의학 연구용 통합 데이터 플랫폼 구현

서제원[1], 지영석[2], 정미경[3], 김주한[4], 김경백[5]
[1]전남대학교 인공지능융합학과 석사과정
[2]전남대학교병원 의료정보센터 교수
[3]전남대학교병원 K-Health 사업단 팀장
[4]전남대학교병원 빅데이터연구센터 교수
[5]전남대학교 인공지능융합학과 교수
cnuh.datateam@gmail.com, yongsok.ji@jnu.ac.kr, mgjung2022@naver.com,
kim@zuhan.com, kyungbaekkim@jnu.ac.kr

Implementation of an Integrated Data Platform for Medical Research

Je-Won Seo[1 4], Yong-Sok Ji[4], Mi-Gyoung Jung[2], Ju-Han Kim[3],
Kyung-Baek Kim[4]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University
[2]Medical Information Center, Chonnam National University Hospital
[3]K-Health Business Group, Chonnam National University Hospital
[4]Bigdata Reseach Center, Chonnam National University Hospital

요 약

본 연구는 의료 연구에서 다양한 형태의 멀티모달 데이터를 통합적으로 관리하고 분석할 수 있는 통합 데이터 플랫폼을 구현하는 것을 목표로 한다. 이 플랫폼은 정형 데이터와 비정형 데이터를 일원화하여 연구자들이 쉽고 빠르게 접근할 수 있도록 설계되었으며, 데이터의 효율적 수집, 처리, 저장, 분석을 지원한다. 또한, 정형 데이터와 비정형 데이터 간 통합 검색 기능, 질환별 데이터 추출, 비식별화 기능 등을 구현하여 연구자들이 안전하고 신뢰성 있는 환경에서 데이터를 활용할 수 있도록 하였다. 이 플랫폼은 기존의 CDW 시스템의 한계를 극복하며, 의학 연구의 다양한 요구를 충족시킬 것으로 기대한다.

1. 서론

의료 데이터의 중요성은 날로 커지고 있으며, IT 기술의 발전에 따라 의료 연구자들은 데이터에 대한 다양한 분석 요구를 가지고 있다. 특히, 기존의 임상연구정보시스템(CDW; Clinical Data Warehouse)은 주로 정형 데이터만을 활용하여 분석을 수행해 왔으나, 이 같은 접근 방식은 복잡해지는 연구 요구를 충족시키기에는 한계가 있었다. 의료 현장에서 생성되는 데이터는 텍스트, 이미지, 영상 등 비정형 데이터를 포함한 멀티모달 형태로 다양화되고 있으며, 이를 효과적으로 통합하고 분석할 수 있는 새로운 플랫폼의 필요성이 부각 되고 있다.

본 연구에서는 이러한 요구에 대응하기 위해 정형 데이터와 비정형 데이터를 통합하여 연구자들이 손쉽게 접근하고 활용할 수 있는 통합 데이터 플랫폼을 구현하여, 멀티모달 데이터를 효율적으로 관리하고, 기존 시스템의 한계를 극복하여 의료 연구의 다양한 요구를 충족시키기 위한 기반을 마련하고자 하였다.

2. 관련 연구

CDW 시스템은 2010년대부터 국내 주요 대형 병원을 중심으로 도입되기 시작하여, 현재 많은 병원에서 운영되고 있다. 대표적으로, 서울대학교병원은 SUPREME이라는 이름으로 구축하여 운영 중이며, 삼성서울병원은 DARWIN-C, 서울성모병원은 nU라는 CDW 시스템을 통해 임상 데이터를 통합 관리하고 있다. 그러나 텍스트, 이미지, 영상 등의 비정형 데이터를 정형 데이터와 함께 통합하여 검색하고 추출할 수 있는 플랫폼 구축 사례는 많지 않다. 2020년 중국 West China Hospital of Sichuan University에서 WCH-BDP[1]를 구축하였고, 국내에는 가톨릭중앙의료원이 2021년에 EDP(Enterprise Data Platform)[2]을 구축하여 운영 중이다.

(그림 1) 통합 데이터 플랫폼 기능 개념도

3. 통합 데이터 플랫폼 설계

전남대학교병원은 1996년부터 수집된 데이터를 기반으로 현재 약 260만 명의 환자 데이터를 관리하고 있으며, 약 5,400개의 테이블과 5억 6천만 장의 의료영상 이미지를 포함하여 총 데이터 용량이 약 400TB에 이른다. 통합 데이터 플랫폼은 이러한 방대한 데이터를 활용한 의료 연구의 효율성을 높이고 연구자들이 다양한 데이터에 쉽게 접근할 수 있도록 크게 데이터의 수집, 포털, 플랫폼 세 가지 주요 모듈로 설계하였다(그림 1).

데이터의 수집 모듈은 전자의무기록(EMR), 의료영상(DICOM), 임상 데이터(CRF), 원무과 데이터(ADT), 검사 접수(ORD), 전자건강기록(EHR) 등 다양한 의료 데이터가 효율적으로 수집되도록 하였다. 수집된 데이터는 ETL(Extract, Transform, Load) 프로세스를 거쳐, 가공, 변환, 저장되어 일원화된 데이터 카탈로그가 형성되도록 하였으며, 이 과정은 데이터를 정제하고, 사용자가 요구하는 형식으로 변환하여 데이터의 활용성을 높일 수 있도록 하였다.

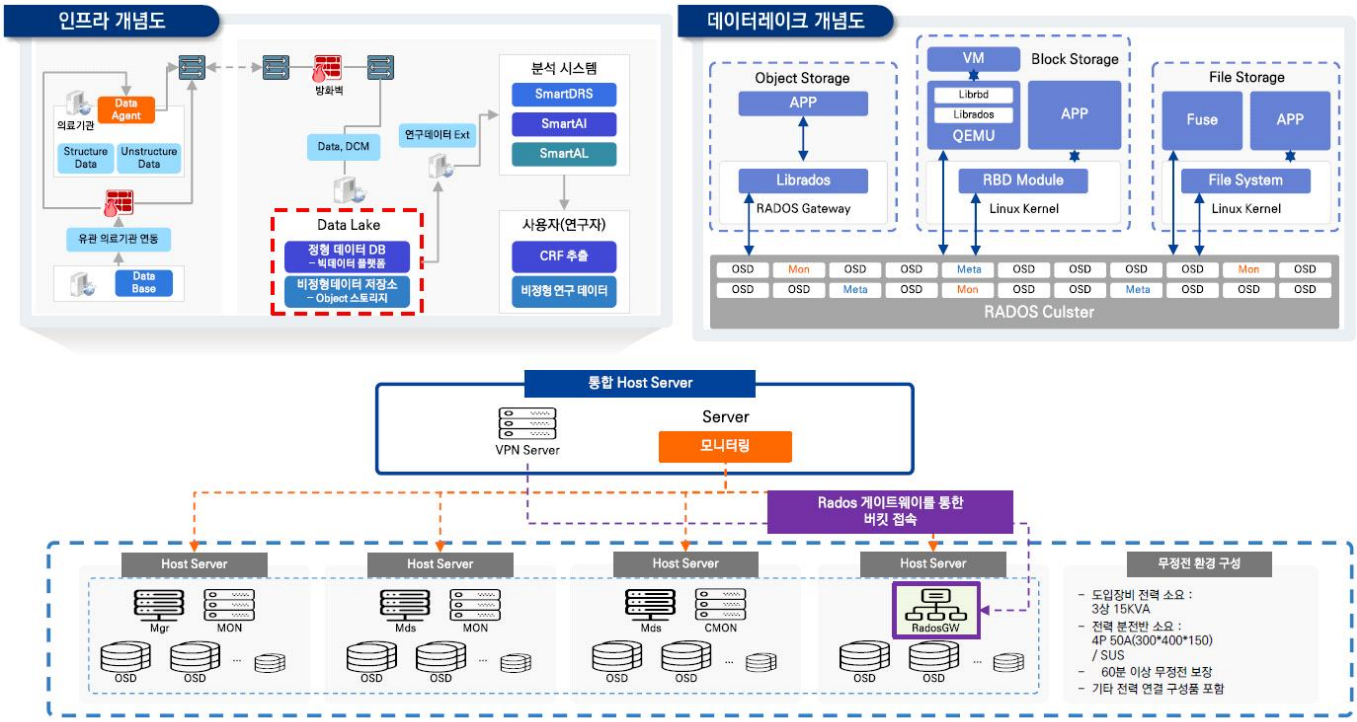
포털 모듈은 연구자들이 데이터에 쉽게 접근하고, 다양한 분석 도구를 활용할 수 있는 사용자 친화적인 인터페이스를 제공하도록 하였다. 이 모듈은 분석 대시보드, 데이터 시각화, 데이터 모니터링 기능을 갖추고 있어 연구자들이 데이터를 직관적으로 이해하고 활용할 수 있도록 하였다. 또한, 데이터 카탈로그 기반의 인터페이스는 연구자가 필요한 데이터를 쉽고 빠르게 검색할 수 있도록 설계하였다.

플랫폼 모듈은 데이터의 가상화와 통합을 담당하며, 정형 데이터와 비정형 데이터를 포함한 멀티모달 데이터를 일원화하여 관리할 수 있도록 하였다. 이 모듈은 데이터 카탈로그 관리, 메타데이터 관리, 비정형 데이터 분석 및 품질 관리, 개인정보 비식별 처리 등 데이터의 안전성과 신뢰성을 보장하기 위한 다양한 기능과 이를 통해 연구자들은 의료영상, 텍스트, 숫자 데이터 등을 하나의 플랫폼에서 통합적으로 활용할 수 있도록 하였다.

위의 세 가지 모듈과 별개로 플랫폼의 폐쇄형 클라우드와 데이터 레이크 구조는 데이터를 안전하게 관리하고, 고성능 연산 자원을 최적화하여 활용할 수 있도록 설계되었다. 폐쇄형 클라우드는 연구용 데이터를 외부 환경으로부터 보호하며, 데이터의 보안과 접근 제어를 강화하여 민감한 의료 데이터를 안전하게 저장하고 처리할 수 있는 환경을 제공할 수 있도록 하였다. 데이터 레이크는 정형 데이터와 비정형 데이터를 통합하여 저장하고 관리하는 핵심 구성 요소로, 다양한 데이터 소스를 통합적으로 처리할 수 있는 유연성을 갖추도록 하였다.

4. 통합 데이터 플랫폼 구현

통합 데이터 플랫폼은 데이터의 수집, 저장, 처리, 관리의 효율성을 극대화하기 위해 고성능 하드웨어 인프라를 기반으로 구현되었다. 플랫폼의 하드웨어 구성은 프라이빗 클라우드 환경과 데이터 레이크, 무정전 전력 시스템을 포함하며, 이를 통해 데이



(그림 2) 하드웨어 인프라 개념도 및 구성도.

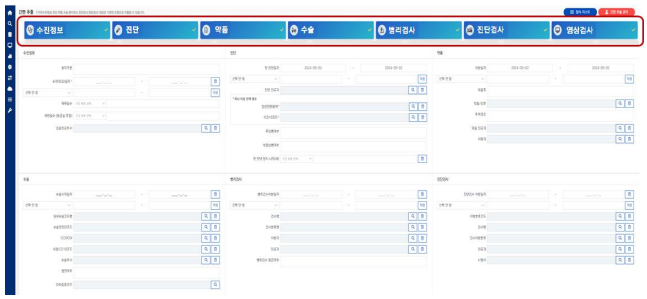
터의 안전한 관리와 고성능 연산을 지원하도록 하였다. 통합 데이터 플랫폼의 핵심은 폐쇄형 프라이빗 클라우드로, 컨트롤러 노드와 4개의 가상화 서버, GPU 서버로 구성하였다(그림 2). 각 가상화 서버는 고성능의 가상 환경을 제공하여 데이터 처리와 분석 작업을 수행하며, 11개의 컨테이너와 5개의 가상머신(VM)을 통해 유연한 자원 활용이 가능하다. GPU 서버는 이미지 내 개인정보에 대한 비식별처리를 수행하도록 하였다.

스토리지는 SSD와 HDD를 조합한 형태로 구성되어 있으며, 가상화 서버 스토리지는 SSD 구성하여 빠른 데이터 접근 속도를 제공하도록 하였고, 통합 저장 장치는 HDD로 구성하여 대규모 데이터 저장을 가능하게 하며, 안정적인 데이터 보관을 지원하도록 하였다.

데이터 레이크는 정형 데이터와 비정형 데이터를 통합 관리할 수 있는 핵심 구성 요소로, RADOS(Reliable Autonomic Distributed Object Store) 클러스터를 통해 다양한 형태의 스토리지 서비스를 제공하도록 하였다. RADOS 클러스터는 VM, 블록 스토리지, 파일 스토리지 등 여러 유형의 저장소를 지원하며, 데이터의 안전한 저장과 빠른 접근을 보장할 수 있다. 클러스터는 다수의 호스트 서버로 구성되어 있으며, 각 서버는 데이터 저장, 모니터링, 게이트웨이 기능을 수행한다. 분산 처리 구조를 채택하여 시스템의 성능과 안정성을 높였으며, 이를 통해 대규모 데이터의 효율적 관리와 신속한 데이터 처리 환경을 구축하였다.

5. 통합 데이터 플랫폼 기능

통합 데이터 플랫폼은 연구자가 데이터를 효과적으로 활용할 수 있도록 조회 조건의 복잡도와 대상 테이블의 범위에 따라서 간편 추출과 고급 추출 기능으로 나누었다.



(그림 3) 간편 추출.

간편 추출은 연구자가 주로 활용하는 7가지 주요 분야를 중심으로 환자 번호를 키로 하여 간단한 조건을 조합해 데이터를 추출할 수 있도록 하였다(그림 3). 이를 통해 연구자들은 복잡한 조건 설정 없이 자주 사용하는 데이터에 빠르게 접근할 수 있으

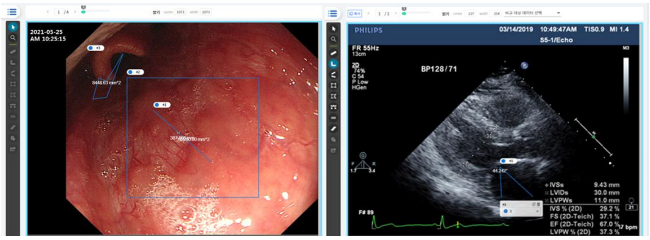
며, 손쉽게 필요한 정보를 얻을 수 있다.

(그림 4) 고급 추출.

고급 추출은 간편 추출보다 훨씬 복잡한 조건을 조합하여 데이터 검색이 가능하도록 구성되었다(그림 4). 연구자가 검색 조건을 세밀하게 설정할 수 있도록 지원하여, 복잡한 데이터 분석 요구에도 유연하게 대응할 수 있으며, 여러 조건을 조합한 검색 결과를 개인 PC로 손쉽게 내보낼 수 있다.

또한, 고급 추출에서는 검색된 결과를 기본으로 관련 환자의 의료영상 이미지를 기존 PACS 시스템이 아닌 플랫폼 환경에서 조회할 수 있으며, 의료영상 분석 기능을 통해 연구자들은 원하는 영상의 종류를 선택하고, 리스트에서 환자들의 영상 이미지 종류를 확인할 수 있다.

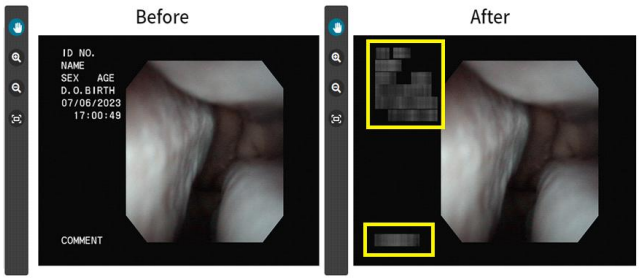
(그림 5) 의료영상 이미지 preview 화면.



(그림 6) 의료영상 이미지 저작기능.

이미지가 있는 환자를 선택하면 해당 영상 이미지를 미리보기(Preview) 할 수 있으며(그림 5), 이후 전용 화면으로 이동하여 길이, 면적, 각도 등의 측정이 가능하다. 또한, 연구자는 영상에 직접 라벨링하여 결과를 저장할 수 있어(그림 6), 데이터의 정밀한 분석이 가능하다.

플랫폼에서는 환자 개인정보 보호를 위해 환자 개인정보는 비식별 처리를 하였다. 의료영상 데이터의 경우 DICOM 헤더 파일의 비식별화와 이미지 내



(그림 7) 의료영상 이미지 비식별화.

개인정보 블러 처리를 병행하여 개인정보가 노출되지 않도록 한다. 이러한 이중 비식별화 방식을 통해 연구자는 안전하게 데이터를 활용할 수 있으며, 의료 연구와 분석 과정에서 개인정보 보호를 철저히 준수할 수 있다(그림 7).

6. 결론 및 향후 연구

통합 데이터 플랫폼의 구현은 기존의 CDW 시스템이 가지고 있던 정형 데이터 활용에만 국한된 한계를 극복하고, 정형 데이터와 비정형 데이터를 통합적으로 관리하고 분석할 수 있는 환경을 제공하게 되었다.

앞으로 연구자들의 경험을 분석하여 통합 데이터 플랫폼 이용 시 데이터 추출의 효율성이 얼마나 향상됐는지 추가 연구를 진행하여 통합 데이터 플랫폼의 실질적 효과를 입증하고자 한다.

Acknowledge

이 논문은 과학기술정보통신부 국가균형발전특별회계의 K-Health 국민의료 AI서비스 및 산업생태계 구축사업의 지원을 받아 수행된 연구 결과입니다. [과제명 : K-Health 국민의료 AI서비스 및 산업생태계 구축사업 과제고유번호 : ITAH0603230110010001000100100]

참고문헌

[1] Wang M, Li S, Zheng T, Li N, Shi Q, Zhuo X, Ding R, Huang Y Big Data Health Care Platform With Multisource Heterogeneous Data Integration and Massive High-Dimensional Data Governance for Large Hospitals: Design, Development, and Application, JMIR Med Inform 2022;10(4):e36481

[2] 가톨릭대학교 서울성모병원 CORD. CDW/EDP 소개. 가톨릭대학교 서울성모병원. 접근 2024년 9월 24일. https://cord.cmcnu.or.kr/openCdwAndEdp