1차 보고서

Source-Free Domain Adaptation을 활용한 전자 의료 기록(EHR) 기반

질병 예측 모델의 병원 간 데이터 공유 한계 극복 및 성능 개선

Overcoming Inter-Hospital Data Sharing Limitations and

Improving Electronic Health Record-Based Disease Prediction Models

with Source-Free Domain Adaptation

**그로쓰 17팀 그린데빌**

박선우(2171015)

유승혜(2190050)

전지윤(2270068)

1. Project-Summary (과제 요약)

1.1. 문제 정의

* **목표 사용자**본 연구의 목표 사용자는 병원 간 데이터 공유 및 활용에 제약이 있는 의료분야 종사자들(병원 관계자, 연구자 등)이다.
* **문제점**

1. **데이터 이질성**  
   병원 간 환자 분포, 측정 기기, 진단 기준 등은 서로 달라서 동일한 질병에 대해서도 데이터 특성이 상이할 수 있다. 이로 인해 모델의 성능에 큰 격차가 발생하게 된다. 예를 들어, A 병원에서 훈련된 모델이 B 병원에서는 성능이 크게 저하될 수 있다.
2. **병원 간 데이터 공유 제약**  
   의료 데이터는 환자 개인정보 보호와 관련된 법적 규제(예: HIPAA[[1]](#footnote-0), GDPR[[2]](#footnote-1), PIPA[[3]](#footnote-2) 등)가 엄격히 적용된다. 이러한 규제로 인해 병원 간 데이터 공유가 어려워, 데이터 공유 없이도 성능을 유지할 수 있는 방법이 필요하다.

* **필요기능**

1. 다양한 병원 간 데이터의 차이를 극복할 수 있는 도메인 적응 기법이 필요하다.
2. 데이터 공유 없이도 기존 모델의 성능을 유지할 수 있는 방법이 필요하다.
3. 의료 텍스트 데이터를 활용한 도메인 적응 연구가 부족하기 때문에, 이에 대한 새로운 접근법이 요구된다.

1.2. 기존 연구의 한계

의료 데이터는 대부분 이미지 또는 생체 신호 데이터를 활용한 연구들이 많다. 특히, Tent 기법은 이미지 데이터에 맞춰 개발된 모델이며, 해당 기법은 모델 불확실성을 측정하고 이를 관리하는 방법에 대해 명확한 방법론을 제시하고 있어 의료 이미지 데이터를 활용한 도메인 적응 연구는 많이 진행되었다. 그러나 기존의 텍스트 데이터를 활용한 도메인 적응 연구에 비해 표준화된 연구 환경 및 데이터 접근성의 어려움이 존재하는 의료 텍스트 데이터를 대상으로 한 연구는 드문 편이다. 따라서, 의료 텍스트 데이터를 활용한 도메인 적응 연구는 아직 초기 단계에 있으며, 이와 관련된 연구를 진행하는 것이 중요한 도전 과제가 된다.

1.3. 제안 내용

* **연구 문제점 및 해결책**  
  현재 데이터 공유 제약 문제와 도메인 적응 필요성에 대한 해결책으로, 데이터 공유 없이도 도메인 적응 기법을 활용하여 공유받은 기존 모델의 성능을 유지할 수 있는 방법을 제시하고자 한다. 구체적으로, 기존 모델을 다른 병원에서 사용하기 위해 데이터 공유 없이도 동일한 성능을 유지할 수 있도록 도메인 적응 기법을 적용하는 것이다.
* **기존 연구와의 차별성**

1. **이미지 데이터가 아닌 텍스트 데이터 기반 도메인 적응**  
   기존의 도메인 적응 연구는 *Model Adaptation: Unsupervised Domain Adaptation without Source Data[[4]](#footnote-3)*와 같은 연구처럼 이미지 데이터나 생체 신호 데이터를 중심으로 연구가 진행되었다. 그러나, 텍스트 데이터는 이미지나 생체 신호와는 다른 특성을 가지고 있기 때문에, 텍스트 데이터에 맞는 새로운 도메인 적응 방법을 제안한다.
2. **의료 텍스트 데이터에 대한 연구 기반 마련**  
   의료 분야에서의 도메인 적응 연구는 *A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records[[5]](#footnote-4) 처럼* 이미지 또는 생체 신호 데이터에 한정되어 있으며, 의료 텍스트 데이터에 대한 연구는 거의 없다. 의사의 진단 결과, 환자 기록, 병원에서 발생하는 의료 정보 등을 포함한 텍스트 데이터를 활용한 도메인 적응 연구가 부족한 상황을 본 연구에서 극복하려고 한다.

* **본 연구의 제안**이 문제를 해결하기 위해 본 연구에서는 Med-BERT라는 기존의 질병 예측 모델을 기반으로 하고, 해당 모델을 다른 병원에서 사용할 수 있도록 Tent 기법을 적용한다. Tent 기법은 원래 이미지 데이터에 맞춰 개발된 기법이지만, 이를 의료 텍스트 데이터에 맞게 수정하여 사용할것이다. 구체적으로, 기존의 BatchNorm parameter를 LayerNorm parameter로 교체하고, 텍스트 데이터에 적합한 방식으로 Tent 기법을 적용하여 예측 성능을 개선하려고 한다.

1.4. 기대 효과 및 의의

* **기대효과**

1. **도메인 차이 극복**

병원, 지역, 진료 환경 등에 따라 의료 텍스트 데이터의 표현 방식이 달라 발생하는 성능 저하 문제를 완화할 수 있을 것이다.

1. **데이터 공유 제약 해소**

원천 데이터를 직접 공유하지 않고도 테스트 시점에서 모델이 자체적으로 적응 가능하므로 데이터 공유 제약 문제를 우회할 수 있으며, 이를 통해 협력 연구나 다기관 연구에서 발생하는 실질적 제약을 줄이고 의료 AI 활용 가능성을 높일 수 있을 것이다.

1. **텍스트 기반 도메인 적응 연구 확장**

임상 기록, 전자의료기록, 환자 보고서 등 실제 활용 가치가 높은 텍스트 자원에서 도메인 적응 가능성을 검증할 수 있을 것이다.

* **연구 의의**

1. **연구 공백 해소**

의료 텍스트 데이터를 활용한 도메인 적응 연구는 상대적으로 드물기 때문에 본 연구가 해당 영역의 공백을 채우는 역할을 할 것이다.

1. **실질적 임상 적용 가능성**

다양한 병원과 환경에서 작성된 의료 기록에 대해 일관된 성능을 발휘할 수 있는 모델 개발은 실제 임상 활용에 필수적이다. 따라서 기관별 데이터 편차를 줄임으로써 의료 ai의 신뢰성을 높이고 임상 의사결정 지원 도구로서의 실용성을 증대할 수 있을 것이다.

1. **도메인 적응 방법론의 일반화 가능성 제시**

본 연구에서 제시하는 방법은 의료 도메인에 국한되지 않고 법률 문서, 소셜 미디어 등 다양한 전문 텍스트 데이터 영역에도 확장 가능하다.

1.5. 주요 기능 리스트

**기능1: Test-Time Entropy Minimization(TENT) 기반 적응 기법 적용**

* 원래 이미지 데이터에 특화되어 제안된 TENT를 텍스트 기반 의료 데이터에 맞게 변형하여 적용한다.
* 사전 학습된 모델(Med-BERT)을 그대로 사용하고, 추가 학습 과정 없이 테스트 시점에서 바로 도메인 적응 가능하도록 수정한다.
* 엔트로피를 최소화하여 source - target 간 도메인 차이를 줄이고 더 신뢰도 높은 질병 예측 결과 도출한다.
* 모델 전체를 학습하지 않고 일부 파라미터만 업데이트함으로써 효율성과 안정성을 확보한다.
* 기존 TENT는 BatchNorm 파라미터만 업데이트하지만, 텍스트 데이터 특성을 고려해 BatchNorm 대신 LayerNorm 파라미터를 조정하도록 수정한다.

**기능2: 라벨에 의존하지 않는 적응 방식**

* 실제 임상 환경에서 라벨이 없는 데이터가 대부분이라는 현실을 반영한다.
* 라벨 없이도 적응 가능한 TENT 구조를 활용하여 실제 적용성을 확인한다.

**기능3: 실제 의료 데이터(eICU) 기반 검증**

* 미국 중환자실 전자의무기록(EMR) 데이터 활용한다.
* source 병원(7,059명 환자)과 target 병원(723명 환자)을 도메인으로 설정하여 분포 차이가 있는 실제 상황 모방한다.
* 텍스트로 기록된 환자 정보와 임상 기록을 활용하여 질병 예측 성능을 평가한다.

**기능4: 비교 모델 설정 및 성능 검증**

* 모델 1: 사전 학습된 Med-BERT 기반 질병 예측 baseline
* 모델 2: 모델 1에 TENT를 적용하여 도메인 적응 성능 보강
* 두 모델의 타겟 병원 데이터 예측 성능을 비교하여 TENT의 효과성을 검증한다.

**최종 목표**

* source 병원 데이터에서 학습된 baseline 모델 (Med-BERT) 대비 target 병원 데이터에서의 예측 AUC를 최소 0.05 ~ 0.1 이상 향상시킨다.
* 기존 baseline 모델 대비 성능 개선을 입증함으로써 의료 텍스트 기반 도메인 적응 연구의 가능성과 임상적 활용성을 제시한다.

1. Project-Design (과제 설계)

2.1. 요구사항 정의

1. **데이터 관련**

* 미국 eICU 데이터셋 활용: source 병원과 target 병원 선정한다.
* 병원 선정시 환자 수 및 지역적 분포 차이를 고려하여 도메인 간 이질성을 확보한다.
* 환자의 개인정보 보호를 준수한다.

1. **모델 관련**

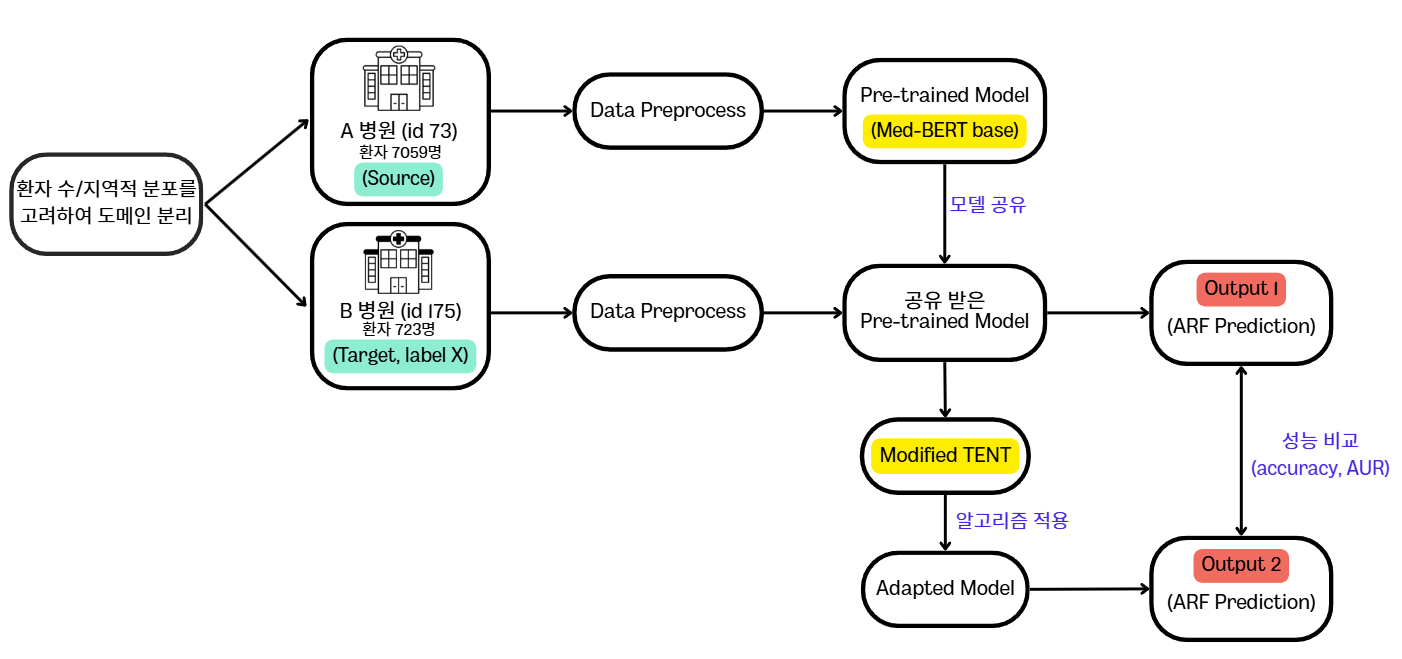
* 사전 학습된 Med-BERT를 질병 예측 baseline 모델로 사용한다.
* 테스트 시점에서 추가 학습 데이터 없이 바로 도메인 적응 가능하도록 한다.
* 기존 TENT 구조를 텍스트 데이터에 맞게 수정: BatchNorm 대신 LayerNorm 파라미터를 업데이트하도록 코드 수정한다.

1. **성능 평가 관련**

* source 병원 데이터를 활용한 baseline 모델 성능을 확인한다.
* target 병원 데이터에서 baseline 대비 TENT 적용 모델의 성능 개선을 검증한다.
* 도메인 차이로 인한 예측 성능 저하를 보완하는지 평가한다.

2.2. 전체시스템 구성

**[실험용 급성 호흡 부전 예측 모델 시스템 구성도]**



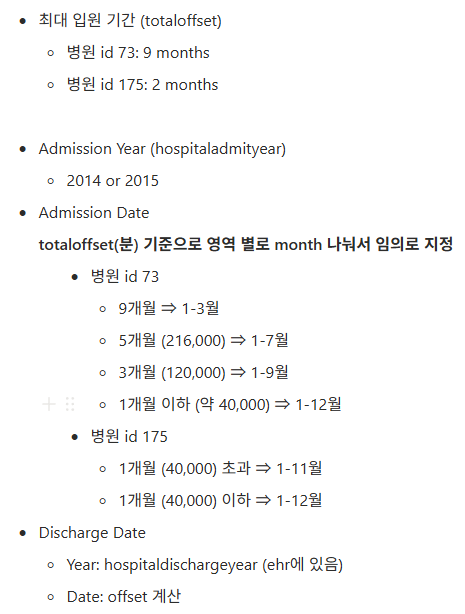
1. **도메인 분리**

사용 데이터셋인 eICU 데이터를 DB 내의 Hospital ID(병원 구별 용 ID)를 기준으로 하여 각 병원의 환자 수, 지리적 위치, 환자의 분포 등을 고려하여 2개의 실험 대상 도메인을 선정하였다. Source 도메인은 병원 ID 73의 병원(이하 73병원)으로 7059명의 환자 데이터를 포함하며, Target 도메인은 병원 ID 175의 병원(이하 175병원)으로 723명의 환자 데이터(source 도메인의 10%)를 포함한다.

1. **Data Preprocess(데이터 전처리)**

각 데이터 별로 다음과 같은 전처리 작업을 수행하였다:

* **Missing Data 처리**: 질병의 진단 코드인 ICD code가 입력되지 않은 경우 해당 데이터를 삭제한다.
* **환자 별 입원 순서 추정**: 시계열 데이터를 처리해야 하기 때문에 환자 별 입원 순서를 정렬해야 한다. 다만 eICU 데이터에는 입/퇴원 날짜 정보가 없기 때문에 입원 시 나이, 퇴원 년도, 퇴원 시 사망 여부 등을 통해 추정 가능한 데이터들만 정렬하며, 불가능한 경우 임의의 한 입원만 포함시킨다.
* **입/퇴원 날짜 입력**: 사용하는 Med-BERT 모델은 입/퇴원 날짜를 필요로 하지만, eICU 데이터는 날짜 정보는 없다. 대신 DB에 퇴원 년도와 총 입원 기간이 존재하기 때문에, 이 둘을 통해 환자 별 첫 입원 날짜는 임의로 지정하였으며, 이후의 입원은 랜덤 간격을 두고 날짜를 지정하였다(입원 간의 간격은 모델의 feature가 아님). 상세 날짜 입력 로직은 아래와 같다.



* **ICD code 통일**: ICD-9 code와 ICD-10 code가 혼용된, 혹은 모두 사용된 eICU 데이터를 ICD-9 code로 통일시키기 위해, 둘 다 사용된 경우 ICD-10 code는 삭제한다. 단, ICD-10 code만 사용된 진단의 경우, 직접 icd-9 code로 전환하는 것은 적절하지 않다고 판단하여 해당 경우에만 ICD-10 code를 유지시킨다.
* **Data Labeling**: 급성 호흡 부전 여부에 대한 라벨링을 진행. 급성 호흡 부전에 해당하는 ICD-9 code(518.51, 518.53, 518.81, 518.83, 518.84)를 진단받은 경우 1, 진단 받은 경험이 없는 경우 0으로 라벨링한다.

1. **Med-BERT Model Pre-training**

73병원의 환자 데이터를 통해 Med-BERT 모델을 학습시킨다. 앞선 2. 데이터 전처리 과정을 바탕으로 BERT 모델의 입력 데이터 형식에 맞는 구조로 변환한 후, Pre-training을 통해 BERT 모델이 환자 별 진단 시퀀스를 학습한다.

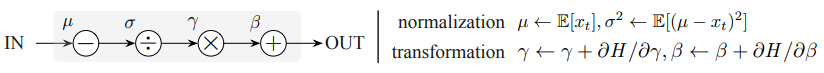
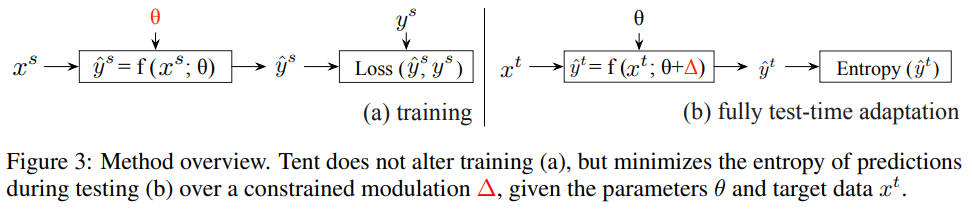
1. **Med-BERT Model Fine-tuning (Baseline 구축)**

Med-BERT로 생성한 임베딩을 입력받아 Logistic Regression 방법으로 73병원의 데이터를 바탕으로 환자의 급성 호흡 부전 발생 여부를 학습한 후, test 단계에서 이를 예측한다. 이 때, test는 73병원의 test dataset(test1)와 175병원의 전체 데이터(test2)를 대상으로 각각 시행하며, 이 때 175병원의 test AUC 값을 본 연구의 기준 성능으로 삼는다.

핵심 output: 73병원 test AUC, 175병원 test AUC(기준 성능)

1. **TENT Modify**

도메인 적응 핵심 솔루션인 TENT 알고리즘을 텍스트 데이터에 맞게 수정한다. TENT는 엔트로피가 낮은 예측일수록 더 정확하지만, 소스-타겟 차이에 따른 데이터 손상은 엔트로피를 증가(예측 불확실성 증가)시킨다는 문제를 엔트로피 최소화(Test-Time Entropy Minimization)를 통해 보정한다는 아이디어의 도메인 적응 기법임. 사전 학습된 모델을 사용하여 학습 과정 없이 테스트 시점에 바로 적용시킬 수 있으며, 라벨이 없는 상황에서 BatchNorm 파라미터만 업데이트하기에 일부 파라미터만으로 빠르고 안정적으로 적응 가능하다.



단, 기존 TENT는 이미지 데이터에 맞춰 고안된 알고리즘이기 때문에 eICU 텍스트 데이터에 맞게 조정할 필요가 있다. 따라서 본 연구에서는 기존 BatchNorm parameter를 대신하여 LayerNorm parameter를 업데이트하도록 수정하고, 텍스트 데이터에 적합한 방식으로 Tent 기법을 적용할 수 있도록 하고자 한다.

1. **Adapted Model Test (최종 실험)**

4에서 구축한 Baseline Model에 수정한 TENT를 적용한 Adapted Model를 구축하여 175병원의 전체 데이터(test2)로 test를 시행하여 AUC를 구하고, 4의 baseline 기준 성능과 비교하여 성능 개선을 증명한다.  
기대 및 목표 성능: AUC 0.05 이상 향상

* **주요 연구 현황**
  + **데이터셋 확보**

eICU 데이터셋 확보 및 초기 데이터 분석 실행하였다. 분석 결과 및 다기관 의료 데이터 특성을 반영하여 연구 실험 설계를 수립하였다.

output: 실험용 급성 호흡 부전 예측 모델 시스템 구성도

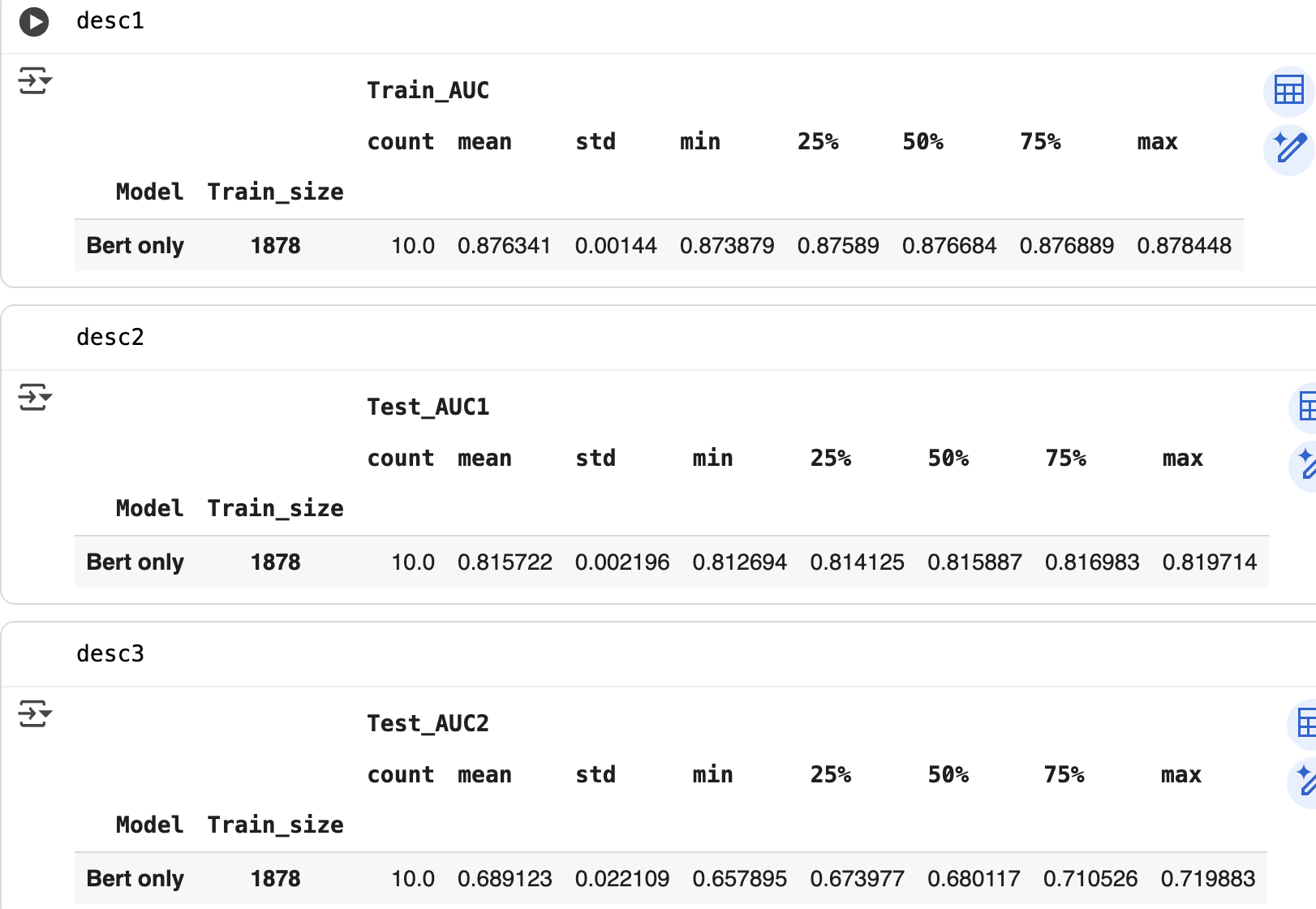
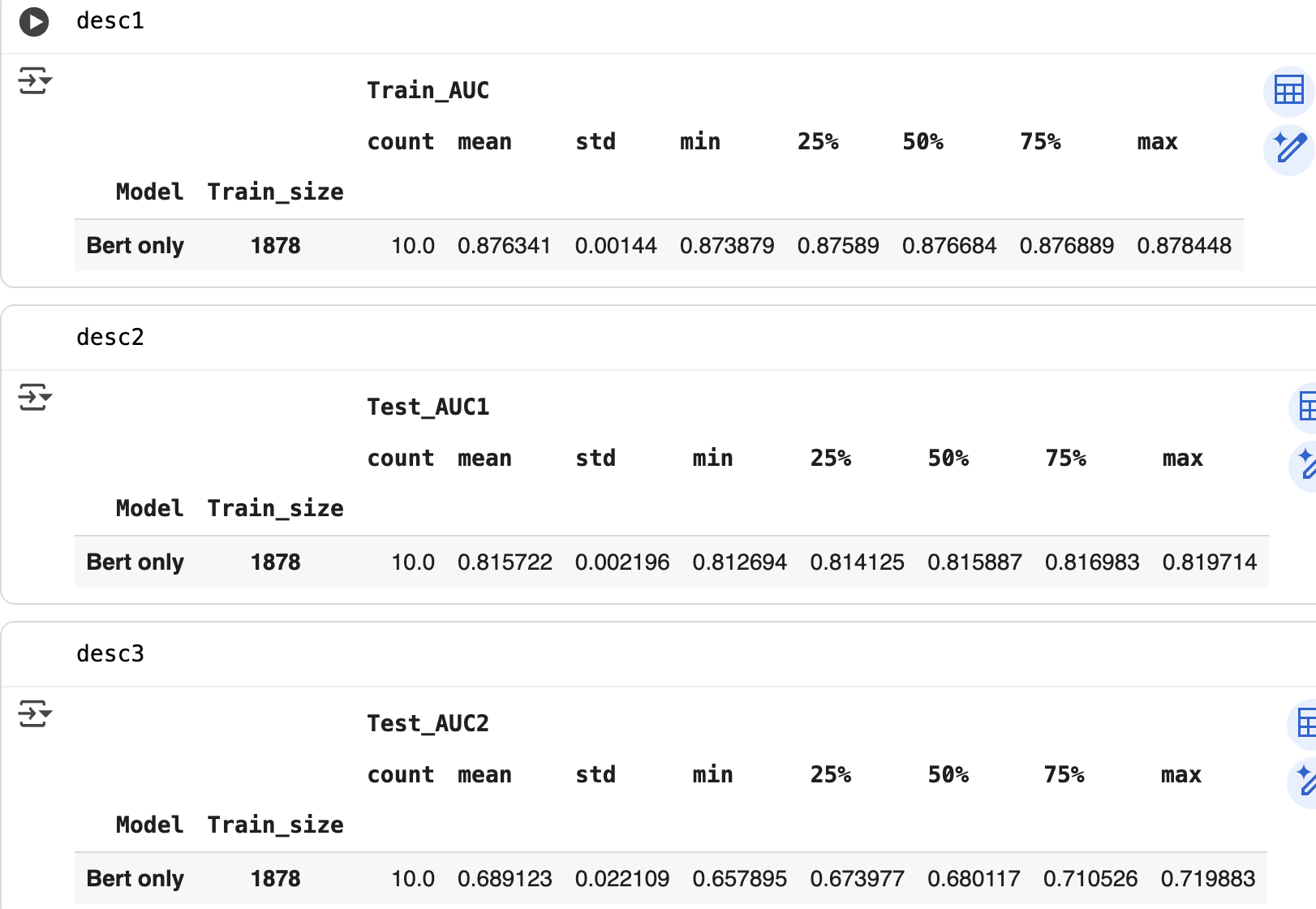
* + **데이터 전처리 시행**

데이터 전처리 파이프라인을 구축하여 주요 변수 및 환자 단위로 데이터를 정리하였다. 해당 파이프라인을 기반으로 실제 전처리를 진행하여 학습 용 데이터 정제 완료한 상태이다.

* + **Med-BERT 기반 Baseline Model 구현**

Med-BERT pretraining 및 fine-tuning 진행을 통해 Baseline 모델 구현을 완료하였다. 해당 Baseline 모델을 활용한 실험을 진행하여 도메인 적응 알고리즘 적용 전 기준 성능을 확인한 결과, 175병원의 AUC가 73병원의 AUC에 비해 약 11% 저조한 것을 확인하였다.

73병원 AUC(Test\_AUC1): 0.82, 175병원 AUC(Test\_AUC2): 0.69



* + **SFDA 알고리즘 적용 (진행중)**

TENT 기반 SFDA 알고리즘 수정 및 적용하여 baseline 대비 성능 비교를 진행하고 있다. 또한 Label smoothing 기법을 적용하여 불균형 데이터 환경에서의 모델 안정성을 개선하고자 시도하였다.

1. US, Health Insurance Portability and Accountability Act [↑](#footnote-ref-0)
2. EU, General Data Protection Regulation [↑](#footnote-ref-1)
3. KR, 한국 개인정보 보호법 [↑](#footnote-ref-2)
4. Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, Si Wu, https://arxiv.org/abs/2502.19316 [↑](#footnote-ref-3)
5. Egoitz Laparra *,* Aurelie Mascio *,* Sumithra Velupillai *,* Timothy Miller, https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0041-1726522 [↑](#footnote-ref-4)