creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

치의과학박사 학위논문

의생명 문헌 기반 특질 추출을 이용한 항암 약물 조합의 시너지 효과 예측 접근 방법 개발에 관한 연구

A study on the development of a method for predicting the synergistic effects of anti-cancer drug combinations using biomedical literature-based feature extraction

2023년 2월

서울대학교 대학원

치의과학과 의료경영과정보학전공

심 용 선

의생명 문헌 기반 특질 추출을 이용한 항암 약물 조합의 시너지 효과 예측 접근 방법 개발에 관한 연구

지도교수 김 홍 기

이 논문을 치의과학박사 학위논문으로 제출함
2023년 1월

서울대학교 대학원
치의과학과 의료경영과정보학전공
심 용 선

심용선의 치의과학박사 학위논문을 인준함
2023년 1월

위 원 장 _____ 류현모 _____ (인)

부위원장 _____ 김홍기 _____ (인)

위 원 _____ 김필종 _____ (인)

위 원 _____ 이수경 _____ (인)

위 원 _____ 안진현 _____ (인)

국문초록

단일 약물 사용에 비하여 약물의 부작용을 최소화하고, 치료효과를 증진시키는데 목표를 지닌 항암 약물 조합 연구는 중요성이 점차 증가하고 있다. 약물 조합의 수는 새로운 약이 발견되거나, 약물 농도의 변화에 따라 기하급수적으로 증가하여 시너지가 있는 약물 조합 발견에 많은 시간과 비용이 소모된다. 이러한 문제를 해결하기 위하여 대용량의 의생명 데이터를 분석하는 *in-silico* 기반의 항암 약물 조합 시너지 예측 연구가 증가하고 있다. *In-silico* 기반 항암 약물 조합의 시너지 예측 모델은 새로운 항암 약물 조합의 시너지를 예측하는데 있어 활용 될 수 있으나, 기존 연구들에서 구축한 예측 모델을 활용하여 미확인된 항암 약물 조합의 잠재적 시너지를 예측한 데이터베이스는 거의 없다. 또한, 대부분의 *in-silico* 기반 항암 약물 조합 시너지 예측 연구에서는 정형데이터를 기반으로 약물과 세포주 특질을 추출하여 예측 모델을 구축하였다. 그러나 정형데이터는 최신의 데이터를 이용하는데 제한이 있으며 데이터 확장 측면에 어려움이 있다.

본 연구에서는 의생명 문헌과 같은 비정형데이터를 기반으로 약물과 세포주 특질을 추출하여 항암 약물 조합의 시너지 예측 모델을 구축하는 방법을 개발하였다. 본 연구에서 개발한 특질 추출 방법을 이용하여 구축한 예측 모델의 성능은 기존 연구들에 비하여 우수한 성능을 보여주었다. 이러한 결과를 바탕으로 미확인된

항암 약물 조합의 잠재적 시너지를 예측하여 DPS(Database of Potential Synergy)를 구축하였으며 DPS를 검색할 수 있는 웹 어플리케이션 FPS(Finder of Potential Synergy)를 구축하였다. FPS는 http://bike-fps.snu.ac.kr에서 이용 가능하다.

본 연구의 과정과 결과물을 통해 정형데이터를 주로 사용하던 기존의 항암 약물 조합 시너지 예측 연구 방향을 비정형데이터를 이용한 연구 방향으로 확장 시킬 수 있을 것으로 기대한다. 또한, 약물과 세포주 키워드만을 이용하여 항암 약물 조합의 잠재적 시너지를 검색 할 수 있는 FPS는 도구 사용에 대한 허들이 낮고 직관적이므로 다양한 분야의 연구자들이 접근하기 용이하여 항암 약물 조합 발견에 있어 다양한 관점 및 방향을 줄 수 있을 것으로 기대한다. 이를 통해 *in-vivo* 또는 *in-vitro* 실험 전 연구자들이 새로운 항암 약물 조합의 후보를 선정하는데 있어 기여 할 수 있을 것으로 기대한다.

주요어 : 의생명 문헌 기반 특질 추출, 항암 약물 조합 시너지 예측, 항암 약물 조합 데이터베이스 확장
학 번 : 2015-22084

목 차

국문초록 ····· i
목 차 ····· iii
표 목 차 ····· v
그 림 목 차 ····· vi
Ⅰ. 서 론 ····· 1
1. 연구 배경 ····· 1
2. 관련 연구 ····· 9
3. 연구 목적 ····· 17
4. 논문 구성 ····· 19
Ⅱ. 연구 방법 ····· 20
1. 개요 ····· 20
2. 의생명 문헌 전처리 ····· 22
3. 항암 약물 조합 시너지 예측 모델 구축 ····· 24
Ⅲ. 연구 결과 ····· 31
1. 의생명 문헌 전처리 결과 ····· 31
2. 항암 약물 조합 시너지 예측 모델 구축 결과 ····· 32
Ⅳ. 연구 활용 ····· 39
1. 개요 ····· 39
2. 데이터리소스 ····· 41
3. 항암 약물 조합 시너지 예측 모델 구축 ····· 45
4. 항암 약물 조합 시너지 예측 모델 성능 비교 ····· 46
5. Database of Potential Synergy (DPS) ····· 53
6. Finder of Potential Synergy (FPS) ····· 54
Ⅴ. 결 론 ····· 59

Ⅵ. 연구의 제한점 및 제언 ····· 62
참고문헌 ····· 64
Abstract ····· 70

표 목 차

표 1. *In-silico* 기반 항암 약물 조합 시너지 예측 연구에서 사용한 데이터 종류 및 타입 ····· 15
표 2. Word2Vec 하이퍼파라메터 셋팅 값 ····· 25
표 3. NCI-ALMANAC 데이터베이스 예시 ····· 26
표 4. ROC-AUC 기반 예측 모델 성능 비교 ····· 34
표 5. AUPR 기반 예측 모델 성능 비교 ····· 35
표 6. F1-score 기반 예측 모델 성능 비교 ····· 36
표 7. DrugComb 데이터베이스 예시 ····· 43
표 8. NCI-ALMANAC 데이터베이스와 DrugComb 데이터베이스의 데이터 수 비교 ····· 47
표 9. ROC-AUC 기반 예측 모델 성능 비교 ····· 48
표 10. AUPR 기반 예측 모델 성능 비교 ····· 49
표 11. F1-score 기반 예측 모델 성능 비교 ····· 50

그 림 목 차

그림 1. 약물 조합 관련 연구의 연도별 출판물 수 ····· 2
그림 2. *In-silico* 기반 약물 조합 관련 연구의 연도별 출판물 수 중요성 증가 ····· 5
그림 3. 정형/비정형데이터의 증가 비교 ····· 7
그림 4. *In-silico* 기반 항암 약물 조합 시너지 예측 연구의 공통 프로세스 ····· 10
그림 5. FFNN 프로세스 ····· 12
그림 6. AE 프로세스 ····· 13
그림 7. 의생명 문헌 기반 특질 추출을 이용한 항암 약물 조합 시너지 효과 예측 방법의 워크플로우 ····· 21
그림 8. FFNN 기반 예측 모델의 혼동행렬 ····· 38
그림 9. FFNN 기반 예측 모델의 ROC 커브 ····· 38
그림 10. DPS 구축의 워크플로우 ····· 40
그림 11. DrugComb+FFNN 기반 예측 모델의 혼동행렬 ·· 52
그림 12. FPS의 인터페이스 ····· 56
그림 13. BEE 검색 결과 예시 ····· 58

Ⅰ. 서 론

1. 연구 배경

암은 사망률이 높은 질병으로, 전 세계 과학자들은 암을 극복하기 위하여 새로운 항암 약물을 발견하고자 많은 노력을 하고 있다 [1]. 새로운 항암 약물 개발 방법에는 크게 두 가지 접근법이 있다. 첫 번째는 이전에 없던 새로운 화학 물질을 개발하거나, 발견하는 약물 발견(drug discovery) 측면이다 [2]. 두 번째는 다른 질병에서 사용하던 약물이 항암효과가 있는 것을 발견하여 약물의 용도를 확장하는 약물 재배치(drug repositioning) 측면이다 [3]. 약물 재배치는 미국 FDA에서 승인된 약을 대상으로 용도를 확장하는 방법이기 때문에 약물 발견 측면에 비하여 약물의 부작용이나 위험성을 예측할 수 있는 장점이 있다.

약물 재배치 방법 가운데 두 개 이상의 약물을 조합하여 질병을 치료하는 약물 조합 연구의 중요성은 점차 증가하고 있다 [4]. 그림 1은 의생명 문헌 저장소인 PubMed에서 drug repositioning, drug repurposing, drug combination 키워드로 검색하여 추출된 출판물의 개수이다. 시간이 지남에 따라 약물 조합 관련 연구가 점차 증가하고 있는 것을 확인 할 수 있다.

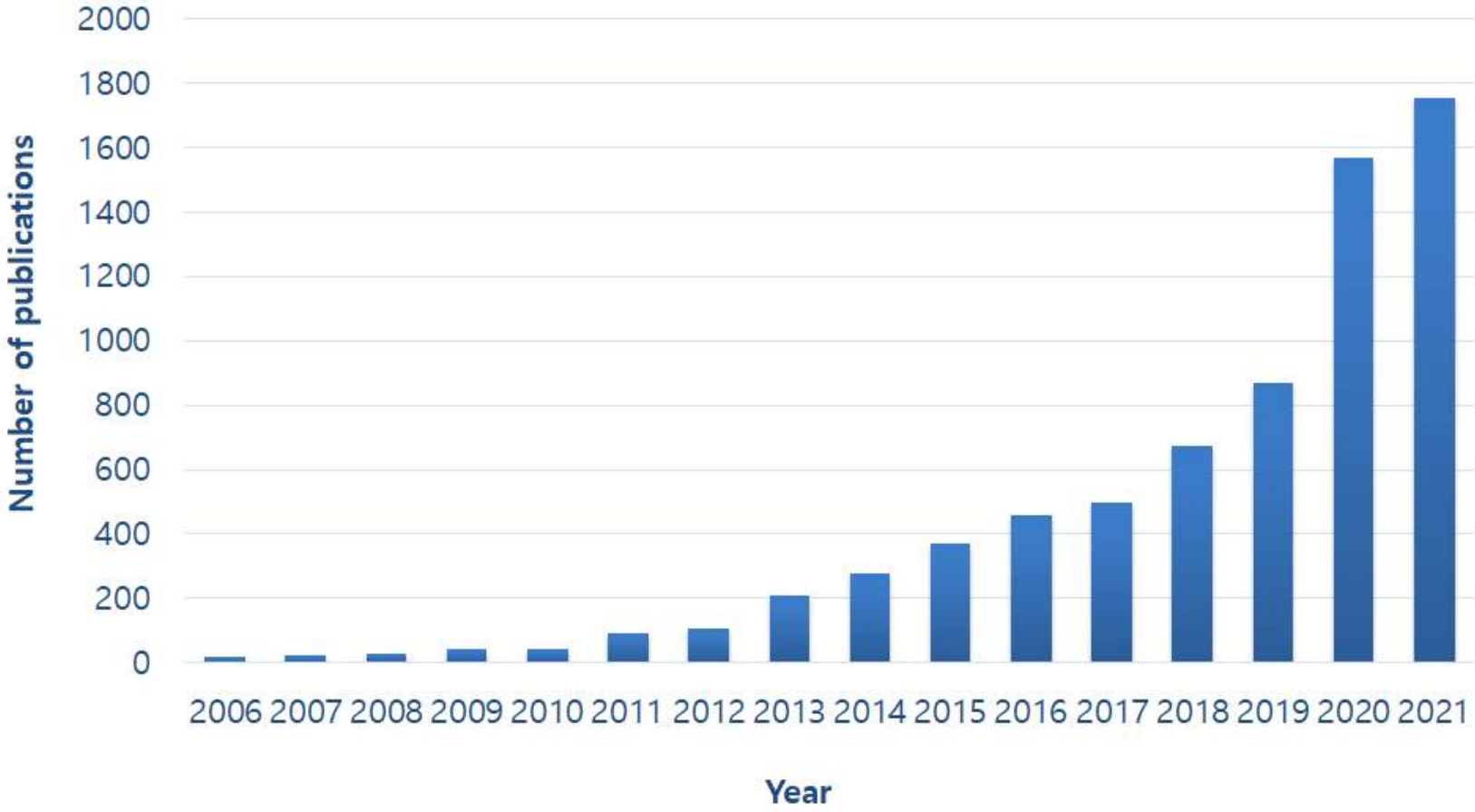

그림 1. 약물 조합 관련 연구의 연도별 출판물 수

약물 조합은 단일 약물 사용에 비하여 투여하는 약물의 용량을 줄여 약물로 인한 부작용을 최소화하는 동시에 치료효과를 증가시키는데 목적을 두고 있다 [5]. 약물 조합 연구에서 시너지(synergy)란 두 개 이상의 약물을 조합하여 치료효과가 발현되는 것을 의미 한다 [6]. 예를 들어, 약물 AZD-3355와 Crizotinib 조합은 *in-vivo* 및 *in-vitro* 실험에서 항암 효과가 있는 것이 증명되었다 [7]. 또한, 약물 Metformin과 Temozolomide 조합을 교모세포종(glioblastoma)에 적용한 결과, *in-vivo* 및 *in-vitro* 실험에서 효능이 입증되어 PhaseⅠ, PhaseⅡ 실험까지 진행되었다 [8].

최적의 약물 조합을 발견하기 위해서는 약물 간의 부작용과 약물 용량과 같은 다양한 조건을 고려하여야 한다. 약물 조합의 수는 새로운 약물이 발견이 되거나, 약물의 개수가 증가하거나, 약물 농도가 변경될 때마다 기하급수적으로 증가한다. 약물 조합은 일반적으로 *in-vivo* 및 *in-vitro* 실험에서 발견된다. 최근 고처리량 스크리닝(high-throughput screening)을 통해 약물 조합 발견 실험이 진행되고 있으나, 잠재적인 약물 조합을 발견하고 최적화하는데 시간과 비용이 많이 소요된다. 이러한 문제를 해결하기 위하여, *in-silico* 방법을 기반에 둔 연구가 증가하고 있다 [9]. *In-silico* 방법은 컴퓨터를 이용하여 대용량의 의생명 데이터 분석을 통해 이전에 발견되지 못한 패턴이나, 숨겨진 의미를 발견하는 분석법으로, 중요성이 점차 증가하고 있다. 이러한 *in-silico* 방법을 통해 연구자들은 의생명 데이터를 분석하여 시너지 효과가 있는 약물 조합을 발견하고자 노력하고 있다. 그림 2는 PubMed 데이터베이스에서 drug repositioning, drug repurposing, drug combination, *in-silico* 키워드로 검색하여 추출된 출판물의 개수이다. 시간이 지남에 따라 *in-silico* 기반의 약물 조합 연구가 점차 증가하고 있는 것을 확인 할 수 있다.

In-silico 기반 항암 약물 조합의 시너지 예측 모델은 새로운 항암 약물 조합의 시너지를 예측하는데 있어 활용 될 수 있다. 이를 통해 항암 약물 조합 연구 실험을 수행하는 연구자들이 실험 후보군을 결정하는데 있어 참고할 수 있다. 그러므로 항암 약물 조합 관련 연구자들이 참고하여 사용할 수 있도록 밝혀지지 않은 항암 약물 조합의 잠재적 시너지 예측 결과를 참조할 수 있는 데이터베이스의 존재는 중요하다. 그러나 이러한 기존 연구들에서 구축한 예측 모델을 활용하여 미확인된 항암 약물 조합의 잠재적 시너지를 예측한 데이터베이스는 거의 없다.

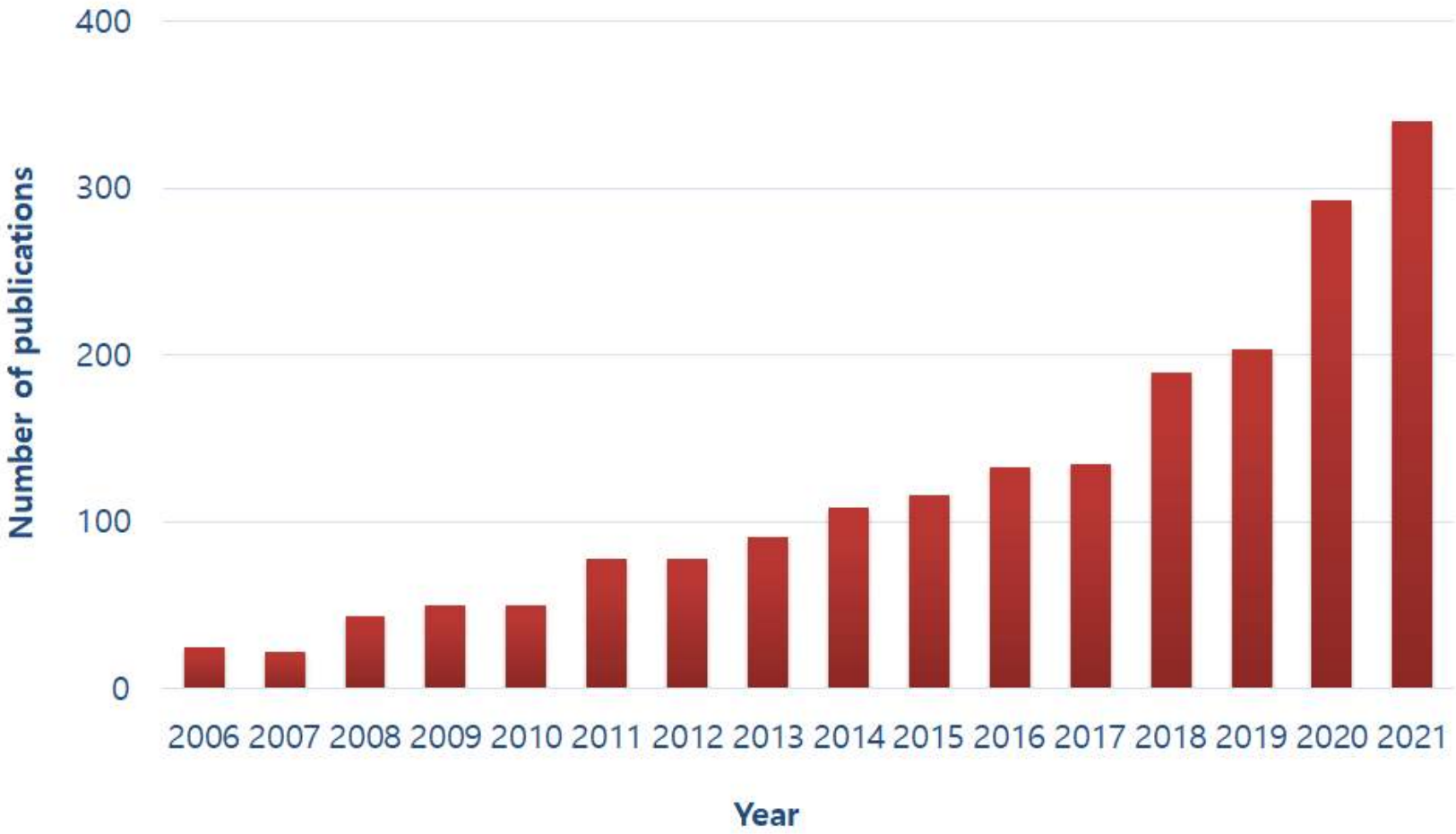

그림 2. *In-silico* 기반 약물 조합 관련 연구의 연도별 출판물 수

대부분의 연구들은 정형데이터를 기반으로 약물과 세포주의 특질을 추출하였다. 정형데이터는 정해진 필드를 기반으로 생성된 데이터로, 대부분 관계형 데이터베이스나 엑셀과 같은 스프레드시트 형태로 존재한다. 정형데이터는 대부분 수작업으로 구축되기 때문에, 새롭게 출간된 연구의 결과가 공개된 데이터베이스에 구조화되기까지 오랜 시간이 소요된다. 약물 관련 데이터베이스인 DrugBank[1)]는 2022년 1월 이후로 업데이트가 되지 않고 있으며, 직전의 업데이트는 2021년 1월로 약 1년의 차이가 있다. 그리고 약물과 유전자 간의 상호작용 관련 데이터베이스인 DGIdb[2)]는 2022년 2월 이후로 업데이트가 되지 않고 있으며, 직전의 업데이트는 2021년 5월로 약 9개월의 차이가 있다. 이처럼 정형데이터 기반의 데이터베이스는 데이터를 최신의 상태로 유지하는데 많은 시간이 소요되어 데이터 확장의 측면에 있어 어려움이 있다.

비정형데이터는 정형데이터와 달리 고정된 필드가 없어, 자유롭게 데이터를 생성할 수 있다. 비정형데이터에는 텍스트, 이미지, 동영상 등의 종류가 있다. 비정형데이터는 구조 변경 및 데이터 생성에 있어 정형데이터에 비하여 자유롭다. 이에 따라 비정형데이터가 정형데이터에 비하여 데이터의 양이 급격하게 증가하고 있다 [10]. 그림 3은 정형데이터와 비정형데이터의 증가폭을 보여준다. 이를 통해, 비정형데이터의 증가 폭이 점차 커지고 있으며, 비정형데이터의 증가 속도가 정형데이터의 증가 속도에 비하여 압도적인 것을 볼 수 있다.

1) https://go.drugbank.com/releases/latest
2) https://www.dgidb.org/downloads

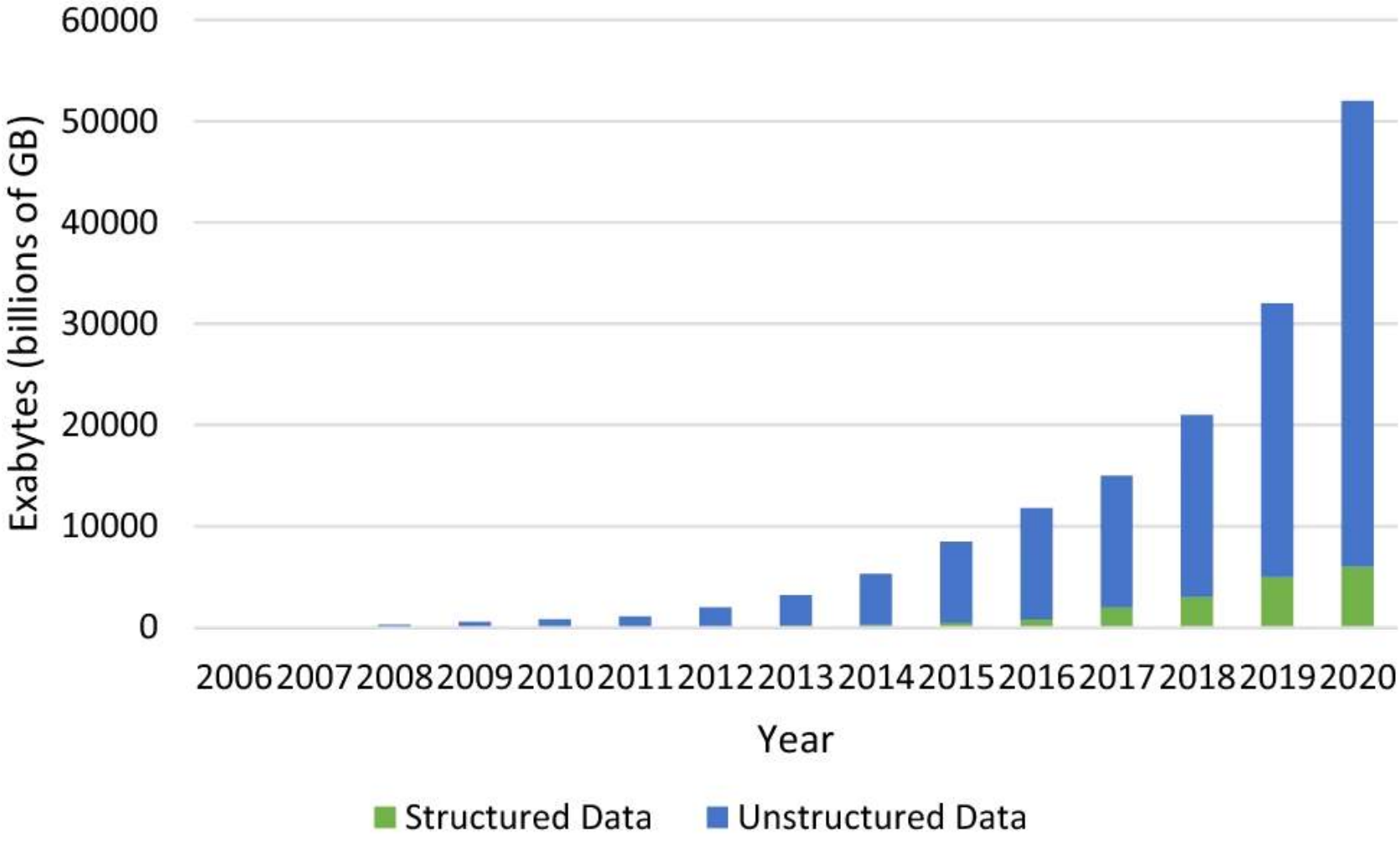

그림 3. 정형/비정형데이터의 증가 비교

의생명 문헌에는 실험 방법, 결과 등 실험에 대한 정보뿐만 아니라 의생명 개체 간의 상호작용 등 다양한 정보가 포함되어 있다. 이와 같은 정보들을 활용하여 각 약물과 세포주가 지닌 고유의 특성을 추출할 수 있고, 약물과 세포주 간의 새로운 관련성을 발견할 수 있다. 그럼에도 불구하고, 의생명 문헌으로부터 특질을 추출하여 항암 약물 조합의 시너지를 예측하는 연구는 거의 없다. 그러므로 의생명 문헌으로부터 약물과 세포주의 특질을 추출하여 항암 약물 조합의 시너지를 예측하는 것은 공개된 의생명 데이터베이스의 의존도를 줄일 수 있고, 대용량의 비정형데이터를 이용할 수 있는 장점이 있다.

2. 관련 연구

In-silico 기반의 항암 약물 조합의 시너지를 예측하는 연구는 공통적인 프로세스가 존재 한다 [11]. 그림 4는 *in-silico* 기반 항암 약물 조합 시너지 예측 연구의 공통 프로세스를 보여준다.

첫 번째 단계인 'Data collection'에서는 약물 조합 데이터베이스 및 의생명 데이터베이스에서 필요한 데이터를 추출한다. 일반적으로 약물 조합 데이터베이스에는 '약물1+약물2+세포주'로 구성된 조합이 있으며, 약물의 농도에 따른 약물 농도-반응성 매트릭스 형태로 구성되어 있으며, 각 조합의 반응성 수치를 계산 모델에 기반 하여 추출된 조합 스코어가 있다. 또한, 약물과 세포주의 특질 추출을 위하여 의생명 관련 데이터베이스에서 필요한 데이터를 추출한다. 의생명 관련 데이터베이스는 연구의 방향성과 연구자의 경험에 따라 달라질 수 있다.

두 번째 단계인 'Data representation'에서는 약물과 세포주 데이터를 기반으로 벡터화(vectorization)하여 각 약물과 세포주의 특질을 추출한다. 이후, 약물 조합 데이터베이스를 기반으로 약물과 세포주 특질을 조합하여 조합 벡터를 구축하고, 조합 스코어를 기반으로 시너지와 비시너지 조합을 분리하여 학습데이터를 구축한다.

세 번째 단계인 'Model fittng' 단계에서는 학습데이터를 학습 시킬 수 있는 다양한 알고리즘 기반의 예측 모델을 구축하고, 예측 모델의 성능을 평가한다. 예측 모델은 머신러닝 및 딥러닝 기반의 알고리즘을 사용하고, 각 알고리즘은 데이터의 양, 데이터의 특성, 연구자의 경험에 따라 결정되며, 하이퍼파라메터 설정 값에 따라 성능 차이가 발생한다.

대부분의 연구들은 이러한 공통 프로세스를 거치며, 각 연구에서 사용하는 데이터나 알고리즘에 따라 예측 모델의 성능 차이를 보여준다.

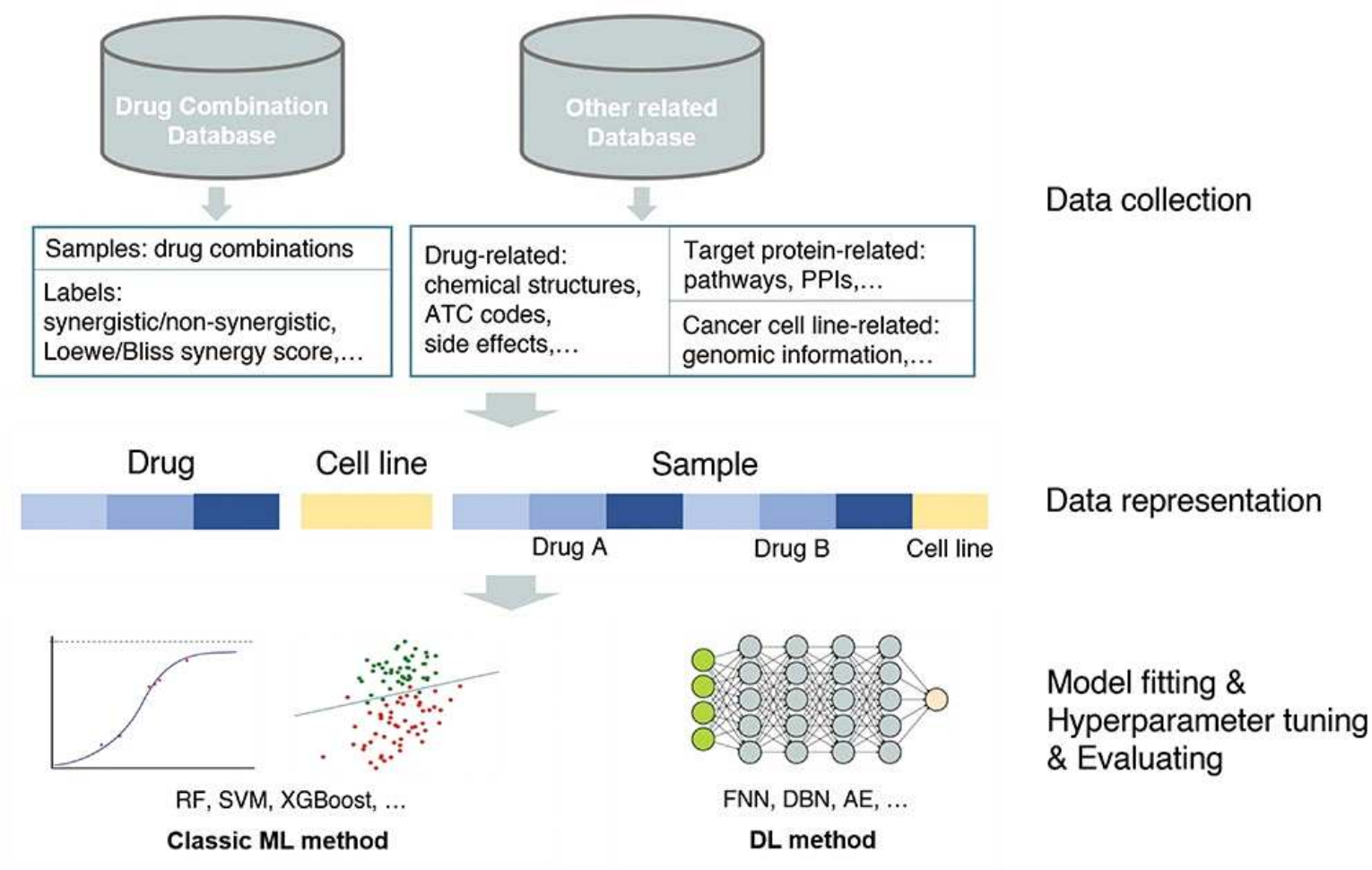

그림 4. *In-silico* 기반 항암 약물 조합 시너지 예측 연구의 공통 프로세스

대부분의 연구들은 머신러닝 및 딥러닝 기반의 알고리즘을 사용하여 예측모델을 구축한다. 다음에서 머신러닝 및 딥러닝 기반의 알고리즘과 이를 활용하여 항암 약물 조합의 시너지를 예측한 연구들을 소개한다.

Feed Forward Neural Network(FFNN)는 여러 층의 퍼셉트론을 순차적으로 붙여 넣고 입력 레이어(input layer), 은닉 레이어(hidden layer), 출력 레이어(output layer)로 구성되어 정방향으로 연결한 네트워크 모델로 딥러닝에서 가장 기본이 되는 모델이다 [12]. 입력 레이어는 데이터를 입력받는 단계로, 다음 레이어로 전달하는 역할을 하고, 레이어의 크기는 입력 벡터의 길이로 정해진다. 은닉 레이어는 입력 레이어와 연결된 전결합 단계로 은닉 레이어의 함수에 따라 모델의 성능이 결정되기 때문에 FFNN 모델에서 가장 핵심적인 부분이라 할 수 있다. 출력 레이어는 은닉 레이어의 값을 받아 출력해 주는 단계로 출력 값은 출력 레이어의 활성 함수에 의하여 결정된다. 그림 5는 FFNN의 대표적인 프로세스를 보여 준다[3]. Preuer 연구진은 화학적(chemical) 및 유전체(genomic) 정보를 정규화하여 입력 데이터의 이질성(heterogeneity)을 식별하고, 원추형 레이어(conical layers)를 사용한 딥러닝 기반의 약물 시너지 모델인 'DeepSynergy'를 발표하였다 [13]. Xia 연구진은 분자적 특질(gene expression, microRNA, proteome)을 인코딩하고, 약물 조합에 대한 세포주의 반응을 예측하기 위하여 딥러닝 기반의 계산 모델을 발표하였다 [14].

---

3) https://mukulrathi.com/demystifying-deep-learning/feed-forward-neural-network/

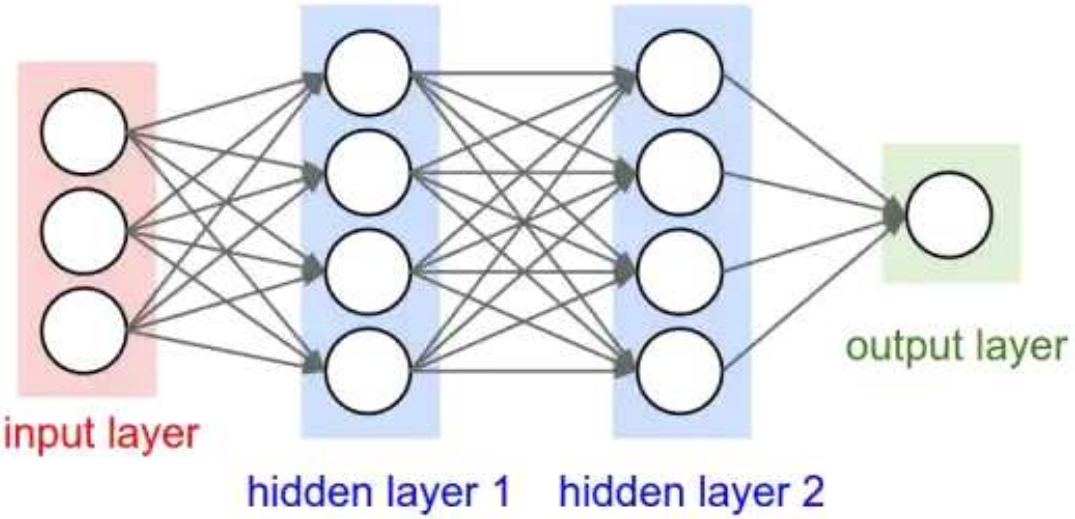

그림 5. Feed Forward Neural Network 프로세스 요약도

Autoencoder(AE)는 입력 데이터를 압축 및 축소한 후 다시 확장하여 출력 데이터를 입력 데이터와 최대한 동일하게 만드는 심층 신경망 모델이다 [15]. 입력 데이터를 차원 축소 한 후 다시 확장함으로써 AE는 입·출력 데이터를 그대로 복사하지 않고, 불필요한 특질은 제거 및 변경하면서 데이터를 효율적으로 표현하는 방법을 학습하도록 제어한다. AE는 크게 인코더(encoder)와 디코더(decoder)로 구성되어 있다. 인코더는 인지 네트워크(recognition network)라고도 하며, 입력 데이터를 차원 축소하여 함축된 데이터를 구축한다. 디코더는 생성 네트워크(generative network)라고도 하며, 함축된 데이터를 확장하여 입력 데이터와 유사한 데이터를 구축한다. 이를 통해, 입력 데이터에 있는 특질 가운데 데이터를 표현하는데 있어 중요한 특질을 표현할 수 있다. 그림 6은 AE의 대표적인 프로세스를 보여 준다[4]. Kim 연구진은 약물의 분자적(molecular) 및 유전체(genomic) 특질과 세포주의 type 및 유전체(genomic) 특질을

4) https://gaussian37.github.io/dl-concept-autoencoder2/

사용하는 멀티태스킹 심층 신경망 기반 약물 조합의 시너지 예측 모델을 발표하였다 [16]. Zhang 연구진은 AE를 사용하여 인코딩된 multi-omics 데이터(gene expression, mutation, copy number 등)를 이용하여 딥러닝 기반 약물 조합 예측 모델인 'AuDNNsynergy'를 발표하였다 [17].

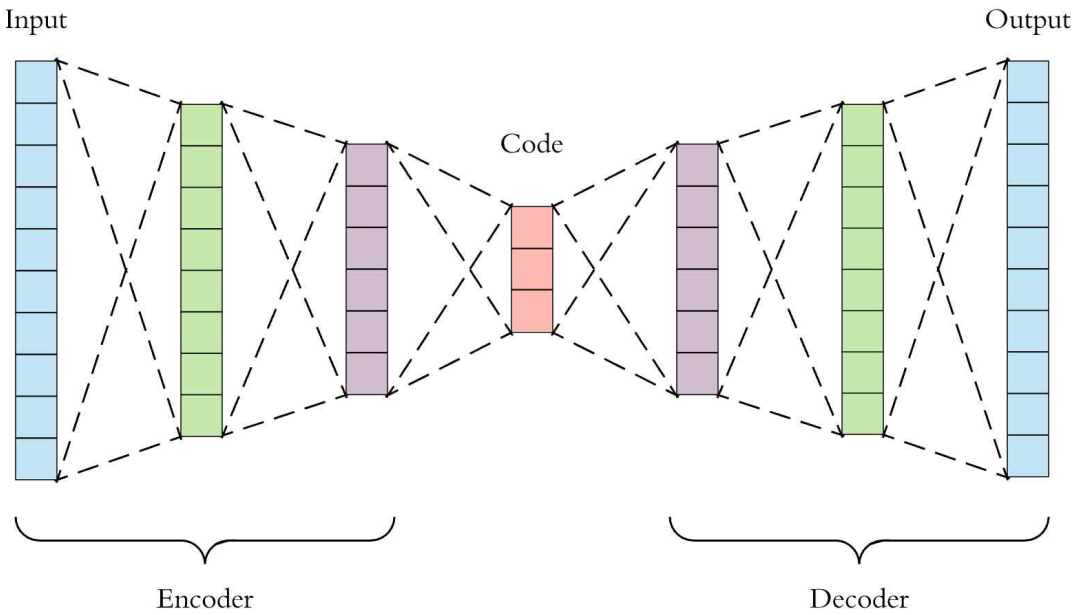

그림 6. Autoencoder 프로세스 요약도

XGBoost(XGB)는 여러 의사 결정 트리(decision tree)를 결합하여 트리 부스팅 알고리즘에서 현 시점 최고 성능을 보여주는 앙상블 알고리즘이다 [18]. Janizek 연구진은 약물의 화학적(chemical) 및 물리적(physical) 특질과 세포주의 유전자 발현 레벨(gene expression level)을 사용하여 XGBoost 기반의 약물 조합 예측 모델인 'TreeCombo'를 발표하였다 [19]. Celebi 연구진은 XGBoost를 사용하여 생물학적으로 관련된 약물 조합 및 세포주 특질을 기반으로 시너지를 예측하는 모델을 발표하였다 [20].

Extremely Randomized Trees(ERT)는 각 포레스트 트리의 후보 속성(candidate attribution)을 무작위로 나누어 랜덤성을 증가시킨 랜덤 포레스트 모델의 변형된 형태이다 [21]. Jeon 연구진은 유전체(genomic) 정보, 약물의 타겟 정보, 약리학적(pharmacological) 정보를 사용하여 ERT를 기반으로 개인 맞춤 약물 조합 예측 모델을 발표하였다 [22].

Logistic Regression(LR)은 종속 변수와 독립 변수 간 관계 기반의 함수를 구축하여 클래스를 분류하는 모델이다 [23]. Li 연구진은 약물의 타겟 정보를 기반으로 구축한 특질과 gene expression, essentiality 기반의 세포주 특질을 사용하여 LR 기반 약물 조합 예측 모델을 발표하였다 [24].

표 1은 관련 연구들에서 사용한 약물 조합 데이터베이스와 약물과 세포주 특질을 구축하기 위하여 사용된 데이터 타입 및 예측 모델을 구축하기 위한 알고리즘을 보여준다.

표 1. *In-silico* 기반 항암 약물 조합 시너지 예측 연구에서 사용한 데이터 종류 및 타입

Ref.	약물 조합 데이터베이스	약물 데이터	세포주 데이터	알고리즘
[13]	O'Neil	• Molecular fingerprints • Physicochemical properties • Binary toxicophore features	• Gene expression	Feed-Forward Neural Network
[14]	NCI-ALMANAC	• Molecular fingerprints • Molecular descriptors	• Gene expression • MicroRNA expression • Protein abundance	Feed-Forward Neural Network
[16]	DrugComb	• Molecular fingerprints • SMILES representation • Drug-target information	• Gene expression • Tissue/cancer type	Autoencoder
[17]	O'Neil	• Molecular fingerprints • Physicochemical properties	• Gene expression • Somatic mutation • CNV	Autoencoder

[19]	O'Neil	• Molecular fingerprints • Binary toxicophore features	• Gene expression	XGBoost
[20]	DREAM	• Molecular fingerprints • Drug-target information • Target protein domains • Targeted pathways • Monotherapy information	• Gene expression • Gene mutation • CNV	XGBoost
[22]	O'Neil	• Drug-target information • Monotherapy information	• Gene expression • Gene mutation • CNV • Synthetic lethality	Extremely Randomized Trees
[24]	NCI-ALMANAC	• Drug-target information	• Gene expression • Gene essentiality • Pathway information	Linear Regression

3. 연구 목적

본 연구에서는 의생명 문헌에 있는 정보를 기반으로 약물과 세포주의 특질을 추출하였다. 본 연구의 특질 추출 방법을 통해 문헌에 있는 정보를 모두 활용함으로써 각 약물과 세포주가 지닌 고유의 특성을 추출할 수 있고, 숨겨진 연관성을 발견할 수 있다. 본 연구에서 개발한 특질 추출 방법은 약물이나 세포주의 정보가 추가되거나 변경됨에 따라 관련 문헌을 추가하여 약물 및 세포주 특질에 반영 할 수 있다. 그러므로 정형데이터 기반의 특질 추출 방법에 비하여 데이터 확장성에 유연하다. 또한, 특질 추출에 있어 정형데이터의 의존도를 감소시킬 수 있다.

본 연구에서는 의생명 문헌을 기반으로 추출된 특질을 이용하여 항암 약물 조합의 시너지 효과를 예측할 수 있는 예측 모델을 구축하였다. 예측 모델은 이전 연구들에서 사용된 다양한 딥러닝, 머신러닝 알고리즘을 이용하여 구축하고, 이전 연구들과 결과를 비교하였다. 본 연구의 소스코드 및 약물과 세포주의 특질은 GitHub에 공개하였고, https://github.com/YongsunShim/anticancer-drug-combination-prediction-using-documents-based-features.git에서 이용 가능하다.

본 연구에서 개발한 방법을 기반으로 공개된 항암 약물 조합 데이터베이스에 포함되어 있지 않은 조합의 시너지를 예측한 잠재적 항암 약물 조합 데이터베이스인 DPS(Database of Potential Synergy)를 구축하였다. 또한, 새로운 항암 약물 조합의 잠재적 시너지를 검색할 수 있는 유저 친화적인 웹 어플리케이션인 FPS(Finder of Potential Synergy)를 구축하였다. FPS는 약물과 세포주 키워드만을 이용하여 항암 약물 조합의 잠재적 시너지를 검색 할 수 있어, 도구 사용에 대한 진입장벽을 낮추었다. 이를 통해 항암 약물 조합 연구와 관련성이 떨어진 연구자들 또한

쉽게 이용 가능하여 다양한 분야의 연구자들이 접근하기 용이하다. FPS는 http://bike-fps.snu.ac.kr/에서 이용 가능하다.

4. 논문 구성

본 논문은 다음과 같이 구성되어 있다. ‘Ⅱ. 연구방법’ 장에서는 본 연구의 의생명 문헌 기반 특질 추출을 이용한 항암 약물 조합의 시너지 효과를 예측하는 방법이 기술되어 있다. 연구 방법은 크게 두 부분으로 나뉘어져 있다. 첫 번째는 약물과 세포주 관련 의생명 문헌을 추출하고 처리하는 의생명 문헌 전처리 단계이고, 두 번째는 전처리된 문헌 기반 추출된 약물과 세포주의 특질을 이용하여 항암 약물 조합의 시너지 예측 모델을 구축하는 단계이다. ‘Ⅲ. 연구 결과’ 장에서는 본 연구에서 구축한 예측 모델의 성능과 기존 연구들에서 구축한 예측 모델의 성능을 비교하였다. ‘Ⅳ. 연구 활용’ 장에서는 본 연구에서 개발한 방법을 기반으로 미확인된 항암 약물 조합의 시너지를 예측한 데이터베이스 DPS를 소개하고, DPS의 내용을 검색 할 수 있는 유저 친화적인 웹 어플리케이션 FPS를 소개한다. ‘Ⅴ. 결론’ 장에서는 본 연구의 요약과 기대효과가 기술되어 있고, ‘Ⅵ. 연구의 제한점 및 제언’ 장에서는 본 연구의 제한점과 제언이 기술되어 있다.

Ⅱ. 연구 방법

1. 개요

본 연구에서 개발한 의생명 문헌 기반 추출된 특질을 이용한 항암 약물 조합의 시너지 예측 방법은 두 가지 단계로 이루어져 있다.

첫 번째는 의생명 문헌 전처리 단계이다. 공개된 항암 약물 조합 데이터베이스에는 약물과 세포주의 단어가 있다. 각 단어에 대한 동의어를 약물 및 세포주 관련 공개 데이터베이스에서 추출하여 말뭉치를 구축하였다. 말뭉치에 있는 단어를 기반으로 해당 단어가 포함되어 있는 의생명 문헌을 추출하였다. 또한, 말뭉치를 이용하여 약물과 세포주의 동의어를 대표단어로 변환하고, 토큰화(tokenization), 표제어 추출(lemmatization), 불용어(stopwords)를 제거하여 추출된 의생명 문헌의 전처리를 수행하였다.

두 번째는 예측 모델 구축 단계로, 전처리된 문헌으로부터 약물과 세포주 특질을 추출하고, 항암 약물 조합 데이터베이스를 기반으로 특질을 조합하여 학습데이터를 구축하였다. 이후, 학습데이터를 이용하여 5-fold 교차 검증으로 딥러닝 및 머신러닝 알고리즘 기반의 예측 모델을 구축하였다.

그림 7은 본 연구에서 개발한 방법에 대한 워크플로우를 보여준다. 각 단계에 대한 자세한 내용은 다음에 기술되어 있다.

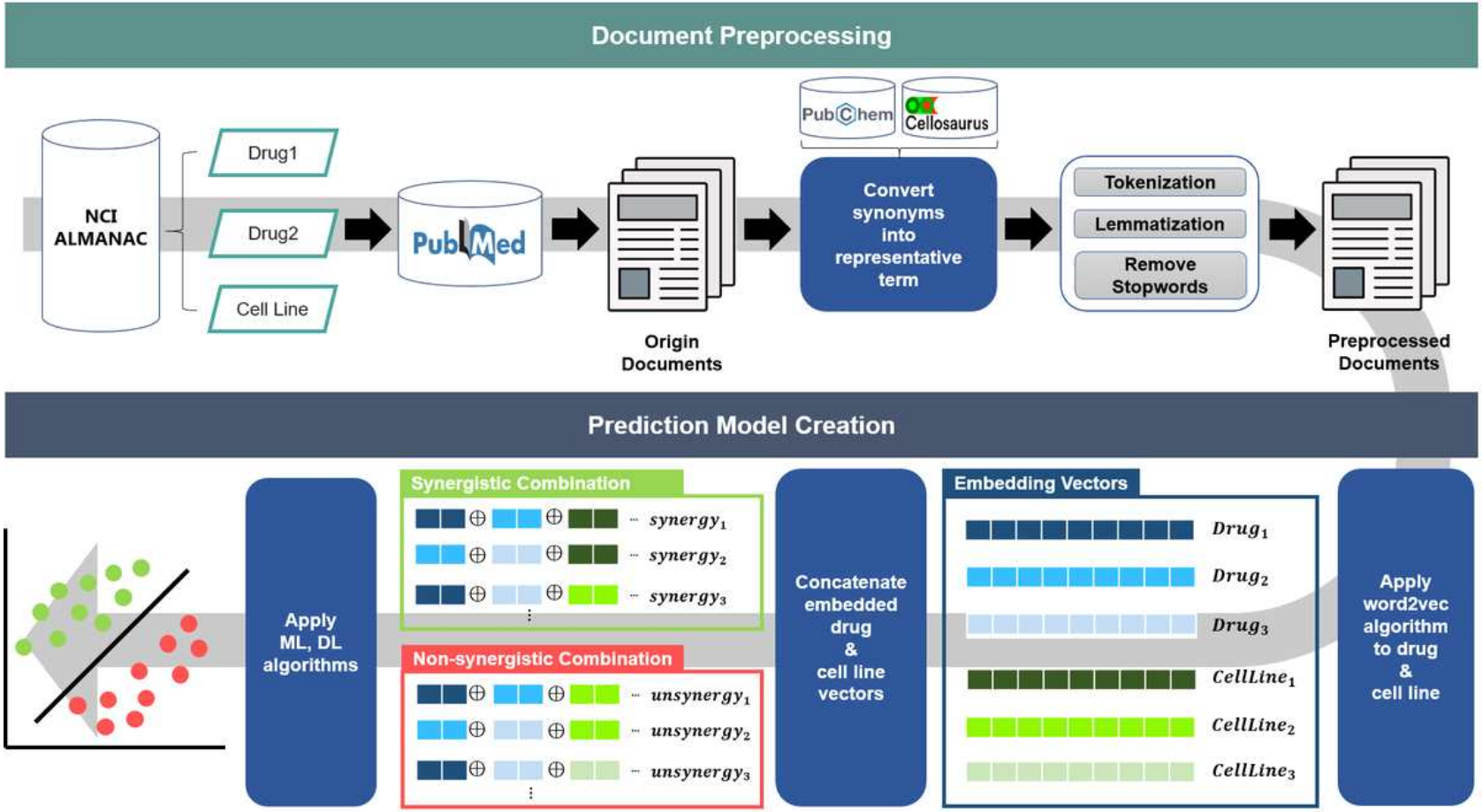

그림 7. 의생명 문헌 기반 특질 추출을 이용한 항암 약물 조합 시너지 효과 예측 방법의 워크플로우

2. 의생명 문헌 전처리

NCI-ALMANAC(A large matrix of anti-neoplastic agent combinations)은 대부분의 항암 약물 조합 연구에서 기준으로 사용되는 데이터베이스로, 미국 국립 암 센터(National Cancer Institute, NCI)에서 발표하였다 [25]. NCI-ALMANAC 데이터베이스는 FDA에서 승인된 약 5,000개 이상의 약물 조합을 60개의 인간 종양 세포주에 투여하여 각 약물 조합의 약물 농도-반응성 매트릭스로 구성되어 있다. 또한, 각 약물 조합의 약물 농도-반응성 결과를 기반으로 조합의 결과를 계산한 ComboScore가 있다. NCI-ALMANAC 데이터베이스에는 총 304,549개의 항암 약물 조합과 각 조합의 ComboScore가 포함되어 있다.

PubMed 데이터베이스는 의생명 관련 문헌을 통합한 전 세계에서 가장 큰 저장소이다. 본 연구에서는 PubChem[26]을 이용하여 NCI-ALMANAC 데이터베이스에서 추출한 약물에 대한 약물 말뭉치를 구축하였고, Cellosaurus[27]를 이용하여 세포주에 대한 세포주 말뭉치를 구축하였다. 이러한 말뭉치에 있는 키워드를 기반으로 해당 키워드가 포함되어 있는 의생명 문헌을 PubMed에서 추출하였다. 의생명 문헌에는 정해진 포맷이 없기 때문에, 동일한 약물과 세포주가 문헌이나 문맥에 따라 다르게 표기되어 있다. 예를 들어, 약물 'Fluorouracil'은 PubMed ID '10026447'에서 'Fluorouracil'로 표기되어 있으나, '10026446'에서는 '5-FU'로 표기되어 있다. 의생명 문헌의 전처리 과정 없이 특질 추출을 한다면 특질 추출 시 동일한 약물이나 세포주의 정보가 분산된다. 이는 약물과 세포주 주변의 다양한 정보를 활용하기 어렵게 만들어, 약물과 세포주가 지닌 고유한 특질을 추출하는데 방해가 된다. 이러한 문제를 해결하기 위하여, 문헌에 있는 약물 및 세포주의 동의어를 대표단어로

변환하는 것이 필요하다. 본 논문에서는 말뭉치를 이용하여 동의어를 대표단어로 변환함으로써 동의어에 분산되는 정보를 대표단어로 집중시켜 특질을 추출하였다. 이후, 추출된 문헌에서 토큰화를 통해 단어를 추출하였다. 추출된 단어는 표제어 추출을 통해 각 단어의 기본형으로 변경하여 분산될 수 있는 정보를 단어의 기본형에 집중시켰다. 또한, 불용어를 제거하여 전처리된 의생명 문헌 데이터 셋을 구축하였다.

3. 항암 약물 조합 시너지 예측 모델 구축

다양한 단어 임베딩 알고리즘들의 가장 큰 목적은 텍스트 기반의 데이터를 숫자로 변형하여 머신러닝과 같은 다양한 모델에 적용하기 위함이다. Word2Vec 알고리즘은 구글에서 발표한 단어 임베딩 알고리즘 가운데 하나로, 주변 단어를 이용하여 타겟 단어를 임베딩한다 [28]. Word2Vec 알고리즘은 타겟 단어의 인덱스 값만 1이고, 나머지 인덱스는 모두 0으로 표현하는 원-핫 벡터를 기반으로 각 단어를 희소 표현(sparse representation) 한다. 이후, 주변 단어를 학습하여 분산 표현(distributed representation)을 통해 각 단어를 임베딩한다. 그러므로 Word2Vec 알고리즘을 이용하여 약물과 세포주 단어를 임베딩하는 것은 의생명 문헌에 있는 모든 정보를 활용하여 약물과 세포주의 특질을 표현할 수 있게 해준다.

본 연구에서는 전처리된 의생명 문헌을 기반으로 Word2Vec 알고리즘을 사용하여 약물과 세포주 특질을 추출하였다. 표 2는 본 연구에서 사용된 Word2Vec 알고리즘의 하이퍼파라메터 셋팅 값을 보여준다. ‘vector size’는 벡터를 구성하기 위한 차원의 수를 의미한다. 본 연구에서는 각 단어를 256차원의 벡터로 구축하였다, ‘window’는 Word2Vec 모델이 단어를 학습할 때, 사용하는 앞뒤 단어의 수이다. 본 연구에서는 5개의 앞뒤 단어를 사용하였다. ‘min count’는 전체 문헌에서 해당 단어가 등장한 수를 의미한다. 본 연구에서는 자주 등장하지 않은 단어 또한 포함 될 것으로 가정하여 1번 이상 포함된 모든 단어에 대하여 학습을 진행하였다. ‘sg’는 Word2Vec 모델을 학습하는데 있어 사용할 알고리즘을 선택하는 옵션이다. 1은 타겟 단어를 이용하여 주변 단어를 학습하는 skip-gram 알고리즘을 의미하고 0은 주변 단어를 이용하여 타겟 단어를

학습하는 cbow 알고리즘을 의미한다. 본 연구에서는 약물과 세포주 단어의 주변 정보를 활용하는 것이 목적이므로 cbow 알고리즘을 사용하였다. ‘epochs’은 Word2Vec 모델을 학습하는 횟수를 의미하고, 본 연구에서는 200회를 사용하였다.

표 2. Word2Vec 하이퍼파라메터 셋팅 값

Hyperparameter name	Value	Type
Vector Size	256	Numeric
Window	5	Numeric
Min Count	1	Numeric
sg	0	boolean
Epochs	200	Numeric

NCI-ALMANAC 데이터베이스에서 ComboScore는 각 약물 조합의 약물 농도-반응성 데이터를 기반으로 계산된 점수이다. 동일한 ‘약물1+약물2+세포주’ 조합일지라도, 약물의 농도에 따라 1개 이상의 ComboScore가 있다. 그러므로 본 연구에서는 각 조합 ComboScore의 평균을 사용하여 데이터 셋을 변환하였다. 표 3은 평균을 적용한 NCI-ALMANAC 데이터베이스의 예시이다.

표 3. NCI-ALMANAC 데이터베이스 예시

Drug 1	Drug 2	Cell line	Drug1 concentration	Drug2 concentration	ComboScore	Average
Methotrexate	Hydroxyurea	SF-295	1.00E-09	2.00E-07	5	14.22
			1.00E-09	2.00E-06	0	
			1.00E-09	2.00E-05	-4	
			1.00E-08	2.00E-07	28	
			1.00E-08	2.00E-06	73	
			1.00E-08	2.00E-05	39	
			1.00E-07	2.00E-07	-4	
			1.00E-07	2.00E-06	0	
			1.00E-07	2.00E-05	-9	
Busulfan	2-Fluoro Ara-A	CAKI-1	5.00E-07	6.00E-08	-1	14.33
			5.00E-07	6.00E-07	4	
			5.00E-07	6.00E-06	12	
			5.00E-06	6.00E-08	-5	
			5.00E-06	6.00E-07	-23	
			5.00E-06	6.00E-06	36	
			5.00E-05	6.00E-08	23	
			5.00E-05	6.00E-07	17	
			5.00E-05	6.00E-06	66	

Azacitidine	Thiotepa	NCI-H460	3.00E-08	1.00E-07	-7	20.44
			3.00E-08	1.00E-06	-4	
			3.00E-08	1.00E-05	4	
			3.00E-07	1.00E-07	-2	
			3.00E-07	1.00E-06	33	
			3.00E-07	1.00E-05	39	
			3.00E-06	1.00E-07	16	
			3.00E-06	1.00E-06	38	
			3.00E-06	1.00E-05	67	
Methotrexate	Dactinomycin	786-0	1.00E-09	1.00E-09	-4	-7.22
			1.00E-09	1.00E-08	0	
			1.00E-09	1.00E-07	3	
			1.00E-08	1.00E-09	-10	
			1.00E-08	1.00E-08	-8	
			1.00E-08	1.00E-07	7	
			1.00E-07	1.00E-09	3	
			1.00E-07	1.00E-08	-7	
			1.00E-07	1.00E-07	-49	
Busulfan	Mercaptopurine	A498	5.00E-07	2.00E-08	-11	-6.11
			5.00E-07	2.00E-07	-6	
			5.00E-07	2.00E-06	-6	
			5.00E-06	2.00E-08	-8	

			5.00E-06	2.00E-07	-9	
			5.00E-06	2.00E-06	-5	
			5.00E-05	2.00E-08	-8	
			5.00E-05	2.00E-07	-11	
			5.00E-05	2.00E-06	9	
Azacitidine	Thiotepa	CAKI-1	3.00E-08	1.00E-07	1	-16.33
			3.00E-08	1.00E-06	0	
			3.00E-08	1.00E-05	-18	
			3.00E-07	1.00E-07	-3	
			3.00E-07	1.00E-06	1	
			3.00E-07	1.00E-05	-2	
			3.00E-06	1.00E-07	-55	
			3.00E-06	1.00E-06	-41	
			3.00E-06	1.00E-05	-30	

예측 모델을 구축하기 위해서는 예측 모델을 학습시킬 학습데이터가 필요하다. 학습데이터는 데이터의 특성에 따라 세 가지 형태로 구축할 수 있다. 학습데이터의 정답이 있는 데이터를 기반으로 학습하는 지도학습(supervised learning), 학습데이터의 정답이 없는 데이터를 기반으로 학습하는 비지도학습(unsupervised learning), 학습데이터의 정답과 정답이 없는 데이터를 혼합하여 학습하는 준지도학습(semi-supervised learning)이 있다.

본 연구에서는 앞서 구축한 Word2Vec 모델을 통해 NCI-ALMANAC 데이터베이스에 있는 약물과 세포주의 임베딩 벡터를 추출하고, '약물1+약물2+세포주' 조합에 근거하여 학습데이터를 구축하였다. 또한, 평균 ComboScore를 기반으로 각 조합의 정답을 구축하였고, 이를 시너지 조합과 비시너지 조합으로 나누어 지도학습 형태의 학습데이터를 구축하였다. 구축된 학습데이터를 이용하여 각 항암 약물 조합의 시너지/비시너지를 분류할 수 있는 딥러닝 및 머신러닝 알고리즘 기반의 예측 모델을 구축하였다. 딥러닝 기반의 모델은 FFNN과 AE 알고리즘을 사용하였고, 머신러닝 기반의 모델은 XGB, ERT, LR 알고리즘을 사용하였다.

Kunjie 연구진은 기존의 *in-silico* 기반 약물 조합 예측 연구들에서 소개된 계산모델이나 구조를 단일화된 프레임워크로 구축하였다 [29]. 일반적으로, 각 연구들에서 사용된 데이터와 알고리즘은 다르기 때문에 각 연구들에서 구축한 예측 모델의 성능을 직접적으로 비교하는 것은 어렵다. 표 1에서처럼 각 연구에 따라 사용한 항암 약물 조합 데이터베이스도 다르고, 특질 추출을 위한 데이터 타입 및 알고리즘도 모두 상이하기 때문에 각 연구에 있는 성능 수치를 직접적으로 비교하는 것은 적합하지 않았다. 그러므로 본 연구에서는 Kunjie 연구진의 프레임워크를 이용하여 동일한 항암 약물 조합 데이터베이스 및 특질 데이터를 기반으로 구

축된 각각의 예측 모델의 성능을 직접 비교하였다. 또한, 프레임워크에서는 이전 연구들에서 사용된 하이퍼파라메터를 제공하기 때문에, 본 연구에서는 동일한 하이퍼파라메터 셋팅 값을 사용하여 항암 약물 조합의 시너지를 분류할 수 있는 예측 모델을 구축하였다.

Ⅲ. 연구 결과

1. 의생명 문헌 전처리 결과

NCI-ALMANAC 데이터베이스에는 약물 101개와 세포주 60개 데이터가 있다. 의생명 문헌은 포맷이 정해져있지 않아 저자에 따라 동일 단어가 다양한 동의어로 표현될 수 있다. 그러므로 본 연구에서는 동의어를 풍부하게 만들기 위하여 약물과 세포주에 대한 말뭉치를 구축하였다. NCI-ALMANAC 데이터베이스는 약물과 세포주 단어의 동의어를 제공하고 있으나, 공개된 약물 및 세포주 데이터베이스에 있는 동의어에 비하여 부족하였다. 그러므로 본 연구에서는 말뭉치 데이터를 구축하기 위하여 약물 관련 동의어는 PubChem에서 추출하였고, 세포주 관련 동의어는 Cellosaurus에서 추출하였다. 그 결과, NCI-ALMANAC 데이터베이스에 있는 단어 1,644개보다 약 8배 증가한 13,962개의 단어를 추출하여 말뭉치를 구축하였다.

본 연구에서는 말뭉치를 이용하여 PubMed에서 각 단어가 포함되어 있는 1,693,709개의 의생명 문헌을 추출하였고, 동의어를 대표단어로 변환, 토큰화, 표제어 추출, 불용어 제거를 통해 936,734개의 전처리 문헌을 구축하였다. 이후, 전처리된 문헌을 이용하여 Word2Vec 모델을 학습시킨 후, 101개의 약물과 60개의 세포주 단어의 임베딩 벡터를 추출하였다.

2. 항암 약물 조합 시너지 예측 모델 구축 결과

NCI-ALMANAC 데이터베이스에서 60개의 세포주와 적어도 1개 이상의 유전자를 타겟하는 68개의 약물을 조합한 130,182개의 조합을 추출하였다 [29]. 이 가운데 시너지 조합 3,548개, 비시너지 조합 126,634개를 추출하여 학습데이터를 구축하고 다양한 딥러닝 및 머신러닝 알고리즘을 적용하였다.

본 연구에서는 Kunjie 연구진의 프레임워크를 사용하여 예측 모델의 성능을 평가하였다. Kunjie 연구진의 결과를 베이스라인으로 하여 본 연구에서 구축한 예측 모델의 결과와 비교하였다. 베이스라인에서 사용된 세포주 특질(expression, mutation 등)은 CellMinerCDB에서 추출되었고, 약물 특질(drug-target, molecular 등)은 Python의 RDKit 패키지를 사용하여 DrugBank[30]로부터 추출되었다 [29].

예측 모델은 5-fold 교차검증(cross validation)으로 각각 5번씩 수행하여 구축되었다. 예측 모델 성능을 검증하기 위한 지표로 ROC-AUC(area under the receiver operating characteristic curve), AUPR(area under the precision-recall curve), F1-score를 사용하였다. ROC-AUC는 True Positive Rate(TPR)과 False Positive Rate(FPR) 비율을 기반으로 평가한다. TPR은 실제 시너지 조합 중 시너지로 예측한 조합의 비율을 의미하고, FPR은 실제 비시너지 조합 중 시너지로 잘못 예측한 조합의 비율을 의미한다. 그러므로 TPR 값이 높고, FPR 값이 낮을수록 예측 모델의 성능이 우수한 것을 의미한다. AUPR은 정밀도(precision)과 재현율(recall)의 비율을 기반으로 평가한다. 정밀도는 시너지로 예측한 조합 중 실제 시너지 조합의 비율을 의미하고, 재현율은 실제 시너지 조합 중 시너지로 예측한 조합의 비율을 의미한다. 그러므로

정밀도와 재현율이 모두 높을수록 예측 모델의 성능이 우수한 것을 의미한다. F1-score는 정밀도와 재현율의 조화 평균을 기반으로 평가하므로 정밀도와 재현율이 모두 높을수록 예측 모델의 성능이 우수한 것을 의미한다.

예측 모델 구축 시, 학습데이터의 학습 셋과 테스트 셋이 랜덤으로 추출되고 5번 반복 수행하였기 때문에, 예측 모델마다 성능의 차이가 생긴다. 그러므로 본 연구에서는 성능 평가 지표의 평균과 표준편차를 사용하여 평가하였다. 표 4~6에는 각각의 성능 평가 지표에 따른 예측 모델의 성능 결과가 있다.

표 4. ROC-AUC 기반 예측 모델 성능 비교

Algorithms	FFNN		AE		XGB		ERT	LR
Ref.	[13]	[14]	[16]	[17]	[19]	[20]	[22]	[24]
Baseline	0.912 ±0.004	0.91 ±0.005	0.914 ±0.004	0.895 ±0.007	0.885 ±0.01	0.895 ±0.005	0.843 ±0.007	0.847 ±0.006
Proposed	**0.924 ±0.001**	0.915 ±0.006	0.923 ±0.003	0.92 ±0.004	0.889 ±0.003	0.892 ±0.004	0.881 ±0.004	0.854 ±0.007

표 5. AUPR 기반 예측 모델 성능 비교

Algorithms	FFNN		AE		XGB		ERT	LR
Ref.	[13]	[14]	[16]	[17]	[19]	[20]	[22]	[24]
Baseline	0.402 ±0.017	0.417 ±0.032	0.408 ±0.025	0.339 ±0.026	0.349 ±0.016	0.381 ±0.026	0.24 ±0.015	0.192 ±0.011
Proposed	0.434 ±0.014	0.427 ±0.025	**0.438 ±0.008**	0.424 ±0.011	0.371 ±0.013	0.381 ±0.016	0.326 ±0.012	0.196 ±0.004

표 6. F1-score 기반 예측 모델 성능 비교

Algorithms	FFNN		AE		XGB		ERT	LR
Ref.	[13]	[14]	[16]	[17]	[19]	[20]	[22]	[24]
Baseline	0.359 ±0.027	0.262 ±0.045	0.325 ±0.039	0.224 ±0.05	0.227 ±0.01	0.24 ±0.019	0.272 ±0.018	0.059 ±0.01
Proposed	0.392 ±0.017	0.313 ±0.052	0.296 ±0.021	**0.428 ±0.028**	0.271 ±0.016	0.259 ±0.02	0.263 ±0.015	0.064 ±0.013

본 연구에서 개발한 특질 추출 방법으로 구축한 예측 모델의 성능이 베이스라인에 비하여 대부분 우수하였다. 예측 모델의 성능은 물론이고, 예측 모델 간의 성능 격차 또한 차이가 적은 것을 확인 할 수 있다. 이에 따라, 본 연구에서 개발한 특질 추출 방법을 기반으로 구축한 예측 모델이 베이스라인에 비하여 안정적이고, 우수한 것을 확인 할 수 있다. 또한, 본 연구에서 개발한 약물과 세포주의 특질 추출 방법이 이전 연구들에서 사용한 특질 추출 방법에 비하여 약물과 세포주의 고유한 특성을 더 반영한 것으로 추론 할 수 있다.

그림 8은 성능이 가장 우수한 FFNN 기반 예측 모델의 혼동행렬(confusion matrix) 결과이다. 혼동행렬을 통하여 실제 시너지 조합 중 시너지로 예측한 조합의 비율(TPR)은 낮으나, 실제 비시너지 조합 중 시너지로 잘못 예측한 조합의 비율(FPR)이 더 낮으므로 ROC-AUC 지표에서 우수한 것으로 나타났다. 그림 9는 FFNN 기반 예측 모델의 ROC 커브 결과이다. ROC-AUC 결과에 비하여 AUPR과 F1-score 지표에서 낮은 성능 결과를 보였다. 이는 시너지로 예측한 조합 중 실제 시너지 조합의 비율(정밀도)에 비하여, 실제 시너지 조합 중 시너지로 예측한 조합의 비율(재현율)이 낮은 이유로 분석할 수 있다.

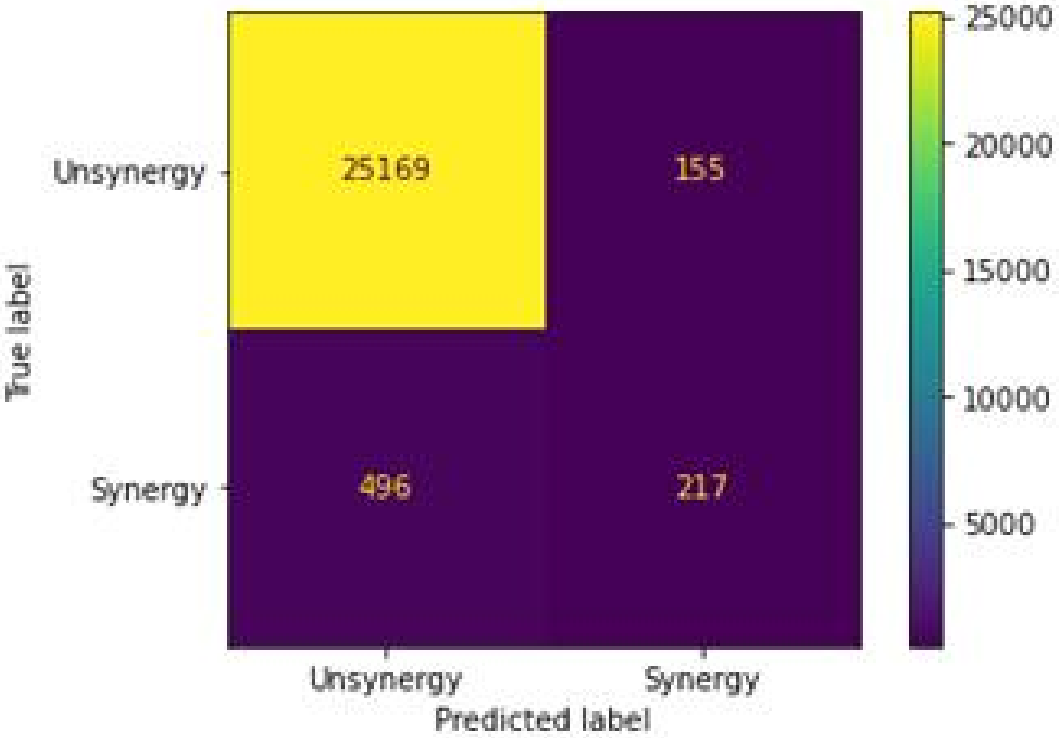

그림 8. FFNN 기반 예측 모델의 혼동행렬

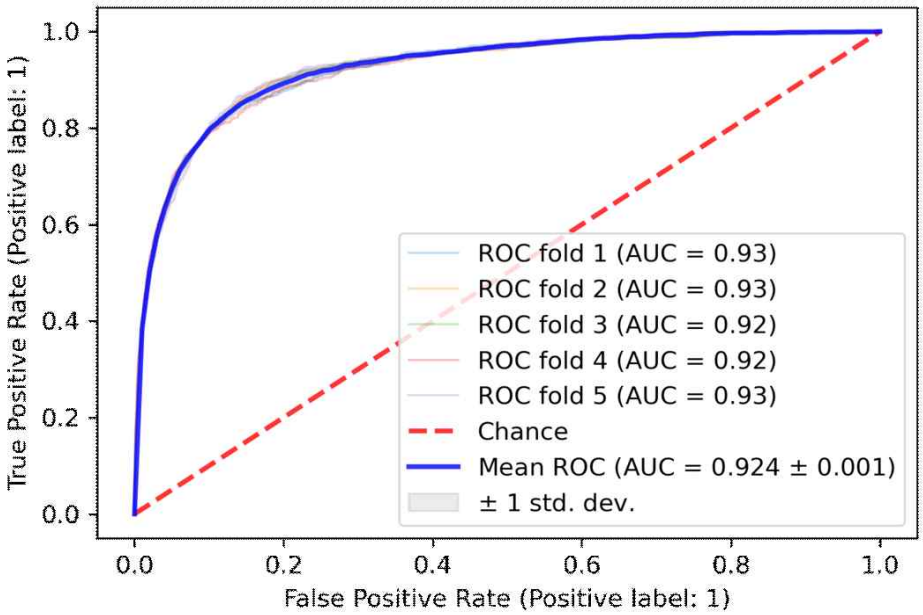

그림 9. FFNN 기반 예측 모델의 ROC 커브

Ⅳ. 연구 활용

1. 개요

본 연구에서는 의생명 문헌으로부터 추출한 특질을 기반으로 항암 약물 조합의 시너지를 예측하는 방법을 개발하였다. 본 방법을 통하여 추출된 예측 모델은 잠재적인 약물 조합의 시너지를 예측하는 연구에 있어 활용 될 수 있다. 그러나 대부분의 기존 연구들은 예측 모델을 구축한 후, 이를 활용하여 미확인된 항암 약물 조합의 잠재적인 시너지를 예측한 연구는 찾아보기 어렵다. 이러한 문제를 해결하기 위하여, 본 연구에서 개발한 방법을 기반으로 대규모 항암 약물 조합 데이터베이스를 이용하여 항암 약물 조합을 확장하고, 각 조합의 잠재적인 시너지를 추출한 데이터베이스 DPS(Database of Potential Synergy)를 구축하였다. 또한, DPS의 내용을 검색할 수 있는 유저 친화적인 웹 어플리케이션 FPS(Finder of Potential Synergy)를 구축하였다. FPS는 약물과 세포주의 키워드만을 이용하여 항암 약물 조합의 잠재적 시너지를 검색이 가능하다. 이를 통해, 항암 약물 조합 연구와 관련성이 낮은 연구자들 또한 쉽게 도구를 사용할 수 있어 다양한 분야에서 진입이 용이하다. FPS는 http://bike-fps.snu.ac.kr/에서 이용 가능하다.

DPS를 구축하기 위한 데이터리소스, 과정 및 결과에 대한 자세한 내용은 다음에 기술되어 있다.

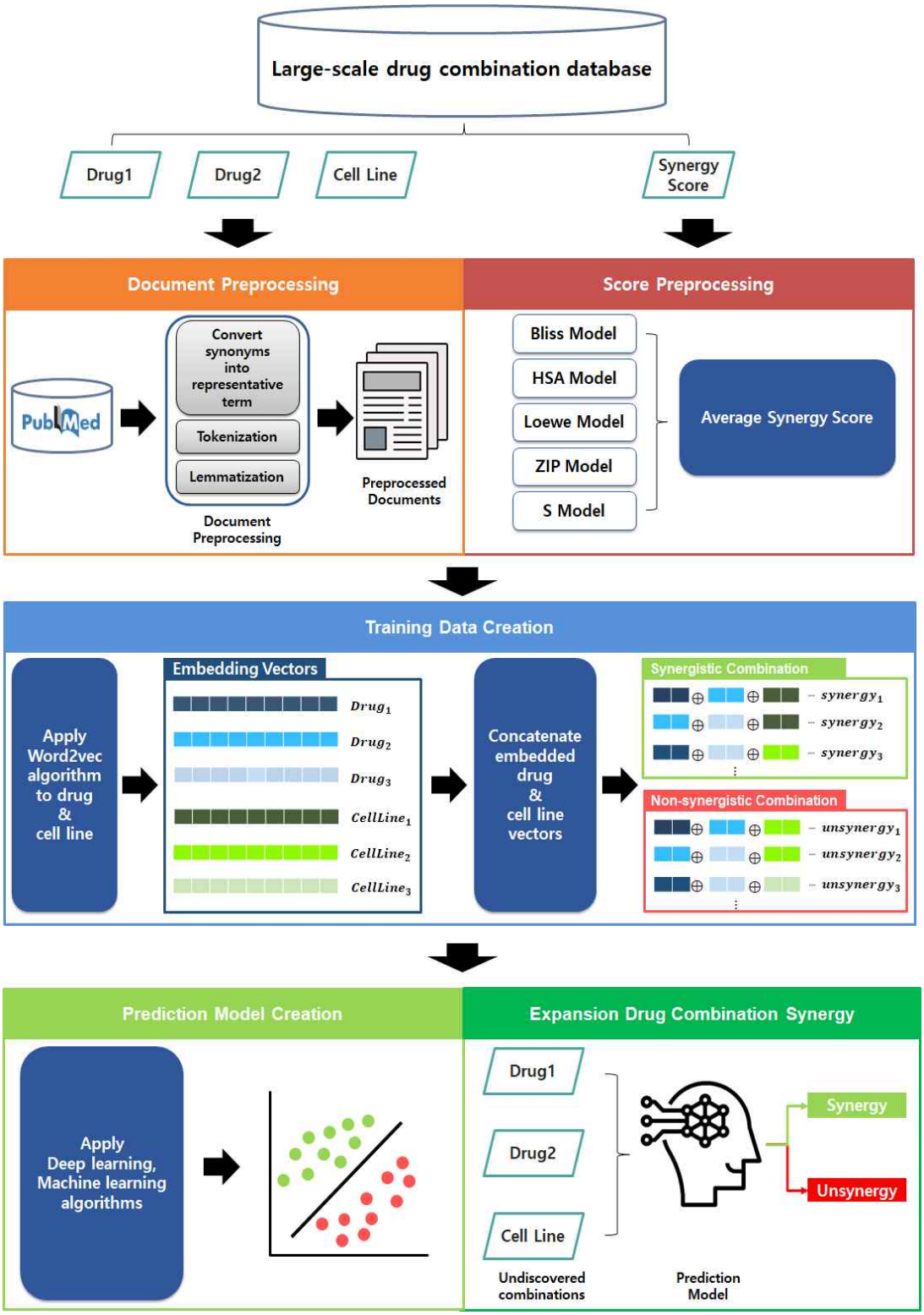

그림 10. DPS 구축의 워크플로우

2. 데이터 리소스

DrugComb는 NCI-ALMANAC, O'NEIL dataset 등 공개된 항암 약물 조합 관련 데이터베이스들을 통합하고 분석한 항암 약물 조합 관련 포털이다 [31]. DrugComb를 통해 사용자는 복합 스크리닝 데이터를 분석하고 시각화할 수 있다. DrugComb 데이터베이스는 항암 약물 조합의 농도-반응성 매트릭스로 구성되어 있고, 각 약물 조합의 반응성 결과를 5개의 계산 모델을 기반으로 계산된 결과가 포함되어 있다.

S 모델은 하나의 약물을 IC50 농도로 고정한 후, 약물 조합의 농도-반응성에 대한 AUC(average area under cover)의 면적인 약물 조합 민감도 점수(drug combination sensitivity score)와 단일 약물 농도-반응성 곡선의 AUC 합계의 차이로 약물 조합의 시너지 효과 스코어를 계산 한다 [31]. ZIP(the zero interaction potency) 모델은 단일 약물이 독립적으로 작용하여 약물 조합 시 서로를 강화하지 않는다는 가정 하에 단일 약물과 그 조합 간의 용량-반응 곡선의 효능(특정 용량 수준에서의 효과) 변화를 비교하여 약물 조합의 시너지 효과 스코어를 계산 한다 [32]. Bliss 모델은 단일 약물 사용 시 발생하는 효과의 확률과 약물을 조합하여 사용 시 발생하는 효과의 확률을 기반으로 약물 조합의 시너지 효과 스코어를 계산 한다 [33]. HSA(highest single agent) 모델은 단일 약물 사용 시 발생하는 효과 가운데 가장 높은 스코어의 약물 농도를 기준으로 약물 조합의 시너지 효과 스코어를 계산 한다 [34]. Loewe additivity 모델은 단일 약물의 농도-반응성 곡선을 기반으로 조합된 약물의 농도가 단일 약물의 농도를 초과하지 않는 다는 것을 가정하여 약물 조합의 시너지 효과 스코어를 계산 한다 [35].

DrugComb 데이터베이스에는 약물 4,268개, 세포주 288개가 있고, 이

들을 조합한 약물 조합 739,652개의 약물 농도-반응성 매트릭스와 계산 값이 있다. 각 항암 약물 조합의 반응성 결과는 S 모델을 기반으로 계산된 'S', ZIP 모델을 기반으로 계산된 'Z', Bliss 모델을 기반으로 계산된 'B', HSA 모델을 기반으로 계산된 'H', Loewe additivity 모델을 기반으로 계산된 'L'로 이루어져있다. 5가지 계산 값은 각 약물 조합의 약물 농도-반응성 데이터를 기반으로 계산되었다. NCI-ALMANAC 데이터베이스와 마판가지로, 동일한 '약물1+약물2+세포주' 조합임에도 불구하고 약물의 농도에 따라 계산 값이 다른 데이터들이 존재한다. 그러므로 농도별 5가지 계산 값의 평균을 적용한 후, 조합에 따라 평균을 사용하였다. 표 7은 각 조합에 대하여 5가지 계산 값과 평균 스코어 예시이다.

표 7. DrugComb 데이터베이스 예시

Drug 1	Drug 2	Cell line	Drug1 CONC	Drug2 CONC	S	Z	B	H	L	Average
Pralatrexate	methotrexate	SF-268	6.77E-05	4.43E-02	27.39	3.53	8.86	6.45	7.14	10.67
AT7519	PF-3758309	EW-8	6.30E-01	5.00E+00	27.14	13.86	13.27	16.79	13.87	19.49
			8.51E-01	1.45E+00	34.07	17.15	15.27	18.78	16.83	
			7.56E-01	2.31E+00	42.31	16.51	15.38	16.62	14.51	
AT7519	PF-3758309	TC-71	4.05E-01	5.00E+00	28.78	14.75	14.32	15.7	13.81	16.44
			3.64E-01	5.00E+00	5.61	7.74	7.25	13.73	11.72	
			4.29E-01	5.00E+00	40.21	17.97	17.03	20.39	17.63	
Alisertib	Trametinib	TC-32	5.00E+00	2.60E-02	-3.67	-0.53	-8.03	-13.47	-37.42	-12.62
Alisertib	ZINC34894448	TC-32	5.00E+00	5.69E-02	0.28	4.16	2.39	-3.37	-26.24	-4.55
Antibiotic AY 22989	Romidepsin	SF-268	9.33E-03	1.00E-02	17.74	-1.95	-8.19	-5.25	-29.07	-5.34

DrugComb 데이터베이스는 PubChem, ChEMBL[36] 등 약물 관련 공개된 데이터베이스들을 기반으로 약물 말뭉치를 구축하였다. 또한, Cellosaurus를 기반으로 세포주 말뭉치를 구축하였다. 말뭉치 데이터는 API 형태로 제공되고 있고, http://api.drugcomb.org에서 이용 가능하다. 본 연구에서는 DrugComb 데이터베이스에서 제공하는 약물 및 세포주 말뭉치를 이용하였다. 이러한 말뭉치에 있는 단어를 기반으로 PubMed에서 해당 단어가 포함되어 있는 의생명 문헌을 추출하였다. 또한, 추출한 의생명 문헌에 있는 약물 및 세포주 동의어를 말뭉치에 있는 대표 단어로 변환하였다. 이후, 토큰화를 통해 단어를 추출하고, 표제어 추출을 통해 각 단어를 단어의 기본형으로 변환하여 전처리된 의생명 문헌 데이터셋을 구축하였다.

3. 항암 약물 조합 시너지 예측 모델 구축

본 연구에서는 앞서 개발한 방법을 기반으로 항암 약물 조합 데이터베이스를 확장하고, 잠재적인 항암 약물 조합의 시너지를 예측하였다. 그러므로 전처리된 의생명 문헌을 기반으로 Word2Vec 모델을 학습하여 약물과 세포주 특질을 추출하였다. 이때, 표 2에서 사용된 하이퍼마라메터 셋팅 값을 동일하게 사용하였다.

학습된 Word2Vec 모델을 통해 DrugComb 데이터베이스에 있는 약물과 세포주의 임베딩 벡터를 추출하고, '약물1+약물2+세포주' 조합에 근거하여 학습데이터를 구축하였다. 또한, 학습데이터를 각 조합의 평균스코어를 기반으로 시너지 조합과 비시너지 조합으로 분리하여 학습데이터를 구축하였다. 예측 모델 구축 시, NCI-ALMANAC 데이터베이스 기반 예측 모델의 성능 결과와 DrugComb 데이터베이스 기반 예측 모델의 성능 결과를 비교하기 위하여 베이스라인에서 사용한 딥러닝(FFNN, AE) 및 머신러닝(XGB, ERT, LR) 알고리즘과 하이퍼파라메터 셋팅 값을 동일하게 사용하였다.

4. 항암 약물 조합 시너지 예측 모델 성능 비교

본 연구에서는 약물과 세포주 말뭉치를 이용하여 PubMed에서 각 단어가 포함되어 있는 11,683,382개의 의생명 문헌을 추출하였다. 또한, 약물 말뭉치를 이용하여 문헌에 있는 약물 동의어를 대표단어로 변환하였고, 세포주 말뭉치를 이용하여 세포주 동의어를 대표단어로 변환하였다. 이후, 토큰화, 표제어 추출을 통해 8,891,419개의 전처리 문헌을 구축하였다. 전처리 문헌을 이용하여 Word2vec 모델 기반 약물과 세포주 단어의 임베딩 벡터를 추출하였다.

PubMed에 의생명 문헌이 없는 약물과 세포주를 제외한 결과, 약물은 3,627개, 세포주는 222개를 추출하고, 해당 약물과 세포주가 포함되어 있는 479,056개의 조합을 추출하였다. 이 가운데 시너지 조합 52,094개, 비시너지 조합 426,962개를 추출하여 학습데이터를 구축하고 딥러닝 및 머신러닝 알고리즘을 적용하였다.

표 8에서는 본 연구에서 사용된 NCI-ALMANAC 데이터베이스와 DrugComb 데이터베이스 기반으로 추출된 약물, 세포주 등의 데이터 수를 비교하였다. DrugComb 데이터베이스가 공개된 항암 약물 조합 데이터베이스들을 통합하였기 때문에 NCI-ALMANAC 데이터베이스에 비하여 약물의 수, 세포주의 수, 조합의 수 등 정량적 데이터 수가 많은 것을 확인 할 수 있다.

표 8. NCI-ALMANAC 데이터베이스와 DrugComb 데이터베이스의 데이터 수 비교

	NCI-ALMANAC		DrugComb	
약물 (개수)	101		3,627	
세포주 (개수)	60		222	
말뭉치 데이터 (개수)	13,962		293,346	
의생명 문헌 (개수)	936,734		8,891,419	
학습데이터 (개수)	시너지 셋	비시너지 셋	시너지 셋	비시너지 셋
	3,548	126,634	52,094	426,962
	130,182		479,056	

예측 모델은 5-fold 교차검증을 각각 5번 수행하여 예측 모델의 성능을 평가하였다. 예측 모델의 성능을 평가하기 위한 지표로 앞서 사용한 ROC-AUC, AUPR, F1-score를 사용하였다. Kunjie 연구진의 결과를 베이스라인으로 하고, NCI-ALMANAC 데이터베이스를 기반으로 구축한 예측 모델의 성능과 DrugComb 데이터베이스를 기반으로 구축한 예측 모델의 성능을 비교하였다. 표 9~11에는 각각의 성능 평가 지표에 따른 예측 모델의 성능 결과가 있다.

표 9. ROC-AUC 기반 예측 모델 성능 비교

Algorithms	FFNN		AE		XGB		ERT	LR
Ref.	[13]	[14]	[16]	[17]	[19]	[20]	[22]	[24]
Baseline (NCI-ALMANAC)	0.912 ±0.004	0.91 ±0.005	0.914 ±0.004	0.895 ±0.007	0.885 ±0.01	0.895 ±0.005	0.843 ±0.007	0.847 ±0.006
Proposed (NCI-ALMANAC)	0.924 ±0.001	0.915 ±0.006	0.923 ±0.003	0.92 ±0.004	0.889 ±0.003	0.892 ±0.004	0.881 ±0.004	0.854 ±0.007
Proposed (DrugComb)	**0.939 ±0.001**	0.933 ±0.002	0.931 ±0.001	0.93 ±0.001	0.822 ±0.001	0.823 ±0.002	0.896 ±0.003	0.82 ±0.001

표 10. AUPR 기반 예측 모델 성능 비교

Algorithms	FFNN		AE		XGB		ERT	LR
Ref.	[13]	[14]	[16]	[17]	[19]	[20]	[22]	[24]
Baseline (NCI-ALMANAC)	0.402 ±0.017	0.417 ±0.032	0.408 ±0.025	0.339 ±0.026	0.349 ±0.016	0.381 ±0.026	0.24 ±0.015	0.192 ±0.011
Proposed (NCI-ALMANAC)	0.434 ±0.014	0.427 ±0.025	0.438 ±0.008	0.424 ±0.011	0.371 ±0.013	0.381 ±0.016	0.326 ±0.012	0.196 ±0.004
Proposed (DrugComb)	**0.763 ±0.003**	0.742 ±0.006	0.737 ±0.002	0.752 ±0.004	0.382 ±0.008	0.376 ±0.005	0.589 ±0.005	0.352 ±0.003

표 11. F1-score 기반 예측 모델 성능 비교

Algorithms	FFNN		AE		XGB		ERT	LR
Ref.	[13]	[14]	[16]	[17]	[19]	[20]	[22]	[24]
Baseline (NCI-ALMANAC)	0.359 ±0.027	0.262 ±0.045	0.325 ±0.039	0.224 ±0.05	0.227 ±0.01	0.24 ±0.019	0.272 ±0.018	0.059 ±0.01
Proposed (NCI-ALMANAC)	0.392 ±0.017	0.313 ±0.052	0.296 ±0.021	0.428 ±0.028	0.271 ±0.016	0.259 ±0.02	0.263 ±0.015	0.064 ±0.013
Proposed (DrugComb)	**0.67 ±0.006**	0.614 ±0.011	0.587 ±0.006	0.655 ±0.009	0.061 ±0.003	0.054 ±0.003	0.392 ±0.003	0.111 ±0.002

예측 모델을 구축하기 위하여 사용된 데이터베이스가 다르기 때문에 예측 모델 간의 직접적인 비교는 적절하지 않으나, DrugComb 데이터베이스를 기반으로 구축한 예측 모델의 성능이 베이스라인과 NCI-ALMANAC 데이터베이스 기반의 예측 모델에 비하여 절대적인 수치에서 가장 우수하고, 각 모델 간의 성능 편차도 크지 않았다. 이를 통해, DrugComb 데이터베이스를 기반으로 구축한 예측 모델의 성능이 가장 안정적이며 성능이 우수한 것을 확인 할 수 있다.

그림 11은 성능이 가장 우수한 DrugComb+FFNN 기반 예측 모델의 혼동행렬 결과이다. NCI-ALMANAC 데이터베이스를 기반으로 구축한 모델의 결과인 그림 8과 그림 11을 비교한 결과, DrugComb 데이터베이스 기반으로 구축된 예측 모델의 성능이 NCI-ALMANAC 데이터베이스 기반으로 구축된 예측 모델의 성능에 비하여 실제 시너지 조합 중 시너지로 예측한 조합의 비율(TPR)이 약 2배 상승한 것을 확인 할 수 있다. 그러므로 DrugComb 데이터베이스를 기반으로 구축한 예측 모델의 시너지 예측률이 더 우수한 것을 확인 할 수 있다. 이러한 결과를 바탕으로 ROC-AUC 지표가 상승한 것을 확인할 수 있다. 또한, 시너지로 예측한 조합 중 실제 시너지 조합의 비율(정밀도)과 실제 시너지 조합 중 시너지로 예측한 조합의 비율(재현율)이 모두 상승하여 AUPR과 F1-score 지표가 큰 폭으로 상승하였다. 이를 통해, 학습데이터의 증가가 예측 모델의 성능 향상 요인으로 작용한 것을 확인 할 수 있다.

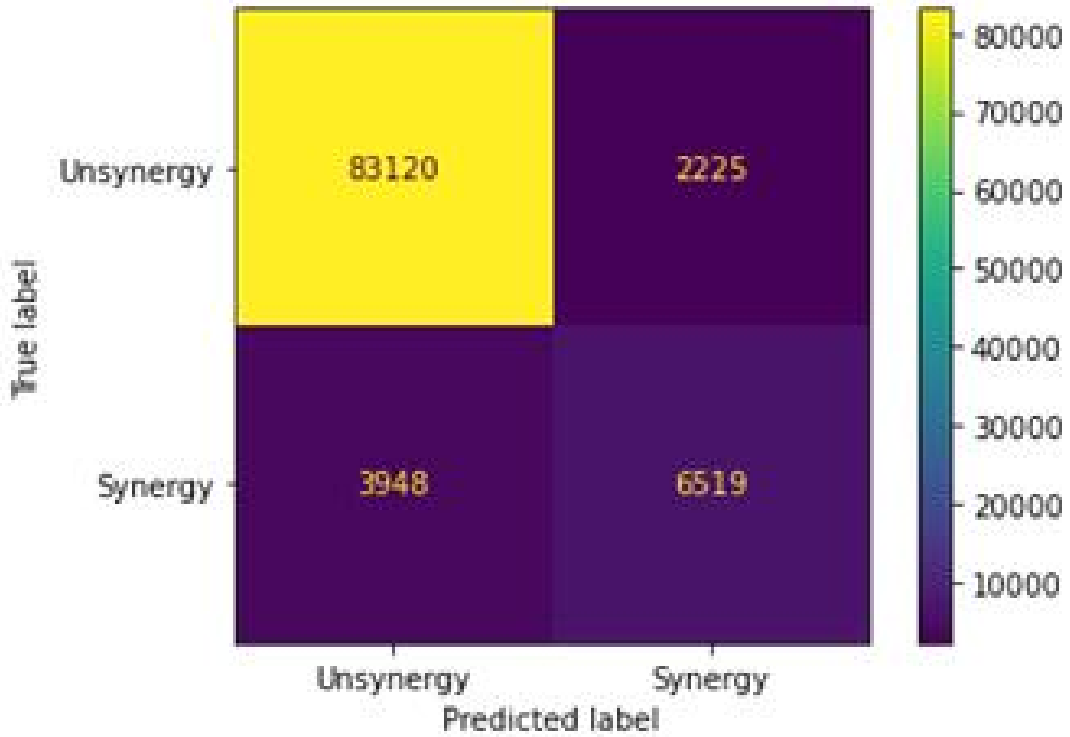

그림 11. DrugComb+FFNN 기반 예측 모델의 혼동행렬

5. Database of Potential Synergy (DPS)

각 연구들에서 구축한 예측 모델은 잠재적인 약물 조합의 시너지를 예측하는 연구에 있어 활용 될 수 있다. 그러나 *in-silico* 기반 항암 약물 조합의 시너지 예측 모델을 활용하여 밝혀지지 않은 항암 약물 조합의 시너지를 예측한 데이터베이스는 찾아보기 어렵다.

본 연구에서는 공개된 항암 약물 조합 데이터베이스에 포함되어 있지 않은 조합의 잠재적인 시너지를 예측한 데이터베이스인 DPS를 구축하였다. DPS를 구축하기 위하여 예측 모델 중 성능이 가장 우수한 DrugComb+FFNN 기반의 예측 모델을 활용하였다. 예측 모델에서 항암 약물 조합의 시너지 효과를 예측하는데 있어 시너지 효과가 있는지 없는지 분류 할 수 있고, 분류 결과에 대한 확률을 보여줄 수 있다.

본 연구에서는 앞서 추출한 약물과 세포주를 조합하여 646,749,492개의 조합을 추출하였다. DrugComb+FFNN 기반의 예측 모델을 활용하여 각 조합에 대하여 잠재적 시너지를 분류하고, 분류 결과에 대한 확률을 구축하였다.

6. Finder of Potential Synergy (FPS)

본 연구에서는 DPS의 데이터를 검색 할 수 있는 유저 친화적인 웹 어플리케이션 FPS를 구축하였다. FPS는 약물과 세포주 키워드만을 이용하여 항암 약물 조합의 잠재적 시너지를 검색 할 수 있어, 도구 사용에 대한 진입장벽을 낮추었다.

메인 페이지에는 약물 조합과 타겟 세포주 조합의 결과를 검색할 수 있는 'Combination' 섹션, 약물을 기준으로 잠재적인 약물 조합의 시너지를 검색 할 수 있는 'Drug' 섹션과 세포주를 기준으로 잠재적인 약물 조합의 시너지를 검색 할 수 있는 'Cell line' 섹션이 있다. 'Combination' 섹션에서 사용자는 확인하고자 하는 약물 조합과 세포주 키워드를 입력한다. 'Drug' 섹션과 'Cell line' 섹션에서는 각각 약물과 세포주 키워드를 입력한다. 그림 12(a)는 FPS의 인터페이스 중 메인 페이지를 보여준다.

결과 페이지에는 사용자가 메인 페이지에서 선택한 약물과 세포주 조합의 잠재적인 시너지 예측 결과가 있다. 'Combination' 섹션을 통해 검색한 경우 선택된 조합에 대한 결과를 보여주고, 'Drug' 섹션과 'Cell line' 섹션을 통해 검색한 경우 선택된 키워드가 포함되어 있는 조합에 대한 결과를 보여준다. 결과 테이블에서는 약물과 세포주가 있고, 해당 조합의 예측 결과가 있다. 또한, 예측 결과를 통해 출력된 약물 조합의 시너지 예측 결과와 확률을 확인 할 수 있다. 그림 12(b)는 FPS의 인터페이스 중 결과 페이지를 보여준다.

FPS는 도메인 전문가뿐만 아닌 비도메인 전문가를 모두 포함한 사용자를 대상으로 구축되었다. 그러므로 약물 및 세포주에 관한 간략한 정보와 자세하게 참고할 수 있는 정보 페이지를 구축하였다. 약물 및 세포주 정보 페이지에 사용자는 데이터베이스에 있는 약물과 세포주에 대한

정보를 볼 수 있다. 해당 페이지를 통해 약물 및 세포주 관련 도메인 전문가뿐만 아닌 비도메인 전문가 사용자의 편의성을 증가시키고 약물 및 세포주 정보에 대하여 참고할 수 있다. 약물 정보 페이지에서는 약물의 이름, 간략한 설명, 약물의 SMILES, 동의어 및 기타 참고할 수 있는 페이지를 맵핑하였다. 세포주 정보 페이지에서는 세포주의 이름, 동의어 및 기타 참고할 수 있는 페이지를 맵핑하였다. FPS의 인터페이스 중 약물과 세포주 정보 페이지는 그림 12(c)와 (d)에서 확인 할 수 있다.

(a)

Potential Synergy Search

Combination Drug Cell line

(+)-BICUCULLINE | ACAMPROSATE | SK-MEL-5

Search

(b)

Prediction Result

DRUG1	DRUG2	Cell Line	Prediction Result	Probability
(+)-BICUCULLINE	ACAMPROSATE	SK-MEL-5	Synergistic effect	0.843

(c)

Drug Information

Name	Description	Smiles	Synonyms	References
(+)-BICUCULLINE	Bicuculline is a benzylisoquinoline alkaloid that is 6-methyl-5,6,7,8-tetrahydro [1,3]dioxolo[4,5-g]isoquinoline which is substituted at the 5-pro-S po [more]	CN1CCC2=CC3=C(C=C2C1C4C5=C(C6=C(C=C5)OCO6)C(=O)O4)OCO3	(+)-Bicuculline; bicuculline; d-Bicuculline; 485-49-4; Bicculine; Bicucullin; Bicuculline (+); UNII-Y37615DVKC; NSC 32192; C20H17NO6; EINECS 207-619-7 [more]	https://pubchem.ncbi.nlm.nih.gov/compound/10237 https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL417990 https://www.genome.jp/entry/C09364

(d)

Cell Line Information

Name	Synonyms	References
22RV1	22Rv1; 22RV1; 22Rv-1; 22rV1; CWR22-Rv1; CWR22R-V1; CWR22-R1; CWR22Rv1; CWR22R	https://web.expasy.org/cellosaurus/CVCL_1045 https://depmap.org/portal/cell_line/ACH-000956 https://cellmodelpassports.sanger.ac.uk/passports?q=SIDM00499

그림 12. FPS의 인터페이스

그림 12(a)의 예시에서는 약물 '(+)-BICUCULLINE', 'ACAMPROSATE'와 세포주 'SK-MEL-5' 조합을 사용하였다. 그림 12(b)의 예시에서는 약물 '(+)-BICUCULLINE', 'ACAMPROSATE'와 세포주 'SK-MEL-5' 조합의 결과를 보여준다. '(+)-BICUCULLINE', 'ACAMPROSATE', 'SK-MEL-5' 조합은 시너지 효과가 있는 것으로 예측되었고, 예측 결과에 대한 확률은 84.3% 인 것을 확인 할 수 있다.

BEE(Biomedical Entity Explorer)는 유전자를 기준으로 의생명 개체 간의 관계를 통합하여 각 개체들 간의 관계를 복합적으로 질의할 수 있는 시스템이다 [37]. BEE에서는 각 의생명 개체와 관련 있는 유전자를 검색할 수 있으며, 이종 의생명 개체들 사이의 관련 있는 유전자 또한 검색 할 수 있다. BEE에서는 세포주 데이터의 뿌리 개념인 질병 데이터를 기반으로 관련 있는 유전자를 검색할 수 있다. 세포주 'SK-MEL-5'는 흑색종(cutaneous melanoma)으로부터 추출되었으므로 SK-MEL-5 대신 cutaneous melanoma 키워드를 이용하여 '(+)-BICUCULLINE', 'ACAMPROSATE', 'cutaneous melanoma' 조합과 관련된 유전자를 검색하였다. 그 결과 '(+)-BICUCULLINE', 'ACAMPROSATE', 'cutaneous melanoma' 조합이 GABRA3(Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha3) 유전자와 공통적으로 관련 있는 것을 확인할 수 있었다. 그림 13은 BEE에서 '(+)-BICUCULLINE', 'ACAMPROSATE', 'cutaneous melanoma' 조합과 관련된 유전자 검색 결과에 대한 예시이다.

이처럼 FPS를 통해 발견되지 않은 항암 약물 조합의 시너지 가능성을 보여줌으로써, 항암 약물 조합 실험 전, 후보군을 선별하는데 있어 활용 될 수 있다. 또한, 기존 도구들과 함께 사용함으로써 후보군 선별의 근거를 마련하는데 활용될 수 있다.

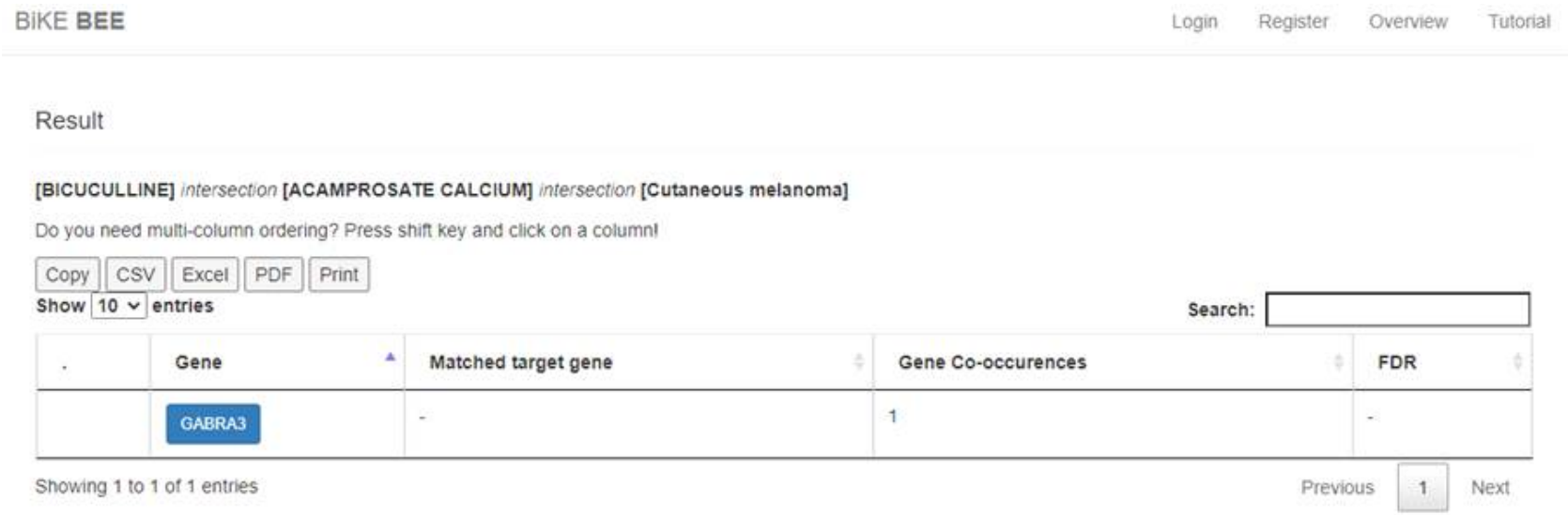

그림 13. BEE 검색 결과 예시

V. 결 론

본 연구에서는 의생명 문헌에서 약물 및 세포주 특질을 추출하여 항암 약물 조합의 시너지 효과를 예측하였다. 본 연구에서 구축한 예측 모델은 기존의 연구들에 비하여 대부분 우수한 성능을 보여주었다. 이러한 결과를 바탕으로 본 연구의 특질 추출 방법이 기존 연구들에서 사용된 특질 추출 방법에 비하여 약물과 세포주 정보를 보다 풍부하게 반영할 수 있는 가능성을 보여주었다. 또한, 본 연구에서 개발한 방법을 활용하여 대규모 항암 약물 조합 데이터베이스 기반의 미확인된 항암 약물 조합의 잠재적 시너지를 예측하여 DPS를 구축하였으며, DPS의 데이터를 검색할 수 있는 유저 친화적인 웹 어플리케이션 FPS를 구축하였다. 본 연구의 과정과 결론을 통해 다음과 같은 학술적 기여를 기대한다.

첫째, *in-silico* 기반의 연구에서 특질 추출 방법 측면에 대한 기여를 기대한다. 정형데이터는 비정형데이터에 비하여 데이터를 파악하거나, 사용함에 있어 직관적이지만, 데이터를 추가하거나 확장하는데 있어 어려움이 있다. 새롭게 발표된 연구 내용을 구조화하는 것은 도메인 전문가 없이 수행하기 어려움이 있다. 그러므로 새로운 연구가 공개된 데이터베이스에 업데이트가 되기까지 오랜 시간이 소요된다. 본 연구에서는 비정형데이터를 사용함으로써 약물과 세포주 관련 새로운 연구 출판 시 문헌을 추가하여 약물 및 세포주 특질 추출에 반영 할 수 있어 데이터 확장성에 유연하다. 또한, 본 연구를 통해 추출된 특질은 기존 연구에서 추출한 정형데이터 기반의 특질과도 함께 사용 할 수 있기 때문에 기존 연구에서 추출한 특질 정보를 이용할 수 있다. 그러므로 본 연구에서 개발한 특질 추출 방법을 기반으로 약물의 타겟을 치주질환, 만성질환 등 다른 질병으로 확장한다면 다양한 의생명 연구 분야에 활용 가능할 것으로 기

대한다.

둘째, 새로운 연구 방향성을 제시할 수 있을 것으로 기대한다. 기존의 연구들은 정형데이터로부터 특질을 추출한 반면, 본 연구에서는 비정형 데이터로부터 특질을 추출하여 *in-silico* 기반 항암 약물 조합 시너지 예측 연구를 진행하였다. 본 연구를 통해 구축된 예측 모델의 성능은 기존 연구들에 비하여 대부분 우수한 성능 지표를 보여주었다. 이를 통해, *in-silico* 기반의 항암 약물 조합 시너지 예측 연구에서 비정형데이터를 기반으로 추출한 특질을 활용하는 것에 대한 근거를 줄 수 있다. 이러한 결과를 바탕으로 *in-silico* 기반 항암 약물 조합 시너지 예측 연구에서 새로운 방향성을 제시할 수 있을 것으로 기대한다.

셋째, 약물 조합의 시너지 예측 도구 측면에 대한 기여를 기대한다. 항암 약물 조합 시너지 예측 모델은 실제 약물 조합 실험 전, 연구자들이 잠재적 시너지를 예측하는데 있어 중요한 역할을 할 수 있다. 그러나 예측 모델을 활용하여 발견되지 않은 항암 약물 조합의 잠재적 시너지를 예측한 데이터베이스들은 거의 없다. 본 연구에서는 예측모델을 기반으로 밝혀지지 않은 항암 약물 조합의 잠재적 시너지를 예측하여 DPS를 구축하였다. 또한, DPS에 있는 데이터를 검색할 수 있는 웹 기반의 도구 FPS를 구축하였다. FPS는 약물과 세포주 키워드만을 이용하여 항암 약물 조합의 잠재적 시너지를 검색 할 수 있어, 도구 사용에 대한 허들이 낮고 직관적이다. 그러므로 다양한 분야의 연구자들이 접근하기 용이하여 항암 약물 조합 발견에 있어 다양한 관점 및 방향을 줄 수 있을 것으로 기대한다.

넷째, 신약 개발 측면에 대한 기여를 기대한다. 새로운 약물이 발견되거나, 약물의 개수가 증가됨에 따라 약물 조합의 수는 급격하게 증가한다. 현실적으로 모든 약물 조합의 사례를 테스트하는 것은 수많은 시간

과 비용이 소요된다. 본 연구에서 개발한 방법을 통해 새로운 항암 약물 조합의 시너지 효과 가능성을 추출함으로써 항암 약물 조합 실험에 있어 시간과 비용을 줄일 수 있다. 이를 통해 *in-vivo* 또는 *in-vitro* 실험 전 연구자들이 새로운 항암 약물 조합의 후보를 선정하는데 있어 기여 할 수 있을 것으로 기대한다.

다섯째, 본 연구의 결과물을 웹에 공개함으로써 의생명 연구 분야 발전에 대한 기여를 기대한다. 본 연구의 소스 코드 및 약물과 세포주 특질은 https://github.com/YongsunShim/anticancer-drug-combination-prediction-using-documents-based-features.git에서 확인 가능하며, FPS는 http://bike-fps.snu.ac.kr에서 이용 가능하다. 본 연구의 방향과 결과물이 국내・외 연구자들의 학술적 목적과 발전에 기여할 수 있기를 기대한다.

Ⅵ. 연구의 제한점 및 제언

본 연구의 과정과 결론을 통해 다음과 같이 제한점과 새로운 연구 방향을 제안하고자 한다.

첫째, 본 연구에서는 PubMed에서 제공하는 의생명 문헌의 초록을 사용하여 약물과 세포주 특질을 추출하였다. 초록에는 의생명 문헌의 핵심 정보가 포함되어 있으나, 문헌 전체의 정보가 포함되어 있지는 않다. 특질 추출에 있어 문헌에 있는 전체적인 내용을 사용하면 더 좋은 특질을 추출할 수 있을 것으로 기대할 수 있으나, 출판사나 저자에 따라 문헌의 구조가 다르고 형식이 달라 전체 문헌을 처리하는데 있어 기술적 한계가 존재한다. 또한, 의생명 문헌의 전체 내용을 제공하는 데이터베이스는 PubMed Central[5], bioRxiv[6] 등이 있으나, PubMed에 비하여 수가 현저히 부족한 한계가 있다. 문헌 전체를 사용할 수 있는 기술적 한계 문제와 데이터 수의 부족 문제가 극복된다면 특질 추출에 있어 더 많은 정보를 포함 할 수 있을 것이라 판단된다.

둘째, 본 연구에서는 NCI-ALMANAC 데이터베이스와 DrugComb 데이터베이스를 기반으로 예측 모델을 구축하고 비교하였다. 그 결과, 데이터의 수가 약 4배 많은 DrugComb 데이터베이스를 기반으로 구축한 예측 모델의 성능이 더 우수하였다. 이를 통해, 약물 조합 관련 데이터의 수가 증가한다면 예측 모델의 성능이 증가할 것이고, 이를 통해 *in-silico* 기반 항암 약물 조합 시너지 예측 모델의 역할이 더욱 증대될 수 있을 것이라 판단된다.

셋째, 본 연구에서는 2가지 약물의 조합으로 이루어진 데이터를 기반

5) https://www.ncbi.nlm.nih.gov/pmc/
6) https://www.biorxiv.org/

으로 예측 모델을 구축하였다. 현재 공개된 항암 약물 조합 관련 데이터베이스는 2가지 약물의 조합으로 이루어져있다. 추후 항암 약물 조합 연구가 발전하고, 데이터가 증대된다면 3가지 이상의 약물 조합의 시너지를 예측하는 연구로 발전 할 수 있을 것으로 판단된다.

참 고 문 헌

[1] Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: a cancer journal for clinicians, 69(1), 7-34.

[2] Tonkens, R. (2005). An overview of the drug development process. Physician executive, 31(3), 48.

[3] Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., & Lu, Z. (2016). A survey of current trends in computational drug repositioning. Briefings in bioinformatics, 17(1), 2-12.

[4] Sun, W., Sanderson, P. E., & Zheng, W. (2016). Drug combination therapy increases successful drug repositioning. Drug discovery today, 21(7), 1189-1195.

[5] Cheng, F., Kovács, I. A., & Barabási, A. L. (2019). Network-based prediction of drug combinations. Nature communications, 10(1), 1-11.

[6] Breitinger, H. G. (2012). Drug synergy - mechanisms and methods of analysis. Toxicity and Drug Testing, 143-166.

[7] Crystal, A. S., Shaw, A. T., Sequist, L. V., Friboulet, L., Niederst, M. J., Lockerman, E. L., Frias, R. L., Gainor, J. F., Amzallag, A., Greninger, P., Dana Lee1, A. K., Gomez-Caraballo, M., Elamine, L., Howe, E., Hur, W., Lifshits, E., Robinson, H. E., Katayama, R., Faber, A. C., Awad, M. M., Ramaswamy, S., Mino-Kenudson, M., Iafrate, A. J., Benes, C.H., & Engelman, J. A. (2014). Patient-derived models of acquired resistance can identify effective drug combinations for cancer. Science 346, 1480 - 1486.

[8] Yu, Z., Zhao, G., Xie, G., Zhao, L., Chen, Y., Yu, H., Zhang, Z., Li, C., & Li, Y. (2015). Metformin and temozolomide act synergistically to inhibit growth of glioma cells and glioma stem cells in vitro and in vivo. Oncotarget, 6(32), 32930.

[9] Ekins, S., Mestres, J., & Testa, B. (2007). In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. British journal of pharmacology, 152(1), 9-20.

[10] Azad, P., Navimipour, N. J., Rahmani, A. M., & Sharifi, A. (2020). The role of structured and unstructured data managing mechanisms in the Internet of things. Cluster computing, 23(2), 1185-1198.

[11] Wu, L., Wen, Y., Leng, D., Zhang, Q., Dai, C., Wang, Z., Liu, Z., Yan, B., Zhang, Y., Wang, J., He, S., & Bo, X. (2022). Machine learning methods, databases and tools for drug combination prediction. Briefings in bioinformatics, 23(1), bbab355.

[12] Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems, 39(1), 43-62.

[13] Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., & Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics, 34(9), 1538-1546.

[14] Xia, F., Shukla, M., Brettin, T., Garcia-Cardona, C., Cohn, J., Allen, J. E., Maslov, S., Holbeck, S. L., Doroshow, J. H., Evrard, Y.A., Stahlberg, E. A., & Stevens, R. L. (2018). Predicting tumor cell line response to drug pairs with deep learning. BMC bioinformatics,

19(18), 71-79.

[15] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[16] Kim, Y., Zheng, S., Tang, J., Jim Zheng, W., Li, Z., & Jiang, X. (2021). Anticancer drug synergy prediction in understudied tissues using transfer learning. Journal of the American Medical Informatics Association, 28(1), 42-51.

[17] Zhang, T., Zhang, L., Payne, P. R., & Li, F. (2021). Synergistic drug combination prediction by integrating multiomics data in deep learning models. In Translational bioinformatics for therapeutic development (pp. 223-238). Humana, New York, NY.

[18] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[19] Janizek, J. D., Celik, S., & Lee, S. I. (2018). Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine. BioRxiv, 331769.

[20] Celebi, R., Bear Don't Walk, O., Movva, R., Alpsoy, S., & Dumontier, M. (2019). In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data. Scientific reports, 9(1), 1-10.

[21] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.

[22] Jeon, M., Kim, S., Park, S., Lee, H., & Kang, J. (2018). In

silico drug combination discovery for personalized cancer therapy. BMC systems biology, 12(2), 59-67.

[23] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5-6), 352-359.

[24] Li, J., Huo, Y., Wu, X., Liu, E., Zeng, Z., Tian, Z., Fan, K., Stover, D., Cheng, L., & Li, L. (2020). Essentiality and transcriptome-enriched pathway scores predict drug-combination synergy. Biology, 9(9), 278.

[25] Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D., Collins, J. M., & Doroshow, J. H. (2017). The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic ActivityNCI ALMANAC of Approved Cancer Drug Combinations. Cancer research, 77(13), 3564-3576.

[26] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic acids research, 49(D1), D1388-D1395.

[27] Bairoch, A. (2018). The cellosaurus, a cell-line knowledge resource. Journal of biomolecular techniques: JBT, 29(2), 25.

[28] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[29] Fan, K., Cheng, L., & Li, L. (2021). Artificial intelligence and machine learning methods in predicting anti-cancer drug combination effects. Briefings in bioinformatics, 22(6), bbab271.

[30] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 36(suppl_1), D901-D906.

[31] Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, J., Jafari, M., Tanoli, M., Pessia, A., & Tang, J. (2019). DrugComb: an integrative cancer drug combination data portal. Nucleic acids research, 47(W1), W43-W51.

[32] Yadav, B., Wennerberg, K., Aittokallio, T., & Tang, J. (2015). Searching for drug synergy in complex dose-response landscapes using an interaction potency model. Computational and structural biotechnology journal, 13, 504-513.

[33] Bliss, C. I. (1939). The toxicity of poisons applied jointly 1. Annals of applied biology, 26(3), 585-615.

[34] Berenbaum, M. C. (1989). What is synergy?. Pharmacological reviews, 41(2), 93-141.

[35] Loewe, S. (1953). The problem of synergism and antagonism of combined drugs. Arzneimittelforschung, 3, 285-290.

[36] Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research,

40(D1), D1100-D1107.

[37] Jung, J., Joe, H., Ha, K., Lim, J. M., & Kim, H. G. (2021). Biomedical Entity Explorer: A Web Server for Biomedical Entity Exploration. Journal of Computational Biology, 28(6), 619-628.

Abstract

A study on the development of a method for predicting the synergistic effects of anti-cancer drug combinations using biomedical literature-based feature extraction

Yongsun Shim
Department of Dental Science
The Graduate School
Seoul National University

To reduce drug side effects and enhance their therapeutic effect compared with single drugs, drug combination research, combining two or more drugs, is highly important. Conducting *in-vivo* and *in-vitro* experiments on a vast number of drug combinations incurs astronomical time and cost. To reduce the number of combinations, researchers classify whether drug combinations are synergistic through *in-silico* methods. Since unstructured data, such as biomedical documents, include experimental types, methods, and results, it can be beneficial extracting features from documents to

predict anti-cancer drug combination synergy. However, few studies predict anti-cancer drug combination synergy using document-extracted features. In addition, there are few databases that predict the potential synergy of undiscovered anti-cancer drug combinations by using the anti-cancer drug combination synergy prediction model.

This study presents a novel approach for anti-cancer drug combination synergy prediction using document-based feature extraction. The use of proposed features yielded higher results compared to the majority of published studies. Based on these results, a DPS (Database of Potential Synergy) was constructed by predicting the potential synergy of an undiscovered anti-cancer drug combination, and a web application FPS (Finder of Potential Synergy) that can search the DPS was created. FPS is available at http://bike-fps.snu.ac.kr.

Using the process and results in this study, researchers can save time and cost on new anti-cancer drug combination discoveries. Additionally, since the proposed feature extraction method does not require structuring of unstructured data, new data can be immediately applied without any data scalability issues.

keywords : Biomedical literature based feature extraction, Anti-cancer drug combination synergy prediction, Anti-cancer drug combination database expansion
Student Number : 2015-22084