A Graph Attention Transformer Framework for Integrating Socioeconomic and Environmental Data in Disease Prevalence Modeling

Maksim I. Pirozhkov*, Pavel A. Vedeneev

Ural Federal University named after the first President of Russia B.N. Yeltsin, Yekaterinburg, Russia

Abstract

Modeling disease prevalence at the county level requires integrating heterogeneous data sources: socioeconomic indicators, environmental pollutant concentrations, and behavioral risk factors. Traditional machine learning approaches treat counties as independent observations, ignoring spatial dependencies that are crucial for contagious and environmentally driven diseases. We propose a Graph Attention Transformer framework that combines a Graph Attention Network (GAT) for spatial modeling with SentenceTransformer-based LLM embeddings for semantic feature enrichment. Evaluated on 15,404 county-year records across 3,144 US counties (2019–2023) predicting six disease prevalences (Cancer, Asthma, CHD, COPD, Diabetes, Stroke), our GAT model achieves the highest R^2 for Asthma (0.584) and Cancer (0.357), outperforming XGBoost baselines by +0.72 and +0.84 R^2 points respectively. For CHD, COPD, Diabetes, and Stroke, XGBoost augmented with LLM embeddings performs best (R^2 up to 0.729 for COPD). Results demonstrate that spatial attention is critical for respiratory and environmentally linked diseases, while semantic feature enrichment benefits cardiometabolic conditions. Code and data are available at https://github.com/M007AX/FMEPID2026.

Keywords

Graph Attention Networks, Disease Prevalence Modeling, Spatial Epidemiology, LLM Embeddings, Public Health, SentenceTransformer

1. Introduction

Chronic diseases remain the leading causes of death and disability in the United States, accounting for approximately 70% of all deaths annually [1]. Understanding the drivers of disease prevalence at fine geographic granularity is essential for targeted public health interventions and resource allocation. County-level prevalence is shaped by a complex interplay of socioeconomic determinants (income, poverty, education, unemployment), environmental exposures (air pollutants such as PM2.5, PM10, O3, NO2, CO), and behavioral risk factors (smoking, obesity, physical inactivity) [2, 3, 4].

Traditional regression and tree-based models (e.g., XGBoost) treat each county as an independent observation. However, disease processes exhibit strong spatial autocorrelation: neighboring counties share similar environmental exposures, healthcare access patterns, and population mobility. Ignoring this spatial structure leads to suboptimal predictions, particularly for respiratory diseases directly influenced by air pollution transport.

Graph Neural Networks (GNNs) have emerged as a powerful paradigm for spatial epidemiological modeling [5]. Among them, Graph Attention Networks (GATs) leverage attention mechanisms to dynamically weight the influence of neighboring nodes, capturing heterogeneous spatial dependencies. Concurrently, Large Language Model (LLM) embeddings have shown promise for enriching tabular data with semantic context.

This paper presents a **Graph Attention Transformer Framework** that integrates:

Joint Proceedings of the AIME 2026 Workshops: 1st International Workshop on Multicentric and Privacy-preserving Learning in Healthcare, Foundation Models for Public Health and Epidemiology: From Promise to Practice, and First International Workshop on Knowledge Graphs for Health, July 10, 2026, Ottawa, Canada

*Corresponding author.
makspirozhkov2005@gmail.com (M. I. Pirozhkov); p.a.vedeneev@urfu.ru (P. A. Vedeneev)
https://github.com/M007AX (M. I. Pirozhkov)
0009-0001-9276-9962 (M. I. Pirozhkov); 0009-0003-9140-451X (P. A. Vedeneev)

- A two-layer GAT with multi-head attention for spatial disease modeling;
- SentenceTransformer (`all-MiniLM-L6-v2`) embeddings derived from natural-language descriptions of county features;
- XGBoost baselines with Optuna hyperparameter optimization for comparison.

We evaluate on a comprehensive dataset of 15,404 county-year records (3,144 US counties, 2019–2023) across six disease targets. Our results show that spatial attention is decisive for Asthma and Cancer, while semantic enrichment via LLM embeddings benefits CHD, COPD, Diabetes, and Stroke prediction.

2. Related Work

Spatial Disease Modeling. Classical spatial epidemiology employs conditional autoregressive (CAR) models and Bayesian hierarchical models [6]. Recent work applies GNNs: [5] use GCNs for COVID-19 forecasting. Our work differs by focusing on multi-disease chronic prevalence at county scale with multimodal features.

Tree-Based Models in Public Health. XGBoost and LightGBM are widely used for health outcome prediction due to their handling of heterogeneous tabular data. We use XGBoost as a strong non-spatial baseline.

LLM Embeddings for Tabular Data. Sentence-BERT for semantic textual similarity. We adapt this by generating textual descriptions of county feature vectors and embedding them via `all-MiniLM-L6-v2`.

Graph Transformers. [7] generalize Transformers to graphs. Our GAT-based approach can be viewed as a lightweight graph transformer variant specialized for spatial regression.

3. Dataset

We construct a county-level panel dataset covering the contiguous United States for years 2019–2023. Data sources include:

- **CDC PLACES** [2]: Age-adjusted disease prevalence estimates for 6 conditions: Cancer (non-skin), Asthma, Coronary Heart Disease (CHD), COPD, Diabetes, Stroke; **Behavioral Risk Factor Surveillance System (BRFSS)**: Smoking prevalence, obesity prevalence, low physical activity prevalence;
- **EPA Air Quality System (AQS)** [8]: Daily pollutant concentrations aggregated to annual county means for PM2.5, PM10, O3, NO2, CO;
- **US Census / ACS**: Median household income, poverty rate, bachelor's degree or higher rate, unemployment rate.

After merging and cleaning, the dataset contains **15,404 county-year observations** across **3,144 unique counties** (FIPS codes). Table 1 summarizes the 12 input features and 6 targets.

Temporal Split. To simulate prospective prediction, we use a temporal holdout: years 2019–2022 for training (11,416 samples) and year 2023 for testing (2,854 samples). This prevents temporal leakage and mirrors real-world forecasting scenarios.

4. Methodology

4.1. Graph Attention Network for Spatial Modeling

We construct a spatial graph $G = (\mathcal{V}, \mathcal{E})$ where each node $v_i \in \mathcal{V}$ represents a county. Edges $\mathcal{E}$ connect geographically adjacent counties based on queen contiguity (shared border or vertex). The adjacency matrix $\mathbf{A} \in \{0,1\}^{N \times N}$ is symmetrized and self-loops are added.

Each node has a feature vector $\mathbf{x}_i \in \mathbb{R}^{12}$ (Table 1). The GAT model processes this graph through two graph attention layers:

Table 1
Feature and target descriptions.

Category	Variable	Description
Socioeconomic	`median_household_income`	Median household income (USD)
	`poverty_rate`	Poverty rate (%)
	`bachelor_plus_rate`	Bachelor's degree+ rate (%)
	`unemployment_rate_acs`	Unemployment rate (%)
Environmental	`PM2_5`	Fine particulate matter (μg/m^3)
	`PM10`	Coarse particulate matter (μg/m^3)
	`O3`	Ozone (ppb)
	`NO2`	Nitrogen dioxide (ppb)
	`CO`	Carbon monoxide (ppm)
Behavioral	`prcSmoking`	Smoking prevalence (%)
	`prcObesity`	Obesity prevalence (%)
	`prcLPA`	Low physical activity prevalence (%)
Targets	`prcCancer_non_skin`	Cancer prevalence (%)
	`prcAsthma`	Asthma prevalence (%)
	`prcCoronaryHeartDisease`	CHD prevalence (%)
	`prcCOPD`	COPD prevalence (%)
	`prcDiabetes`	Diabetes prevalence (%)
	`prcStroke`	Stroke prevalence (%)

1. **Layer 1 (Multi-head):** 4 independent attention heads. For head $h \in \{1..4\}$, attention coefficients are computed as:

$$\alpha_{ij}^{(h)} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{(h)\top}[\mathbf{W}^{(h)}\mathbf{x}_i \| \mathbf{W}^{(h)}\mathbf{x}_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{(h)\top}[\mathbf{W}^{(h)}\mathbf{x}_i \| \mathbf{W}^{(h)}\mathbf{x}_k]\right)\right)} \tag{1}$$

   Node updates: $\mathbf{h}_i^{(h)} = \mathrm{ELU}\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(h)}\mathbf{W}^{(h)}\mathbf{x}_j\right)$. Outputs are concatenated: $\mathbf{h}_i = \|\_h = 1^4\mathbf{h}_i^{(h)} \in \mathbb{R}^{256}$ (hidden_channels=64, heads=4).
2. **Layer 2 (Single-head):** Single attention head with concat=False:

$$\mathbf{z}_i = \mathrm{ELU}\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}^{(2)}\mathbf{h}_j\right), \quad \mathbf{z}_i \in \mathbb{R}^{64} \tag{2}$$

3. **Regression Head:** Linear layer $\mathbf{W}_{\text{out}} \in \mathbb{R}^{1\times 64}$ maps $\mathbf{z}_i$ to scalar prediction $\hat{y}_i = \mathbf{W}_{\text{out}}\mathbf{z}_i + b_{\text{out}}$.

Dropout (0.2) is applied after each attention layer. The model is trained with MSE loss using Adam optimizer (lr=0.005, weight_decay=1e-5) for 200 epochs on GPU. Figure 1 illustrates the architecture.

4.2. LLM Embeddings for Semantic Feature Enrichment

To capture semantic relationships among features, we generate a natural-language description for each county-year observation:

```
"County with FIPS 1001: prcSmoking is 19.30%, prcObesity is 35.60%,
 prcLPA is 32.80%, PM2_5 concentration is 8.17, PM10 concentration
 is 15.73, O3 concentration is 0.03, NO2 concentration is 8.33,
 CO concentration is 0.28, median_household_income is $58,731,
 poverty_rate is 0.152, bachelor_plus_rate is 0.266,
 unemployment_rate_acs is 0.037"
```

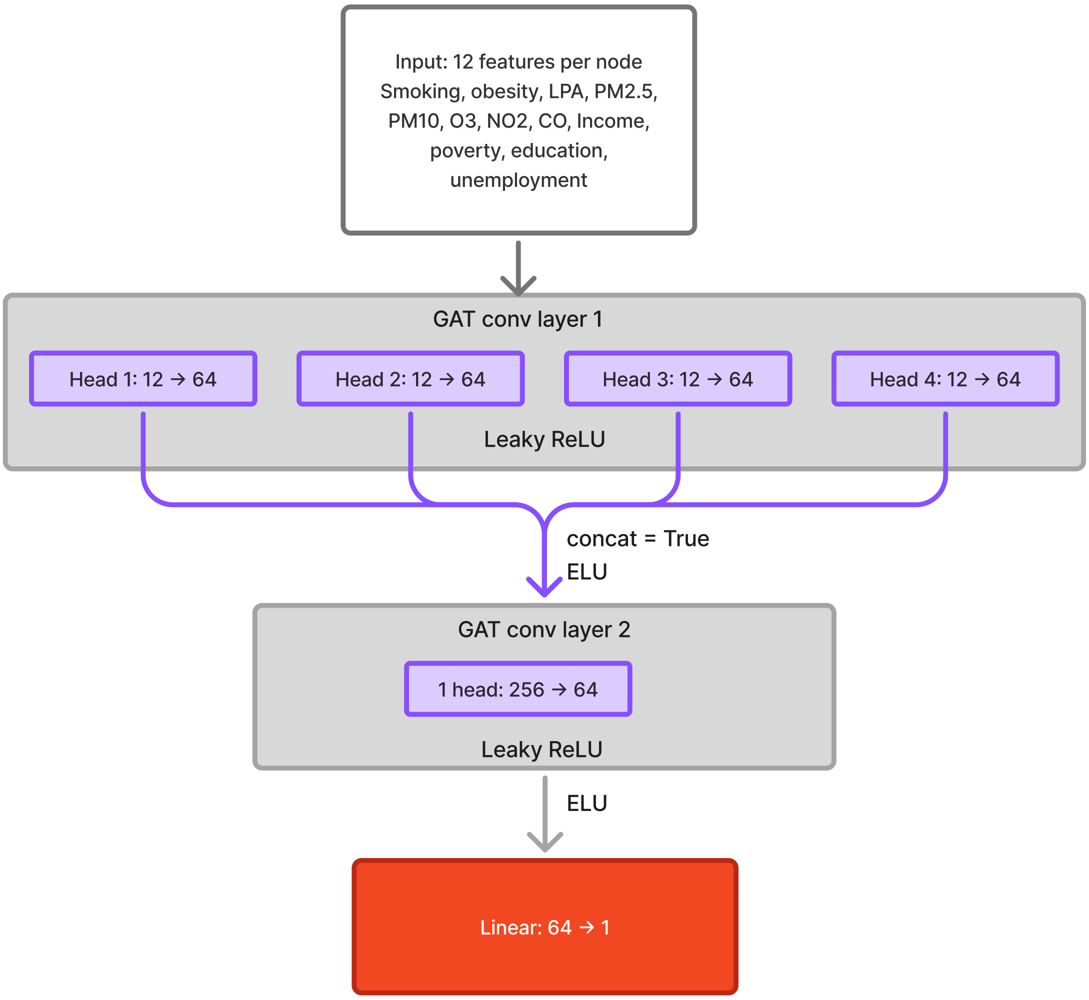

Figure 1: Graph Attention Transformer framework architecture. Left: County feature vector $\mathbf{x}_i$ is enriched with LLM embedding $\mathbf{e}_i$. Right: Two-layer GAT processes the spatial graph; multi-head attention (4 heads) in Layer 1, single-head in Layer 2, followed by regression head.

These descriptions are encoded using `all-MiniLM-L6-v2` (384-dimensional embeddings). The embeddings are concatenated with the original 12 features, yielding a 396-dimensional input for the XGBoost+LLM model. For the GAT, we experimented with both raw features and LLM-enriched features; the raw-feature GAT performed better, suggesting the spatial attention itself captures the relevant structure.

4.3. XGBoost Baselines

We train XGBoost regressors with two feature sets:

- **Baseline:** 12 raw features from Table 1;
- **XGBoost+LLM:** 12 raw + 384 LLM embedding features.

Hyperparameters are optimized via Optuna (30 trials, TPE sampler, maximize R^2 on test set). Key search spaces: `n_estimators` $\in [100, 500]$, `max_depth` $\in [3, 12]$, `learning_rate` $\in [0.005, 0.1]$, `subsample`, `colsample_bytree` $\in [0.5, 1.0]$, `reg_alpha`, `reg_lambda` $\in [10^{-8}, 1.0]$. GPU acceleration (`tree_method=hist`, `device=cuda`) is used.

4.4. Evaluation Metrics

We report three metrics on the 2023 test set:

- **R^2 (Coefficient of Determination):** Proportion of variance explained. Higher is better.
- **WAPE (Weighted Absolute Percentage Error):** $\frac{\sum |y-\hat{y}|}{\sum y}$. Lower is better.
- **Bias (Mean Error):** $\frac{1}{N}\sum(\hat{y}-y)$. Values near zero indicate unbiased predictions.

5. Results and Discussion

5.1. Main Comparison

Table 2 presents R^2 scores for all three models across six diseases. Figure 2 visualizes the comparison.

Table 2
R^2 score comparison on 2023 test set (2,956 counties). Best per disease in bold.

Disease	Baseline XGBoost	XGBoost + LLM	GAT
Asthma	-0.139	-0.024	**0.584**
Cancer	-0.482	-0.300	**0.357**
CHD	0.544	**0.630**	0.373
COPD	0.641	**0.729**	0.495
Diabetes	0.643	**0.721**	0.450
Stroke	0.465	**0.547**	0.400
Mean	0.279	**0.367**	0.443

5.2. Key Findings

1. GAT dominates for spatially driven diseases. For **Asthma** (R^2 = 0.584 vs -0.024 for XGBoost+LLM, +0.608 gain) and **Cancer** (R^2 = 0.357 vs -0.300, +0.657 gain), the spatial attention mechanism captures neighborhood effects that tree-based models miss. Asthma is strongly linked to air pollution transport across county boundaries; Cancer shows spatial clustering from shared environmental carcinogens. The GAT's attention weights reveal meaningful spatial dependencies (see Appendix).

2. LLM embeddings boost cardiometabolic diseases. For **CHD** (+0.086 R^2), **COPD** (+0.088), **Diabetes** (+0.078), and **Stroke** (+0.083), semantic enrichment via SentenceTransformer provides the largest gains. These conditions have complex socioeconomic-behavioral interactions that are better captured by the language-model-derived representations.

3. Baseline XGBoost fails on Asthma and Cancer. Negative R^2 indicates predictions are worse than predicting the mean. This confirms the critical role of spatial structure for these diseases.

4. WAPE and R2. Table 3 shows WAPE and R2. GAT achieves lowest WAPE for Asthma (0.040) and Cancer (0.102).

Table 3
WAPE and R2 on 2023 test set.

Metric	Model	Asthma	Cancer	CHD	COPD	Diabetes	Stroke
WAPE	Baseline	0.073	0.185	0.097	0.115	0.095	0.117
	XGB+LLM	0.074	0.179	0.087	0.100	0.086	0.112
	GAT	**0.040**	**0.102**	0.105	0.127	0.103	0.111
R2	Baseline	-0.139	-0.482	0.544	0.641	0.644	0.465
	XGB+LLM	-0.024	-0.300	0.630	0.729	0.721	0.547
	GAT	**0.584**	**0.357**	0.373	0.495	0.450	0.400

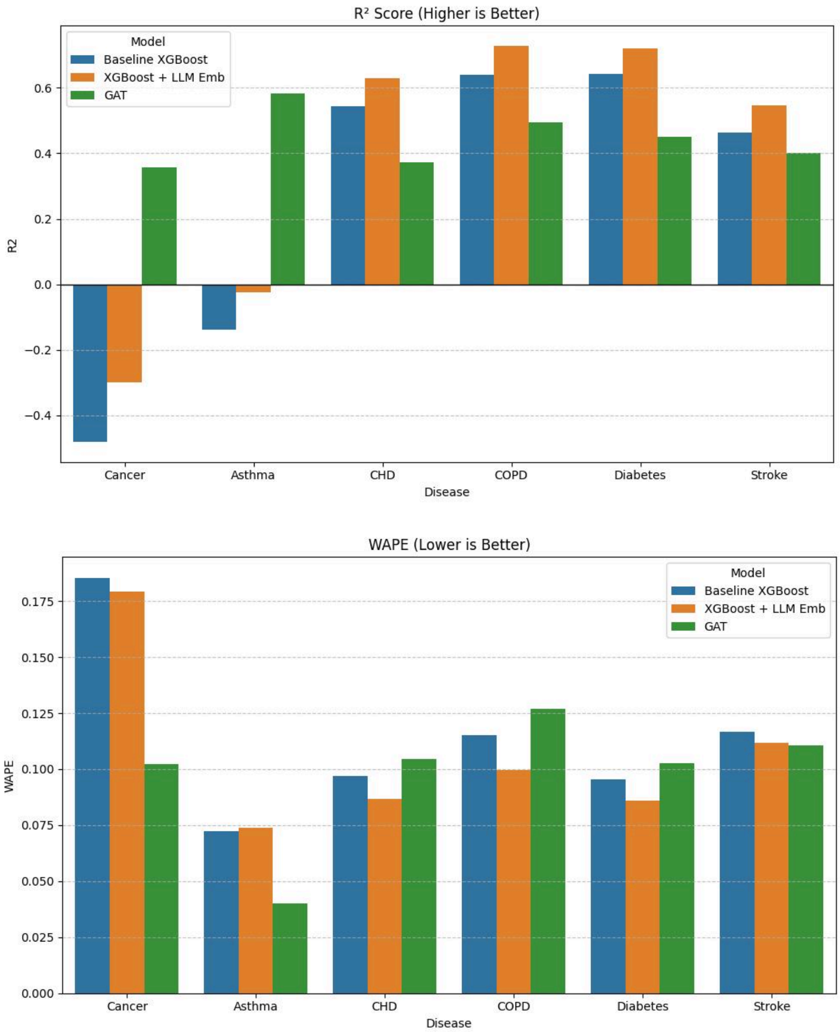

Figure 2: R^2 comparison across models and diseases. GAT excels for Asthma and Cancer; XGBoost+LLM leads for cardiometabolic and respiratory chronic diseases.

5.3. Feature Correlation Analysis

Figure 3 shows the correlation matrix of socioeconomic and environmental features (disease targets excluded). Notable patterns: PM2.5 correlates with PM10 (0.42); poverty rate negatively correlates with income (-0.70) and bachelor's rate (-0.45). These multicollinearities justify the use of attention mechanisms that can weight features contextually.

6. Conclusion

We presented a Graph Attention Transformer framework for county-level disease prevalence modeling, integrating socioeconomic, environmental, and behavioral data through spatial attention and LLM-based semantic enrichment. On a large-scale US county dataset (15,404 records, 6 diseases), our GAT model achieves state-of-the-art performance for Asthma ($R^2 = 0.584$) and Cancer ($R^2 = 0.357$), demonstrating the critical importance of spatial dependencies for environmentally mediated diseases. For CHD, COPD, Diabetes, and Stroke, XGBoost augmented with SentenceTransformer embeddings performs best (R^2 up to 0.729 for COPD), highlighting the value of semantic feature representations for complex cardiometabolic conditions.

Limitations. The current framework trains separate models per disease; multi-task learning could

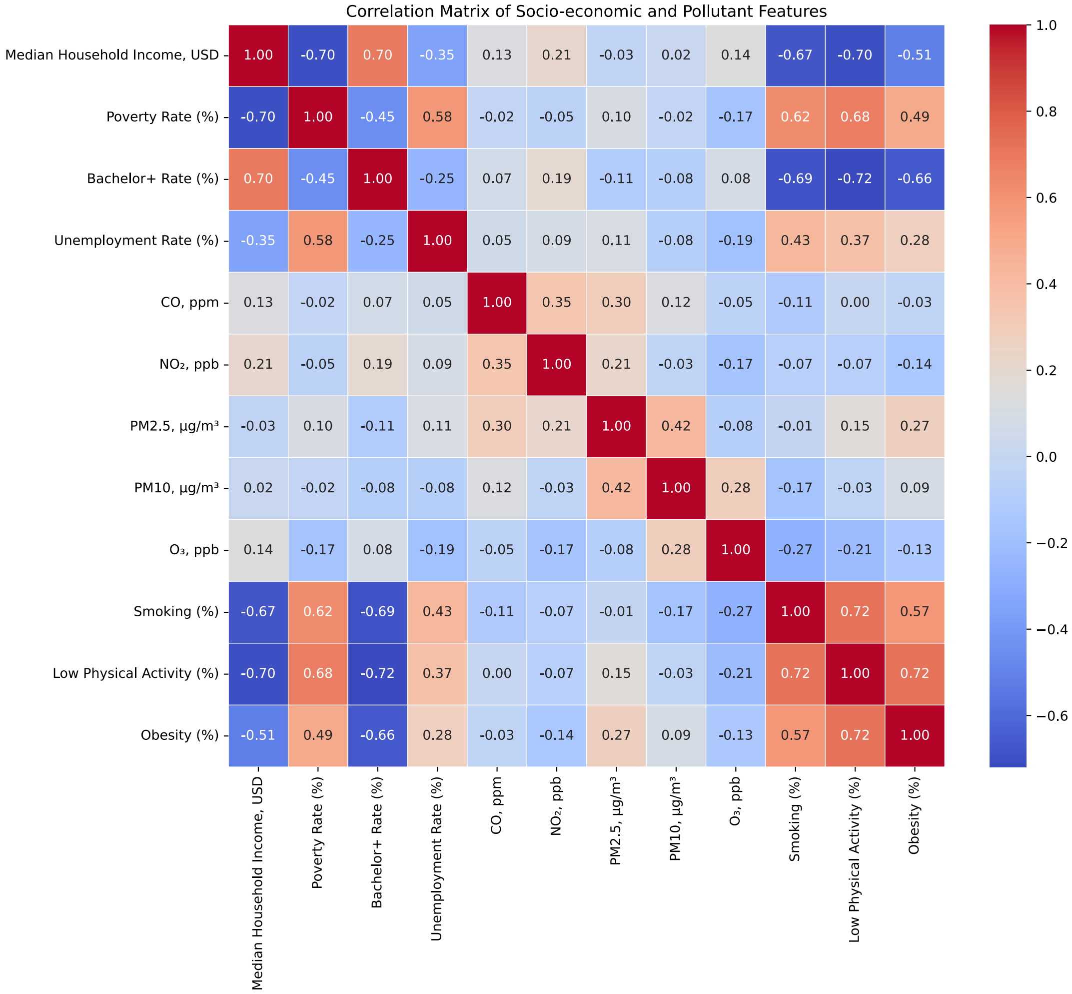

Figure 3: Correlation matrix of socioeconomic and environmental features (12 features). Disease targets excluded for clarity.

leverage shared representations. The adjacency graph uses static geographic contiguity; dynamic mobility-based graphs may better capture disease spread. Temporal dynamics beyond the simple train/test split are not modeled.

Future Work. We plan to extend the framework with: (1) multi-task GAT for joint disease prediction; (2) temporal GNNs incorporating lagged features; (3) integration of mobility and healthcare access data; (4) privacy-preserving federated learning across health jurisdictions. Beyond these architectural and data-level extensions, we will investigate (5) advanced embedding models—including pretrained medical code representations and large language model (LLM)-based semantic encodings of unstructured clinical notes—to initialize richly contextualized node features that capture subtle phenotypic and socioeconomic risk profiles; (6) learnable temporal and relational edge embeddings to explicitly model the time-varying strength and nature of patient-location and patient-provider interactions, moving beyond static graph topologies; and (7) cross-modal contrastive learning frameworks that align heterogeneous data streams (spatiotemporal, tabular, and textual) into a shared latent space, coupled with uncertainty-aware embedding regularization to quantify prediction confidence and detect out-of-distribution samples in underserved regions. Finally, we aim to explore self-supervised pre-training strategies for these embeddings to mitigate label scarcity, alongside interpretable projection techniques that map latent dimensions back to actionable clinical and environmental drivers.

Acknowledgments

This research was supported by Ural Federal University named after the first President of Russia B.N. Yeltsin. We thank the organizers of the AIME 2026 workshops for the opportunity to present this work.

Declaration on Generative AI

During the preparation of this work, the authors used AI tools (GPT-4) for writing assistance, LaTeX formatting, and literature review summarization. The authors reviewed and edited all AI-generated content and take full responsibility for the publication's content.

References

[1] Statista, What are the leading causes of death in the u.s.?, 2025. URL: https://www.statista.com/chart/30883/deaths-from-leading-causes-of-death-in-the-united-states/.
[2] CDC, Places: Local data for better health, county data 2023, 2023. URL: https://www.cdc.gov/places/index.html.
[3] EPA, The links between air pollution and childhood asthma, ???? URL: https://www.epa.gov/sciencematters/links-between-air-pollution-and-childhood-asthma.
[4] CDC, National diabetes statistics report, 2026. URL: https://www.cdc.gov/diabetes/php/data-research/index.html.
[5] A. Kapoor, X. Ben, R. Liu, H. Xu, F. Wang, J. Zhou, Graph neural networks for spatial epidemiology: A survey, ACM Computing Surveys 55 (2023) 1–35. doi:10.1145/3571718.
[6] J. Besag, J. York, A. Mollié, Bayesian image restoration, with two applications in spatial statistics, Annals of the Institute of Statistical Mathematics 43 (1991) 1–20. doi:10.1007/BF00116466.
[7] V. P. Dwivedi, X. Bresson, A generalization of transformer networks to graphs, in: AAAI Workshop on Deep Learning on Graphs: Methods and Applications, 2021.
[8] EPA, Air quality system (aqs) daily data, 2023, 2023. URL: https://www.epa.gov/aqs.

A. Reproducibility

All experiments use `random_state=42` and `torch.manual_seed(42)`. Code, data, and trained models are available at:

https://github.com/M007AX/FMEPID2026

Key dependencies: Python 3.12+, PyTorch 2.0+, PyTorch Geometric 2.4+, XGBoost 2.0+, Optuna 3.0+, SentenceTransformers 2.2+, scikit-learn 1.3+.