의료 합성데이터 적정성 검증 사례 연구

김성현[1], 신신애[1], 조연제[1]
[1]한국지능정보사회진흥원 지능데이터본부
Kimcon@nia.or.kr, Sashin@nia.or.kr, Yjyjyjcho@nia.or.kr

A Case Study on Medical Synthetic Data Sets Evaluation

Sung Hyun Kim[1], SinAe Shin[1], Yeon-je Cho[1]
[1]Dept. of AI Data, National Information Society Agency

요 약

과학기술정보통신부와 한국지능정보사회진흥원은 2022년부터 인공지능 학습용 데이터를 구하기 힘든 의료 분야의 AI 학습용 합성데이터를 구축하여 제공하고 있다. 하지만 구축한 합성데이터에 대해 안전성 측면의 검증을 수행하지 않아 다운로드가 불가능한 AI허브의 온라인 안심존을 통해서만 제공하였다. 데이터 활용성의 향상을 위해 합성데이터는 자유로운 활용이 가능한 형태로 개방 되어야 한다. 본 연구에서는 개인정보위원회에서 발간한 합성데이터 생성 참조모델(2024.5)에 따라 검증한 2개 데이터 사례를 제시하여 검증 내용에 대한 구체적인 정보를 제공하고 다른 합성데이터 생성과 검증에 대한 가이드를 제시하였다.

1. 서론

의료분야 AI 개발을 위해서는 양질의 학습 데이터 활용이 중요하나, 민감정보가 포함된 의료데이터 특성상 전면적 개방에 제약이 있다. 이러한 제약을 피하기 위해서는 현행 개인정보보호법이나 생명윤리 및 안전에 관한 법률(생명윤리법)에 저촉되지 않는 데이터의 구축과 보급이 필요하다. 합성데이터는 원본 데이터와 특성이 유사하지만 실제 원본 데이터와는 다른 데이터로 가상으로 재현된 데이터를 의미한다[1]. 의료데이터는 관련 법령에 의해 데이터 가공과 활용을 위한 기관 연구윤리위원회(IRB, Internal Review Board), 데이터 심의위원회(DRB; Data Review Board)의 심의 필요하지만, 가상으로 생성된 합성데이터는 현행법에 저촉될 여지가 적다. 하지만 개인의 의료정보를 활용하였으므로 이에 대한 검증또한 필요한 상황이다. 본 연구에서는 한국지능정보사회진흥원에서 지원사업으로 구축한 2종의 데이터를 개인정보보호위원회의 『합성데이터 생성 참조모델(2024.5)』에 기반하여 검증하고 배포한 사례를 제공하여 관련 연구에 가이드라인을 제공하고자 한다[2].

2. 합성데이터(Synthetic Data) 개요

합성데이터 기술은 민감 데이터의 개인정보 보호, 국가 안보 등 접근하기 어려운 데이터의 법적 제약 문제 해결, 데이터의 희소성을 극복하기 위한 데이터 증강 기법의 일환으로 활용되고 있다. 합성데이터는 생성형 AI를 활용한 개인정보 보호 강화기술(PET, Privacy-Enhanced Technology)로서 개인정보를 보호하면서도 데이터의 효용을 보장하는 기술로서 '23년 6월 ICO(영국 개인정보감독기구)를 통해 'PET 가이드라인'에서 소개된 바 있다[3].

생성형 AI를 활용한 이미지 합성데이터 모델 알고리즘은 GAN(Generative Adversarial Network)과 SytelGAN 등 GAN 기반모델부터 Diffusion 기반 모델에 이르기까지 많은 발전이 있었다. 이미지·영상 모달리티 합성데이터 분야에서는 최근에는 Stable Diffusion 모델을 활용하여 Text-to-Image 등 멀티모달 모델을 활용한 합성데이터 생성 또한 이루어지고 있다[4]. 자연어 모달리티 합성데이터 분야에서는 최근 LLM을 활용하여 프롬프트 엔지니어링을 활용한 합성데이터 생성이 연구된 바 있다[5]. 이외에도 음성데이터, 시계열 데이터 모달리티(Modality)별로 Transformer 기반의 합성데이터 모델들이 연구되고 있다. 합성데이터를 생성하는데 쓰이는 StyleGAN과 Diffusion 계열 알고리즘은 비가역적인 특성을 지니고 있으며, 임의의 난수값에서 시작해 원본 데이터의 스타일을 학습하여 이미지를 생성하거나, 디노이징 과정을 통해 이미지를 생성한

다. 이 과정에서 원본과 동일한 이미지를 얻을 가능성은 거의 없다. StyleGAN 사용자는 Style 파라미터를 조정하여 원하는 방향으로 학습을 조정할 수 있다. 이 알고리즘들은 고해상도의 합성 이미지를 생성할 수 있다는 장점이 있다[6, 7].

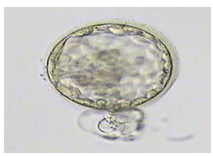
3. 의료분야에서의 합성데이터와 공유 문제

의료분야의 인공지능 학습을 위한 데이터 생산을 위해서는 다양한 임상학적 데이터 확보가 필요하다. 의료영상 데이터의 다양한 분석 및 모델 편향 문제 해결을 위한 레이블링 적용한 데이터 및 다양한 질병군의 상세 데이터 필요하며, 실제 임상 영상데이터의 경우 법적 제한으로 인해 일반적인 사용자 중심의 데이터 공유 체계구축이 어렵다. 이러한 문제해 대해 합성데이터는 유력한 대안이 될 수 있다. 의료 분야 합성데이터는 초음파, CT, MRI, X-ray와 같은 의료 이미지·영상 EMR(Electronic Medical Records), EHR(Electoronic Health Records)와 같은 정형데이터 및 자연어 데이터 분야 그리고 ECG(Electro Encephalo Gram) 데이터, EEG (Electro Encephalo Graphy)와 같은 시계열 데이터와 같이 모달리티 분야별로 합성데이터 생성 모델이 연구되고 있다. 이미지 데이터에서는 GAN을 이용하여 기존 데이터의 분포를 학습해 새로운 데이터를 만드는 사례가 대표적이다. Waheed et al.은 GAN으로 생성된 영상이 실제영상와 큰 차이가 없으며 영상이 비식별화 되어 공유가 용이하다는 장점이 있고, 실제 흉부 X-ray를 GAN으로 합성하여 COVID-19를 자동식별하여 95%의 정확도를 선보인 사례가 있다[8]. 이스라엘의 SMC는 2018년 간 병변 진단 AI에 합성데이터 추가 사용으로 정확도를 78.6%에서 85.7%로 향상하였고, 국내 스타트업인 CNAI는 위 내시경의 암 이미지 검출에 비지도학습으로합성데이터를 활용하여 기존 모델의 성능을 80%에서 89%로 향상 시킨 바 있다. 의료분야에서 개인정보가 포함된 데이터는 개인정보보호와 윤리적 관점에서의 데이터 공유 한계가 있다. 대한민국에서는 법적으로는 개인정보보호법에 의한 IRB와 생명윤리법에 의한 DRB로 인해 데이터 사용의 진입 장벽이 매우 높다. 적절한 절차를 거쳤다고 해도 개인의 의료정보를 가공하였다고 오해할 수 있어 검증의 대상이 된다. 이러한 문제로 인해 의료분야의 데이터 이용은 제한적인 상황이며, 합성 알고리즘의 비가역성으로 인해 개인정보 문제가 없다고 볼 수 있는 합성데이터 또한 개인정보 문제가 없다는 대내외적인 검증과 증명이 필요하다.

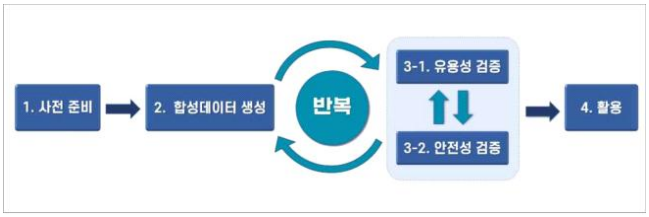
4. NIA 합성데이터

검증의 대상이 되는 데이터는 2023년 인공지능 학습용 데이터 구축사업의 일환으로 제작되었다. 아주대학교 컨소시움과 서울대학교 치과병원 컨소시움은 정부의 지원을 받아 합성데이터 구축사업을 수행하였다. 개별 합성데이터는 기 구축된 원시 데이터의 사용허가를 받아 합성데이터 알고리즘을 활용해 생성되었으며 전문의 이상의 자격을 지닌 전문가가 검수하였다. 품질 전문기관인 한국정보통신기술협회가 품질 검증을 총괄하였다. 데이터의 개요는 <표 1>에 기술하였다.

<표 1> 합성데이터 개요

데이터명	배아 이미지 합성데이터	구강 이미지 합성데이터
주관기관 참여기관	아주대학교병원 (카이헬스)	서울대학교치과병원 (한양대학교)
내용	배아의 현미경 및 타임랩스 이미지	구강의 전면부를 비롯한 상하좌우의 촬영 이미지
샘플		
구축건수	61,133건	145,140건

5. 검증방법론 및 검증결과



(그림 1) 개인정보보호위원회 참조모델[2]

그간 구축된 데이터는 AI 데이터 공개 포털인 AI 허브(www.aihub.or.kr)에서 데이터 다운로드가 불가능한 온라인 안심존을 통해 제공되어 왔는데, AI허브 개방을 통한 데이터의 접근성 및 활용성 향상을 도모하고자 지난 6월 개인정보보호위원회가 발간한 「합성데이터 생성 참조모델」의 절차에 따라 적정성을 심의하였다. 심의는 합성데이터에 대한 전문성이

있는 정보보호, 법률 전문가와 인공지능 전문가, 의료와 금융 전문지식이 있는 전문가가 참여하여 사전준비, 생성과정, 유용성 및 안전성을 평가하였다(그림 1)은 합성데이터 생성과정을 도식화한 것이며 <표 2>는 이에 따른 검토 지표를 설명한 것이다.

<표 2> 합성데이터 검증지표

생성절차	검토 내용
사전준비	(1) 합성데이터의 활용 목적 및 범위가 적절하게 설정되었는가?
	(2) 합성데이터 생성에 활용된 원본데이터가 적법하게 확보되었는가?
합성데이터 생성	(3) 합성데이터 생성 방법이 적절하였는가?
	(4) 합성데이터 생성 과정에서 위험성은 없었는가?
유용성 검증	(6) 생성된 합성데이터는 목적 달성이 가능한가?
안전성 검증	(6) 생성된 합성데이터에는 개인식별 위험이 없는가?

심의 대상인 합성데이터들은 이상치를 제거하고 비식별 조치된 데이터들을 기관연구윤리심의회 등의 승인하에 기관내 폐쇄망에서 인공지능 모델을 학습하고, 원본 데이터를 추론할 수 없는 알고리즘을 적용하여 적절하게 생성되었음을 확인하였다. 품질전문기관이 주관한 성능비교 테스트를 통과해 실제데이터와 유사하고 인공지능 모델의 성능 향상도를 확인하여 유용성을 증명함에 따라 공개가 가능할 정도로 적정하게 구축되었음을 승인하였다.

<표 3> 합성데이터 심의내용

데이터명	배아 이미지 합성데이터	구강 이미지 합성데이터
사전준비	보건의료데이터 가이드라인 준수	기승인 IRB활용 신규 DRB 적용
합성데이터생성	식별자 제거, Deffusion 적용	식별자 제거 후 DDPM→ SDM→ SANTA→ESRGAN의 단계별 적용
유용성 검증	VTT 0.45 FID 16.47 AUROC 차이: 0.05	치아경계(mAP) 차이: 0.059 충치판단모델/ (accuracy) 차이: 0.056
안전성 검증	비가역적인 Diffusion 모델 사용	개별적인 치아 합성을 수행하고 이를 조합하여 완성

안전성 검증에서는 원본데이터와 합성데이터의 구조적 유사성을 측정하는 SSIM(Structural Similarity Index Measure)과 같은 지표는 사용하지 않고 비가역적인 알고리즘의 사용, 원본을 전혀 추론할 수 없는 분할정복(Divide and conquer) 방법론의 적용을 설명하여 인정을 받을 수 있었다. 본 사례는 개인정보보호위원회의 자체시행 사례를 제외한 공공부문 최초 적용이며 심의 내용은 <표 3>에 성과는 <표 4>에 정리하였다.

<표 4> 합성데이터 활용 성과

심의 전		심의 후
·합성데이터의 유용성에 대한 우려 존재 ※ 진짜 데이터와 동일, 혹은 우월한 성능 보장 필요 ·합성데이터의 안전성에 대한 불안 존재 ※ 합성데이터의 알고리즘 학습에 활용된 개인정보 탑재 원본 추론에 대한 우려 존재	⇨	·데이터의 유용성과 알고리즘의 비가역성을 전문가들이 검증하여 시장의 우려 불식 ※ VTT(Visual Turing Test) 등 인공지능분야에서 널리 공인된 방법으로 유용성 검증 ※ 개인정보보호를 주관하는 정부기관에서 발행한 참조모델을 기반으로 안전성 검증
·폐쇄된 환경(온라인 안심존)에서만 제공 ※ 연구자, 산업계 등에서 활용이 불편		·AI허브를 통한 개방 및 활용 ※ 접근성 및 활용성이 크게 향상

6. 결론 및 제언

본 연구에서는 일반적으로는 접하기 어려운 의료분야의 데이터를 활용하기 위한 합성데이터 검증 사례를 다루었다. 배아 이미지 합성데이터, 구강 이미지 합성데이터는 과학기술정보통신부가 주관하는 2024 ICT기금사업 우수사례로 선정되어 대외적으로도 성과를 인정받았다. 개인정보보호 이슈가 없는 합성데이터는 중소, 벤처, 스타트업, 학계 등 민간의 인공지능 기술개발, 인공지능 응용서비스, 제품 개발의 촉진에 기여 할 것이다. 또한 의료영상에 대한 대중의 이해도 증대 및 생태계 활성화에도 기여할 것으로 기대된다. 개방되는 합성데이터는 최근 급격하게 발전하고 있는 초거대 모델의 Medical Foundation Model의 이미지 학습 및 신규 서비스 창출에 유용하게 활용 될 수 있다. 본 연구는 합성영상데이터 학습 및 활용에 대한 가이드라인 제공을 통해 기존 인공지능 모델의 성능 향상에 기여할 것이다. 또한 특정지표 없이도 비가역성을 설명하여

안전성을 설명할 수 있다는 점도 성과로 볼 수 있다. 합성데이터는 비정형 데이터 뿐만 아니라, 센서스 데이터 같은 정형데이터에서도 널리 쓰이고 있는데 이에 대한 검증사례는 향후 사례 연구가 더 필요할 것이다. 또한 재난재해 상황, 교통사고 등 꼭 필요하지만 구하기 힘든 내용의 합성데이터의 생성과 검증시의 사례 공유도 더 필요할 것으로 판단된다.

Acknowledgment

저자들은 자료와 인터뷰를 제공해주신 데이터 합성 기관관계자들께 감사드립니다. 본 논문의 내용은 연구자들의 개인적인 의견일 뿐 소속기관과는 무관함을 알려드립니다.

참고문헌

[1] National Information Society Agency (2023). '가짜' 데이터가 만드는 '진짜' 인공지능 시대. IT & Future Strategy, Seoul: National Information Society Agency.

[2] 개인정보보호위원회, (2024) 합성데이터 생성 참조모델. 대한민국정부

[3] ICO, Privacy-enhanced technologies guideline, 2023. pp.17

[4] Yonglong et al. “StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners”. NeurIPS, 2023.

[5] Lin Long et al. “On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey”. 2024.. 6.

[6] Karras, T et al., Progressive growing of gans for improved quality, stability, and variation. arXiv preprint. 2017

[7] Karras, T., Alias-free generative adversarial networks. Advances in neural information processing systems, 34, 852-863. 2021.

[8] Waheed et al. “CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection, IEEE Access, vol. 8, pp. 91916-91923, 2020