Incorporating Hierarchical Information into Multiple Instance Learning for Patient Phenotype Prediction with scRNA-seq Data

https://www.biorxiv.org/content/10.1101/2025.02.10.637389v1.full.pdf

ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS LAB
명지대학교 융합소프트웨어학부 김민

* Our implementation is available at https://github.com/minhchaudo/hier-mil.

Experiments and Results

- DATA

1. *Cardio: 심근병증 환자 데이터를 활용한 다중 분류 (DCM, HCM, 정상; 세 가지 분류)*
2. **ICB: 면역항암제 치료 반응 여부를 예측하는 이진 분류**
3. COVID: COVID-19 감염 여부를 예측하는 이진 분류

https://singlecell.broadinstitute.org/single_cell/study/SCP1289/impaired-local-intrinsic-immunity-to-sars-cov-2-infection-in-severe-covid-19#study-download

20210701_NasalSwab_MetaData.txt

Columns (27개)

컬럼명	설명
NAME	세포 또는 샘플의 고유 이름 (예: 바코드, cell ID 등)
donor_id	세포를 제공한 환자(또는 샘플 제공자)의 ID
Peak_Respiratory_Support_WHO_Score	WHO 기준에 따른 최대 호흡기 보조 수준 (0: 없음 ~ 7: 인공호흡기)
Bloody_Swab	채취한 면봉에 혈액이 있었는지 여부 (`Yes` / `No`)
Percent_Mitochondrial	세포 내 미토콘드리아 유전자 발현 비율 (품질 지표로 사용됨)
SARSCoV2_PCR_Status	환자 단위의 PCR 감염 여부 (`pos` / `neg`)
SARSCoV2_PCR_Status_and_WHO_Score	PCR 감염 여부 + WHO 점수를 합친 값 (예: `pos_5`)
Cohort_Disease_WHO_Score	질병 유형 + WHO 점수를 합친 값 (예: `COVID_WHO_5`)
biosample_id	동일한 샘플(코 swab 등)의 고유 ID
SingleCell_SARSCoV2_RNA_Status	개별 세포 내 SARS-CoV-2 RNA 유무 (`pos` / `neg`)
SARSCoV2_Unspliced_TRS_Total_Corrected	splicing되지 않은 바이러스 RNA 총량
SARSCoV2_Spliced_TRS_Total_Corrected	splicing된 바이러스 RNA 총량
SARSCoV2_NegativeStrand_Total_Corrected	음성가닥 바이러스 RNA 양 (복제 과정 중 생성됨)
SARSCoV2_PositiveStrand_Total_Corrected	양성가닥 바이러스 RNA 양
SARSCoV2_Total_Corrected	전체 바이러스 RNA 총합 (위 항목들 합산?)

species	종(species) ID (예: `NCBITaxon_9606` = 인간)
species__ontology_label	종 이름 (예: `Homo sapiens`)
sex	환자의 생물학적 성별 (`male` / `female`)
disease	질병 코드 (예: `PATO_0000461` = 정상?)
disease__ontology_label	질병 이름 (`normal`, `COVID-19`, 등)
organ	조직 ID (예: `UBERON_0001728`)
organ__ontology_label	조직 이름 (`nasopharynx` = 비인두)
library_preparation_protocol	실험에서 사용된 library prep 프로토콜의 ID
library_preparation_protocol__ontology_label	프로토콜 이름 (`Seq-Well`, `10x`, 등)
age	환자의 나이대 (`50-59`, `30-39` 등 범주형)
Coarse_Cell_Annotations	세포의 대분류 cell type (예: `Ciliated Cells`, `Goblet Cells`)
Detailed_Cell_Annotations	세포의 세분류 cell type
covid_label	너가 새로 만든 COVID 감염 여부 (PCR 기준): `1` = 감염, `0` = 비감염

ICB 데이터셋 :

label : 면역항암제 반응 결과 (Combined_outcome) 이진 분류.
"Favourable" → 1(양성), "Unfavourable" → 0(음성)

- AnnData object with n_obs × n_vars = 9292 × 824
- **총 세포 수 : 9,292개** * (9292, 197) DataFrame
- 유전자 정보 : 824개 *(824,) Index

```
(9292, 824)
   HAVCR2  CTLA4  PDCD1  IDO1  CXCL10     CXCL9   HLA-DRA     STAT1  IFNG  CD3E  ...  EZH2  TP53      CALR  STAG2  CEBPA  CUX1  U2AF1  EP300  PHF6  KRAS
0     0.0    0.0    0.0   0.0     0.0  0.000000  0.000000  0.000000   0.0   0.0  ...   0.0   0.0  0.000000    0.0    0.0   0.0    0.0    0.0   0.0   0.0
1     0.0    0.0    0.0   0.0     0.0  0.000000  0.000000  0.000000   0.0   0.0  ...   0.0   0.0  0.000000    0.0    0.0   0.0    0.0    0.0   0.0   0.0
2     0.0    0.0    0.0   0.0     0.0  0.000000  3.111702  3.111702   0.0   0.0  ...   0.0   0.0  0.000000    0.0    0.0   0.0    0.0    0.0   0.0   0.0
3     0.0    0.0    0.0   0.0     0.0  2.892357  0.000000  2.892357   0.0   0.0  ...   0.0   0.0  2.892357    0.0    0.0   0.0    0.0    0.0   0.0   0.0
4     0.0    0.0    0.0   0.0     0.0  0.000000  0.000000  0.000000   0.0   0.0  ...   0.0   0.0  3.905236    0.0    0.0   0.0    0.0    0.0   0.0   0.0

[5 rows x 824 columns]
```

- **세포 정보** * (9292, 197) DataFrame
  - 환자 수 : 57명
  - cell_type_annotation : 23개

```
                                        sample_id                        cell_id orig.ident  nCount_RNA  ...  pANN_0.25_0.21_35 Study_name_cancer label cell_type_annotation
Row.names                                                                                                ...
Breast_previous_Breast_BIOKEY_10_Pre_AAAGCAAAGC...  BIOKEY_10  BIOKEY_10_Pre_AAAGCAAAGCGTCTAT-1     BIOKEY         407  ...               NaN       Bassez:TNBC     1     Mesangial cells
Breast_previous_Breast_BIOKEY_10_Pre_AAATGCCGTT...  BIOKEY_10  BIOKEY_10_Pre_AAATGCCGTTAGGGTG-1     BIOKEY         411  ...               NaN       Bassez:TNBC     1                 HSC
Breast_previous_Breast_BIOKEY_10_Pre_AACTCTTGTA...  BIOKEY_10  BIOKEY_10_Pre_AACTCTTGTAACGTTC-1     BIOKEY         466  ...               NaN       Bassez:TNBC     1     Mesangial cells
Breast_previous_Breast_BIOKEY_10_Pre_AACTGGTAGT...  BIOKEY_10  BIOKEY_10_Pre_AACTGGTAGTACATGA-1     BIOKEY         587  ...               NaN       Bassez:TNBC     1           B-cells
Breast_previous_Breast_BIOKEY_10_Pre_AACTTTCAGG...  BIOKEY_10  BIOKEY_10_Pre_AACTTTCAGGATCGCA-1     BIOKEY         411  ...               NaN       Bassez:TNBC     1         Adipocytes

[5 rows x 197 columns]
```

```
def get_df(adata, patient_id_key="patient", label_key="label",
cell_type_annot_key="cell_type_annotation", no_label=False):
    try:
        df = pd.DataFrame(adata.X.toarray())
    except:
        df = pd.DataFrame(adata.X)
    df.index = adata.obs.index

    if not no_label:
        df[["patient","cell_type_annotation","label"]] =
adata.obs[[patient_id_key,cell_type_annot_key, label_key]]
    else:
        df[["patient","cell_type_annotation"]] =
adata.obs[[patient_id_key,cell_type_annot_key]]
        df["label"] = -1
    return df
```

run.py

Task 번호	실험 이름	목적
0	train_and_tune	**모델 학습 + 하이퍼파라미터 튜닝**
1	predict_and_save	학습된 모델로 예측값 저장
2	repeated_k_fold	**10번 반복된 k-fold 교차검증**으로 모델 성능 평균화
3	vary_train_size	학습 데이터 크기를 줄여가며 성능 비교 (0.25, 0.5, 0.75)
4	vary_cell_count	셀 수를 줄였을 때 성능 변화 분석
5	randomize_cell_annot	셀 타입 정보를 랜덤으로 섞어서 모델 의존도 확인
6	get_p_val_cell_type	permutation test로 중요한 세포 타입 찾기 (biological insight용)

utils.py : get_data(df, all_ct, samples, …)

- Xs : 각 환자에 대한 모든 세포의 유전자 발현 행렬

```
sample_df = df[df["patient"]==sample]
        x = sample_df.iloc[:,:df.shape[-1]-3].to_numpy()
```

- ys : 환자 샘플들의 진단 결과 (label: 0 or 1)

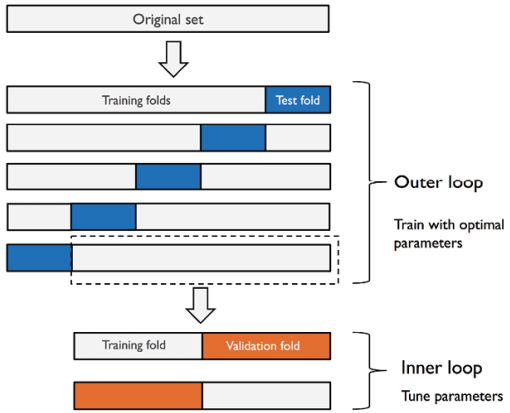
```
ys = torch.tensor(samples["label"].to_list(), dtype = …)
```

- batches : 환자 × 셀타입 조합에 해당하는 그룹 번호

```
batch = [(idx * len(all_ct) + ct_dict[ct]) for … ]
```

  - idx: 환자 index
  - ct_dict[ct]: cell_type을 숫자로 인코딩한 값

train.py : repeated_k_fold(df, meta, args)



1. 반복 루프 (args.n_repeats번 반복)
   - 각 반복마다 다른 seed를 설정해서 K-fold를 새롭게 섞음 **(Outer Cross Validation, Outer CV)**

```
for i in range(args.n_repeats):
        skf = StratifiedKFold(args.n_folds, shuffle=True, random_state=i)
```

2. 각 fold에 대해: 튜닝 → 학습 → 평가
   - 각 fold마다 Optuna를 이용해 하이퍼파라미터 튜닝 수행 **(inner CV)**

```
    for train_idx, test_idx in skf.split(samples, samples["label"]):
        ...
        study = optuna.create_study(direction="maximize", sampler=sampler)
        study.optimize(objective_wrapper(…), n_trials=args.n_tune_trials)
        best_params = study.best_params
```

* Optuna : 자동 하이퍼파라미터 튜닝
* `samples = df[["patient", "label"]].drop_duplicates()`

model.py

1. multi-layer perceptron (MLP)

```
X = self.lin(X)
```

* layers.extend([torch.nn.Linear(curr_in, curr_out), torch.nn.ReLU(), torch.nn.Dropout(dropout)])

2. 셀 수준 (attn1)

```
if self.attn1:
    w_c = softmax(self.w_c(X).squeeze(), batch)
    X = global_add_pool(X * w_c.unsqueeze(dim=-1), batch, size=ct_size)
```

* batches : 환자 × 셀타입
* w_c : batch 별 attention weight 계산
* 세포 표현 X에 각 attention weight를 곱 적용, batch 단위로 pooling

model.py

3. 셀타입 수준 (attn2)

```
X = self.lin2(X)
if self.attn2:
    w_ct = torch.nn.Softmax(dim=1)(self.w_ct(X))
    X = torch.sum(X * w_ct, dim=1)
```

* w_ct : 각 환자 내 셀 타입 중요도 계산
* 각 환자마다 하나의 벡터로 요약

4. 예측

```
X = self.lin_out(X)
```

icb.py (Preprocessing)

```
# meta는 cell 단위의 메타데이터
meta = df.iloc[:,:195]
# X는 cell × gene 유전자 발현 행렬
X = df.iloc[:,195:]

# 세포 타입 주석 추가
# ct = pd.read_csv("singler_icb_pre.csv", index_col=0)
ct = pd.read_csv("singler_icb.csv", index_col=0)    # 작성
meta["cell_type_annotation"] = ct["pruned.labels"]
meta = meta[meta["cell_type_annotation"].notna()]
```

COVID 데이터셋 (20210701_NasalSwab_MetaData.txt)

- **총 샘플 수: 32,588개**
- **총 환자 수(donor_id): 58명**
  - **SARSCoV2_PCR_Status**
    - pos 18073
    - neg 14515
  - **disease__ontology_label**
    - COVID-19 18073
    - normal 8874
    - respiratory failure 3335
    - long COVID-19 2306
  - Coarse_Cell_Annotations (18개)