

기말 프로젝트 제안서

2024191010 정호준

1. 주제

태아 질병 예측에서 임상적 스크리닝 (CTG/초음파/혈청)과 유전 기반 (cfDNA-NIPT) 중 무엇이 더 효과적일까?

2. 설명

저는 질병을 통제 가능한 수준으로 만드는 임상의가 되는 것을 목표로 하는 의대생입니다. 특히 예측을 통해 개입 시점을 앞당기는 일이 가장 큰 임상적 가치를 만든다고 믿습니다. 임상 현장에서는 동일한 질환이라도 예측·판단 방식이 사례마다 달라지고, 데이터 처리 기준도 일관되지 않은 경우가 많아 표준화된 데이터 기반 의사결정의 필요성을 자주 체감해 왔습니다.

이 프로젝트의 출발점은 태아기라는 특수한 시기에 있습니다. 성인과 달리 태아는 생활습관·환경 요인의 변동성이 상대적으로 작기 때문에 유전·발생학적 신호의 설명력이 더 클 수 있다는 가설을 세울 수 있습니다. 그렇다면 동일한 예측 문제에서 **임상 스크리닝(초음파·혈청·CTG)**과 **유전 기반(cfDNA-NIPT)** 방법 가운데 무엇이 더 일관적이고 재현 가능한 성능을 보이는지, 그리고 제가 지향하는 표준화된 데이터 처리 파이프라인에서도 그 우월성이 유지되는지 확인해 보고자 합니다.

이에 본 연구는

1. 공개 임상 데이터(CTG)로 객관적 분류 성능을 재현·보고하고,
2. 문헌 기반으로 cfDNA-NIPT 및 혈청 스크리닝의 질환별 성능을 동일 지표 체계로 구조화하며,
3. 질환 유형(염색체 수적 이상 vs. 생리·기능 사건)에 따라 어떤 접근이 더 효과적인지 통일된 데이터 처리 기준에서 비교·해석하는 것을 목표로 합니다.

궁극적으로 예측 파이프라인의 표준화가 임상 선택지(유전 vs. 임상 스크리닝)의 상대적 가치를 어떻게 바꾸는지를 검증하고, 태아기 질병 예측에서 개입 전략의 체계화에 기여하고자 합니다.

3. 가설

1. 염색체 수적 이상 — 정확도 가설: T21/T18/T13 에 대해서는 cfDNA-NIPT 의 민감도, 특이도, PPV/NPV 가 전통적 혈청/NT 기반 선별보다 우수할 것이다. 이때 양성 시 확진 검사를 전제로 평가한다.

2. 구조적 이상 — 적합도 가설: 심장·중추신경계 등 구조적 기형의 1 차 탐지는 중기 정밀 초음파가 가장 효과적이며, NIPT 만으로는 대체 불가하다(필요 시 NIPT 양성·음성과 무관하게 초음파는 보편 시행).

3. 생리적 이벤트 — 분류력 가설: 저산소증/태아곤란 같은 분만 전후 생리 상태의 실시간 분류·의사결정 지원에는 CTG 기반 신호 특징/ML 모델이 더 설명력이 높다

최종적으로는, 위 가설들에서 얻은 결과들을 기반으로 통합 파이프라인을 어떻게 최적화하여 구축해나갈 것인지를 논의해보고자 합니다. (ex. 보편적 초음파(+혈청) → 고위험 분류 시 NIPT → 양성/고위험 시 확진으로 가는 계층형(Contingent) 경로)

4. 데이터 소개 및 전처리, 생성 계획

(1) 임상 데이터로 실제 분류 성능을 재현·보고하고

(2) 유전 기반 선별의 질환별(특히 T21/T18/T13) 성능을 동일 지표 체계로 요약·비교

하는 것이 주 목적입니다.

-임상 스크리닝 데이터셋: Fetal Health Classification (Cardiotocography, Kaggle)

(<https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification>)

사용 데이터는 Kaggle 의 Fetal Health Classification CSV이며, 21 개 feature + 1 개 target 으로 구성되어 있는 2126 개 행의 데이터입니다. target 인 fetal_health 는 1=정상, 2=의심, 3=병적의 3 분류이며, 정상 비율이 높고 병적이 상대적으로 적은 클래스 불균형 특성을 보입니다. 입력 특징은 태아 심박·자궁수축 신호로부터 표준화된 수치 요약치로, (a) 기저 심박수 및 사건 빈도 (b) 감속 유형 빈도 (c) 단·장기 변이성으로 묶을 수 있습니다. 변수들은 임상 해석 가능성이 높고, CTG 해석 가이드라인의 개념(가속/감속, 변이도)을 수치화한 특징들이므로, 태아 저산소증·태아곤란과 같은 생리적 사건의 데이터 기반 분류에 적합합니다.

분석에서는 결측·이상치 점검 후 표준화/정규화를 적용하고, 학습·검증 분할을 통해 AUC, PR-AUC, F1 을 일관 지표로 산출하겠습니다. 클래스 불균형은 클래스 가중치로 우선 대응하고자 합니다. 학습시킬 모델 설정에 관한 부분은 더 알아보고자 합니다.

-유전 기반 성능표

cfDNA-NIPT는 환자 단위의 원시 CSV가 공개로 거의 유통되지 않으므로, 최신 리뷰, 원문, 가이드라인의 성능 표(민감도, 특이도, PPV, NPV, 표본수 등)를 수집·정리하여 CSV로 생성하겠습니다. (ex. <https://www.mdpi.com/2673-3897/6/2/15>) 스키마는 condition(T21/T18/T13), method(cfDNA / 1st-tri combined / 2nd-tri quad 등), sensitivity, specificity, ppv, npv, n_studies(또는 N)로 정의하여, 출처별 수치를 동일 척도로 정규화하고 중앙값, IQR를 함께 보고하겠습니다. 이 표는 임상 CTG 분류에서 얻은 실측 성능과 직접 동일 코호트 비교가 아니라는 점을 명확히 밝힐 것입니다.

5. 최종 파이프라인

1. 임상 스크리닝(CTG) 데이터셋: 실제 학습·검증을 수행하여 재현 가능한 성능(AUC/PR-AUC/F1)을 산출하고, 표준화된 전처리·평가 파이프라인 하에서 모형 간 일관 비교를 제시합니다.
2. 유전 기반(NIPT) 성능표 CSV: 문헌에서 추출한 질환별 성능 분포를 동일 포맷으로 정리하여, 염색체 수적 이상 영역에서의 방법 우월성을 지표 수준에서 비교·해석합니다.

통계처리를 위해 크게 해소해야 하는 부분들은, 1. 다집단 간 비교라는 점 (Kruskal-Wallis 다집단 분석을 이용해 볼 예정입니다) 2. 연구마다 유병률이 달라 PPV, NPV 비교가 왜곡될 수 있다는 점 (공통 가정 유병률을 정해 베이즈 갱신을 해볼까 합니다) 정도라 생각합니다. 이 부분에 관해서는 더 공부해볼 예정입니다.