Research Plan

연구 제목:**전혈구 및 PCR 데이터와 다중 오믹스 정보 통합 기반의 딥러닝을 활용한 주요 암종 조기 진단 스크리닝 모델 개발 및 검증**

Background

전혈구 데이터(CBC)와 PCR을 활용하여 정확도 높고 비용 부담 적은 **암 조기 진단 스크리닝 검사 개발**

최종 목표 : 전혈구 데이터와 PCR 데이터를 기반으로 oral cancer, pancreatic cancer, gastric cancer, Esophageal and lung cancer, kidney cancer 암종에 대해서 AUC 0.9 이상의 조기 스크리닝 모델 개발 및 CBC 및 Omics 데이터와 문진 데이터 간의 주요 상관성 파악.

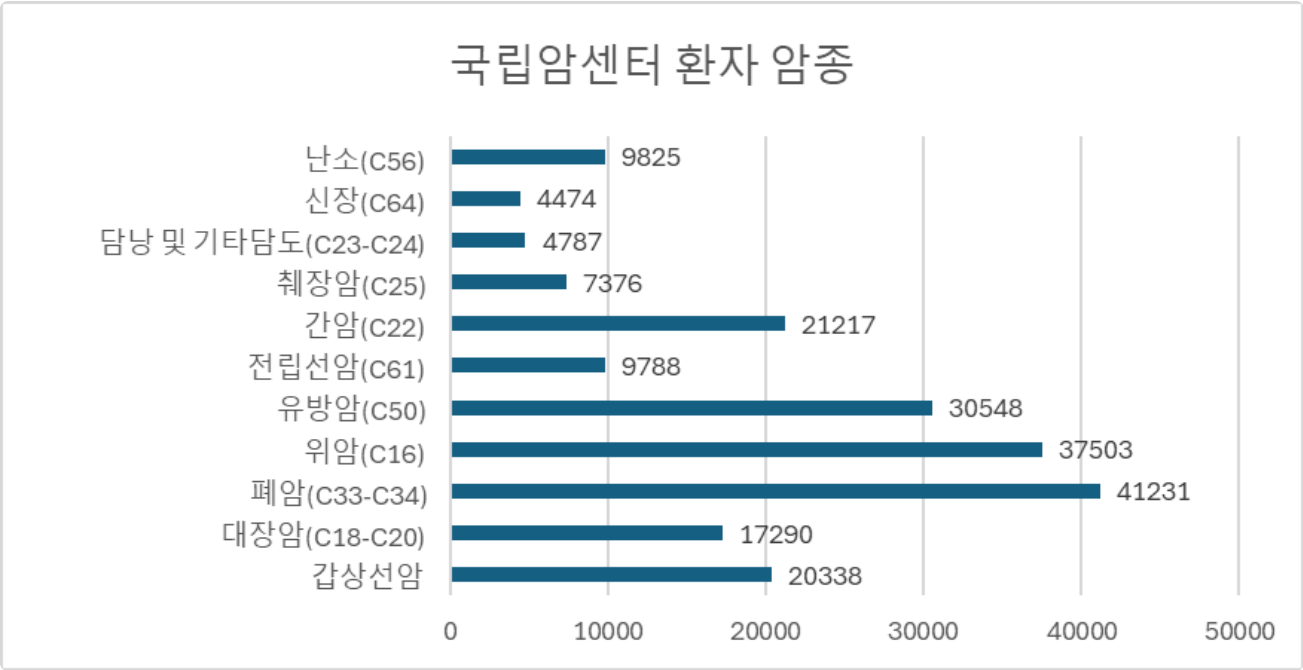
(1) 국립암센터의 대규모 암 환자 코호트 데이터와 (2) 암역학과에서 보유한 특정 암종 환자군(CBC, 문진, 다중 오믹스 데이터 보유) 및 대조군(CBC, 다중 오믹스 데이터 보유) 데이터를 통합적으로 활용하여, 정확도가 높고 비용 효율적인 다중 암종 조기 진단 스크리닝 모델을 개발하고자 합니다.

특히, Deep Metric Learning(DML)과 전이학습(Transfer Learning)을 적용하여 데이터의 이질성을 극복하고, 복잡한 생체 정보 속에서 암 특이적 패턴을 효과적으로 학습하여 기존 모델의 성능을 뛰어넘는 것을 목표로 합니다.

최종적으로는 CBC와 PCR 데이터를 기반으로 하는 실용적인 스크리닝 모델 구축의 가능성을 제시하고자 합니다.

연구 대상자

- 국립암센터 내원한 난소암, 갑상선 암, 폐암, 위암, 유방암, 전립선 암, 간암, 췌장암, 담낭암, 신장암, 대장암으로 진단 받은 만 20세 이상 환자 158808명.



- 기존 위에서 CBC+문진 데이터를 활용하였을 때 난소암의 경우 AUC가 0.7로 가장 높고 나머지는 그 이하의 성능.
- 암역학과에서 보유한 1315명과 1041명의 대조군
    - 이 중 동일한 암종은 oral cancer, pancreatic cancer, gastric cancer, Esophageal and lung cancer, kidney cancer로 총 5종으로 972명의 CBC+문진 및 Omics 와 대조군 1041명의 CBC 및 Omics 데이터 활용

연구 목표

- **세부 목표 1:** 암역학과 환자-대조군 코호트(5대 암종 환자 972명 및 대조군 1,041명)의 CBC, 문진(환자군), 다중 오믹스 데이터를 활용하여, 각 암종에 대한 고성능 예측 모델(AUC > 기존 연구 결과) 개발 및 검증.
- **세부 목표 2:**, (1) 일반적인 암 관련 특징 또는 암종 간 구분 특징을 학습하는 딥러닝 기반 표현 학습(representation learning, DML 임베딩) 모델 사전학습(pre-training) 및 (2) 이 사전학습 모델을 세부 목표 1에서 개발된 암종별 특화 모델에 전이학습(transfer learning)하여 성능을 향상시키는 방안 연구.
- **세부 목표 3:** 세부 목표 1과 2의 연구 결과를 바탕으로, 실제 스크리닝 환경에 적용 가능한 CBC 및 PCR 데이터 기반의 비용 효율적인 다중 암종 조기 진단 모델 프로토타입 개발 및 성능 평가.

연구 흐름도

Stage 1

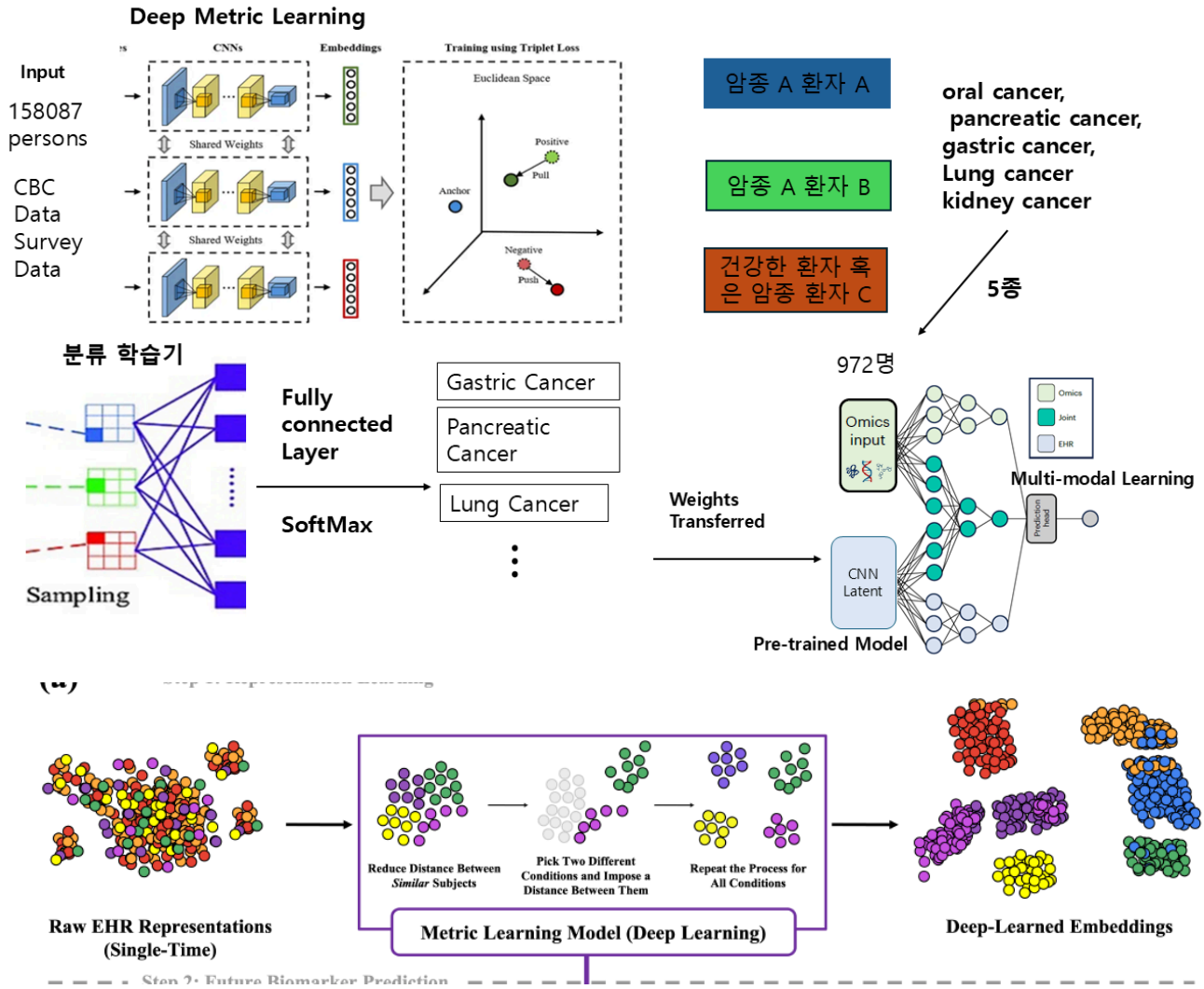
- Deep Metric Learning (DML) 네트워크와 Triplet Loss를 통한 암종별 embedding 공간 학습
    - Triplet Loss
        - 앵커: 특정 암종 A 환자 데이터

- 포지티브: 동일한 암종 A의 다른 환자
- 네거티브: 다른 암종 B 환자의 데이터 혹은 Health Control

Output: 각 환자 데이터에 대한 저차원 임베딩 벡터. 이 공간에서는 같은 암종끼리는 모여 있고, 다른 암종과는 떨어져 있도록 학습

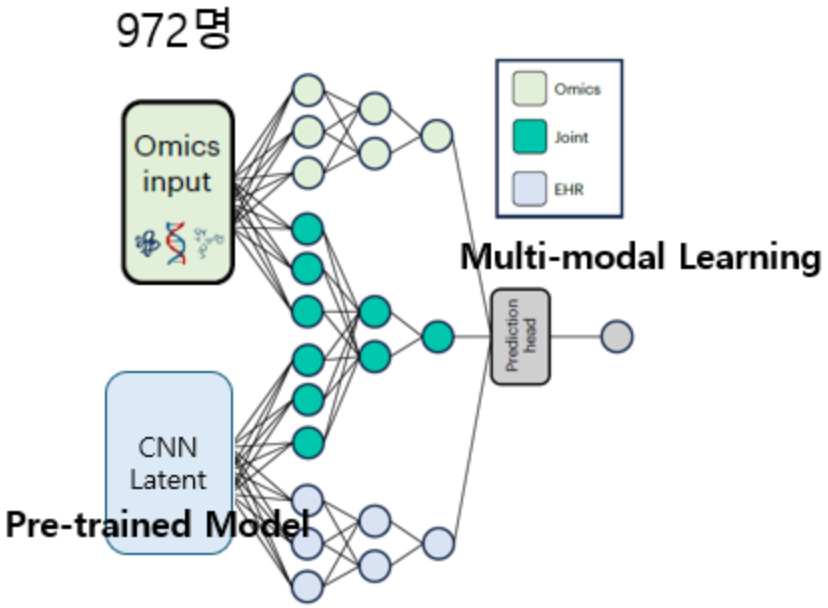
Stage 2

- 암센터 환자 2,356명의 CBC+문진 데이터를 Stage 1에서 학습된 DML 모델에 통과시켜 각 환자별 임베딩 벡터 추출
- Omics 데이터 별도의 전처리 및 특징 공학/추출 과정을 거쳐 유의미한 특징 벡터
- **1) Stage 1 임베딩 벡터**와 **(2) Omics 특징 벡터**를 결합(concatenate)
- **다중 모달리티 학습 (Multi-modal Learning):**
  - 초기 결합(Early fusion), 중간 결합(Intermediate fusion), 후기 결합(Late fusion) 등 다양한 방식으로 CBC/문진, Omics, (만약 여기서 통합한다면) PCR 데이터를 결합하여 모델 학습.



(Lifestyle-Informed Personalized Blood Biomarker Prediction via Novel Representation

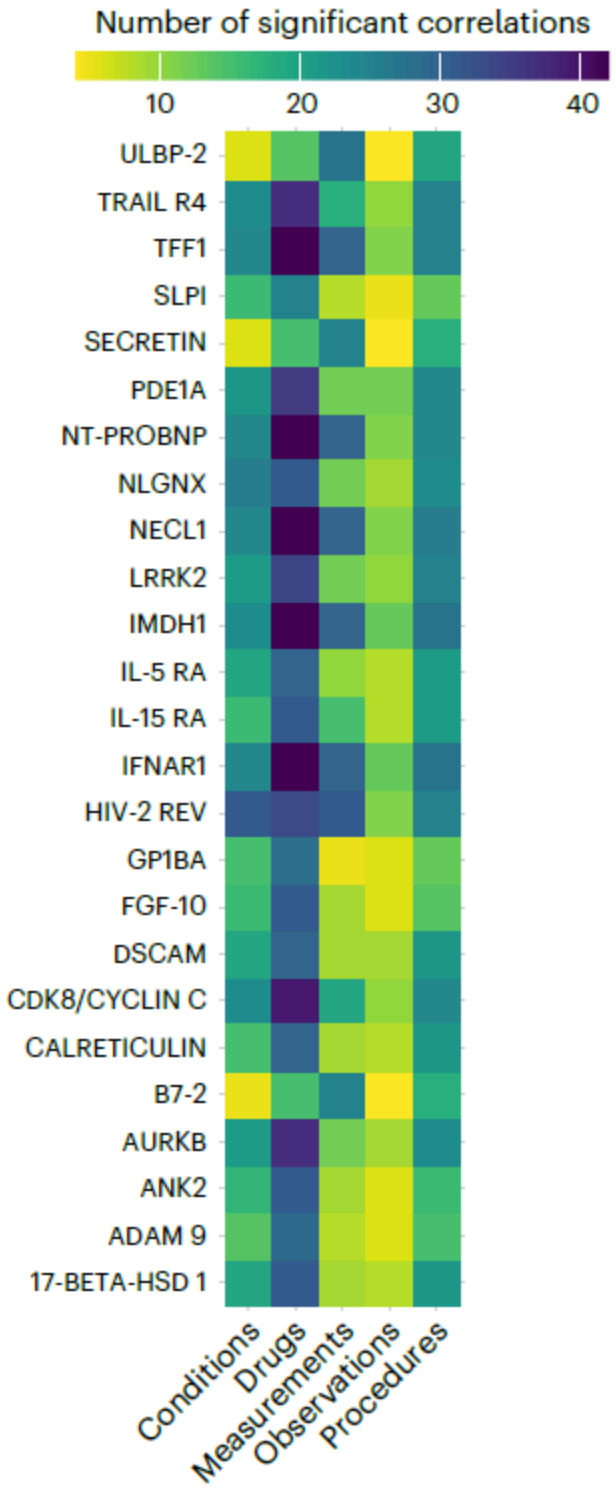
Learning)



모델 최종 Output: 각 샘플 별 특정 암종일 확률

모델 성능 지표 : ROC Curve, AUC, 민감도, 특이도, 정밀도, F1-score

Stage 3

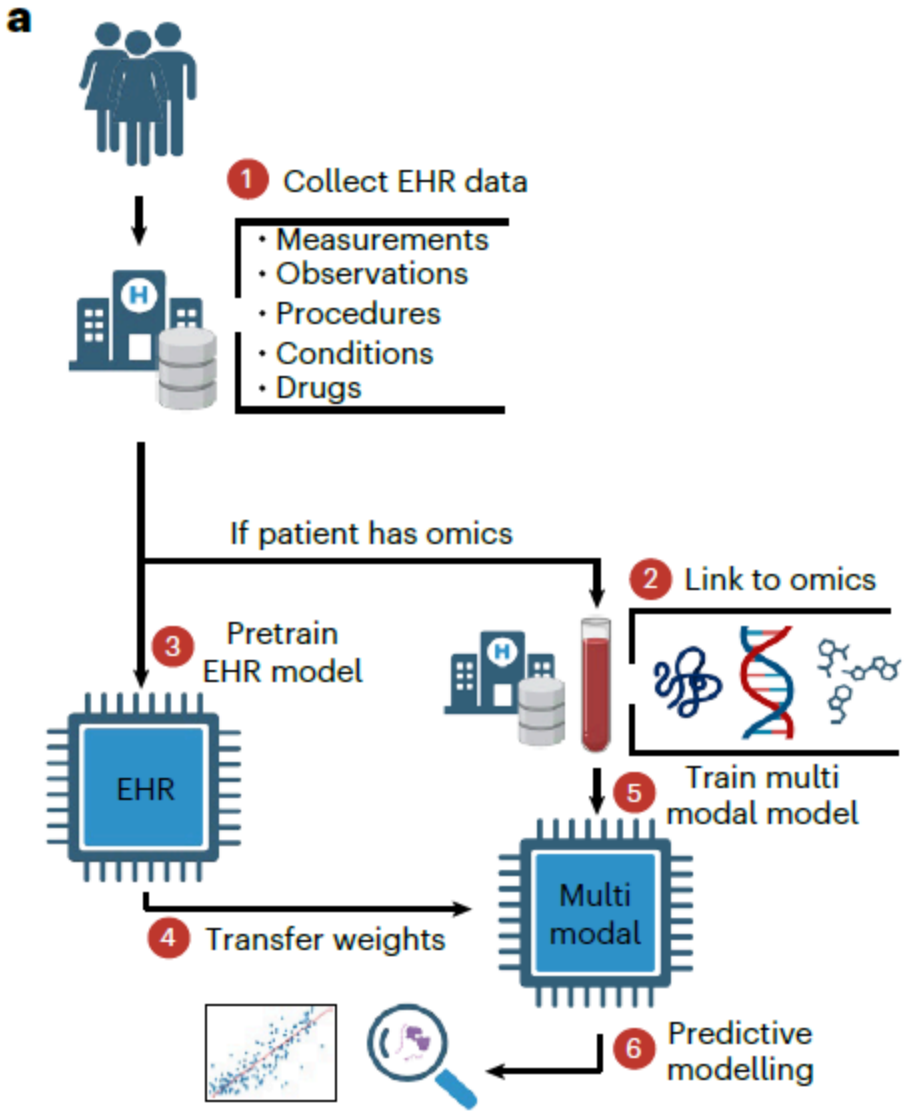


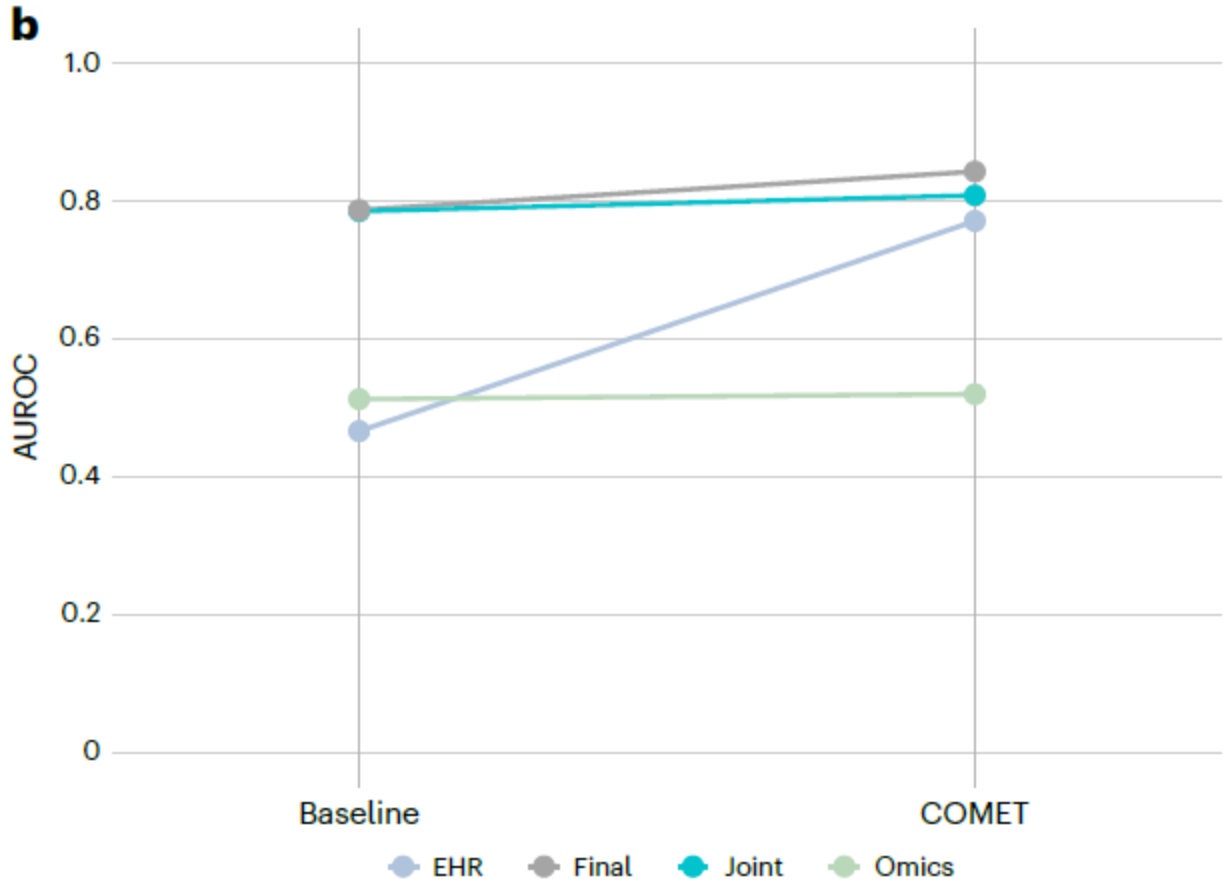
위와 같은 방식으로 문진 데이터와 주요 microbiota,metabolites 혹은 CBC marker 간의 상관성 파악

Reference:

A machine learning approach to leveraging electronic health records for enhanced omics analysis. ***Nature Machine Intelligence*** (2025)

- COMET effectively integrates EHR data from large cohorts with omics data from smaller sub-cohorts.
- The model's pretraining on EHR data enhances its predictive capabilities for labor onset and cancer mortality.
- In cancer prognosis, COMET predicted three-year mortality in 36,901 cancer patients from the UK Biobank.
- The model demonstrated superior performance (**AUROC = 0.842**) compared to baseline models in cancer mortality prediction.





Lifestyle-Informed Personalized Blood Biomarker Prediction via Novel Representation Learning

***IEEE-EMBS International Conference on Biomedical and Health Informatics**(BHI))*(2024)