㈜한국데이터테크 자체 생성AI 모델이 만들어낸 문장과 내부 AI 데이터구축팀이 직접 작성한 문장으로 구성
기대 효과 및 활용 방안	o 쉽고 자연스럽게 대화할 수 있는 AI 챗봇을 개발하여 사회 소외계층, 디지털 소외계층도 함께 누릴 수 있는 챗봇 생태계 조성 o 한국어 대화 맥락 파악 고도화를 통한 맞춤형 상담·검색 기능 확대

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

개인정보 유형 분류표				
연번	**항목명**	**구분**	**설명**	**비고**
1	채팅앱 내 고객 간 일상대화 데이터	① 이름	▪ 대화에서 언급된 대화 당사자 또는 제3자의 이름	텍스트 (비정형 데이터)
		② 생년월일	▪ 대화에서 언급된 대화 당사자 또는 제3자의 생년월일	
		③ 주민등록번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 주민등록번호	
		④ 연락처	▪ 대화에서 언급된 대화 당사자 또는 제3자의 연락처 정보	
		⑤ 카드번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 카드번호	
		⑥ 계좌번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 계좌번호	
		⑦ 차량번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 차량번호	
		⑧ 학번	▪ 대화에서 언급된 대화 당사자 또는 제3자의 학번	
		⑨ 여권번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 여권번호	
		⑩ 운전면허번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 운전면허번호	
		⑪ 외국인등록번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 외국인등록번호	
		⑫ 건강보험증 번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 건강보험증 번호	
		⑬ 상세 주소	▪ 대화에서 언급된 대화 당사자 또는 제3자의 상세 주소 정보	
		⑭ 아이디	▪ 대화에서 언급된 대화 당사자 또는 제3자의 아이디	
		⑮ 비밀번호	▪ 대화에서 언급된 대화 당사자 또는 제3자의 비밀번호	
		⑯ 이메일	▪ 대화에서 언급된 대화 당사자 또는 제3자의 이메일 주소	
		⑰ URL	▪ 대화에서 언급된 대화 당사자 또는 제3자와 관련 있을 수 있는 URL 주소	
		⑱ 나이	▪ 대화에서 언급된 대화 당사자 또는 제3자의 나이 정보	
1-1	대화 사용자 계정정보		▪ 대화 사용자에 대해 고유하게 부여되는 사용자 계정정보 ▪ 해당 대화가 어떤 사용자와 연결되는지 파악하기 위해 사용	메타 데이터 (정형 데이터)

< ※ (참고) 데이터 처리목적 및 활용 방식을 고려한 가명처리 대상 선정 >

▶ ㈜한국데이터테크가 개발하고자 하는 한국어 대화 AI 챗봇은 가명정보를 통해 학습하여 만들어진 '학습용 데이터베이스'와 실제 AI 챗봇의 답변 대화 구성에 활용되는 '답변 데이터베이스'가 완전히 분리되어 운영되므로 가명처리된 텍스트데이터 내의 문장이 AI 챗봇의 답변에 그대로 노출될 확률이 없음

▶ 즉, 가명처리한 텍스트 데이터는 AI 언어모델을 학습하는데만 활용되고 실제 AI챗봇 답변을 통해 발화되지 않으므로, **이름, 생년월일, 연락처 등 개인식별성이 높은 항목들만 철저히 필터링하여 삭제·대체**한다면 이용자 프라이버시가 침해될 확률을 크게 낮출 수 있음

활용데이터 요구 수준표

연번	항목명	구분	요구 수준	비고
1	채팅앱 내 고객 간 일상대화 데이터	① 이름	▪ 대화상대 구분 등 연구목적에 필요하여 삭제하면 안 되며, 임의의 다른 이름으로 치환하여 활용 필요	텍스트 (비정형 데이터)
		② 생년월일	▪ 대화상대 연령에 맞는 대화문 생성 등 연구목적에 필요하여 삭제하면 안 되며, 임의의 다른 값으로 치환하여 활용 필요	
		③ 주민등록번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		④ 연락처	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑤ 카드번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑥ 계좌번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑦ 차량번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑧ 학번	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑨ 여권번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑩ 운전면허번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑪ 외국인등록번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑫ 건강보험증 번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑬ 상세 주소	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑭ 아이디	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑮ 비밀번호	▪ 연구목적에 필요 없어 삭제 가능(해당 정보 포함 전체 문장 삭제)	
		⑯ 이메일	▪ 이메일이 포함된 대화패턴 인식이 필요하여 토큰으로 대체 필요	
		⑰ URL	▪ URL이 포함된 대화패턴 인식이 필요하여 토큰으로 대체 필요	
		⑱ 나이	▪ 발화자 나이에 따른 대화문 생성 등 연구목적에 필요하여 삭제하면 안 되며, 비슷한 나이대로 범주화하여 활용 필요	
	1-1	대화 사용자 계정정보	▪ 대화 사용자와 일상대화 데이터의 연결은 불필요하지만 대화 데이터 발화자의 구분은 필요하므로 랜덤ID로 대체 필요	메타 데이터 (정형 데이터)

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

가명정보 활용목적	한국어 대화가 가능한 AI 챗봇 언어모델 개발	
가명처리 대상 데이터 항목	○ 2018~2020년동안 자사 채팅앱 내에서 1,500명의 고객들 간 이루어진 20,000개의 대화 텍스트 데이터셋 파일 *<연번 1>* ○ 대화 텍스트 데이터셋의 메타정보(대화 사용자 계정정보) *<연번 1-1>*	
데이터 위험성	**식별성 유무**	○ 이름, 주민등록번호, 연락처, 카드번호, 계좌번호, 운전면허번호, 여권번호, 외국인등록번호, 건강보험증번호, 상세주소 등은 개인식별 위험성이 상당히 높음 ○ 아이디, 생년월일, 학번, 나이 등도 다른 대화내용 및 메타데이터와 연계될 시 개인식별 위험성이 존재 ○ 개인정보로 선정한 항목 외에 일반대화내용도 상황에 따라 다른 대화내용과 연계될 시 개인식별 위험성이 생길 수는 있으나, AI 언어모델 학습에만 활용되고 AI챗봇의 답변을 통해 발화되지 않는다면 개인식별 위험성이 미미함 ○ 또한, 메타데이터(대화 사용자 계정정보)를 삭제하여 특정 개인과의 연결성을 제거한 형태로 활용되면 개인식별 위험성이 더욱 낮아질 것임
	특이정보 유무	○ 일반대화내용에 포함된 대화 맥락 및 특수 상황에 대한 언급이 존재할 시 개인식별 위험이 높아질 수 있음
	재식별 시 영향도	○ 일상대화의 내용이 다양하게 담겨있고 상황에 따라 사생활의 내밀한 내용이 포함될 수 있어 재식별 시 영향도가 높은 편임
처리 환경 위험성	**이용 및 제공 형태**	○ 동일 개인정보처리자, 동일 부서 내 활용
	처리 장소	○ 회사 내부의 철저한 폐쇄망 개발환경 활용
	다른 정보 결합 가능성	○ 관리자 승인하에 데이터(프로그램 패키지, 라이브러리, 코드설명서 등) 반입이 가능함 ○ 기존 가명정보와 결합가능성 있는 데이터는 반입이 제한됨
최종 검토의견	▪ 대화내용 내 개인식별성이 있는 항목들을 빠짐없이 선별하여 삭제 또는 치환하여야 함 - 특히 고유식별정보인 주민등록번호, 운전면허번호, 여권번호, 외국인등록번호는 반드시 삭제하여야 함 - 나이정보는 개인식별 위험이 있으나, 발화자 나이에 따른 대화문 생성에 필요한 정보이므로 5세 단위로 범주화하여 활용하는 것이 적절 ▪ 메타데이터인 대화 사용자 계정정보는 파기하고 랜덤 ID로 대체하여 활용함으로써	

	특정 개인과의 연결성을 제거하여 활용하여야 함 ▪ 가명처리된 텍스트데이터가 활용되는 학습데이터베이스와 AI 챗봇을 통해 발화되는 답변데이터베이스가 분리되어 있어, 가명정보가 AI 챗봇의 답변으로 그대로 발화되지 않기에 가명정보의 직접적인 노출로 인한 개인식별 위험은 발생하지 않을 것으로 판단됨

④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

개인정보 유형 분류표

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
1	채팅앱 내 고객 간 일상대화 데이터	① 이름	☑ 치환	▪ 인공지능 모델(Personal Name Recognition) 기반으로 문장 내에서 이름이라고 판단되는 단어가 발견되면, 해당 이름을 관련 통계에 따른 한국인 이름 분포에 따라 랜덤하게 생성한 임의의 이름으로 치환
		② 생년월일	☑ 치환	▪ 생년월일 패턴이 발견되면 해당 정보를 랜덤하게 생성한 임의의 생년월일로 치환함
		③ 주민등록번호	☑ 삭제	▪ 주민등록번호 패턴(000000-0000000)을 포함하는 문장 전체를 삭제
		④ 연락처	☑ 삭제	▪ 전화번호 패턴(00 또는 000-000-0000 등)을 포함하는 문장 전체를 삭제 ▪ 숫자를 한글로 표현(예: 일이삼, 영, 공 등)한 경우까지 고려하여 삭제하며, 한글 숫자와 아라비아 숫자가 병용된 패턴 역시 문장 삭제
		⑤ 카드번호	☑ 삭제	▪ 카드번호 패턴을 포함하는 문장 전체를 삭제
		⑥ 계좌번호	☑ 삭제	▪ 계좌번호 패턴을 포함하는 문장 전체를 삭제
		⑦ 차량번호	☑ 삭제	▪ 차량번호 패턴을 포함하는 문장 전체를 삭제
		⑧ 학번	☑ 삭제	▪ 학번 등의 패턴을 포함하는 문장 전체를 삭제
		⑨ 여권번호	☑ 삭제	▪ 여권번호 패턴을 포함하는 문장 전체를 삭제
		⑩ 운전면허번호	☑ 삭제	▪ 운전면허번호 패턴을 포함하는 문장 전체를 삭제
		⑪ 외국인등록번호	☑ 삭제	▪ 외국인등록번호 패턴을 포함하는 문장 전체를 삭제
		⑫ 건강보험증 번호	☑ 삭제	▪ 건강보험증번호 패턴을 포함하는 문장 전체를 삭제
		⑬ 상세 주소	☑ 삭제	▪ 지역명 주소로 구성된 읍·면·동·리 상세주소 및 도로명 주소로 구성된 길 이름이 포함된 문장 전체를 삭제 ▪ 아파트 동호수 등의 패턴이 포함된 문장 전체를 삭제 ▪ 길안내, 네비게이션 앱 메시지 삭제
		⑭ 아이디	☑ 삭제	▪ 숫자, 영문, 특수문자 등으로 구성된 7자리 이상의 단어가 포함되면 해당 문장 전체를 삭제 ▪ '아이디', 'ID' 등의 문구가 포함되면 해당 문장 전체를 삭제
		⑮ 비밀번호	☑ 삭제	▪ 숫자, 영문, 특수문자 등으로 구성된 7자리 이상의 단어가 포함되면 해당 문장 전체를 삭제 ▪ '비밀번호' '패스워드' 등의 문구가 포함되면 해당 문장 전체를 삭제 ▪ 아파트 비밀번호 등에 자주 쓰이는 특수문자(#, *) 등으로 시작하거나 끝나는 패턴이 있으면 해당 문장을 삭제
		⑯ 이메일	☑ 치환	▪ '@' 기호가 포함되어 있는 이메일 패턴을 토큰 [MAIL]로 치환
		⑰ URL	☑ 치환	▪ http:// 및 .com 등 URL 패턴 부분을 토큰 [URL]로 치환
		⑱ 나이	☑ 범주화	▪ 5세 단위 범주화 (16~20세, 21~25세, 26세~30세, 31세~35세 등)
1-1	대화 사용자 계정정보		☑ 대체	▪ 사용자 계정정보를 파기하고 랜덤한 ID로 대체

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

○ 비정형 데이터 가명처리 기술의 적절성·신뢰성 관련 근거(예시)

연번	항목명	세부 항목	처리 기술명	처리 기술의 적절성·신뢰성 관련 근거 또는 배경
1	채팅앱 내 고객 간 일상대화 데이터	① 이름 ② 생년월일 ⑯ 이메일 ⑰ URL	치환	▪ '이름', '생년월일', '이메일', 'URL' 항목은 자체 필터링 솔루션을 활용하여 임의의 값 또는 토큰으로 대체 - 자체 필터링 솔루션은 자연어 처리 기준으로 검출 정확성은 89%, 처리 정확도는 97%로 측정 *※ 검출 정확도, 처리 정확도(오류율) 증빙자료 별첨* - 검출 정확도가 100%가 아니므로, 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사를 수행
		③ 주민등록번호 ④ 연락처 ⑤ 카드번호 ⑥ 계좌번호 ⑦ 차량번호 ⑧ 학번 ⑨ 여권번호 ⑩ 운전면허번호 ⑪ 외국인등록번호 ⑫ 건강보험증 번호 ⑬ 상세 주소 ⑭ 아이디 ⑮ 비밀번호	삭제	▪ 자체 필터링 솔루션을 활용하여 등록된 패턴을 기준으로 해당 항목을 검출하고, 검출된 항목이 포함된 문장 전체를 삭제 처리 - 자체 필터링 솔루션은 자연어 처리 기준으로 검출 정확성은 89%, 처리 정확도는 97%로 측정 *※ 검출 정확도, 처리 정확도(오류율) 증빙자료 별첨* - 검출 정확도가 100%가 아니므로, 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사를 수행
		⑱ 나이	범주화	▪ 자체 필터링 솔루션을 활용하여 나이값을 검출하고, 5세 단위로 범주화하여 토큰으로 대체 - 자체 필터링 솔루션은 자연어 처리 기준으로 검출 정확성은 89%, 처리 정확도는 97%로 측정 *※ 검출 정확도, 처리 정확도(오류율) 증빙자료 별첨* - 검출 정확도가 100%가 아니므로, 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사를 수행

비정형데이터 가명처리 결과에 대한 자체 검증 결과서

<table>
<tr><td rowspan="7">검증 대상
데이터 명세</td><td colspan="2">개요</td></tr>
<tr><td colspan="2">㈜ 한국데이터테크는 채팅앱을 통해 수집된 고객들 간 일상대화 데이터 중 개인식별 위험성이 높은 항목들을 자체 필터링 솔루션을 활용하여 가명처리(치환·삭제·범주화)하였음</td></tr>
<tr><td>데이터 유형</td><td>텍스트</td></tr>
<tr><td>원본 데이터 형식(파일 포맷)</td><td>JSON</td></tr>
<tr><td>처리 결과 데이터 형식(파일 포맷)</td><td>JSON</td></tr>
<tr><td>데이터 규모 및 크기</td><td>고객 1,500명이 대화한 일상대화 데이터 20,000개 파일</td></tr>
<tr><td>대상 데이터 항목명</td><td>일상대화 데이터 텍스트 파일 <연번 1></td></tr>
<tr><td>가명처리 적용 기술</td><td>- ㈜한국데이터테크에서 자체 개발한 필터링 솔루션을 활용하여 주요 개인식별 위험 항목을 검출한 뒤 치환, 삭제, 범주화 처리
- 일반대화정보 내 개인식별 가능성이 있는 단어, 문장은 직접 검수하여 삭제 또는 대체</td></tr>
<tr><td>자체 검증
기간</td><td colspan="2">2023년 3월 15일 ~ 2023년 3월 25일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="2">㈜한국데이터테크 데이터센터</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="2">(검증방법)
- 개인정보보호팀 담당 팀장 주도하에 총 4명의 검수자가 Workday 10일간 전수조사행

(검증 시 확인사항)
① 치환되어야 할 항목(이름, 생년월일, 이메일, URL)들이 정해진 텍스트나 정해진 범주 내 랜덤값으로 치환되었는가?
➜ 치환되지 않았을 시 수작업으로 변환 실시
② 삭제되어야 할 문장(주민번호, 연락처 등 포함)들이 삭제되었는가?
➜ 삭제되지 않았을 시 수작업으로 삭제 실시
③ 나이정보가 범주화되어 토큰으로 대체 되었는가?
➜ 대체되지 않았을 시 수작업으로 범주화 및 토큰화 실시
④ 가명처리되어야 할 항목들이 인식·검출되지 않아 제대로 처리가 되지 않은 문장이 존재하는가?
➜ 해당 문장 발견시 수작업으로 추가 가명처리 실시
⑤ 가명처리 항목 외에도 일반대화정보에서 개인식별 가능성이 있는 단어나 문장이 존재하는가?
➜ 확인 시 팀장 주도 최종회의에서 삭제·대체여부 결정하여 추가 가명처리 실시</td></tr>
<tr><td>자체 검증
결과</td><td colspan="2">20,000개 파일 중 4,765개 파일에 대해 오류 발견되어 추가 가명처리 실시 완료</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>개인정보보호 팀장</td><td>가관리</td><td></td></tr>
<tr><td>개인정보보호팀 직원</td><td>나정보</td><td></td></tr>
<tr><td>AI 데이터 구축팀 직원</td><td>다보호</td><td></td></tr>
<tr><td>AI 데이터 구축팀 직원</td><td>라보호</td><td></td></tr>
<tr><td>AI 데이터 구축팀 직원</td><td>마보호</td><td></td></tr>
</table>

시나리오 ❼ : 콜센터 직원 실습용 가상상담 시나리오 생성 AI 개발

대화·교육 분야 (음성, 텍스트)

- 인터넷통신 기업 ㈜코리아인터넷은 인터넷·통신 상품 관련 고객상담 시 분쟁방지 및 서비스 품질 향상 목적으로 통화내용이 녹음됨을 고객에게 공지하고 상담 음성정보를 수집·보관하고 있다. ㈜코리아인터넷은 해당 음성정보를 가명처리한 뒤, 자사 콜센터 상담직원들에게 상황별·업무별 상담실습 교육을 진행하기 위한 목적으로 가상상담 시나리오를 생성하는 AI를 개발하고자 한다.

☑ 데이터의 이용 목적

○ 콜센터 직원 상담 실습교육용 가상상담 시나리오 생성 AI 모듈 개발

☑ 데이터 특징

○ 상담사와 고객 간 질의-응답으로 이루어진 음성 데이터(WAV 포맷)

- 상담 목적(❶단순 문의, ❷A/S, ❸결제, ❹교환, ❺설치, ❻반품, ❼기타), 성별, 연령, 주거 지역별로 샘플링한 상담음성 녹취파일 총 10,000개 파일
- 각 파일 당 평균 3분 분량으로 총 500시간 (60GB)
- 고객 및 상담사의 음성이 녹음되어 있으며 대화 내용에 고객의 이름, 생년월일, 주소, 연락처 등 개인식별성 있는 정보가 포함

☑ 데이터의 이용 환경

○ (망분리 폐쇄형 개발 환경 활용) ㈜코리아인터넷 빅데이터본부에서 구축한 Private 클라우드 기반의 폐쇄형 AI개발 환경에서만 데이터를 활용

- 취급자에게만 접근권한 부여, 접근통제 솔루션으로 비인가자 접근 통제

○ (자료 반입) ㈜코리아인터넷 보안팀의 검토 및 승인 절차에 따라 운영 (관리자가 자료 확인 후 반입)

○ (자료 반출) 분석결과 반출 시 관리자에게 요청(관리자 자료 확인 후 제공)

① 사전준비 단계에서 필요서류가 법/제도 목적에 적합하게 작성되었는지 검토

가명정보 이용•제공 신청서

접수번호	SG-20230110085	접수일	2023년 1월 10일

신청자	조직/부서명	㈜코리아인터넷 빅데이터본부 음성AI팀		
	담당자 직위	책임	담당자 성명	박안전
	전화번호	03-1234-5678	이메일주소	park_security@kor_internet.kr

처리목적		☐ 통계작성 ☑ 과학적 연구 ☐ 공익적 기록보존
		콜센터 직원 실습용 가상상담 시나리오 생성 AI 모듈 개발
활용 형태		☑ 내부이용 ☐ 제3자 제공 ☐ 결합전문기관을 통한 결합
이용 주체		☑ 동일 개인정보처리자 ☐ 제3자(제공받는 자) 제공
처리 환경		☑ 내부 ☐ 외부
처리 장소		☑ 폐쇄 환경 ☐ 제한없음
반복 제공여부		☑ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)
제공 방법		☑ 온라인 ☐ 오프라인
제공 받는 자		(내부) ☑ 동일부서 ☐ 타부서 (외부) ☐ 제3자, 결합전문기관 제공
신청 명세	가명처리 대상 명칭	콜센터 녹취 파일(WAV)
	데이터 내역	2022년 녹취통합 시스템에 저장되어 파일 중 학습용으로 샘플링(상담목적, 성별, 연령, 주거지역 고려)한 10,000개 파일
	가명정보 이용기간	2023년 3월 16일 ~ 2024년 3월 16일(1년)

위와 같이 가명정보 이용·제공을 신청합니다.

2023년 1월 10일

신청인(부서장명) 박 안 전 (서명 또는 인)

첨부서류
3. 연구계획서
4. 개인정보 유형 분류표

② 보호법에서 정한 목적 중 가명정보 처리 목적을 명확히 설정하였는지 검토

과학적 연구 계획서

연구명	콜센터 직원 실습교육용 가상상담 시나리오 생성 AI 모듈 개발
연구진	㈜코리아인터넷 빅데이터본부 음성AI팀 (연구책임: 박안전 책임)
연구 배경 및 목적	o ㈜코리아인터넷에서는 자사 콜센터 직원을 대상으로 상담 업무 능력 향상을 위한 교육에 활용할 교육용 음성생성 AI 기술개발을 진행하고 있음 o 다양한 상담 목적, 고객 유형에 따라 단순 상담부터 난이도 높은 상담까지 문제없이 대응할 수 있도록 AI를 통해 상황별 고객상담 음성을 생성하여 상담직원 교육에 활용하고자 함
예상 연구 기간	2023년 3월 15일 ~ 2024년 3월 16일 (1년)
연구 대상자 수	**다양한 발화자(성별, 연령, 지역 등)의 특징을 고려하여 샘플링한 10,000건의 녹취 파일**
연구 방법	언어모델 개발과 언어모델 학습을 통한 상담 데이터 생성 모델 개발
연구 내용	o **상담음성을 텍스트로 변환(STT)한 뒤**, 개인식별성이 있는 단어, 문장 등을 제거 또는 치환하여 개인식별 위험성을 제거 o **가명처리된 텍스트 데이터**를 기반으로 고객 질문 의도 분류 모델 및 상담사 답변 추천 모델 개발 - 고객 질문 입력 이후 고객 질문 의도를 분류하는 합성곱 신경망 기반 분류 모델 - 모델이 분류한 결과를 실제 고객 질문의도 범주와 비교하여 오차를 계산하고, 오차 역전파(backpropagation) 기법을 통해 모델 학습 진행 - 성능평가 지표를 통해 학습된 모델을 평가 o 개발된 AI 모델에 의해 **새롭게 생성된 콜센터 상담 시나리오(텍스트) 생성** o 해당 상담 시나리오를 **음성파일로 재변환(TTS)**하여 교육실습자료로 활용
기대 효과 및 활용 방안	o 다양한 분야의 상담 데이터와 상담 사례별 데이터를 통해 상담 지원 대상 교육 효과 증대 o 상담 만족도 향상을 위한 상담업무 효율 및 비즈니스 효과 증대

③ 가명처리 단계에서 데이터의 자체 식별 위험성, 처리 환경의 식별 위험성 등 판단 항목을 누락 없이 검토하였는지 개인식별 위험성 체크리스트 및 결과보고서 기반으로 검토

< ※ 연구목적에 따른 연구방법 설정 및 데이터 전처리 (예시) >

- (연구목적 검토) 가상상담 시나리오 생성 AI 개발에는 상담 목적, 고객 특성(성별, 연령, 지역)에 따른 요청·문의 내용 및 질의-응답에 따른 대화 흐름이 중요하며 **실제 고객 및 상담사의 음성 자체는 필요하지 않음**
- (데이터 전처리) **음성변환**(STT, Speech To Text) **엔진**을 통해 **상담음성 녹음파일의 음성을 모두 텍스트로 변환**하고, 변환된 텍스트에서 제대로 변환되지 않은 단어·문장, 비문 등을 **전문 작업자가 보완·수정**하여 텍스트 데이터셋을 구축

※ 음성데이터를 텍스트데이터셋으로 변환하여 활용할 시 상담음성 데이터 원본에 포함된 고객 및 상담사의 실제 음색, 억양, 발음 등을 통한 **개인식별 위험성을 제거할 수 있음**

⇒ **(음성→텍스트)로 변환된 데이터를 가명처리하여 AI 학습에 활용**하고자 함

< 데이터 전처리(STT) 완료된 텍스트 데이터셋 (예시) >

발화자	대화 텍스트
(고객)	아 저희가 지금 인터넷TV 그 사용하고 있는데 속도가 좀 너무 안 나와서요
(상담원)	아 설치 시보다 인터넷 속도가 너무 느리시다는 말씀이신 거죠
(고객)	네
(상담원)	이용 많이 불편하실 텐데 너무 죄송하구요 우선 고객 확인 먼저 하겠습니다. **홍길동** 고객님 본인 맞으실까요?
(고객)	맞습니다
(상담원)	가입해주신 고객님 혹시 **1987년 7월 22일생** 맞으실까요?
(고객)	네 맞습니다
(상담원)	확인 감사드리구요. 인터넷 속도저하 부분이 와이파이나 인터넷 모두 동일한 현상인가요?
(고객)	와이파이가 조금 많이 느려요.. 그래서 TV가 자꾸 끊기거든요
···(중략)···	
(상담원)	**인터넷 가입하신지 12년 5개월**되셨는데, **가족 세 분**과 가족할인 요금제로 사용하고 계시거든요. **다음 달 8일부터 신규 상품으로 변경**하시면서 **배우자분이랑 아드님**도 가족결합할인으로 추가하시겠다는거 맞으시죠? 상품먼저 배송해드릴게요. 주소가 어떻게 되시죠?
(고객)	**서울시 신뢰동 신뢰아파트 백일동 구백삼호**입니다.
(상담원)	네 고객님 **서울시 신뢰동 신뢰아파트 백일동 구백삼호**로 배송해드리겠습니다
(고객)	찾기가 좀 어려우실텐데 **신뢰역 3번출구 앞에서 오른쪽으로 쭉 올라오시면** 돼요
(상담원)	네 기사분께 전달드려놓겠습니다. 기사분 내방전에 전화드리고 갈 건데 **휴대폰 번호 가운데가 팔천오백**이니까 모르는 연락처라도 통화 잘 부탁드리겠습니다
(고객)	네 알겠습니다. 혹시 안 받으면 저희 회사쪽으로 연락주시면 됩니다. **공이 일이삼사 오육칠팔**입니다.
(상담원)	네 혹시 또 다른 문의사항 있으실까요?
(고객)	아니요 없어요.
(상담원)	하하하 네 이상 상담원 **이선정**이었습니다

개인정보 유형 분류표

연번	항목명	구분	설명	비고
1	고객상담 음성파일	고객-상담사 간 상담녹취음성	▪ 상담사와 고객 간 질의-응답으로 이루어진 음성 데이터	음성 (비정형데이터) **⇒ STT 변환후 삭제**

(변환: STT) ↓

※ 음성변환(STT, Speech To Text) 기술을 통해 텍스트로 변환하여 활용
각 음성파일에 고객 관련 메타데이터(상담목적, 성별, 연령, 주거지역) 포함

2	고객상담 텍스트 파일	① 이름	▪ 고객 및 상담사의 이름	텍스트 (비정형데이터)
		② 생년월일	▪ 고객의 생년월일과 관련된 정보	
		③ 주소정보	▪ 고객의 상세 주소와 관련된 정보	
		④ 거주정보	▪ 상세 주소는 아니나 고객이 거주하는 공간과 관련된 정보	
		⑤ 일반 전화번호	▪ 고객의 집·회사 연락처	
		⑥ 휴대폰번호	▪ 고객의 휴대폰 연락처	
		⑦ 일반대화정보	▪ 고객과 상담사 간 질의-응답 내 대화의 맥락을 통해 고객, 상담사, 또는 다른 특정인을 식별할 위험이 있는 대화정보	
	2-1	상담 목적	▪ 고객의 상담 목적 (1.단순문의, 2.A/S, 3.결제, 4.교환, 5.설치, 6.반품, 7.기타)	각 상담파일의 메타데이터 (정형데이터)
	2-2	성별	▪ 고객의 성별 (1. 남성, 2. 여성)	
	2-3	연령	▪ 고객의 연령 (1세단위, 17세~74세까지 분포)	
	2-4	주거지역	▪ 고객의 주거지역 (1.서울, 2.부산, 3.대구, 4.인천, 5.광주, 6.대전, 7.울산, 8.세종, 9.경기, 10.강원, 11.충북, 12.충남, 13.전북, 14.전남, 15.경북, 16.경남, 17.제주)	

개인정보 유형별 예시

연번	항목명	구분	예시
2	고객상담 텍스트 파일	① 이름	☞ “홍길동 고객님 본인 맞으실까요?” ☞ “네 감사합니다. 이상 상담원 이선정이었습니다”
		② 생년월일	☞ “가입해주신 고객님 혹시 1987년 7월 22일생 맞으실까요?”
		③ 주소정보	☞ “고객님 서울시 신뢰동 신뢰아파트 백일동 구백삼호로 배송해드리겠습니다”
		④ 거주정보	☞ “찾기가 좀 어려우실텐데 신뢰역 3번출구 앞에서 오른쪽으로 쭉 올라오시면 돼요”
		⑤ 일반 전화번호	☞ “저희 회사쪽으로 연락주시면 됩니다. 공이 일이삼사 오육칠팔입니다”
		⑥ 휴대폰번호	☞ “기사분 내방전에 전화드리고 갈 건데 휴대폰 번호 가운데가 팔천오백이니까”
		⑦ 일반대화정보	☞ “인터넷 가입하신지 12년 5개월되셨는데, 가족 세 분과 가족할인 요금제로 사용하고 계시거든요”, “다음 달 8일부터 신규 상품으로 변경하시면서 배우자분이랑 아드님도 가족결합할인으로 추가하시겠다는거 맞으시죠?”
	2-1	상담 목적	▪ ‘5’ (설치 목적)
	2-2	성별	▪ ‘2’ (여성)
	2-3	연령	▪ ‘25’ (25세)
	2-4	주거지역	▪ ‘1’ (서울)

활용데이터 요구 수준표

※ 가명정보 처리 목적 달성에 필요한 데이터 항목과 가명처리 요구수준을 검토

연번	항목명	구분	요구 수준
2	고객상담 텍스트 파일	① 이름	연구목적 달성에 필요 없으므로, 다른 이름으로 대체 가능
		② 생년월일	연구목적 달성에 필요 없으므로, 다른 값으로 대체 가능 (고객 연령대는 필요하나, 메타데이터를 통해 확인 가능)
		③ 주소정보	연구목적 달성에 필요 없으므로, 다른 값으로 대체 가능
		④ 거주정보	연구목적 달성에 필요 없으므로, 다른 값으로 대체 가능
		⑤ 일반 전화번호	연구목적 달성에 필요 없으므로, 다른 값으로 대체 가능
		⑥ 휴대폰번호	연구목적 달성에 필요 없으므로, 다른 값으로 대체 가능
		⑦ 일반대화정보	상담 목적 판단, 대화 맥락 인식 등에 필요 다른 항목들을 충분히 가명처리하였을 경우 그대로 활용하거나 훼손 최소화 필요
2-1	상담 목적		연구목적 달성을 위해 그대로 사용 필요
2-2	성별		연구목적 달성을 위해 그대로 사용 필요
2-3	연령		고객 연령대별 구분이 필요하므로, 10~20대, 30~40대, 50~60대, 70대 이상으로 범주화 가능
2-4	주거지역		연구목적 달성을 위해 그대로 사용 필요

식별 위험성 검토 결과보고서

※ 식별 위험성 검토 점검표를 기반으로 식별 위험성 검토 결과보고서 작성

<table>
<tr><td>가명정보
활용목적</td><td colspan="2">○ 콜센터 직원 교육용 가상상담 시나리오 생성 AI 모듈 개발</td></tr>
<tr><td>가명처리 대상
데이터 항목</td><td colspan="2">○ 고객상담텍스트 파일 내 개인식별 위험성이 있는 정보 일체
(이름, 생년월일, 주소정보, 거주정보, 일반 전화번호, 휴대폰번호, 일반대화정보 등)</td></tr>
<tr><td rowspan="3">데이터
위험성</td><td>식별성
유무</td><td>○ 이름, 주소, 휴대폰 번호는 개인식별 위험성이 상당히 높음
○ 거주정보, 일반 전화번호는 다른 대화내용 및 메타데이터와 연계될 시 개인식별 위험성이 존재
○ 일반대화내용도 상황에 따라 다른 대화내용 및 메타데이터와 연계될 시 낮은 확률로 개인식별 위험성이 생길 수 있음</td></tr>
<tr><td>특이정보
유무</td><td>○ 메타데이터 내 ‘연령’은 아주 낮거나 높을 시 특이치 존재 가능
○ 일반대화내용에 포함된 상품명, 상품변경날짜 및 내용, 가족관계 등으로 인해 특이정보가 존재할 수 있음</td></tr>
<tr><td>재식별시
영향도</td><td>○ 콜센터 일반상담 업무 및 업종 특성을 고려하면 관련 대화내용은 재식별 시 영향도가 미미함</td></tr>
<tr><td rowspan="3">처리 환경
위험성</td><td>이용 및
제공 형태</td><td>○ 동일 개인정보처리자, 동일 부서 내 활용</td></tr>
<tr><td>처리 장소</td><td>○ 빅데이터본부 내 Private 클라우드를 구축하여 망분리된 폐쇄형 개발환경에서 처리
○ 개인정보(가명정보)처리시스템에 대한 ISMS-P인증 취득</td></tr>
<tr><td>다른 정보
결합 가능성</td><td>○ 관리자 승인하에 데이터(프로그램 패키지, 라이브러리, 코드설명서 등) 반입이 가능함
○ 기존 가명정보와 결합가능성 있는 데이터는 반입이 제한됨</td></tr>
<tr><td>최종
검토의견</td><td colspan="2">▪ 해당 가명정보는 기관 내 연구자에게 제공될 뿐만 아니라, 클라우드 기반의 폐쇄연구 분석환경(외부 인터넷 이용, 다른 데이터의 반입 및 결합, 데이터 외부반출 등이 제한)에서만 접속하여 분석 가능하므로 식별 가능성이 낮은 편이며 안전한 처리 환경을 고려할 때 다음과 같은 조치가 필요함
- (고객상담 텍스트 파일) 실제 고객과 상담을 진행하고 관련 원본 음성데이터를 보유하고 있는 한 동일 부서에서 활용되므로, 특정 고객이 식별되지 않도록 개인식별정보(이름, 주소, 휴대폰 번호)뿐만 아니라 일반 대화정보까지 특정 고객이 식별될 여지가 없도록 보다 철저한 가명처리 필요
(1) 이름 : 랜덤 이름 또는 교육용으로 설정한 가상의 이름으로 치환
(2) 생년월일 : 범주화된 연령값 중에 랜덤한 값으로 치환
(3) 주소정보 : 가상의 주소값으로 치환
(4) 거주정보 : 가상의 거주정보로 치환하고, 대체가 어려운 경우 삭제</td></tr>
</table>

	(5) 일반 전화번호 : 가상의 번호로 치환 (6) 휴대폰 번호 : 가상의 번호로 치환 (7) 일반대화정보 : 그대로 사용하되, 전수검사하여 대화 맥락이나 특이정보 등을 고려하여 특정 개인이 식별될 가능성이 있다고 판단되는 문장은 삭제 - **(메타데이터)** 상담목적, 성별, 주거지역(시 단위)은 개인식별 위험이 거의 없고 연구목적 달성에 필요하므로 그대로 활용하되, 연령정보는 다른 정보와 결합하여 개인식별 위험이 있으므로 범주화하여 활용

④ 가명처리 단계에서 위험성 검토 결과를 반영하여 가명처리 방법 및 수준을 적정하게 정의하였는지 확인

항목별 가명처리 계획

연번	항목명	세부 항목	처리방법	세부방법 및 처리수준
1	고객상담 음성파일	고객-상담사 간 상담녹취음성	☑ 삭제	▪ 음성변환(STT)을 통해 텍스트로 변환한 후 원본 음성파일은 삭제

(변환: STT) ↓ **※ 음성변환(STT, Speech To Text) 기술을 통해 텍스트로 변환**

2	고객상담 텍스트 파일	① 이름	☑ 치환	▪ 고객 이름은 [NAME_CUSTOMER] 토큰 정보로 변환한 뒤 '김행복'으로 변경 ▪ 상담원 이름은 [NAME_CENTER] 토큰 정보로 변환한 뒤, '김신뢰'로 변경
		② 생년월일	☑ 치환	▪ 생년월일은 [BIRTH] 토큰 정보로 변환한 뒤, 메타데이터 중 연령대(10~20대, 30~40대, 50~60대, 70대 이상) 범주 내에서 랜덤한 값으로 치환
		③ 주소정보	☑ 치환	▪ 주소는 [ADDR] 토큰 정보로 변한한 뒤, 메타데이터 내 주거지역(시단위) 내 가상의 주소로 대체
		④ 거주정보	☑ 삭제 (또는 대체)	▪ 검수자(시나리오 작가)가 직접 가상의 문장으로 각색하고, 대체가 어려운 경우 해당 질의-응답 문장 삭제
		⑤ 일반 전화번호	☑ 치환	▪ 일반전화번호는 [PHONE] 토큰 정보로 변환한 뒤, 실재하지 않는 가상 전화번호로 대체
		⑥ 휴대폰번호	☑ 치환	▪ 휴대폰번호는 [CELLPHONE] 토큰 정보로 변환한 뒤, 실재하지 않는 가상 휴대폰번호로 대체
		⑦ 일반대화정보	☑ 그대로 사용 (또는 삭제)	▪ 연구 목적에 필요하므로 그대로 사용하는 것을 원칙으로 하되, 검수자가 전수 검사하여 대화 맥락이나 특이정보 등을 고려하여 특정 개인이 식별될 가능성이 있다고 판단되는 문장은 삭제 처리
2-1	상담 목적		☑ 그대로 사용	▪ 연구 목적 달성을 위해 별도 처리하지 않음
2-2	성별		☑ 그대로 사용	▪ 연구 목적 달성을 위해 별도 처리하지 않음
2-3	연령		☑ 범주화 (20세 단위)	▪ 고객 연령대 별 분석이 필요하므로, 20세 단위 범주화 처리하고 70세 이상은 상단코딩 적용 (10~20대, 30~40대, 50~60대, 70대 이상)
2-4	주거지역		☑ 그대로 사용	▪ 연구 목적 달성을 위해 별도 처리하지 않음

⑤ 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 확인

주요 비정형데이터 가명처리 수행 결과 (예시)

순번	항목명	세부 항목	가명처리 전	가명처리 후
2	고객상담 텍스트 파일	① 이름	▪ 홍길동 고객님 본인 맞으실까요? <고객이름 '김행복'으로 치환>	▪ '[NAME_CUSTOMER]' 고객님 본인 맞으실까요? ⇒ '김행복' 고객님 본인 맞으실까요?
			▪ 네 감사합니다. 이상 상담원 이선정이었습니다 <상담원 이름 '김신뢰'로 치환>	▪ 네 감사합니다. 이상 상담원 '[NAME_CENTER]' 이었습니다" ⇒ 네 감사합니다. 이상 상담원 '김신뢰'였습니다
		② 생년월일	▪ 가입해주신 고객님 혹시 1987년 7월 22일생 맞으실까요? <메타데이터의 범주화된 연령(30~40대) 내 랜덤값으로 치환>	▪ 가입해주신 고객님 혹시 '[BIRTH]'생 맞으실까요? ⇒ 가입해주신 고객님 혹시 '1991년 6월 8일' 생 맞으실까요?
		③ 주소정보	▪ 고객님 서울시 신뢰동 신뢰아파트 백일동 구백삼호로 배송해드리겠습니다 <메타데이터의 주거지역(서울) 내 가상주소로 치환>	▪ 고객님 '[ADDR]'로 배송해드리겠습니다 ⇒ 고객님 '서울시 신뢰동 신뢰아파트 1동 1호'로 배송해드리겠습니다.
		④ 거주정보	▪ (질의) 찾기가 좀 어려우실텐데 신뢰역 3번출구 앞에서 오른쪽으로 쭉 올라오시면 돼요 ▪ (응답) 네 기사분께 전달드리겠습니다.	<삭제> ⇒ 문맥상 다른 문장으로 대체가 어려우므로, 질의-응답문 세트 전체를 삭제 * 시나리오 작가가 가상의 문장으로 각색하는 것도 가능
		⑤ 일반 전화번호	▪ 저희 회사쪽으로 연락주시면 됩니다. 공이 일이삼사 오육칠팔입니다 <일반전화번호 가상 전화번호로 치환>	▪ 저희 회사쪽으로 연락주시면 됩니다. '[PHONE]'입니다 ⇒ 저희 회사쪽으로 연락주시면 됩니다. '03-111-1111'입니다
		⑥ 휴대폰 번호	▪ 기사분 내방전에 전화드리고 갈 건데 휴대폰 번호 가운데가 팔천오백이니까 <휴대폰번호 가상 전화번호로 치환>	▪ 기사분 내방전에 전화드리고 갈 건데 휴대폰 번호 가운데가 '[CELLPHONE]'이니까 ⇒ 기사분 내방전에 전화드리고 갈 건데 휴대폰 번호 가운데가 '일일일일'이니까
		⑦ 일반 대화정보	▪ 인터넷 가입하신지 12년 5개월되셨는데, 가족 세 분과 가족할인 요금제로 사용하고 계시거든요 ▪ 다음 달 8일부터 신규 상품으로 변경하시면서 배우자분이랑 아드님도 가족결합할인으로 추가하시겠다는거 맞으시죠?	<그대로 사용> ▪ 인터넷 가입하신지 12년 5개월되셨는데, 가족 세 분과 가족할인 요금제로 사용하고 계시거든요 ▪ 다음 달 8일부터 신규 상품으로 변경하시면서 배우자분이랑 아드님도 가족결합할인으로 추가하시겠다는거 맞으시죠? * 단, 전수검사하여 대화 맥락이나 특이정보를 고려하여 개인식별 위험이 있는 문장은 삭제·변경 필요

○ 비정형 데이터 가명처리 기술의 적절성・신뢰성 관련 근거(예시)

연번	항목명	세부 항목	처리 기술명	처리 기술의 적절성·신뢰성 관련 근거 또는 배경
2	고객상담 텍스트 파일	① 이름 ② 생년월일 ③ 주소정보 ⑤ 일반전화번호 ⑥ 휴대폰번호	치환	▪ '이름', '생년월일', '주소정보', '일반전화번호', '휴대폰번호'는 'A사의 텍스트 비식별화 솔루션'을 활용하여 정해진 텍스트나 정해진 범주 내 랜덤텍스트로 대체 - A사의 텍스트 비식별화 솔루션은 자연어 처리 기준으로 텍스트 데이터셋에서 93%의 검출 정확성으로 측정되었으며, 검출된 텍스트에 대한 처리 정확도는 99%로 측정 *※ 검출 정확도, 처리 정확도(오류율) 증빙자료 별첨* - 검출 정확도가 100%가 아니므로, 솔루션 적용 후 처리결과에 대해 추가적인 자체 전수검사를 수행
		④ 거주정보	삭제 (또는 대체)	▪ 정형화된 패턴이 없어 솔루션으로 검출이 어려우므로, 자체 육안 전수검사를 통해 검출하여 삭제 처리 ⇒ 대화 맥락상 중요하고, 개인식별성이 없도록 변경이 가능한 경우 시나리오 작가가 각색 수행)
		⑦ 일반대화정보	삭제	▪ 대화맥락이나 특이정보 등을 통해 개인식별 위험이 생기는 문장에 대해서는 정형화된 패턴이 없어 솔루션으로 검출이 어려우므로, 자체 육안 전수검사를 통해 검출하여 삭제 처리 ⇒ 대화 맥락상 중요하고, 개인식별성이 미미하다고 판단되는 경우에는 삭제하지 않고 그대로 활용)

비정형데이터 가명처리 결과에 대한 자체 검증 결과서

<table>
<tr><td rowspan="8">검증 대상
데이터 명세</td><td colspan="3">개요</td></tr>
<tr><td colspan="3">㈜ 코리아인터넷은 고객상담 음성데이터를 음성변환(STT)을 통해 텍스트데이터로 변환한 뒤, 개인식별 위험성이 있는 대화문장에 대해 대체·삭제처리</td></tr>
<tr><td>데이터 유형</td><td colspan="2">텍스트</td></tr>
<tr><td>원본 데이터 형식(파일 포맷)</td><td colspan="2">WAV ⇒ JSON</td></tr>
<tr><td>처리 결과 데이터 형식(파일 포맷)</td><td colspan="2">JSON</td></tr>
<tr><td>데이터 규모 및 크기</td><td colspan="2">녹취파일(WAV) 10,000개(각 3분 분량, 총 500시간, 60GB)를 텍스트로 변환한 데이터셋</td></tr>
<tr><td>대상 데이터 항목명</td><td colspan="2">고객상담 텍스트 파일 <연번 2></td></tr>
<tr><td>가명처리 적용 기술</td><td colspan="2">- A사 ‘텍스트 비식별화 솔루션’을 활용하여 이름, 생년월일, 주소, 연락처 정보를 검출하여 토큰 정보로 변환한 뒤 정해진 텍스트나 정해진 범주 내 랜덤값으로 치환
- 일반대화정보 내 개인식별 가능성이 있는 단어, 문장은 직접 검수하여 삭제 또는 각색</td></tr>
<tr><td>자체 검증
기간</td><td colspan="3">2023년 2월 10일 ~ 2023년 2월 20일</td></tr>
<tr><td>자체 검증
장소</td><td colspan="3">㈜코리아인터넷 녹취통합시스템 분석PC</td></tr>
<tr><td>자체 검증
과정
및 방법</td><td colspan="3">(검증방법)
- 개인정보보호팀 담당 팀장 주도하에 총 2명의 검수자가 Workday 10일간 전수조사
- 각색이 필요한 문장을 선별하여 시나리오 작가 2인이 각색 수행

(검증 시 확인사항)
① 치환되어야 할 항목들이 정해진 텍스트나 정해진 범주 내 랜덤값으로 치환되었는가?
➜ 치환되지 않았을 시 수작업으로 변환 실시
② 치환대상 외에 일반대화정보 내 개인식별 가능성이 있는 단어, 문장이 포함되어 있는가?
➜ 발견 시 팀장과 논의 하에 삭제할지 가상의 내용으로 각색할지 결정
➜ 각색 대상으로 선별된 문장은 2명의 시나리오 작가가 각색 수행</td></tr>
<tr><td>자체 검증
결과</td><td colspan="3">- 수작업으로 이름, 생년월일, 주소, 연락처 정보 추가 변환 : 6,582개 파일
- 가상의 내용으로 일부 문장 각색 : 231개 파일
> 10,000개 파일 전체 전수검사 완료 및 이상 없음 확인</td></tr>
<tr><td rowspan="6">자체 검증자</td><td>소속 및 직위</td><td>성명</td><td>서명(인)</td></tr>
<tr><td>개인정보보호 팀장</td><td>가관리</td><td></td></tr>
<tr><td>음성 AI팀 인턴</td><td>나정보</td><td></td></tr>
<tr><td>음성 AI팀 인턴</td><td>다보호</td><td></td></tr>
<tr><td>음성 AI팀 시나리오 작가</td><td>라작가</td><td></td></tr>
<tr><td>음성 AI팀 시나리오 작가</td><td>마작가</td><td></td></tr>
</table>