

ISSN 2636-0403

발표논문집

2025
Digital
Contents
Society

(사)한국디지털콘텐츠학회
하계종합학술대회 및 대학생논문경진대회



디지털 라이프,
융합 기술과 콘텐츠

주최 | 한국디지털콘텐츠학회 Digital Contents Society 주관 | SW중심대학사업단 SKIAC MOKWON UNIVERSITY

후원 | JININFRA 위상인프라 GNSOFT 지에프티 GFT 지에프티 SJ정보통신 아이씨티웨이(주) ICTWAY 아이씨티웨이(주) CODABLE JWATS JWATS OSKO SIGONGtech (주)하이게이컨설팅 OPENLINKSYSTEM SI & ITOL AI & BIG DATA CONSULTING (주)에이저투테크 BonC Innovators 대진정보통신(주) NEOBRIX (주)그린텍아이앤씨 PINTEL 지엔미래교육 D PARTNERS (주)에이저투테크

사단
법인 한국디지털콘텐츠학회
Digital Contents Society

□ 논문발표 세션 WE2 (15:00~15:50)		○ 장소 : 카멜리아 홀
좌장	김진술 교수(전남대학교)	
WE2-01. 프라이버시 보호를 위한 다중 생체신호 기반 연합학습 시스템 설계 및 실증 연구 나필우, 김진술, 박수형(전남대학교) 601		
WE2-02. VR 쇼핑 환경의 UI/UX 디자인 요소 분석 나은경, 최인영, 강정운, 신춘성(전남대학교) 604		
WE2-03. 전통 의학에서 천연물 및 화합물의 다약리학 효과 식별 연구 유혜진, 이재인, 유선용(전남대학교) 606		
WE2-04. 딥러닝 기반 Cytochrome P450 2D6 유전자 다형성과 약물 특이적 대사 기능 표현형 예측 연구 나예빈, 정선우, 최희석, 유선용(전남대학교) 608		
WE2-05. 트랜스포머 기반 EO/IR 다중센서 융합을 통한 산불 조기 탐지 모델 최정환, 김진술, 김찬수(전남대학교) 610		
WE2-06. 대형 멀티 모달 모델별 배 질병 이미지 캡션 작성 성능 비교에 관한 연구 이진, 르황안(전남대학교), 배지훈(와세다대학교), 김민경, 김진영(전남대학교) 613		
WE2-07. LLMPatchTST: 대형언어모델 융합 스마트팜 시계열 예측 모델 르황안, 라흐만자히드우르, 유광현, 김진영(전남대학교) 615		
WE2-08. 클래스 밸런싱을 통한 YOLO 11 기반 배 결함 검출 성능 향상 당탄부, 장진호, 김순천(주에이아이시드 기업부설연구소), 박수혁, 김진영, 이현주(전남대학교) 617		
WE2-09. 벡터 데이터베이스 기반 얼굴인증 시스템에 관한 연구 이현주, 당탄부(주에이아이시드 기업부설연구소), 김진영(전남대학교), 이선관(고스트패스), 유광현(전남대학교) 619		
WE2-10. RTLS와 ESL의 위치추적 및 정보 디스플레이를 통합한 시스템 아키텍처 설계 연구 권광철, 김진술, 김명진(전남대학교) 621		

딥러닝 기반 Cytochrome P450 2D6 유전자 다형성과 약물 특이적 대사 기능 표현형 예측 연구

나예빈(*), 정선우(**), 최희석(***), 유선용(****)*

(*, **) 전남대학교 지능전자컴퓨터공학과, naybean990209@gmail.com, swjung@jnu.ac.kr

(***) 전남대학교 빅데이터융합학과, noanlove1@gmail.com

(****) 전남대학교 전자컴퓨터공학부(교신저자), syyoo@jnu.ac.kr

† 교신저자 (Corresponding Author) : 유선용

Prediction of Drug-Specific Metabolic Phenotypes Based on Cytochrome P450 2D6 Genetic Variation Using Deep Learning

Na Yeabeon(*), Jung Sunwoo(**), Choi Hui-seok(***), Yoo Sunyong(****)*

(*, **) Chonnam National University, Department of Intelligence Information Convergence

(***) Chonnam National University, Department of Bigdata Convergence

(****) Chonnam National University, Department of Compute Engineering

요약

Cytochrome P450 2D6 (CYP2D6)은 임상에서 사용되는 약물의 25%를 대사한다. 높은 다형성을 특징으로 하는 CYP2D6의 유전적 변이는 약물 대사에서 개인 간 큰 차이를 초래할 수 있으며, 이는 치료 반응의 차이와 부작용으로 이어질 수 있다. 기존의 CYP2D6 약물 대사 표현형 분류 방식은 CYP2D6에 의해 대사되는 몇 가지 약물의 임상결과를 바탕으로 CYP2D6 변이체에 점수를 매기고, 이를 통해 모든 약물의 대사 능력을 예측하는 방법이었다. 하지만 이 방법은 약물마다 다른 특성을 반영하지 못하기에 모든 약물에 일괄적으로 적용하기에는 한계가 있다. 따라서 본 연구에서는 CYP2D6 변이체와 약물에 대한 임상결과를 직접 활용하여 데이터 라벨링을 수행하고, 딥러닝을 활용한 약물 대사 표현형 예측 모델을 개발하였다.

1. 서론

체내에 약물이 들어오면 흡수, 분포, 대사 그리고 배설의 과정을 거친다. Cytochrome P450(CYP)은 간에서 발현되는 효소그룹 중 하나로, 초기 산화 반응을 촉매하여 약물에 극성기를 부여해 비활성 대사물로 변화시키거나 배설에 용이한 형태로 전환한다. 사람의 CYP는 57개의 동형 단백질로 이루어져 있고, 현재 시판되고 있는 약물의 80%가 CYP를 통해 대사된다.

유전자 내 단일 염기 다형성(SNP)이나 염기의 삽입·삭제(INDEL) 같은 유전자 다형성이 때로는 표현형의 변이를 초래한다. CYP2D6는 가장 다형성이 높은 CYP로 유전자형에서 관찰되는 다형성이 약물 대사

표현형의 개인 간 차이를 초래할 수 있다.

그렇기에 기존에는 CPIC 지침에 따라 CYP2D6의 약물 대사 표현형을 분류해왔다. 이 방법은 CYP2D6에 의해 대사되는 대표적인 약물들의 임상 결과를 바탕으로 CYP2D6 변이체별 점수를 부여하고, 개별 대립유전자 값을 합산하여 총점을 산출한 후 네 가지 범주로 구분하는 방식이다 [1]. 하지만 제한된 수의 약물에서 얻은 임상 데이터에 기반한 이 접근법은 CYP2D6으로 대사되는 모든 약물에 적용하기에는 한계가 있다. 본 연구에서는 각 약물에 특화된 대사 기능을 예측하고자 CYP2D6 변이체와 약물의 조합별 임상 데이터를 직접 수집하고 약물 대사 기능 라벨링을 수행하였다.

2. 데이터 수집 및 전처리

PharmGKB를 통해 CYP2D6 변이체와 약물 조합별 대사 기능 임상 데이터를 수집하였다. 수집한 데이터는 샘플별로 Normal, Decreased, No function 세 가지 대사 기능 범주로 분류하였고, 최종적으로 69개 약물과 117개 CYP2D6 변이체의 조합으로 이루어진 641개 샘플을 얻었다. 전체 데이터 샘플 중 168개는 테스트 데이터로 분리하였으며, 473개 샘플에 대해서는 CYP2D6 하위 변이체 정보를 이용해 데이터 증강을 수행하여 최종적으로 3,311개의 샘플로 모델 학습을 진행하였다.

3. 딥러닝 모델 설계

3.1 모달리티별 데이터 학습

CYP2D6 변이체 데이터는 DNA 시퀀스의 4개 염기(A, T, G, C)와 SNP, INDEL, Exon 여부를 포함하여 총 7개의 특성을 가진 원핫인코딩 매트릭스로 변환하였다. 이렇게 구성된 매트릭스는 CNN(Convolutional Neural Network) 모델의 입력되어 학습되었다.

약물 구조 데이터는 SMILES(Simplified Molecular Input Line Entry System) 형태로 변환하여 처리하였다. 각 약물의 SMILES 표현은 FG-BERT 모델에 입력되어 약물의 특성을 나타내는 벡터로 변환되었다 [2]. 이 모델은 대규모 화합물 데이터셋으로 사전 훈련된 모델이다. 본 연구에서는 FG-BERT를 우리 데이터에 최적화하기 위해 5개 CYP 동형 단백질(CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4)에 대한 약물 저해(inhibition) 여부 데이터를 활용하여 파인튜닝을 수행하였다.

3.2 모델 구조 및 훈련 방식

late fusion 방식을 사용하여 각기 다른 두 모달리티를 독립적으로 학습한 뒤, 동일 차원의 특성으로 변환하여 결합하고 Fully connected layer를 통해 최종 예측을 수행하는 딥러닝 모델을 설계하였다.

모델의 예측 성능에 대한 신뢰성 있는 평가를 위해 10-fold cross-validation을 통한 교차 검증 학습 과정을 수행하였다.

4. 모델 성능평가

모델 성능의 정량적 평가를 위해 분류 정확도(Accuracy), ROC 곡선 하 면적(AUROC), 정밀도-재현율 곡선 하 면적(AUPR)을 주요 평가 메트릭으로 활용하였다. 교차 검증을 통한 성능 평가 결과, Accuracy 0.7877, AUROC 0.9276, AUPR 0.8604로 측정되어 높은 분류 성능을 확인하였다.

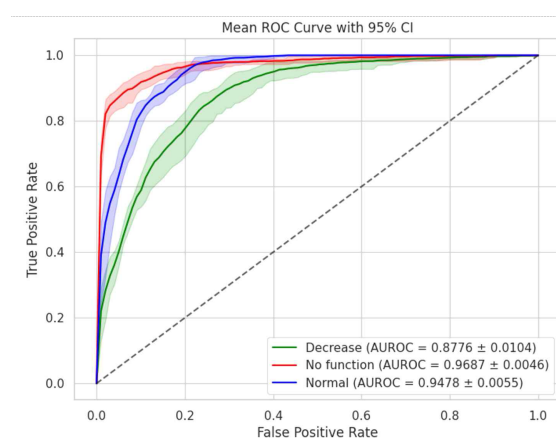


그림 1. 약물 대사 기능 예측 딥러닝 모델의 ROC 곡선

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업(IITP-2025-RS-2022-00156287) 및 식품의약품안전처 연구개발비(RS-2024-00332003, RS-2025-02215961)로 수행되었으며 이에 감사드립니다.

참고 문헌 (참고자료)

- [1] Gaedigk, A., et al., "The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype." *Clinical Pharmacology & Therapeutics*, 2008. 83(2): p. 234-242.
- [2] Li, Biaoshun, et al. "FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction." *Briefings in Bioinformatics* 24.6 (2023): bbad398..