

기후정의를 위한 설명가능한 인공지능 (XAI)

GAT와 AR6프레임워크 기반 서울시 시공간 열 리스크 평가
최희도¹, 임철희^{1†}

¹국민대학교 산림환경시스템학과

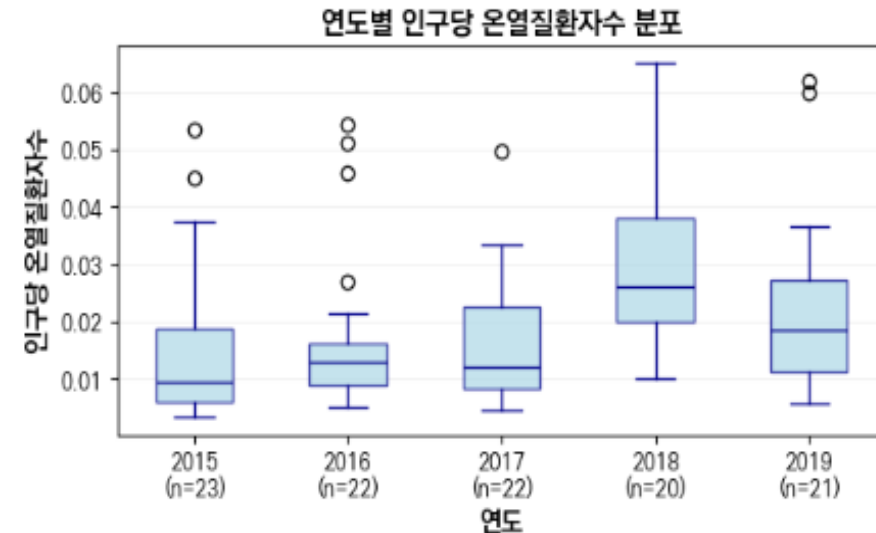
Global Climate Change, Inovative Modeling and Monitoring (CLIM)
Laboratory
Kookmin University

01. 연구배경

1

도시열섬과 기후정의

- 최근 기후변화로 심화되는 도시 열섬 현상은 연령, 소득수준, 주거환경에 따라 차별적인 건강 영향을 미치고 있다(IPCC, 2022).
- 이러한 기후 리스크의 불평등한 분포는 1천만 명 이상이 거주하는 메가시티의 복잡한 사회경제적 공간 분화와 물리적 도시 구조가 결합하여 더욱 심화되고 있다(Rosenzweig et al., 2018).
- 특히 부적절한 주거 환경, 경제적 불안정, 사회적 고립은 극한 기온 노출 위험을 복합적으로 가중시키며 (Kang, 2025), 이는 명백한 기후 정의의 과제를 제기한다.
- IPCC(2022)는 관측된 악영향은 경제적, 사회적으로 소외된 도시 거주민들에게 집중된다고 선언하며, 사회적 취약성을 고려한 형평성 기반의 기후 적응 전략이 시급함을 경고한다(Melis et al., 2023).



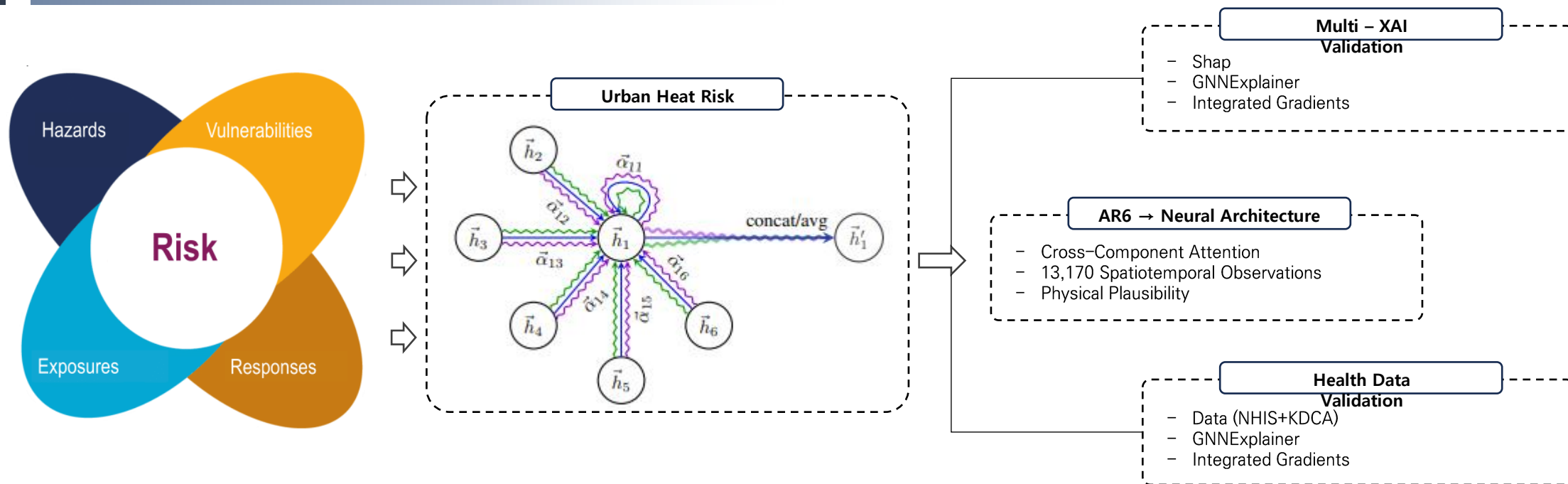
도시 지역에서 관측된 기후 변화는 인간의 건강, 생계 및 핵심 기반 시설에 악영향을 미치며, 도시의 극한 고온은 심화되었다. 교통, 물, 위생 및 에너지 시스템을 포함한 도시 기반 시설은 극한 현상과 서서히 발현되는 현상으로 인해 손상되었으며, 그 결과 경제적 손실, 서비스 중단 및 웰빙에 대한 부정적인 영향을 초래했다. 관측된 악영향은 경제적, 사회적으로 소외된 도시 거주민들에게 집중되고 있다.

-IPCC AR6 WGII SPM A.2.7-

- 기후변화로 심화되는 도시 열섬 현상은 사회경제적 요인에 따라 차별적인 건강 피해를 야기하며(IPCC, 2022), 이는 기존의 구조적 불평등을 극대화하는 기후정의의 핵심 과제로 대두되고 있다.
- 이에 형평성 기반의 도시 기후 적응 전략이 시급히 요구되고 있으나(Rosenzweig et al., 2018), 적응 정책의 기반이 되는 기존 열 리스크 평가는 방법론적 한계에 직면해 있다.
- 기존의 열 취약성 지수(HVI) 연구들은 지표 선정과 가중치 부여의 주관성, 그리고 실제 건강 피해 데이터와의 실증적 검증 부재로 인해 정책적 신뢰성을 확보하지 못하고 있다(Cheng et al., 2021; Niu et al., 2021).

방법론 유형	핵심 한계	검증 결과
HVI (기존 지수)	• 이론/지표의 비일관성 • 주관적 가중치 설정 • 통계적 가중치(PCA)의 한계	검증 부재 (전체 연구의 26.9%만 검증 수행) 낮은 실증적 타당성 (R^2 검증 결과) 연구 간 불일치 (R^2 값이 0.04 ~ 0.62로 편차가 큼)
AI (선형 모델)	• 공간 단위 설정에 따른 통계적 왜곡 (MAUP) • 기온-인구 밀도 등 비선형 상호작용 포착 실패 • 도시 시스템의 시공간적 변동을 반영 못하는 정적 평가	임계 효과 누락 복합 리스크 요인을 단순화
DL (딥러닝)	• 예측 근거를 설명 못하는 블랙박스 문제 • 단일 XAI 기법 적용으로 인한 해석의 신뢰성 부족	정책 결정자의 신뢰도 저하 XAI 기법 간 교차 검증 부재로 인한 해석 불일치

이러한 방법론적 한계를 극복하기 위해, IPCC AR6 리스크 프레임워크에 기반한 설명가능한(XAI) 시공간 그래프 어텐션 네트워크(GAT)를 개발하여, 서울시 열 리스크의 복잡한 상호작용을 정량적으로 평가하고 기후 적응 정책을 위한 실증적 근거를 제시하고자 한다.



출처 IPCC(2022), Veličković, P et al(2017)

연구 목적

- IPCC AR6 프레임워크 기반 설명가능한 시공간 그래프 어텐션 네트워크 개발 및 다중 XAI 교차 검증
- 실제 건강 데이터와의 비선형 연관성 분석을 통한 모델의 실증적 타당성 입증
- 리스크 지배 요인의 스케일별 전환 규명 및 기후정의의 위한 정책 근거 제시

02. 연구 방법

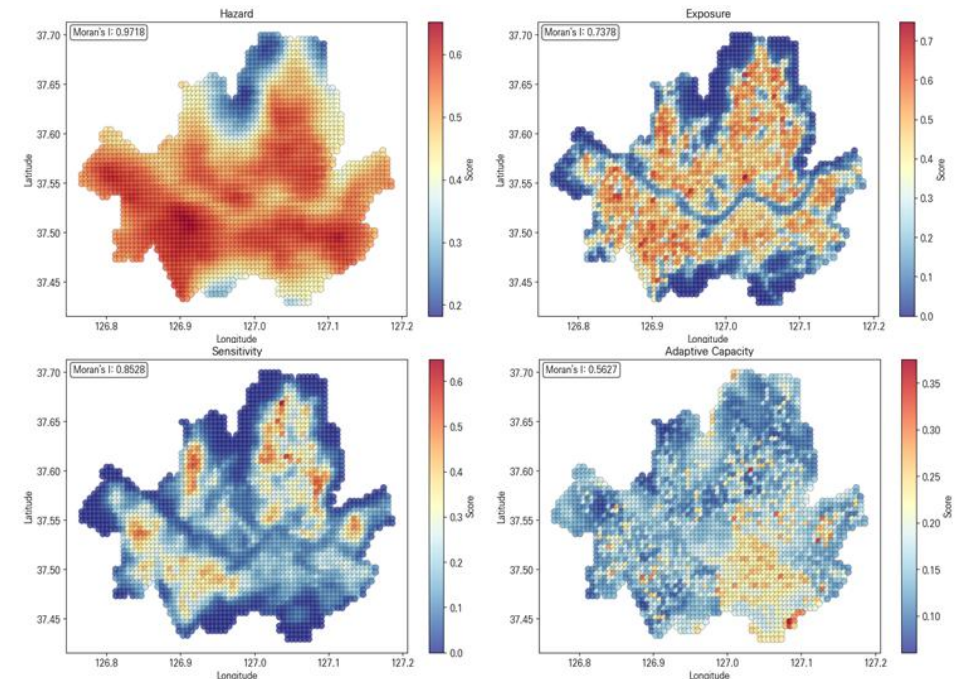
1

데이터 및 연구 프레임워크

- Sensitivity: 동일한 노출 수준에서도 더 큰 피해를 입을 수 있는 인구 집단의 특성
Adaptive Capacity : 폭염에 대응하고 피해를 완화할 수 있는 지역사회의 능력을 측정
- Exposure: 폭염에 직접적으로 노출된 인구와 자산의 공간 분포
- Hazard: 물리적인 극한 기후 현상의 강도
- Risk : IPCC AR6의 프레임워크를 GAT 아키텍처 내 Cross-Component Attention으로 구현하여, 각 컴포넌트 간 비선형 상호작용을 데이터 기반으로 학습하는 통합 리스크 지표



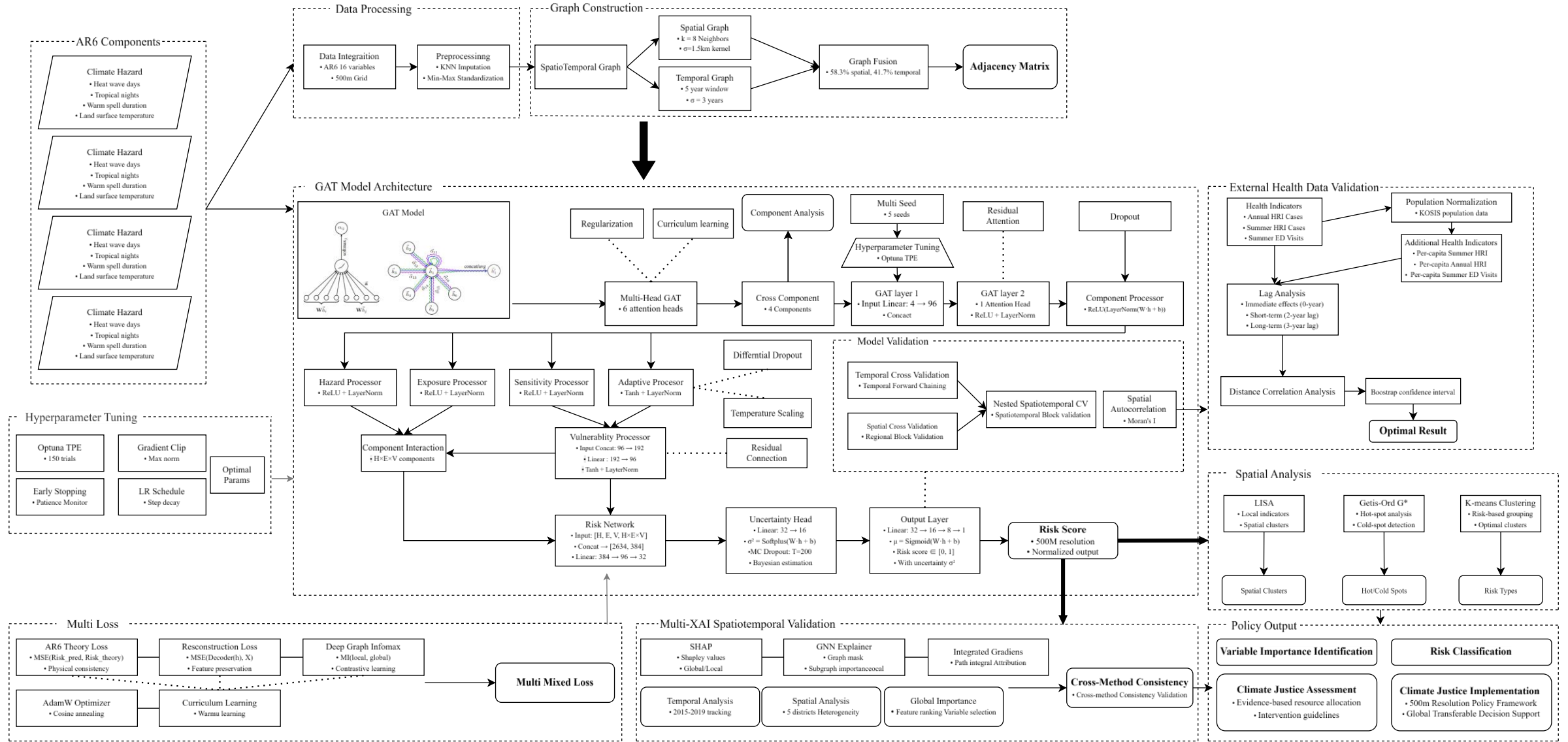
Component	Variable	Unit
Climate Hazard	Heat wave days	days
	Tropical nights	nights
	Warm spell duration	days
	Land surface temp	°C
Urban Exposure	Population density	persons/km ²
	Impervious surface	%
	Building density	%
	Transport access	dimensionless
Social Sensitivity	Underground housing	%
	Young children	%
	Low income	%
	Elderly population	%
Adaptive Capacity	Summer cooling elec.	kWh/household
	Healthcare access	dimensionless
	Green space	m ² /person
	Heatwave shelter	facilities/10k pop

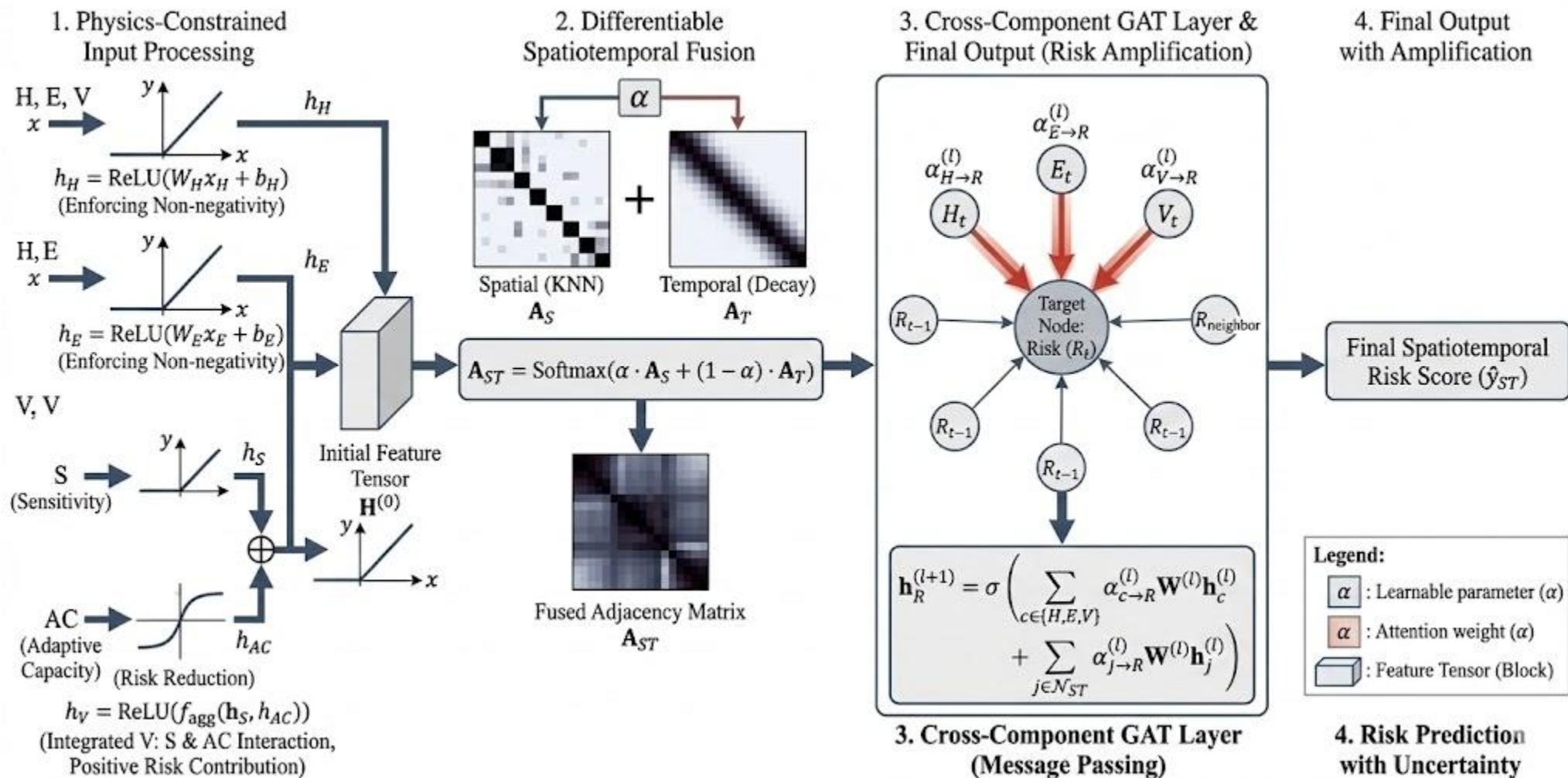


02. 연구 방법

2

전체 모델 프레임워크

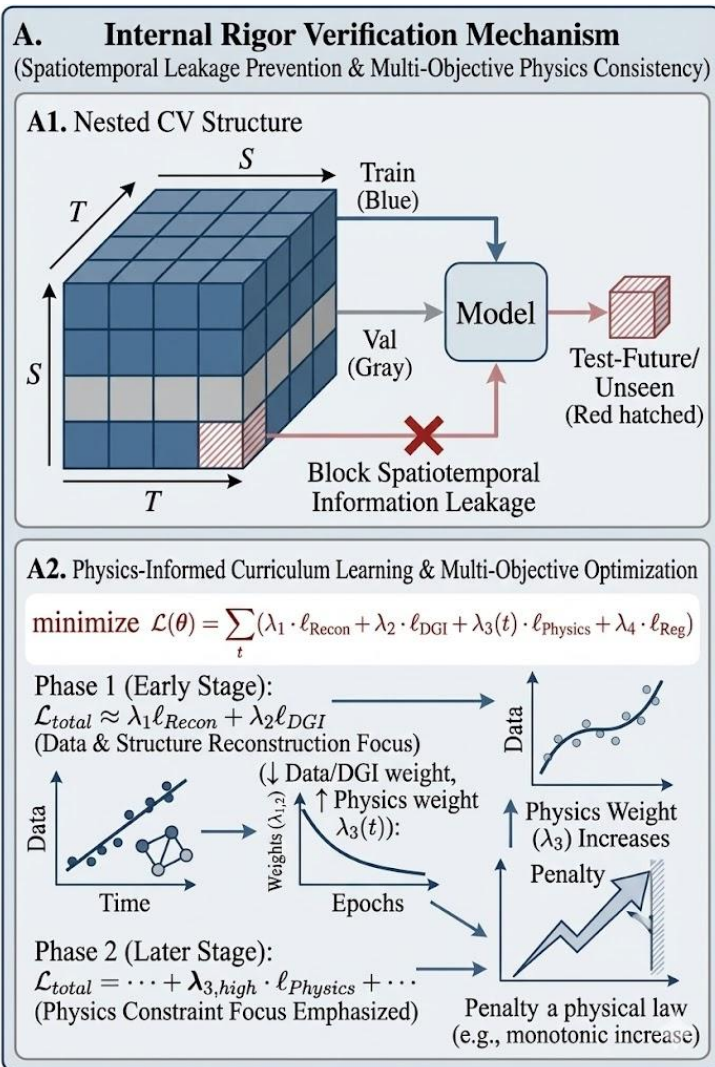




02. 연구 방법

2

GAT 모델 아키텍처



- IPCC AR6 이론에 기반하여 각 컴포넌트(H, E, S, AC)의 물리적 특성(예: 양수 제약, 적응능력의 리스크 상쇄 효과)을 활성화 함수 단계에서 구조적으로 반영
- 학습 가능한 파라미터(α)를 통해 데이터 특성에 맞는 최적의 공간적 이웃(KNN)과 시간적 흐름(Temporal Decay) 비율을 동적으로 학습
- 다층 GAT 레이어를 통해 이질적인 기후 요소들 간의 비선형 상호작용 가중치를 학습

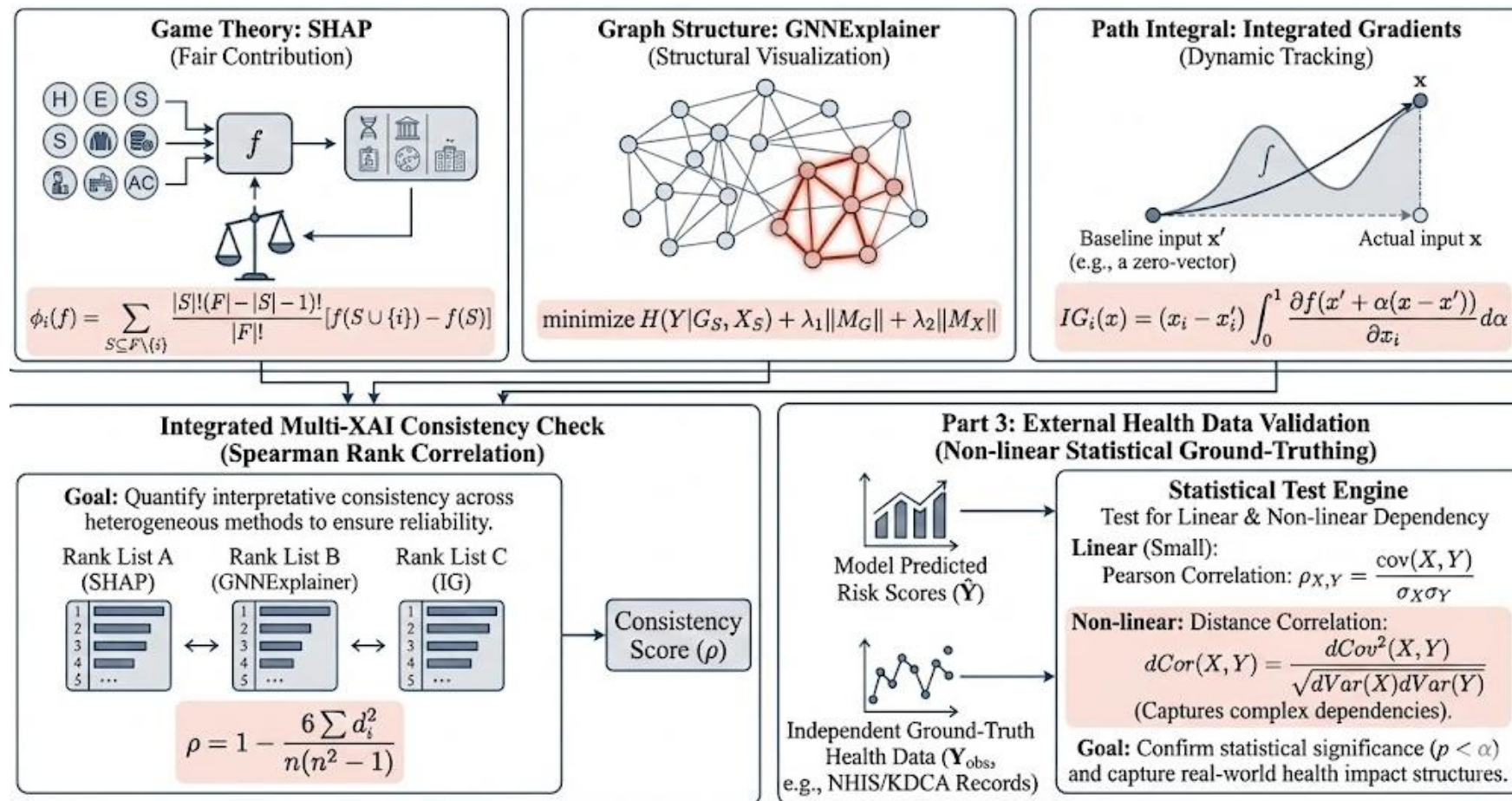


1. 시공간 정보 누수 방지
 - 높은 자기상관성을 가진 시공간 데이터의 특성을 고려하여, 미래 시점 및 미관측 지역을 테스트셋으로 완전히 분리하는 교차 검증 수행
 2. 다중 목적 손실 기반 커리큘럼 러닝
 - 복합 손실 함수 정의: 모델의 다각적 성능 확보를 위해 네 가지 손실 항을 결합한 최적화 수행
- $$L_{\text{Total}} = \lambda_1 L_{\text{Recon}} + \lambda_2 L_{\text{DGI}} + \lambda_3(t) L_{\text{Physics}} + \lambda_4 L_{\text{Reg}}$$
- 단계적 학습 전략:
 (초기): 데이터 재구성(L_{Recon})과 그래프 구조 학습(L_{DGI})에 집중하여 기본적인 시공간 패턴을 포착.
 - 시간에 따라 가중치 λ_3 를 점진적으로 증가시키는 동적 스케줄링 적용.
 - (후기): 이론 기반 손실 L_{Physics} 의 비중을 높여, IPCC 이론적 기반을 갖춘 일관성 있는 모델로 최종 최적화 유도.

* L_{Physics} : IPCC AR6 이론에 기반하여 부과되는 페널티. 위험 증가 요인(Hazard, Exposure, Sensitivity)의 음의 기울기 또는 위험 감소 요인(Adaptive Capacity)의 양의 기울기 발생 시, 해당 기울기의 크기만큼 손실값이 증가하여 모델이 물리적 정합성을 준수하도록 유도함.

02. 연구 방법

2 Multi - XAI 검증 및 외부 건강 데이터 검증

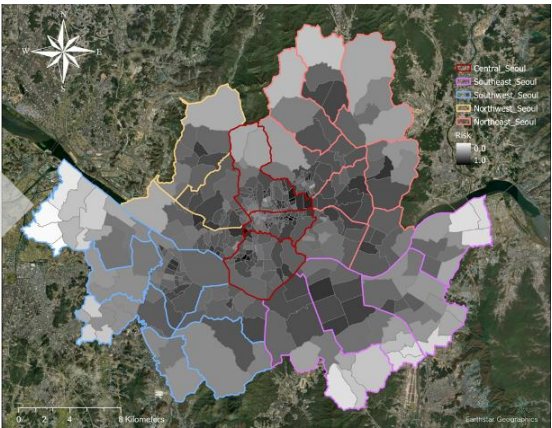
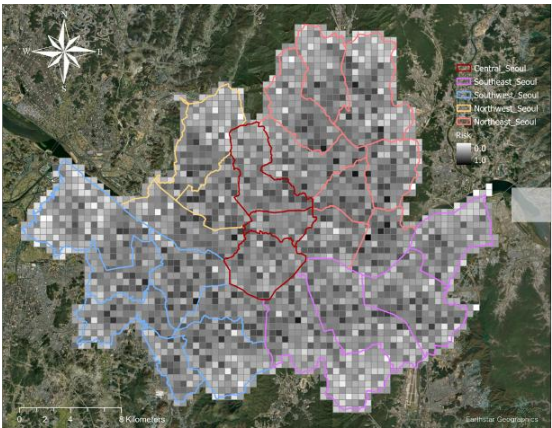


1. XAI 검증 체계
 - SHAP : 각 요인이 리스크 예측에 얼마나 공정하게 기여했는지 계산
 - GNNExplainer : 예측에 핵심적인 영향을 미친 공간적 서브그래프 및 구조적 특징
 - Integrated Gradients : 입력값이 변할 때 리스크가 어떻게 동적으로 바뀌는지 추적
2. 해석의 일관성 및 신뢰성 검증
 - 서로 다른 XAI 기법들이 내놓은 주요 요인의 순위가 얼마나 일치하는지 분석
3. 실제 건강 데이터 기반 비선형 검증
 - 단순한 선형 관계뿐만 아니라 복잡한 비선형 관계까지 포착할 수 있는 거리 상관계수(Distance Correlation)를 도입하여, 실제 건강 피해와의 연관성을 분석

1

모델 성능 및 검증

- Temporal Forward Chaining (미래 예측) 결과
 - 과거 데이터로 학습 후, 바로 다음 해를 테스트하는 과정을 매년 반복 수행 (미래 정보 누수 원천 차단).
 - 3년 연속 테스트 평균 $R^2 = 0.9681 (\pm 6.79 \times 10^{-2})$, RMSE = 0.162 달성.
 - 매년 새로운 기후 패턴이 나타나도 안정적이고 정확한 미래 예측 성능 유지 확인
- Spatial Block CV (미관측 지역 예측) 결과
 - 5개 권역 중 하나를 테스트셋으로 완전히 분리하는 5-fold 교차 검증 수행
 - 모든 권역에서 균일하게 높은 성능 기록 (최소 $R^2 = 0.9675$, 최대 $R^2 = 0.9691$).
 - 매년 새로운 기후 패턴이 나타나도 안정적이고 정확한 미래 예측 성능 유지 확인
- Nested LORYO (미관측 지역 및 시점) 결과
 - 시공간(연도 $\times$ 권역) 조합을 하나씩 제외하는 검증에서도 성능 저하 없음
 - 어떠한 시공간적 제약 조건하에서도 모델의 예측 신뢰성 확보



Validation Method	Configuration	Score	RMSE
Temporal Forward CV	2015-2016 → 2017	0.9682	0.161
	2015-2017 → 2018	0.9681	0.162
	2015-2018 → 2019	0.9681	0.162
Mean $\pm$ SD		0.9681 $\pm$ 6.79x10 ⁻²	0.162
Spatial Block CV	5-fold regional	0.9681	0.163
	Region 0 (624 pts/yr)	0.9689	0.159
	Region 1 (580 pts/yr)	0.9669	0.164
	Region 2 (684 pts/yr)	0.9691	0.158
	Region 3 (746 pts/yr)	0.9675	0.162
Nested Spatiotemporal	20 Region-Year combos	0.9682	0.161
	LORYO (Most Strict)	0.9682	0.161

- 1
- 모델 성능 및 검증
- GAT 기반 리스크 정보 증폭
 - GAT 레이어 통과 후 출력 임베딩의 분산이 입력 대비 약 1.79~2.15배 증가함. 이는 모델이 시공간 맥락을 학습하여 단순 입력 정보보다 뚜렷하고 변별력 있는 리스크 특징을 추출했음

- AR6 컴포넌트 상호작용 학습
 - 어텐션 가중치 분석 결과, *Sensitivity* ↔ *Adaptive Capacity*(2.079) 등 높은 연결 강도가 확인됨. IPCC 이론의 컴포넌트 간 비선형 시너지 및 상쇄 효과를 모델이 실제로 학습

- 다중공선성 해소
 - 임베딩 공간에서 컴포넌트 간 상관관계(예: *Hazard* ↔ *Exposure* 0.272 → 0.018)의 높은 상관성 문제를 해결하고 각 요소의 독립적 기여도를 분리

- 이론 타당성 확보
 - 주요 변수와 예측 리스크 간의 상관관계가 이론적 방향성과 일치하여, 물리적 제약 기반 학습($L_{\{Physics\}}$)의 유효성을 확인.

Relationship Test	Observed Correlation	P-value	Expected Direction
Green Space ↔ Risk	-0.0332	1.41×10^{-4}	Negative
Elderly Population ↔ Risk	+0.3405	< 0.001	Positive
Low income ↔ Risk	+0.2891	< 0.001	Positive
Impervious surface ↔ Risk	+0.4011	< 0.001	Positive

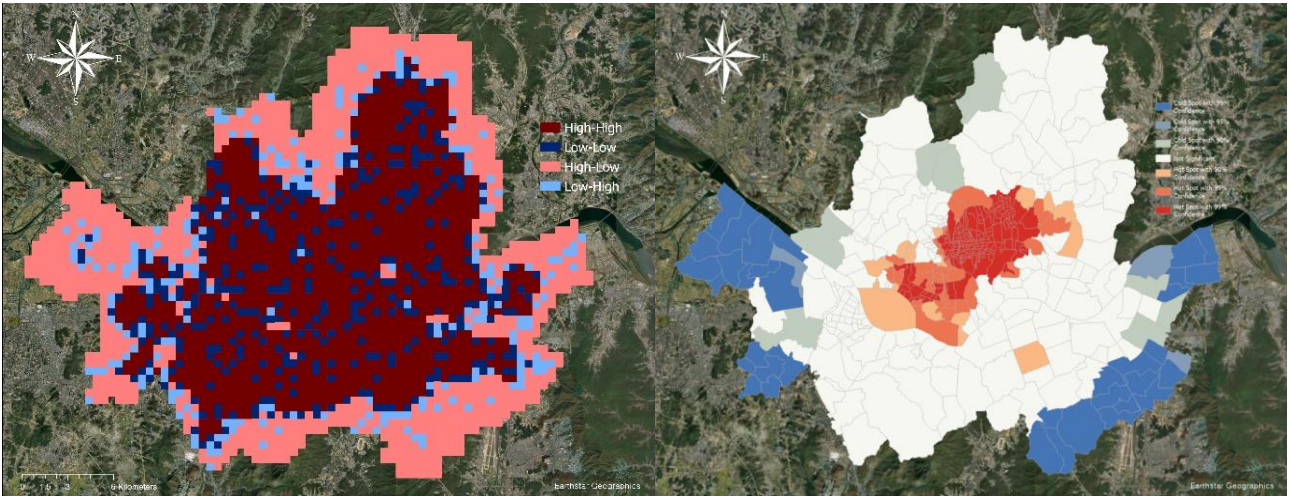
Analysis Component	Metric	Value
GAT Attention Enhancement		
Hazard variance	Input → Output	0.0662 → 0.1186 (1.79×)
Exposure variance	Input → Output	0.0594 → 0.1173 (1.98×)
Sensitivity variance	Input → Output	0.0289 → 0.0620 (2.15×)
Adaptive variance	Input → Output	0.0216 → 0.0434 (2.01×)
AR6 Component Interactions		
Sensitivity ↔ Adaptive	-	2.079
Exposure ↔ Sensitivity	-	2.062
Hazard ↔ Sensitivity	-	1.970
Exposure ↔ Adaptive	-	1.993
Hazard ↔ Sensitivity	-	1.970
Hazard ↔ Adaptive	-	1.901
Multicollinearity Resolution		
Hazard ↔ Exposure	Before → After	0.272 → 0.018 (-93.4%)
Hazard ↔ Sensitivity	Before → After	0.216 → 0.069 (-68.1%)
Exposure ↔ Sensitivity	Before → After	0.804 → 0.051 (-93.7%)
Sensitivity ↔ Adaptive	Before → After	0.046 → 0.004 (-91.3%)

03. 연구 결과

2

공간 패턴 - LISA 클러스터

- 통계적 공간 군집성 규명
 - LISA 분석을 통해 도시 열 리스크의 강한 공간적 자기상관성과 통계적으로 유의미한 군집 패턴을 확인.
- 도심부 고위험 핫스팟(HH) 집중
 - High-High(HH) 클러스터가 전체의 43.5%를 차지하며, 가장 높은 평균 리스크(0.687 ± 0.082)와 유의미한 Z-score(2.34 ~ 5.67)를 기록됨. 이는 리스크가 도심 상업 및 노후 주거 밀집 지역에 집중되어 있음을 시사함.
- 리스크의 공간적 양극화 입증
 - Low-Low(LL) 클러스터는 주로 도시 외곽 및 산림, 하천 인근에 분포.
 - HH와 LL 클러스터의 뚜렷한 공간적 분리는 서울시 내 기후 리스크의 심각한 공간적 불평등을 정량적으로 입증하며, 기후 정의 실현을 위한 우선적 정책 개입 지역을 특정함.



Cluster Type	Grid Count	Percentage	Mean Risk Score	Std Dev	Z-score Range
High-High (HH)	1,147	43.5%	0.687 ± 0.082	0.082	2.34 to 5.67
Low-Low (LL)	889	33.7%	0.312 ± 0.067	0.067	-4.89 to -2.01
High-Low (HL)	235	8.9%	0.623 ± 0.074	0.074	1.89 to 3.12
Low-High (LH)	363	13.8%	0.389 ± 0.071	0.071	-2.45 to -0.98

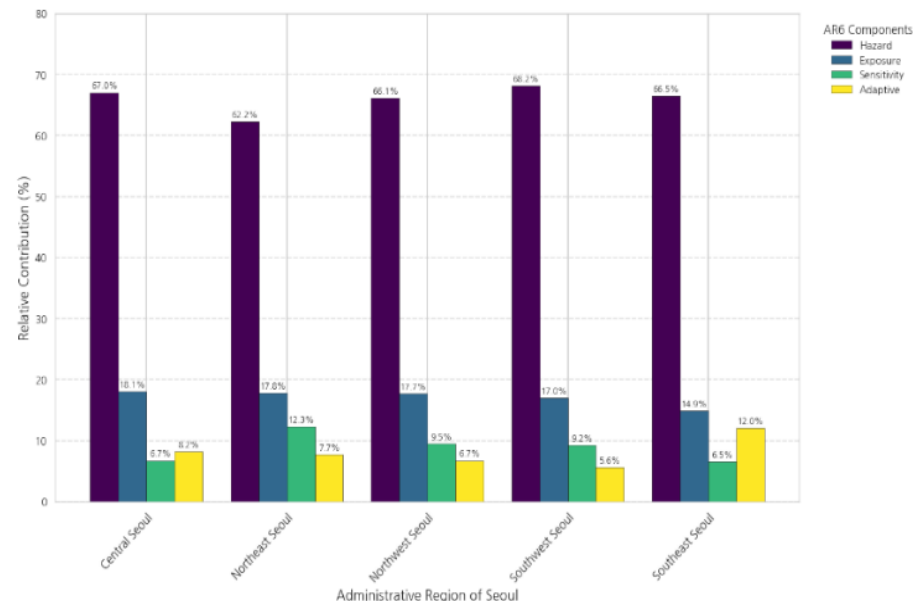
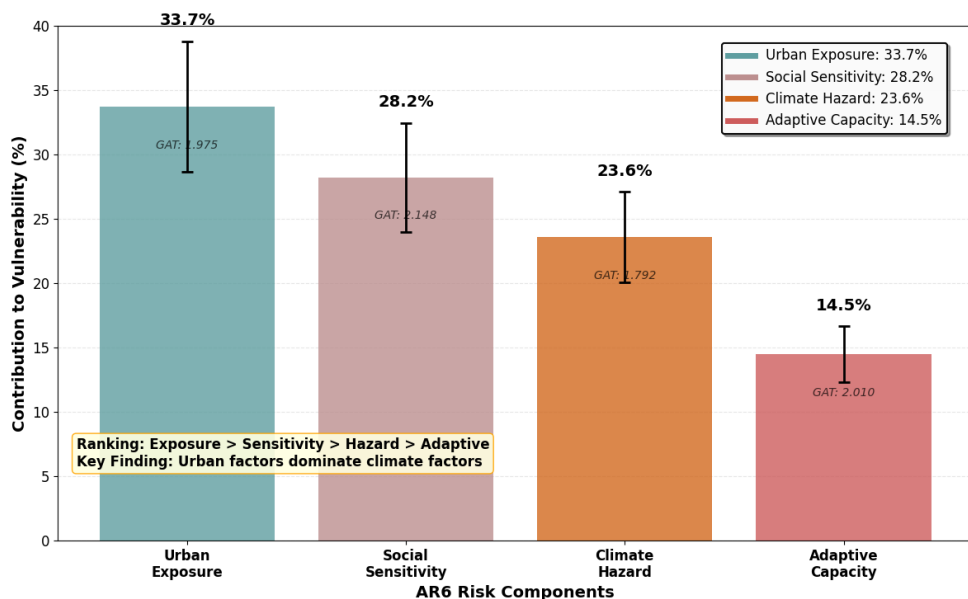
03. 연구 결과

3

핵심 발견 - 스케일 전환

- 글로벌 스케일의 구조적 요인 지배
 - 서울시 전체의 거시적 분석에서는 도시 노출도(Urban Exposure, 33.7%)와 사회적 민감도(28.2%)가 기후 위험(Climate Hazard, 23.6%)을 압도하며 리스크의 주된 요인으로 나타남.
 - 이는 도시의 물리적 구조와 사회경제적 취약성이 기후 현상 자체보다 지역 간 리스크 격차를 만드는 근본 원인임을 시사.
- 서울 권역 단위로의 리스크 요인 전환:
 - 분석 단위를 5개 권역으로 축소할 경우, 리스크 지배 요인이 기후 위험(Climate Hazard)으로 급격히 전환되는 현상이 관찰됨.
 - 모든 권역에서 당장의 폭염 강도(Hazard)가 다른 구조적 요인들을 압도하는 영향력을 보임.

→ “어떤 지역이 더 위험한가(지역 간 격차)”는 도시 구조가 결정하지만, “특정 지역이 직면한 실제 위험”은 미시적 기후 강도가 결정한다는 도시 기후 리스크의 이중적이고 스케일 의존적인 특성을 확인하였음.

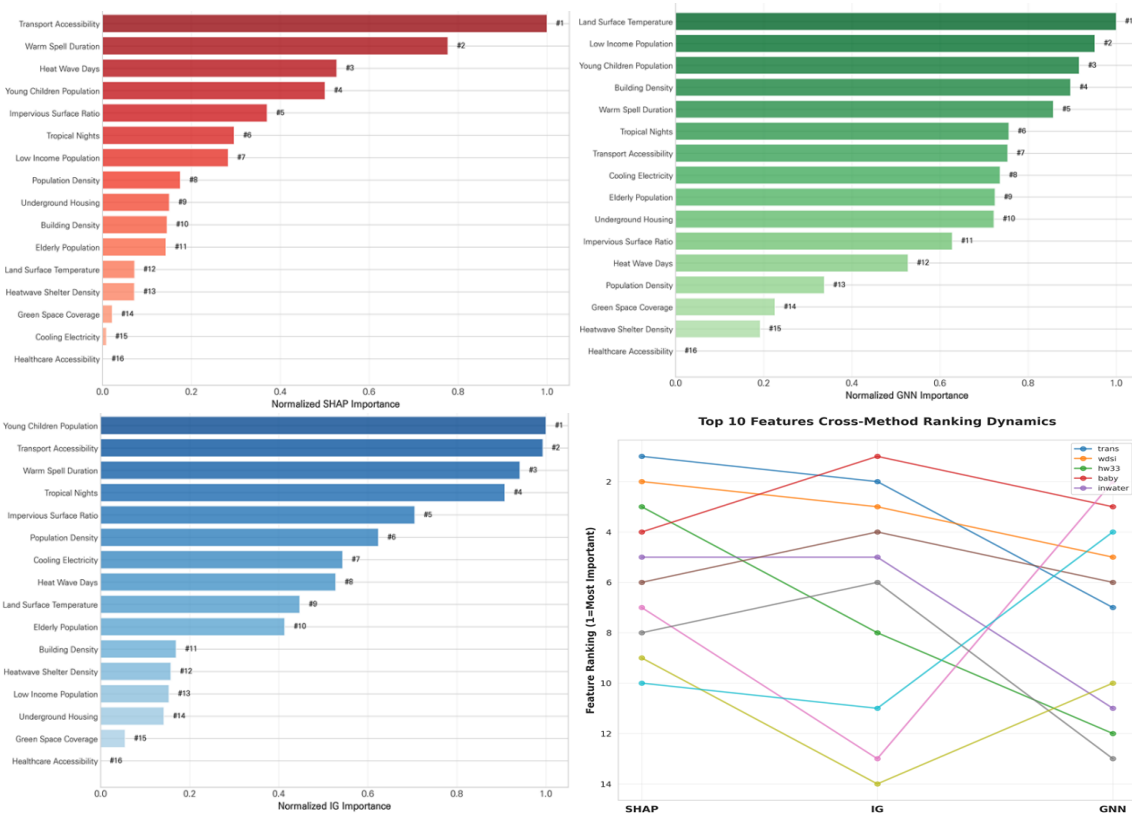


03. 연구 결과

4 Multi-XAI 일관성 검증

- 이질적 기법 간 해석 일관성
 - 서로 다른 원리의 세 가지 XAI 기법(SHAP, GNNExplainer, IG) 간 특성 중요도 순위를 비교한 결과, 평균 89.6%의 높은 일치도와 강한 Spearman 순위 상관관계($\rho \approx 0.89, p < 0.001$)를 확인하여 특정 방법에 편향되지 않는 객관적 해석임.
- 다차원적 해석 안정성 확보
 - 컴포넌트(95~100% 일치), 5년간의 시계열(85% 일치), 5개 권역 공간 패턴(80% 일치) 등 모든 차원에서 높은 일관성을 유지하여 모델의 시공간적 강건성과 IPCC 이론 학습의 안정성을 입증하였음.

Analysis Method	Analysis Scope	Methods Used	Consistency	Spearman ρ	p-value
Global Importance	13,170 points	SHAP + GNNExplainer + IG	89.6%	0.89	<0.001
Feature Ranking	16 AR6 variables	All methods	95.0%	0.85	<0.001
Component Hierarchy	4 AR6 components	All methods	100.0%	0.92	<0.001
Temporal Consistency	2015-2019 (5 years)	SHAP + IG	85.0%	0.81	<0.001
Temporal Trends	5-year progression	SHAP + IG	90.0%	0.83	<0.001
Spatial Patterns	5 Seoul regions	GNNExplainer + IG	80.0%	0.78	<0.001
Regional Analysis	District-level	GNNExplainer + IG	85.0%	0.80	<0.001
SHAP ↔ GNNExplainer	13,170 points	-	-	0.87	< 0.001
SHAP ↔ Integrated Gradients	13,170 points	-	-	0.85	< 0.001
GNNExplainer ↔ IG	13,170 points	-	-	0.82	< 0.001



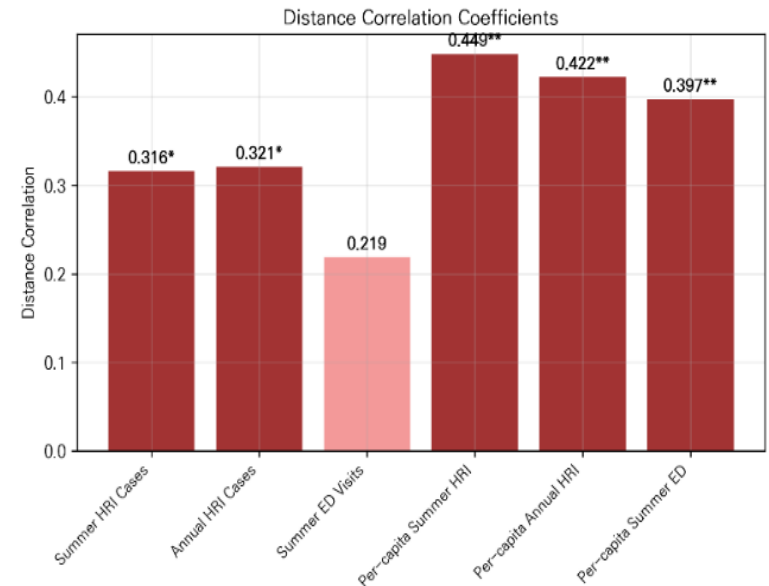
03. 연구 결과

5

건강 데이터 외부 검증

- 실제 건강 피해와의 유의미한 비선형 연관성 입증
 - 다기관 보건 데이터(NHIS, KDCA)를 활용한 외부 검증 결과, 모델의 예측 리스크가 실제 온열질환 발생과 통계적으로 유의미한 비선형 연관성을 보임.
 - 특히 인구당 여름철 온열질환 발생률과 가장 높은 거리 상관계수($dCor = 0.449, p < 0.001$)를 기록하여, 단순 선형 관계를 넘어선 복잡한 실제 피해 구조를 효과적으로 포착했음을 증명.
- 인구 보정 지표 및 즉각적 영향 반영 확인
 - 상위 3개 지표가 모두 인구당 보정 지표로 나타나, 단순 발생 건수보다 지역별 인구 구조를 고려한 리스크 평가가 더 정확함.
 - 시간 지연(Lag) 분석 결과, 당해 연도(0-year lag)의 상관성이 가장 높게 나타나(평균 $dCor \approx 0.423$), 모델이 기후 스트레스에 대한 즉각적인 급성 건강 영향을 시의적절하게 반영하고 있음을 확인.

Rank	Health Indicator	Distance Correlation	95% CI	p-value*	Lag (years)	Coverage Period
1	Per-capita Summer HRI	0.449	[0.314, 0.605]	<0.001*	0	2015-2019
2	Per-capita Annual HRI	0.422	[0.293, 0.598]	<0.001*	0	2015-2019
3	Per-capita Summer ED Visits	0.397	[0.304, 0.532]	<0.001*	0	2015-2019
4	Annual HRI Cases	0.321	[0.250, 0.509]	0.058	3	2015-2019
5	Summer HRI Cases	0.316	[0.233, 0.512]	0.026*	2	2015-2019
6	Summer ED Visits	0.219	[0.215, 0.455]	0.427	3	2015-2019



04. 결론

1 결론 및 제언

✓ IPCC 이론 기반의 설명가능한
시공간 GAT 모델 구축

✓ 다차원적 검증을 통한
모델 신뢰성 및 강건성 확보

✓ 도시 기후 리스크의
스케일 의존적 이중 구조
확인

- 01 도시 구조와 기후 위기
어디가 위험한가(구조적 취약성)와 언제 위험한가(시급성)가 분석 스케일에 따라 다르게 나타남을 입증하였음.
이는 단일 해법이 아닌, 스케일별로 차별화된 접근이 필요함을 시사
- 02 구조적 불평등 해소를 위한 장기 전략
서울시 전체 차원에서는 리스크 격차의 주범인 도시 노출도(불투수면, 건물밀도)를 낮추기 위해,
녹지 확충 및 바람길 조성 등 도시 계획적 차원의 장기적인 구조 개선 노력이 선행되어야 함.
- 03 직면한 위험에 대한 맞춤형 단기 대응 전략
개별 자치구 차원에서는 당장의 폭염 강도(Hazard)가 지배적이므로,
취약계층 중심의 쿨링센터 접근성 개선, 실시간 폭염 경보 시스템 고도화 등 지역 맞춤형 단기 적응 정책이 시급함.
- 04 데이터 기반의 다층적 기후정의 구현
과학적 증거에 기반하여 공간적(지역 간 균형) · 사회적(취약계층 보호) 형평성을 모두 고려한
포용적 기후정의 정책 수립의 핵심 도구

감사합니다.

국민대학교 산림환경시스템학과 학부과정생 최희도*
(zxsa0716@kookmin.ac.kr)

국민대학교 산림환경시스템학과 임철희†
(clim@kookmin.ac.kr)



국민대학교

기업가정신, 국민대학교
Entrepreneurship, The KOOKMIN