앙상블모델 기반 원발암, 순환암, 전이암 분류 유전자 마커 예측

Identification of gene markers classifying primary, circulating, and metastasis cancers based on ensemble machine-learning models

김유림†, 민지인†, 정진명*

수원대학교 데이터과학부

*† Equal contribution , * Corresponding author*

1. Introduction

연구에 따르면, 암에 의한 죽음 중 약 90% 정도가 **암의 전이**(metastasis)에 의한 것으로 보고되고 있다. 암 전이 과정는 크게 1) 원래 암이 혈관에 침투하는 invasion, 2) 암세포가 혈관 내에서 순환하는 circulation, 3) 혈관에서 새로운 기관으로 침투하여 정착하는 colonization으로 나눌 수 있고, 그 과정에 따라 암의 종류를 **원발암, 순환암, 전이암**으로 나눌 수 있다 (그림1).

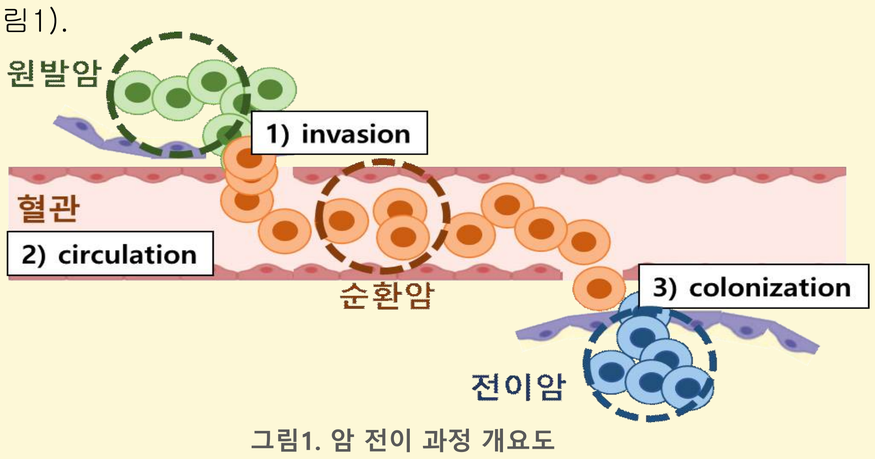

그림1. 암 전이 과정 개요도

본 연구에서는 가장 많이 발생하는 암 중에 하나인 **유방암**(breast cancer)을 분석 대상으로 정하였다. 그리고 유방암의 **원발암, 순환암, 전이암의 유전자 발현 데이터**를 수집한 후, 수집한 **발현데이터를 feature로 하여 세 그룹 (원발암, 순환암, 전이암)을 분류하는 세 개의 앙상블 머신러닝 모델**(Random Forest, XGBoost, AdaBoost)을 구축하였다. 각 모델의 성능을 확인하고, **성능에 중요한 역할을 한 feature들을 선정**하여, **전이의 유전자 마커**로 규명하였다. 규명된 마커들은 기존 **문헌 정보 분석**을 통하여 검증해보았다.

2. Methods

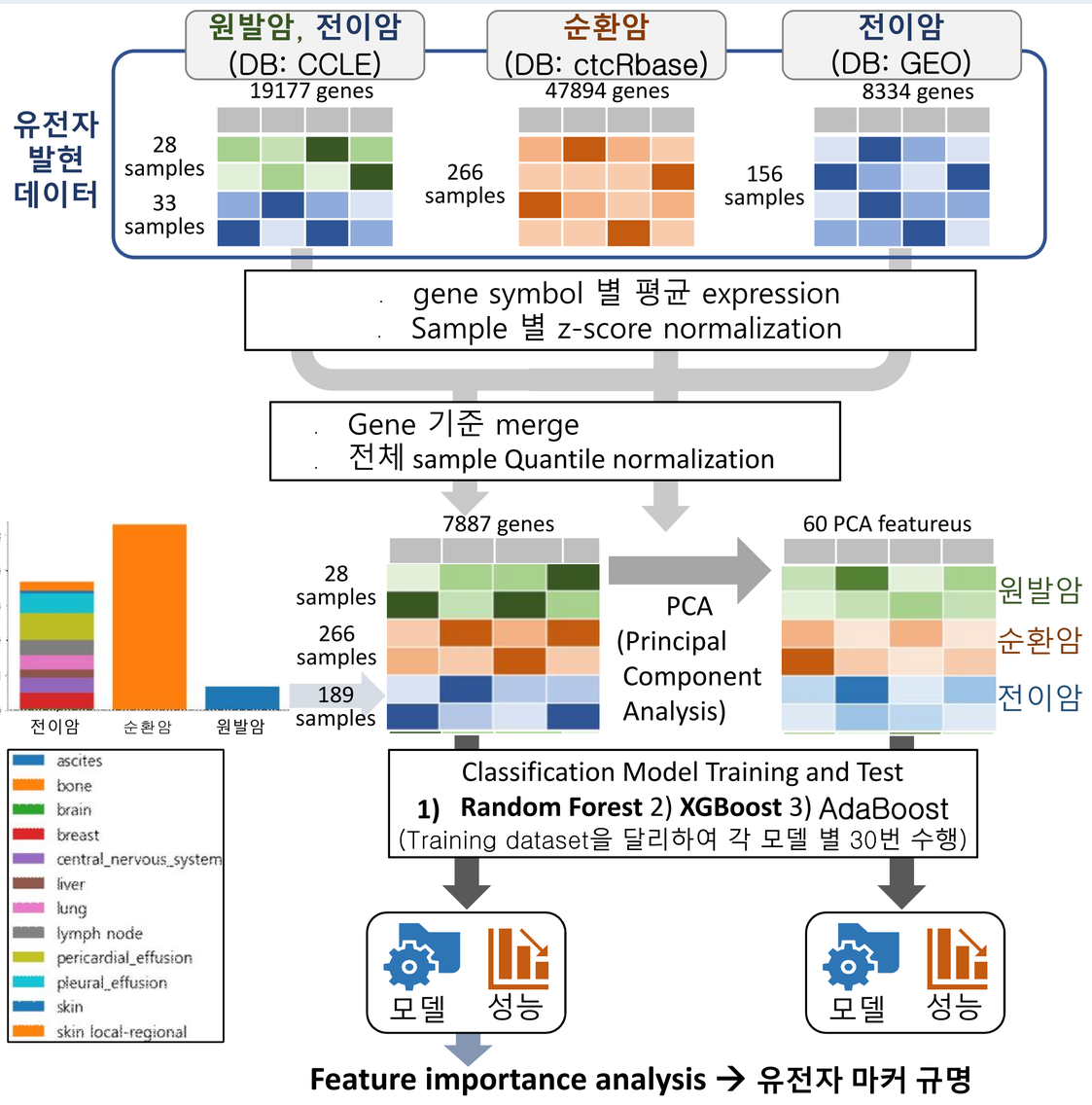

그림2. Method Overview

- CCLE는 원발암과 전이암이 포함되있기 때문에 전처리를 통해 분류
- GEO는 gene 이름 대신 ID_REF로 구분 되어있어 ID_REF를 gene으로 매핑
- 각 데이터에 존재하는 공통 gene으로 Groupby를 진행해 평균 유전자 발현량 추출
- 각 데이터의 column을 gene으로 통일한 후 서로 다른 세 가지 데이터 merge 수행
- 3개의 data set이 유전체 발현 측정 환경이 달라 분포가 다른 것을 확인
  → 변인이 가진 값의 크기에 따라 설명 가능한 분산량이 왜곡될 수 있기 때문에 Quantile normalization으로 값 보정을 진행한 데이터를 생성
- original data는 feature의 개수가 많아 PCA를 통해 차원 축소를 진행 → PCA data를 생성 (explained variance: 70%)
- PCA data와 original data에 **Random Forest (RF), XGBoost (XGB), AdaBoost (ADB)**에 학습 (Training dataset을 달리하여 각 모델 별 30번 수행)

3. Results

- ***머신러닝 모델 성능***

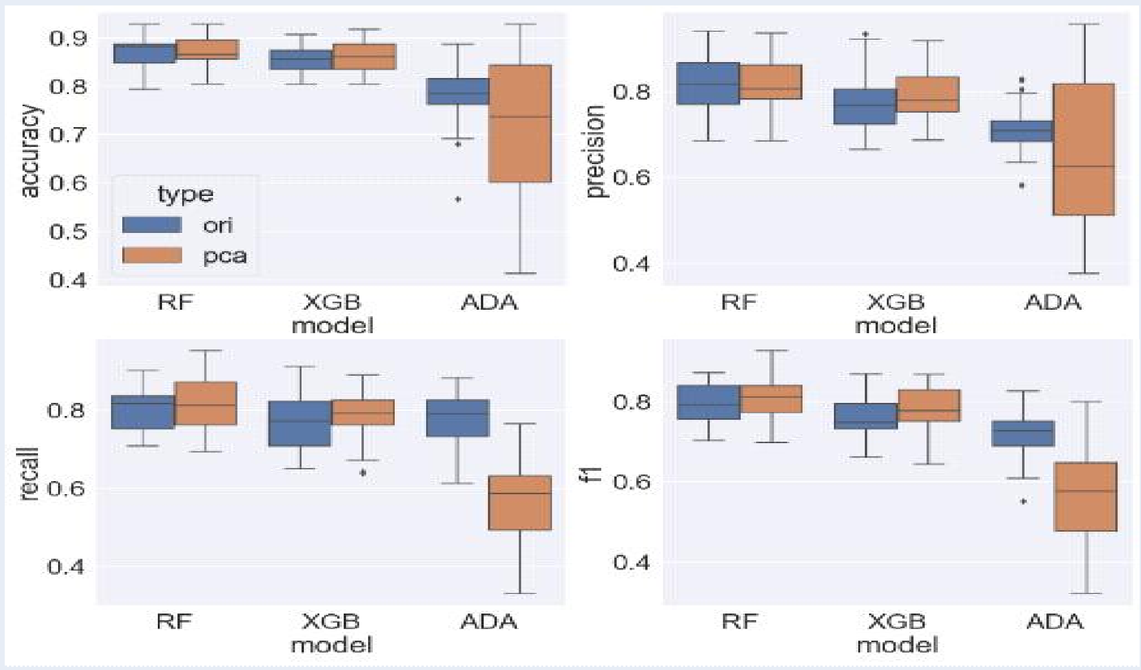

그림3. model 성능 비교

- 세 모델 성능 → accuracy: 0.8~0.9, precision:0.6~0.85, recall: 0.55~0.85, f1-score: 0.55~0.85
- Original data와 PCA data를 학습한 결과 accuracy, recall, precision, F-score 평가에서 대부분 **PCA data를 학습한 모델**의 성능이 더 높은 것을 확인
- **XGB 모델의 성능이 가장 높은 것을 확인** → XGB 모델을 이용하여 원발암, 순환암, 전이암을 분류하는 유전자 마커 예측 진행

- ***유전자 마커***

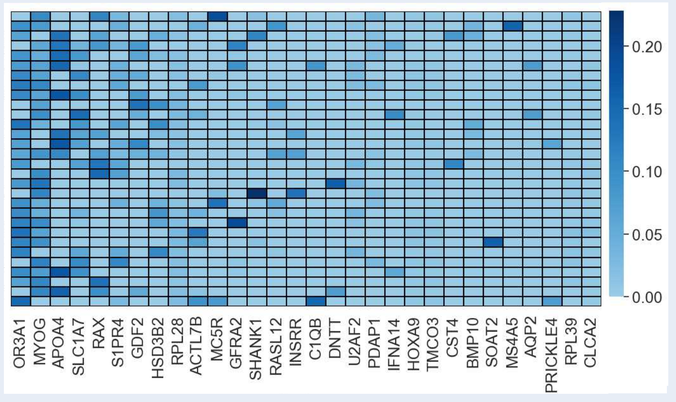

그림4. Feature importance heatmap of XGB

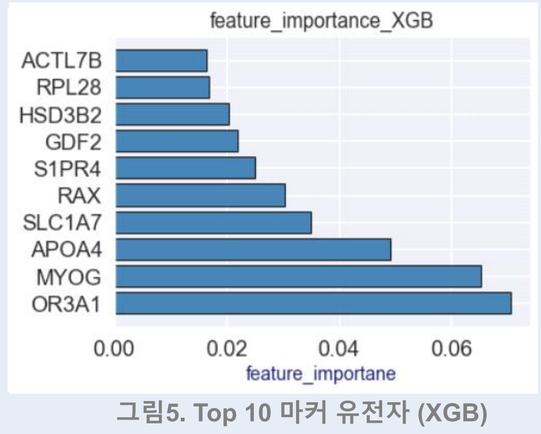

그림5. Top 10 마커 유전자 (XGB)

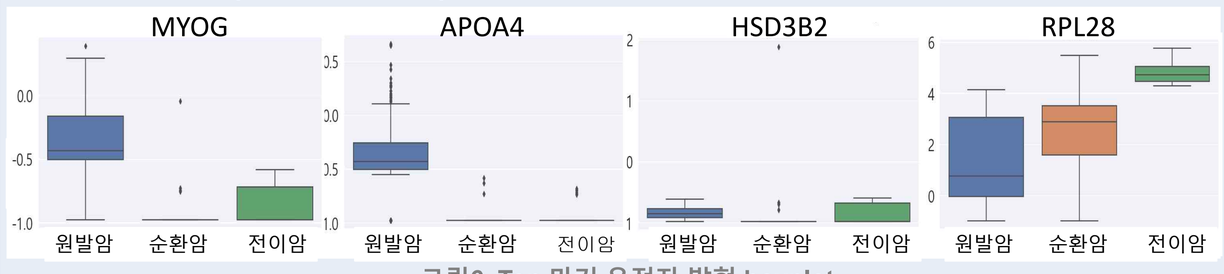

그림6. Top 마커 유전자 발현 boxplot

- XGB 모델 성능에 중요한 역할을 한 gene을 알기 위해 feature importance 상위 30개를 추출해 Heatmap으로 확인
- feature importance Top 10개를 추출해 평균 feature importance 값 확인
- Top10 유전자 중 4개(MYOG, APOA4, HSD3B2, RPL28)유전자의 발현을 그룹(원발암, 순환암, 전이암)별로 확인
- RPL28 유전자 경우 전이암, 순환암, 원발암 순으로 발현량이 많고, MYOG 유전자의 경우 원발암, 전이암, 순환암 순으로 발현량이 많음

- ***Literature evidence***

- MYOG 유전자는 FOXO 단백질의 target 유전자로 breast cancer의 진행과 전이에 큰 역할을 한다는 연구 결과를 확인함 (2022, IJMS, Dilmac *et al.*)
- RPL28 유전자가 breast cancer의 치료에 중요한 marker라는 연구 결과를 확인함 (2019, American Journal of Biomedical Research, Houngue *et al.*)

4. Conclusion

이번 연구에서는 암의 전이 과정의 각 단계 별 암(원발암, 순환암, 전이암)의 발현 데이터를 처리하고 합친 후, 각 단계를 분류하는 머신러닝 모델(RF, XGB, ADB)을 학습을 시켜보았다. feature(gene)의 수가 많아 차원축소(PCA)를 통해 original data보다 더 높은 성능을 얻었다. 적절하게 차원을 줄인다면 보다 좋은 성능을 얻을 수 있을 것이다. 가장 성능이 좋은 모델인 XGB를 이용해 TOP feature importance에 해당하는 유전자들을 선정하고 (MYOG, RPL29 등), 그들의 그룹별 발현량을 boxplot으로 시각화 하고 문헌에서도 관련 연구들을 확인해보았다.