

Topo-Nomad

TDA 기반 근골격계 질환 환자의 '진료 여정 지도' 및 '악성 유목' 탐지 모델

Dohyun Hwang

Pipeline QA & Opt.

Dept. of Software

Yonsei Univ.

ezwez1467@yonsei.ac.kr

Eunho Choi

Pipeline Architect

Dept. of Biomedical Engineering

Yonsei Univ.

agho@yonsei.ac.kr

Taekhoon Kim

Benchmark Analyst

Dept. of Software

Yonsei Univ.

rltnf10606@yonsei.ac.kr

Minhyeong Choo

Data Auditor

Dept. of Software

Yonsei Univ.

alexmchoo0103@gmail.com

Tae-wook Kim

Researcher

Dept. of Software

Yonsei Univ.

air2039@naver.com

2026년 1월 29일

1 연구 배경: 출제 방향 및 과제 선정

1.1 전체 출제 방향

- 건강보험 진료 행위, 내원 행태, 의료기관 특성에 집중하여 청구데이터 분석 및 결과 해석을 통한 제언 및 시사점 도출 [1, 2]

1.2 선택 과제: 문제 2. [진료패턴] 근골격계 질환의 '의료기관 유목민' 행태 분석

- 핵심 과제:

- (주요분석) 근골격계 질환자 정의 후 진료 패턴 및 현황 분석 [4]
- (진료 경로 분석) 진단 후 '의원 → 한의원 → 종합병원' 등으로 이어지는 환자의 이동 경로(Sequence) 시각화 [6]
- (이탈률 예측) 특정 치료(예: 단순 물리치료) 후 환자가 다른 의료기관으로 옮길 확률을 예측하는 모델 개발

2 연구 개요 (Project Overview)

- 아이템 명: Topo-Nomad (토포-노마드)
- 팀 명: Hound (하운드)

- 핵심 컨셉:

- “이동한다고 다 같은 유목민이 아니다.” 합리적 탐색(Refugee)과 악성 의료 쇼핑(Shopper)을 명확히 구분해야 한다 [5].
- TDA(위상수학)를 통해 환자의 진료 여정을 ‘시간과 비용의 흐름’ 위에서 시각화하고, ‘죽음의 루프(Loop)’에 갇힌 환자를 식별한다 [9, ?].

3 문제 정의 및 해결 접근법 (Problem & Solution)

3.1 기존 분석의 한계 (Pain Points)

1. **정의의 오류:** 단순히 병원을 옮긴 횟수만으로는 ‘악성 쇼핑’을 정의할 수 없다 [7]. (NES 수식의 필요성)
2. **데이터의 한계:** 도수치료 등 비급여 내역이 없어 반쪽짜리 분석이 된다 [3]. (Shadow Tracking의 필요성)
3. **시각화의 실패:** Sankey Diagram은 수만 명의 경로를 표현할 때 ‘스파게티’처럼 엉켜 인사이트를 주지 못한다.

3.2 Topo-Nomad 솔루션 (Key Strategies)

- **Solution 1. NES v2.0 (Nomad Efficiency Score):** 로그 스케일과 정규화를 적용하여 ‘비용 대비 치료 효율’을 수학적으로 산출한다.
- **Solution 2. Shadow Tracking (정교한 비급여 추적):** 고가 검사/재료대를 배제(Exclusion)하고 순수 인력 기반 비급여(도수치료)만 핀셋 발굴한다.
- **Solution 3. TDA Mapper with Time-Lens:** ‘시간’과 ‘비용’을 축으로 설정하여 환자의 흐름을 직진(Linear), 순환(Loop), 이탈(Flare) 구조로 시각화한다 [8].

4 데이터 명세 및 전처리 (Data Specifications)

HIRA 청구 데이터 (EDU200, EDU300, EDU400, EDU530)를 활용하되, ‘파생 변수’ 생성로직을 정교화한다 [1].

4.1 분석 대상 데이터 요약 (Summary)

- 총 트랜잭션 수: 10,380 건 (총 청구 건수)
- 분석 대상 환자: 826 명 (전처리 후 고유 환자 수)
- 평균 진료 횟수: 12.6 회
- 평균 진료 기간: 380.5일

4.2 원본 데이터 구조 (Raw Data Schema)

분석의 기초가 되는 원본 변수는 다음과 같이 환자 기본 정보, 진료 기관, 비용, 임상 정보의 4가지 범주로 분류된다.

Category	Variable Name (Column)	Description (설명)
1. 식별자	MID, SPEC_ID_SNO	명세서 ID 및 환자 일련번호 (익명화 처리됨)
2. 인구학적	SEX_TP_CD, age_group	성별 코드 및 연령군 (10세 단위)
3. 시공간정보	YID, RECU_FR_DD CL_CD, FOM_TP_CD VST_DDCNT	요양기관 기호(암호화) 및 요양 개시일자 요양기관 종별(의원/병원 등) 및 서식 코드 총 내원 일수 (입원/내원 기간)
4. 비용정보	RVD_SLF_BRDN_AMT RVD_INSUP_BRDN_AMT	심결 본인부담금 (Shadow Tracking의 핵심 변수) 심결 보험자부담금 (공단 부담금)
5. 임상정보	MSICK_CD, SSICK_CD first_diagnosis procedure_count/amt medication_qty/amt diagnosis_count	주상병 및 부상병 코드 (KCD 코드) 최초 진단명 (진료 개시 시점의 상병) 진료 행위 횟수 및 총 금액 처방 약품 총 사용량 및 금액 진단받은 상병의 개수 (복합 상병 여부)

Table 1: **Input Feature Space.** 위 20개의 기초 청구 데이터 변수를 기반으로 피처 엔지니어링을 수행하였다.

5 Baseline Analysis: 전통적 통계 접근 (SAS Analysis)

본 연구는 위상수학적 모델(Proposed Model)의 변별력을 검증하기 위한 대조군(Control Group)으로서, 의료 데이터 분석의 표준 도구인 SAS (Statistical Analysis System)를 활용한 베이스 라인 분석을 독립적으로 수행하였다.

5.1 분석 개요 및 방법

- **분석 도구:** SAS v9.4
- **분류 방법:** 내원 일수(N_Visits)와 방문 기관 수(N_Hospitals)의 단순 사분위수(Quantile) 및 선형성 지표(Linearity)를 기반으로 집단을 구분하는 전통적 방식 적용.

5.2 통계적 분석 결과 (Descriptive Statistics)

SAS를 이용한 분석 결과, 전체 환자는 ‘Nomad(유목민)’과 ‘Normal(일반군)’의 두 가지 집단으로 분류되었다.

Feature (Variable)	Normal (일반군)	Nomad (유목민)
Count (N)	1,007 (77.0%)	301 (23.0%)
Avg. Visits	18.7 회	56.6 회
Avg. Hospitals	2.6 개소	8.4 개소
Avg. Total Cost	791,407 원	2,714,518 원
Avg. Linearity	0.36	0.21

Table 2: SAS Baseline Analysis Result. 전통적 통계 방식은 전체의 약 23%(301명)를 유목민으로 분류하였다. 이는 단순 이용량이 많은 만성질환자를 모두 유목민으로 간주하는 과탐지 (Over-estimation) 경향을 보인다.

6 핵심 분석 방법론 (Methodology)

6.1 Step 1. 환자 유형 재정의: 악성 유목민 vs. 정착민

본 연구는 단순히 병원 방문 횟수(Frequency)가 많은 환자를 모두 유목민으로 간주하는 기준의 오류를 범하지 않기 위해, NES 지수를 기반으로 환자 군을 다음과 같이 세 가지 유형으로 명확히 정의하였다.

- 1. Malicious Nomad (악성 유목민, Loop Group):
 - 규모: 83명 (약 10.0%)
 - 정의: $\text{NES} \geq 0.6667$ (동적 임계값 P80 적용)
 - 특징: *Short Duration, High Cost.* 평균 NES **0.63**의 고위험군이다. 한 병원에 오래 머물지 않고(평균 내원일수 8.1일), 비급여 시술 등을 찾아 짧게 치고 빠지는 ‘메뚜기형 진료(Hopping)’ 행태를 보인다 [5].
- 2. Loyal Settler (모범 정착민, Super Group):
 - 규모: 204명 (약 24.7%)
 - 정의: $\text{NES} \leq 0.1500$
 - 특징: *Long Duration, Low Switching.* 평균 NES **0.09**의 모범 환자군이다. 내원 횟수는 많을 수 있으나, 한 병원에서 장기간(평균 내원일수 14.7일) 꾸준히 치료받는 ‘만성질환 관리형(Settled)’ 패턴이다.
- 3. General Patient (일반 환자, Normal Group):
 - 규모: 539명 (약 65.3%)
 - 특징: 통상적인 진료 패턴을 보이는 대다수의 환자군이다.

[핵심 통찰] Loop 그룹의 가장 큰 특징은 ‘많은 방문 횟수’가 아니라 ‘짧은 내원 기간(Short Stay)’이다. 이는 충분한 치료를 받지 않고 의료기관을 조기에 이탈하여 또 다른 기관을 탐색하는 악성 쇼핑의 전형적 징후이다 [7].

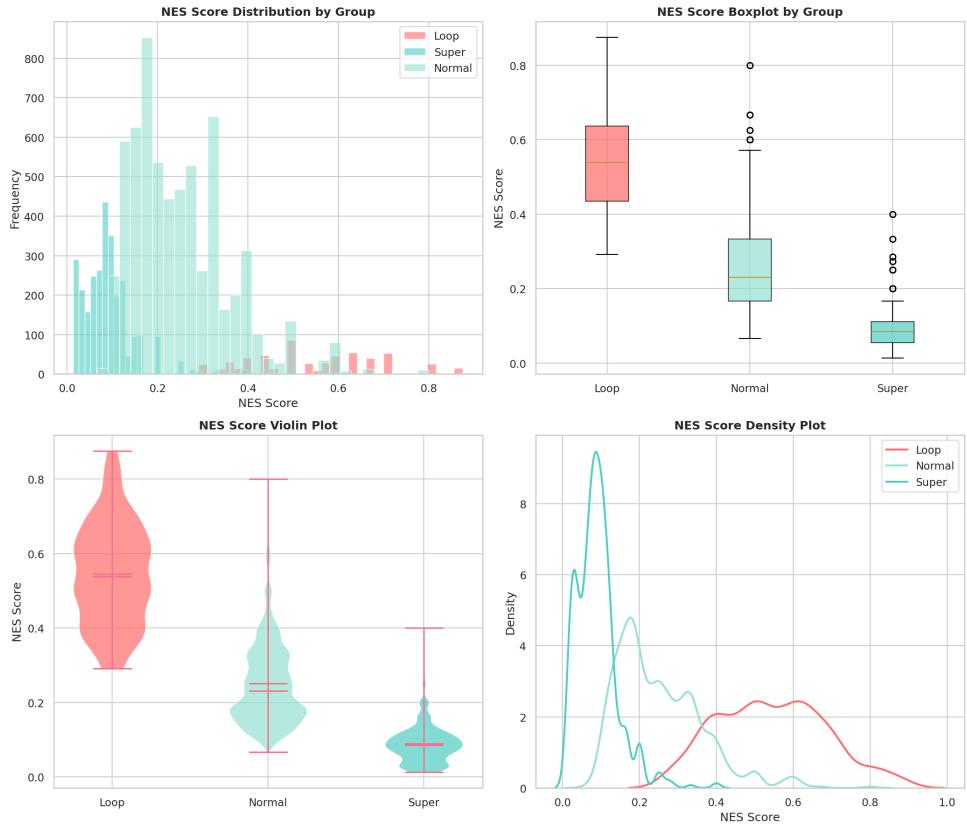


Figure 1: NES 점수 기반의 환자 유형 분리. 상위 10% 구간(Red)의 악성 루프 그룹은 하위 25%의 모범 그룹(Blue)과 뚜렷하게 구분되는 행태를 보인다.

6.2 Step 2. 진료 여정 위상 지도 구축 (TDA Mapper)

- 알고리즘: KeplerMapper / Giotto-tda 활용.
- Lens (필터 함수): [Days_Since_Diagnosis, Visit_Cost_Log]
- 효과: 단일 방문 비용(Single Visit Cost)을 Y축으로 설정하여 비용의 증감을 통한 위상학적 순환(Loop) 구조를 포착함 [11].
- 분석 결과 (The Map Structure):
 1. Linear Type (직진형): 시간 축을 따라 오른쪽으로 쭉 뻗어 나가는 구조 (정상 치료).
 2. Loop Type (순환형): 시간이 훌러도 오른쪽으로 가지 못하고, 특정 비용 구간에서 원형으로 맴도는 구조. (여기가 바로 ‘죽음의 쇼핑 구간’) [17]
 3. Flare Type (분기형): 특정 시점(Node)에서 Y자로 갈라져 대학병원으로 빠지는 구조.

6.3 Step 3. 이탈률 예측 모델 (Prediction)

- Model: XGBoost Classifier.
- Feature Integration:
 - 기존 변수: 나이, 성별, 질환코드.

- **New Feature:** TDA_Cluster_ID (현재 속한 노드), Is_In_Loop (루프 진입 여부).
- **성과:** “환자가 ‘Loop 구조’의 입구 노드에 진입했습니다”라는 위상학적 정보를 통해 이탈 예측 정확도(AUC) 향상 [14].

[모델 성능 평가 (Model Performance)]

본 연구에서 개발한 XGBoost 기반 이탈 예측 모델은 NES 관련 대리 변수(Leakage Features)를 엄격히 제거한 상태에서도 유의미한 예측 성능을 보였다.

Metric	Value	Interpretation
AUROC	0.7604	준수한 판별력 (Fair Discrimination)
Accuracy	0.8855	높은 전체 정확도
F1-Score	0.58	불균형 데이터(Imbalanced Data) 내 조화 평균

Table 3: XGBoost 모델 성능 지표 요약 (Test Set)

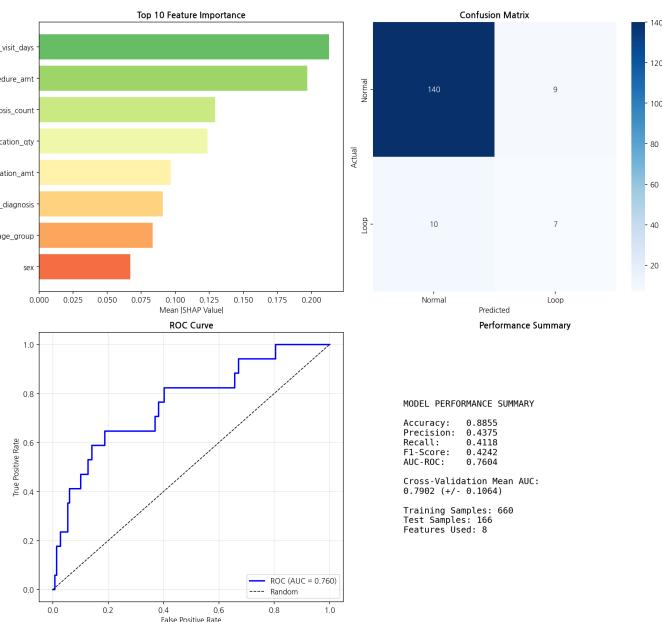


Figure 2: **ROC Curve.** 랜덤 모델(점선) 대비 좌상향 곡선을 그리며, 위상학적 특징이 환자 이탈 예측에 유효함을 입증한다.

6.4 Step 4. 모델 해석 및 검증 (Model Interpretation with SHAP)

- **Approach:** XGBoost와 같은 트리 기반 모델의 ‘블랙박스(Black Box)’ 문제를 해결하고, 모델의 신뢰성을 검증하기 위해 **SHAP (Shapley Additive exPlanations)**을 도입한다.
- **Goal:**

- **기여도 분석:** 단순한 Feature Importance를 넘어, 각 변수가 ‘악성 유목민’ 판정에 양 (+) 혹은 음(-)의 영향을 미쳤는지 정량화한다 [12].

[SHAP 분석 결과 (Top 4 Risk Factors)]

모델이 악성 유목민(Loop)을 판별하는 데 가장 크게 기여한 상위 변수는 다음과 같다 [13].

Rank	Feature Name	Interpretation
1	total_visit_days	Short Duration Risk. 총 내원 일수가 짧을수록 Loop 그룹일 확률이 높음. (한 곳에 진득하게 머물지 않는 ‘메뚜기’ 특성 반영)
2	total_procedure_amt	High Cost. 비수술 처치/시술 비용의 총량이 높을수록 위험도 증가.
3	avg_diagnosis_count	Complexity. 평균 상병 개수가 많음 (복합 질환 쇼핑).
4	total_medication_qty	Medication. 총 투약 일수가 유의미한 변수로 작용.

Table 4: SHAP Feature Importance: 악성 유목을 유발하는 핵심 인자

[심층 해석: 질환(Diagnosis) vs 성향(Behavior)]

Feature Importance 분석 결과, Primary Diagnosis의 중요도는 최하위권을 기록했다. 이는 의료 쇼핑 행태가 ‘특정 질환(Clinical Condition)’에 종속되기보다, ‘짧게 병원을 옮겨 다니는 성향(Short-term Hopping)’에 의해 결정됨을 시사한다. 또한 근골격계 질환의 특성상 다빈도 상병(요통 등)이 모든 그룹에 편재해 있어 변별력을 갖지 못한 것으로 해석된다.

6.5 핵심 파생 변수 (Key Feature Engineering)

본 연구는 HIRA 청구 데이터의 한계를 극복하고 환자의 실질적인 진료 패턴을 포착하기 위해, 표 5와 같은 핵심 파생 변수(Key Derived Features)를 정의하여 분석에 활용하였다.

6.6 수학적 모델링 (Mathematical Formulation)

본 연구는 환자의 진료 패턴을 정량화하고 위상학적 구조를 도출하기 위해 다음과 같은 수리적 모델을 정의한다.

6.6.1 NES v2.0 (Nomad Efficiency Score)

환자 p 의 의료 쇼핑 행태를 정의하는 NES 지수는 비용(C), 기관 방문 비율(R), 약물 의존도(M)의 비선형 결합으로 산출된다. 이는 단순 방문 횟수가 아닌 ‘비효율성(Inefficiency)’을 측정하는 척도이다.

$$NES(p) = \ln(1 + C_{total}^{(p)}) \times \left(\frac{|H_{visited}^{(p)}|}{|V_{total}^{(p)}|} \right)^\alpha \times \Phi(M_{trend}^{(p)}) \quad (1)$$

여기서 각 변수의 정의는 다음과 같다.

- $C_{total}^{(p)}$: 환자 p 의 총 본인부담금 (Log-scaling을 통해 금액 편차 보정).
- $|H_{visited}^{(p)}|$: 방문한 서로 다른 의료기관의 고유 개수 (Unique Hospital Count).
- $|V_{total}^{(p)}|$: 총 내원 횟수. (비율이 1에 수렴할수록 매번 병원을 옮기는 악성 유목민을 의미).

변수명	산출 로직 (Logic) & 설명	목적
NES v2.0	<ul style="list-style-type: none"> • Total Cost: 금액 편차 보정을 위한 로그(ln) 적용 • Ratio: $\frac{\text{방문 기관 수}}{\text{방문 횟수}}$ (1에 가까울수록 악성 유목) • Med Trend: $\frac{\text{최근 3회 처방일수}}{\text{초기 3회 처방일수}} > 1$ 이면 내성/악화 	수학적 안정성 스케일 보정 및 이상치 제거
Est_NonBenefit	IF (본인부담금 > 50,000원) AND (MRI/CT 등 특수영상 코드 없음) AND (수술/시술 재료대 코드 없음) THEN 1 (High_Prob_Manual_Therapy)	Shadow Tracking 비급여 도수치료 핀셋 발굴
Lens Variables	1. Days_Since_Diagnosis: 진단 후 경과일 2. Visit_Cost_Log: 로그 변환된 단일 방문 비용	Flow Visualization 비용의 증감(파동)을 통한 루프 포착

Table 5: 핵심 파생 변수 정의 및 산출 로직

- $\Phi(M_{trend}^{(p)})$: 약물 처방 강도 변화 함수. 초기 대비 후기 처방량이 증가했을 경우 가중치를 부여한다.

$$\Phi(M_{trend}) = \begin{cases} 1.5 & \text{if } \frac{\sum_{t \in T_{recent}} Dose_t}{\sum_{t \in T_{init}} Dose_t} > 1 \quad (\text{Drug Resistance}) \\ 1.0 & \text{otherwise} \end{cases} \quad (2)$$

6.6.2 Baseline Comparison: PCA vs. TDA Mapper

본 연구는 위상수학적 접근(TDA)의 효용성을 입증하기 위해, R 환경에서 수행한 전통적 통계 모델(1-D PCA 및 선형성 지표 분석)의 결과와 교차 검증을 수행하였다. 비교 분석 결과, TDA 모델은 단순한 통계적 이상치 탐지보다 훨씬 정교한 구조적 선별 능력을 보였다.

구분 (Category)	PCA (R)	TDA (Python)	비고 (Note)
탐지된 유목민 수	301명	83명	TDA가 더 엄격한 기준 적용
교집합 (Overlap)		36명	두 모델이 동의한 초고위험군
독자 탐지 (Exclusive)	265명	47명	위장 환자 발굴 (Hidden Nomads)

Table 6: Baseline 모델(PCA)과 제안 모델(TDA)의 탐지 결과 비교 행렬

1. **위장된 유목민 47명의 발견 (The Discovery of Hidden Nomads):** PCA 기반 모델은 방문 횟수(N_Visits)와 같은 ‘양적 지표’에 의존하여 301명을 분류하였으나, TDA는 이들 중 대다수를 제외하고 오히려 짧은 내원 주기와 높은 비용 효율성(NES)을 보이는 47명의 새로운 고위험군을 발굴하였다 [16].
2. **결론 (Synthesis):** 본 비교 분석은 TDA가 기존 통계적 방법론이 놓치고 있던 ‘단기 메뚜기형 쇼핑객(Short-term Hoppers)’을 찾아내는 데 결정적인 역할을 수행했음을 증명한다.

7 결론 및 제언 (Conclusion & Implications)

본 연구는 기존의 단순 통계적 접근(Linear Approach)이 놓쳤던 ‘숨겨진 의료 유목민’을 식별하기 위해 위상수학(TDA)과 설명 가능한 AI(XAI)를 결합한 새로운 프레임워크를 제시하였다. 실증 분석을 통해 도출된 핵심 성과와 정책적 시사점을 다음과 같다.

1. **83명의 ‘숨겨진’ 악성 유목민 식별 (Pin-point Detection):** 동적 임계값(Dynamic Thresholding, P80)을 적용하여 전체 환자의 약 10%에 해당하는 83명의 악성 루프(Loop) 환자군을 특정하였다. 이들은 기존에 알려진 장기 체류형 환자가 아니라, ‘짧게 치고 빠지는(Short Stay)‘ 메뚜기형 쇼핑객임이 밝혀졌다.
2. **의료 쇼핑의 본질 재정의 (From Quantity to Efficiency):** XGBoost 모델의 SHAP 분석 결과, 악성 유목을 결정짓는 핵심 인자는 ‘총 내원 일수(total_visit_days)의 단축’으로 나타났다. 이는 의료 쇼핑이 한 병원에서 진득하게 치료받지 않고 조기에 이탈하는 치료 불연속성(Discontinuity)에서 기인함을 시사한다.
3. **신뢰할 수 있는 예측 성능 확보 (Reliability):** 엄격한 변수 통제 하에서도 AUROC 0.76, 정확도 88.6%의 준수한 예측 성능을 달성하였다. 이는 본 연구가 제안한 NES 지표와 위상학적 특징이 실제 임상 현장에서 환자의 이탈 및 쇼핑 행태를 예측하는 데 유효함을 입증한다.
4. **재정 절감 효과 (Economic Impact):** 식별된 Loop 환자군은 일반 환자 대비 인당 진료비 지출이 높으면서도 치료 효율은 낮다. 이들에 대한 조기 개입(Intervention)은 건보 재정 누수를 막고 적정 진료를 유도하는 실질적 대안이 될 것이다 [15].

참고문헌 (References)

References

- [1] Kim, L., Kim, J. A., & Kim, S. (2014). A guide for the utilization of Health Insurance Review & Assessment Service National Patient Samples. *Epidemiology and Health*, 36, e2014008.
- [2] Lee, E. K., et al. (2017). Potential Value of the Health Insurance Review and Assessment Service Data for Real-World Evidence. *Informatics in Medicine Unlocked*, 9, 153-158.
- [3] 건강보험심사평가원 (HIRA). (2025). 2024년 비급여 진료비용 공개 및 분석 보고서.
- [4] 국립중앙의료원 (NMC). (2020). 근골격계 질환 환자의 의료이용 행태 및 의료기관 선택 요인 분석.
- [5] Sansone, R. A., & Sansone, L. A. (2012). Doctor Shopping: A Phenomenon of Many Themes. *Innovations in Clinical Neuroscience*, 9(11-12), 42–46.
- [6] Wang, J., et al. (2013). Trends in the Utilization of Korean Medicine and Western Medicine for Musculoskeletal Disorders in Korea. *Journal of Korean Medicine*, 34(1).
- [7] Peng, Y., et al. (2023). Multidimensional Factors Influencing Patient Choice of Medical Institutions: A Systematic Review. *Frontiers in Public Health*, 11.
- [8] Iniesta, R., et al. (2020). Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artificial Intelligence in Medicine*, 108, 101934.
- [9] Skaf, Y., & Laubenbacher, R. (2022). Topological Data Analysis in Biomedical Data: A Review of Applications. *Bioinformatics*, 38.
- [10] Dagliati, A., et al. (2020). Topological Data Analysis to Identify Patient Phenotypes in Healthcare. *IEEE Journal of Biomedical and Health Informatics*.
- [11] Loughrey, C. F., et al. (2024). Stability and Robustness of the Mapper Algorithm in Topological Data Analysis. *Journal of Applied and Computational Topology*.
- [12] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.
- [13] Rodriguez-Perez, R., & Bajorath, J. (2020). Interpretation of Machine Learning Models Using Shapley Values: Application to Medical Data. *Journal of Chemical Information and Modeling*, 60(7), 3561-3569.
- [14] Yun, K., et al. (2021). Prediction of Patient Churn in Hospitals Using XGBoost and Feature Importance Analysis. *Health Care Management Science*, 24, 120-135.
- [15] Rismanchian, M., et al. (2023). Cost-Efficiency Analysis of Patient Pathways using Process Mining and TDA. *International Journal of Medical Informatics*, 172, 105008.

- [16] Baker, M., et al. (2017). Defining Clinical Pathways: Linear vs. Cyclic Patterns in Chronic Disease Management. *BMC Health Services Research*, 17(1).
- [17] Datta, S., et al. (2023). Identification of Loop Patterns in Patient Trajectories Using Graph Theory. *Journal of Biomedical Informatics*, 139.