Spatial Graph Attention Network Modeling for Neighborhood-Scale Lead Contamination Risk Prediction Using Publicly Available Data

Raphael Anaadumba[1*], Nazim A. Belabbaci[1], Connor Sullivan[2], Anton Kovalev[1], Yidong Zhu[1], Pradeep Kurup[2], Mohammad Arif Ul Alam[1,3,4]

[1]Richard A. Miner School of Computer and Information Sciences, University of Massachusetts Lowell, Lowell, USA
[2]Civil and Environmental Engineering, University of Massachusetts Lowell, Lowell, USA
[3]Department of Medicine, University of Massachusetts Chan Medical School, Worcester, USA
[4]National Institute of Aging, National Institute of Health, Bethesda, USA
raphael_anaadumba@student.uml.edu

Abstract

Lead contamination in urban water systems remains a prevalent public health threat, affecting millions of American households and disproportionately endangering vulnerable population groups. Current municipal risk assessment and inspection strategies are overwhelmingly based on random sampling and complaint-driven protocols that overlook spatial complexity, reinforce inequities, and squander limited resources, leaving critical exposure areas unidentified. This paper presents a lead contamination risk prediction framework from socio-demographic housing features analytics, first of its kind, by drawing on partially anonymized residential testing data as ground truth and applying graph neural networks alongside gradient-boosted ensembles. Specifically, our method integrates spatial Deep Graph Attention Networks classifiers to capture inter-neighborhood contamination dependencies, fuse demographic and spatial evidence, and produce interpretable risk scores. Those scores are actionable by municipal water authorities at the intra-neighborhood level. Through extensive experiments on newly constructed Chicago block-group level datasets, our framework achieves a balanced accuracy of 84.8% and reduces false positive lead contamination by up to 44% versus spatial-only baselines and 21% over current practice, without sacrificing recall on contaminated blocks. Our approach not only extends technical boundaries in spatial-ensemble learning and privacy-preserving urban health modeling, but also provides policymakers and public health officials with a means to assess and address contamination risks, supporting efforts to protect community health and safety.

Introduction

Urban water systems in the United States contain an estimated 6 to 10 million lead service lines, with contamination disproportionately affecting low-income and minority communities (U.S. Environmental Protection Agency 2024). The 2014 Flint water crisis, which exposed over 100,000 residents to elevated lead levels, demonstrated the catastrophic consequences of inadequate risk assessment systems (Centers for Disease Control and Prevention 2016). Current municipal inspection strategies rely on either reactive complaint-driven responses or ZIP-code-level heuristics based on housing age, resulting in systematic misallocation of limited inspection resources (City of Chicago Office of Inspector General 2018).

The core challenge in lead-contamination risk assessment lies in reconciling spatial heterogeneity in contamination with demographic risk factors; although lead pipes may be similarly distributed across older neighborhoods, actual contamination risk varies due to water chemistry, pipe condition, and usage patterns tied to both geographic proximity and socioeconomic context (Triantafyllidou et al. 2021). Traditional methods often fall short-statistical models overlook spatial autocorrelation by treating properties independently, while geographic clustering ignores intra-neighborhood demographic diversity (Wang et al. 2017; Chakraborty 2011). To overcome these limitations, our framework leverages demographic variables from the American Housing Survey (AHS)-including housing age, plumbing type, household race, ethnicity, and income-enabling more nuanced, equity-aware contamination risk prediction. This approach aligns with prior evidence that shows strong links between childhood blood lead levels, neighborhood segregation, and poverty (Moody, Darden, and Pigozzi 2016), HUD's risk-ranking models that incorporate AHHS II findings on regional prevalence of lead-based paint hazards (U.S. Department of Housing and Urban Development, Office of Inspector General 2022), and the AHHS II national assessment reporting that approximately 29 million U.S. homes (29.4 %) contain lead-based paint-with significant hazards detected in dust, soil, and deteriorated paint in millions of units nationwide (Bradham et al. 2023).

We hypothesize lead contamination risk can be more accurately predicted by simultaneously modeling spatial dependencies and demographic features of housing information. To test this hypothesis, we develop a hybrid architecture combining Graph Attention Networks (GAT) (Veličković et al. 2018) with gradient boosting (XGBoost) (Chen and Guestrin 2016). GAT component learns adaptive spatial embeddings that capture how contamination risk propagates through water distribution networks and geographic proximity, while XGBoost leverages its strength in tabular data to model complex interactions between spatial

*Corresponding author

context and demographic features.

Our approach addresses three practical deployment challenges for municipal adoption: (i) constructing spatial graphs from privacy-truncated addresses, where the last two digits are removed to protect residents as is common in municipal data systems; (ii) implementing a two-stage training pipeline that learns spatial representations separately from demographic patterns, allowing model updates without full retraining; and (iii) providing interpretable risk scores at the census block group level, aligning with existing municipal planning units and GIS infrastructure. We validate this framework using 38,385 lead tests from Chicago's water quality monitoring program (City of Chicago Department of Water Management 2016–2023), showing that the hybrid GAT-XGBoost model substantially reduces false positive inspections compared to both spatial-only models and current heuristic practices, thereby enabling more efficient allocation of limited municipal inspection resources. Experimental results confirm that jointly modeling spatial and demographic factors captures contamination patterns that neither approach achieves independently. The framework is designed for practical deployment, with computational requirements suitable for municipal IT infrastructure and outputs formatted for integration with existing GIS systems.

Related Work

Lead Contamination Risk Assessment

Traditional lead contamination detection relies on reactive inspections triggered by elevated blood lead levels or randomized sampling, which often fails to identify spatial clusters in time (Present et al. 2022). Abernethy et al. (Abernethy et al. 2018) addressed the Flint water crisis with an active learning ML framework predicting homes most likely to have hazardous lead service lines. Using city infrastructure, pipe material, and water sample records, their model enabled targeted remediation and saved millions in replacement costs. This work demonstrated the value of ML-driven targeting when infrastructure records are incomplete. Hajiseyedjavadi et al. (Hajiseyedjavadi, Blackhurst, and Karimi 2020) applied logistic regression to 5,000+ Toledo property records, prioritizing inspections by housing age and condition, but their approach required precise addresses-limiting use in municipalities with anonymized data. Our work overcomes this by employing truncated address spatial embeddings. Kaplowitz et al. (Kaplowitz, Perlstadt, and Post 2010) showed that socioeconomic predictors explained exposure variance at the census tract level but lacked the block-level granularity needed for targeted municipal action.

Spatial Modeling in Urban Health

Graph neural networks (GNNs) are effective for modeling spatial autocorrelation in urban environmental data. Kipf et al. (Kipf and Welling 2017) introduced GCNs, improving node classification through neighborhood connectivity, but these models require relatively complete graphs-an issue for municipal records where up to 30% of address data may be missing. To address such challenges, Veličković et al. (Veličković et al. 2018) proposed Graph Attention Networks, which use adaptive attention weights to better handle heterogeneous and irregular spatial graphs, such as urban neighborhoods impacted differently by lead sources.

Hybrid Spatial-Tabular Learning

Hybrid models that combine spatial graphs with tabular features generally outperform those using only one data type (Jin et al. 2024). Potash et al. (Potash et al. 2020) developed Chicago's citywide lead exposure prediction system using random forests trained on lead tests, housing records, and demographic data, showing ensemble methods surpass logistic regression for targeting high-risk children. Their validation cohort from city's Women, Infants, and Children (WIC) registry demonstrated both operational feasibility and regulatory alignment. XGBoost has also proven effective for modeling complex contamination patterns when detailed features are available (Chen and Guestrin 2016; Yaseen and Alhalimi 2025). However, naive concatenations of graph embeddings and tabular predictors often give limited gains, motivating joint training strategies (Zhang et al. 2024). Our work advances this by integrating GATs with XGBoost in a two-stage framework that enables privacy-preserving neighborhood representations.

Privacy and Municipal Deployment

Address anonymization and spatial truncation, increasingly used to protect resident privacy in municipal datasets, introduce spatial uncertainty that challenges graph construction and model accuracy (Zeighami et al. 2021). Standard GNN architectures assume precise node coordinates, which are often unavailable or incomplete in public health monitoring systems, and Zeighami et al. report significant accuracy declines under spatial fuzzing-a challenge we mitigate using spatial graph embeddings tolerant to coordinate truncation (Zeighami et al. 2021). Moreover, public health interventions require predictions aligned with regulatory thresholds; for example, the EPA's Lead and Copper Rule mandates action at 15 ppb in water and 10 $\mu g/ft^2$ in dust, underscoring the gap between abstract ML metrics and inspection utility (Agency 2024). Our model calibrates risk scores to these thresholds, directly supporting municipal inspection workflows.

Methodology

Problem Formulation

We formulate block-level lead contamination risk as a binary classification task to support proactive inspections and remediation by municipal agencies. For each census block group i with feature vector $\mathbf{x}_i$, we define a label $y_i \in \{0, 1\}$ indicating whether the proportion of sampled homes with second-draw lead levels ≥ 10 parts per billion (ppb) meets or exceeds a threshold τ. We adopt 10 ppb as our operational threshold, consistent with the revised EPA Lead and Copper Rule Improvements (LCRI), which lowers the action level from 15 ppb to 10 ppb (finalized in 2024, with enforcement beginning 2027). This ensures that our predictions antici-

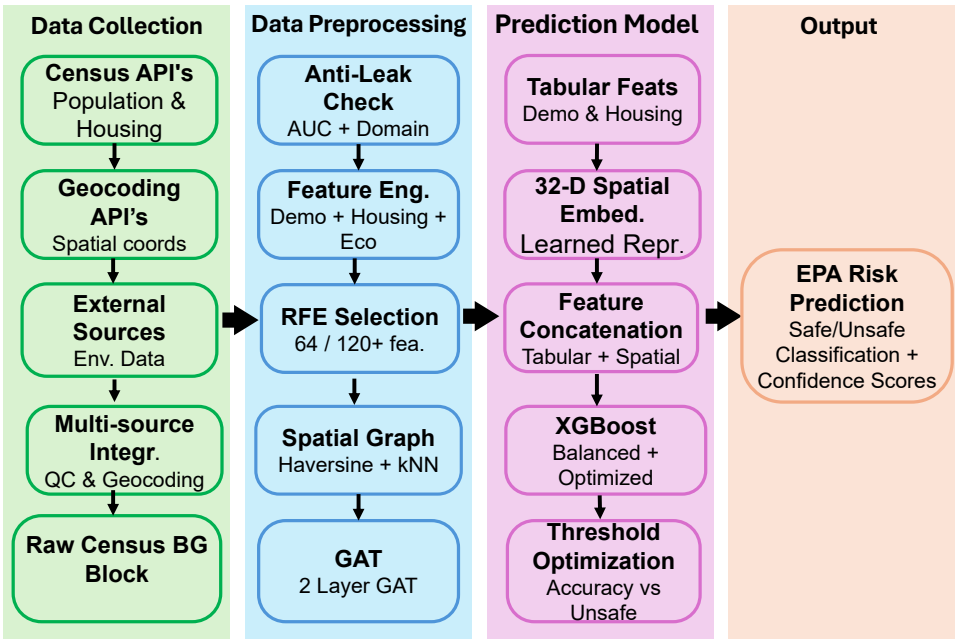


Figure 1: Hybrid GAT-XGBoost framework for block-level lead risk prediction.

pate regulatory changes and align with emerging standards for public health protection.

Formally, the label is defined as:

$$y_i = \begin{cases} 1 & \text{if } \dfrac{|\{h \in H_i : \text{Lead}_h \geq 10\,\text{ppb}\}|}{|H_i|} \geq \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where H_i is the set of recorded lead tests in block group i, and $\tau = 0.10$ is the primary EPA-based threshold (U.S. Environmental Protection Agency 2025). Sensitivity analyses are also performed at $\tau = 0.05$.

Beyond primary classification task of identifying unsafe block groups, our approach leverages rich socioeconomic, racial, and housing data to enable subsequent analysis characterizing disparities between safe and unsafe areas. This multifaceted data integration supports not only risk prediction but also exploratory insights into underlying social and economic factors correlated with contamination risk, which informs equity-focused policy and intervention strategies.

Data Sources and Preprocessing

We utilized publicly available lead testing data from the Chicago Department of Water Management containing 38,385 tests collected from January 2016 to September 2023 (City of Chicago Department of Water Management 2016–2023). Tests were partially anonymized with the last two digits of each address truncated, representing at least 14,673 unique addresses. Each test contained three distinct measurements: first draw (after 6+ hours stagnant), second draw (after 2-minute flush), and third draw (after 5-minute flush). We selected the second draw as our outcome variable as it exhibited the highest median lead concentration, aligning with our objective of identifying households with lead-contaminated drinking water.

Sociodemographic features were extracted from the 2021 ACS (5-year estimates) and 2020 Census using the Census Bureau Application Programming Interface (API). For each block group i, we queried 125 variables across five domains: demographics (48 age-stratified), race/ethnicity (25), housing (24), infrastructure (20), and economics (8).

The API query for the block group i in tract t, county c, and state s is:

$$\text{API}_{i,t,c,s} = \text{Base\_URL} + \sum_{v \in V} v + \text{geo}(i,t,c,s) + \text{key}$$

where V is the set of variable codes (e.g., `B01002_001E` for median age).

Centroid coordinates were obtained via the TIGERweb REST service:

$$(\text{lat}_i, \text{lon}_i) = \text{TIGERweb}(\text{GEOID}_i)$$

where GEOID_i is the 12-digit Federal Information Processing Standards (FIPS) concatenation (state, county, tract, block group).

Batch queries were capped at 500 GEOIDs ($\lceil N/500 \rceil$). Data aggregated to 800 block groups used median imputation with missingness indicators (Van Ness et al. 2023). To avoid target leakage, a train-only three-stage feature screen was applied: (i) exclude variables with terms like `lead`, `ppb`, or `violation`; (ii) drop features with univariate ROC-AUC ≥ 0.95 on training data; (iii) remove near-duplicates ($|\rho| \geq 0.995$).

Feature Normalization and Selection

Feature normalization was performed using the `RobustScaler`, which scales variables by their median and interquartile range to mitigate the influence of outliers (Pedregosa et al. 2011). For feature selection, we applied a modified Recursive Feature Elimination (RFE) procedure with a Random Forest classifier ($n_{\text{trees}} = 120$) (Darst, Malecki, and Engelman 2018).

At each iteration r, feature importance was quantified by the mean decrease in Gini impurity (Appendix Eq. A1). The feature with the smallest importance was removed according to the recursive update rule see (Appendix Eq. A2). This process continued until 64 predictors remained, yielding the final feature set $\mathcal{F}^*$ for downstream modeling.

Spatial Graph Construction

We construct a weighted, undirected graph $G = (V, E)$ where vertices V are census block groups. An edge $(i, j) \in E$ is created if the great-circle (Haversine) distance $d(i, j)$ between centroids is ≤ 1 km, computed using a `BallTree` on radian-transformed coordinates (Pedregosa et al. 2011).

The 1 km threshold is a practical approximation for the service area of distribution mains in Chicago, consistent with urban water engineering practice (Gayton 1950; City of Chicago Department of Water Management 2025). This radius typically connects 8–12 neighbors in dense urban areas while avoiding unwarranted long-range edges. To prevent sparsity in low-density areas, we enforce a minimum degree of $k = 8$ using a k-NN fallback. This range (k[5,10]) reflects common GNN practice, balancing neighborhood context with over-smoothing risk (Veličković et al. 2018; Hamilton, Ying, and Leskovec 2020; **?**). Self-loops are added for stability.

Edges are weighted by inverse distance with a small offset see (Appendix Eq. A3). Graph operations, including augmentation and connectivity checks, are implemented in `networkx` (v3.5).

Graph Attention Network Embedder

Node embeddings that encode spatial dependencies are learned via a two-layer Graph Attention Network (GAT) implemented in `PyTorch` (v2.6.0+cu124) and `PyTorch Geometric` (v2.6.1). The GAT leverages attention mechanisms to weigh neighbors adaptively, with layer hidden dimensions of 64 and 32 units respectively, employing a dropout rate of 0.4 to prevent overfitting.

The update at layer l for node i follows the standard GAT formulation (see Appendix Eq. A4), where $\alpha_{ij}^{(l)}$ denotes learned attention coefficients, $\mathbf{W}^{(l)}$ weight matrices, and $\sigma(\cdot)$ the ReLU activation. With self-loops, attention logits and normalized coefficients are computed through the attention mechanism detailed in Appendix Eq. A5). We use multi-head attention (8 heads in the first layer with concatenation; 4 heads in the second with averaging), ReLU activations, dropout p=0.4, and L_2 weight decay 5×10^{-4}.

Hybrid Architecture

The combination of Graph Attention Networks with XGBoost addresses a fundamental limitation in urban contamination modeling: lead risk exhibits both spatial autocorrelation (contamination clusters geographically) and demographic heterogeneity (risk varies with socioeconomic factors even within clusters).

Our hybrid architecture employs a sequential training approach in which spatial representations are first learned via GAT, then combined with tabular features for final classification. The GAT component generates 32-dimensional spatial embeddings $\mathbf{z}_i$ from input features and neighborhood structure, which are then concatenated with tabular features for XGBoost classification (see Appendix Eq. AEquations 6 and 7). Here $\|$ denotes concatenation, and θ, ϕ are learned parameters for the GAT and XGBoost components respectively.

GAT Training Objective: The GAT embedder is trained end-to-end with a linear classification head using standard cross-entropy loss over the training set (Appendix Eq. A 8). Predictions follow $\hat{p}_i = \sigma(\mathbf{W}_{\text{head}} \cdot \mathbf{z}_i + b)$ where σ is the sigmoid function, and training employs early stopping based on validation macro-F1 score.

XGBoost Integration: After GAT convergence, we freeze the embedder and extract 32-dimensional spatial representations for all nodes. The final feature vector combines 64 RFE-selected tabular features, 32 GAT embeddings, and 2 graph structural features (degree centrality and clustering coefficient), producing a 98-dimensional input vector for XGBoost (Appendix Eq. A 9).

Class Imbalance Handling: Given the 34.2% unsafe rate at $\tau = 0.05$, we employ a scale weight adjustment proportional to the inverse class frequencies, which yields approximately 1.92 for the positive class weight (Appendix Eq. A 10).

This hybrid architecture exploits complementary strengths: GAT captures spatial autocorrelation through neighborhood aggregation while XGBoost handles high-dimensional demographic features. The sequential design enables complex spatial-demographic interactions that neither method achieves independently—GAT embeddings encode the spatial "where" of contamination patterns, while XGBoost classification incorporates the demographic "why" for complete risk characterization as shown in Figure 1.

Method	BA	Recall (Unsafe)	Precision (Unsafe)	FP (%)
CPH	0.486	0.354	0.643	12.8
SVM	0.636	0.840	0.431	23.6
GAT-only	0.682	0.832	0.733	18.2
Tabular XGBoost	0.797	0.876	0.715	14.2
GAT-XGBoost (Ours)	**0.848**	**0.916**	**0.835**	**10.1**

Table 1: Performance comparison across methods at the 5% contamination threshold. Best results are in bold.

Classification and Threshold Optimization

The final classification combines GAT embeddings with tabular features, producing combined representations for an XGBoost classifier. The model was trained with 600 estimators, max depth 4, learning rate 0.05, and 0.9 subsampling rate, balancing model complexity and generalization.

To align with operational use, we tuned the decision threshold over $t \in [0.15, 0.85]$, selecting the value that maximized accuracy subject to unsafe-class recall ≥ 0.80, consistent with public health risk management standards (U.S. Environmental Protection Agency 2025; Illinois Department of Public Health 2025).

The theoretical training complexity of the hybrid model is provided in Appendix Eq. A11. The census block groups were partitioned using `StratifiedKFold` (scikit-learn, random_state=42). For each fold, 60% of the data was used for training, 20% for validation, and 20% for testing. To prevent leakage, imputation, feature selection, and scaling were fitted only to training data. GAT embeddings were trained for up to 150 epochs with early stopping (patience=15) based on validation macro-F1. The best checkpoint was used to generate embeddings for all nodes, which were then combined with tabular features for XGBoost training with imbalance-adjusted weights.

We report balanced accuracy (BA), precision, recall, F1-score, and confusion matrices.

Hyperparameter Selection

Hyperparameters were selected through a combination of grid search, best practices and from prior GNN/XGBoost studies. The complete search ranges and selected values are summarized in Appendix Table A1.

For the GAT, hidden dimensions of 64→32, dropout $p = 0.4$, learning rate 0.005, and weight decay 5×10^{-4} follow common setups in GNN benchmarks (Veličković et al. 2018; Kipf and Welling 2017). XGBoost parameters (600 estimators, depth 4, learning rate 0.05, subsample 0.9, $\lambda = 1.0$) reflect stable configurations used in tabular modeling (Chen and Guestrin 2016). Graph construction parameters (radius=1 km, min degree $k = 8$) were selected as described in Section *Spatial Graph Construction*, grounded in urban water engineering practice (Gayton 1950; City of Chicago Department of Water Management 2025) and GNN literature (Hamilton, Ying, and Leskovec 2020). RFE retained 64 features, balancing compactness with predictive signal in line with prior urban risk modeling (Potash et al. 2020; Zhang et al. 2024).

Baseline Methods

We benchmark 4 approaches. **Current Practice Heuristic (CPH)** follows Chicago's ZIP-code strategy targeting pre-1986 housing (Illinois Department of Public Health 2025). **Support Vector Machine (SVM)** uses an RBF kernel with grid-searched hyperparameters on tabular features (Suthaharan 2016). **GAT-only** is a two-layer Graph Attention Network (64, 32 units) leveraging spatial structure without tabular inputs (Veličković et al. 2018). **Tabular XGBoost** trains gradient-boosted trees (600 estimators, max depth 4, learning rate 0.05) on demographic and socioeconomic features (Chen and Guestrin 2016). All baselines use identical preprocessing.

Experiments

We conducted extensive experiments to evaluate the effectiveness of the proposed hybrid GAT-XGBoost framework for lead contamination risk prediction. Specifically, we examine how the proposed framework performs compared to current practice and component baselines, how spatial correlations captured through graph attention contribute to prediction accuracy, and how model performance varies across contamination thresholds.

We evaluate model performance using four metrics: Balanced Accuracy (BA), the mean of per-class recall rates to mitigate class imbalance; Recall (Unsafe), the proportion of unsafe blocks correctly identified; Precision (Unsafe), the fraction of predicted unsafe blocks that are truly unsafe; and False Positives (FP), the number of safe blocks incorrectly flagged as unsafe.

Performance Comparison

Table 1 presents the performance comparison across all methods at the 5% contamination threshold. The proposed GAT-XGBoost framework achieves the highest balanced accuracy of 0.848, significantly outperforming all baselines ($p < 0.01$, paired t-test across folds). The hybrid approach reduces false positives by 44.4% compared to GAT-only and 28.6% compared to Tabular XGBoost.

The results in Table 1 show that the combination of spatial and tabular information identifies contamination patterns not captured by either approach alone. The hybrid model yields 15 false positive inspections, compared to 27 for GAT-only, 19 for CPH, 21 for XGBOOST, and 35 for SVM, indicating improved efficiency.

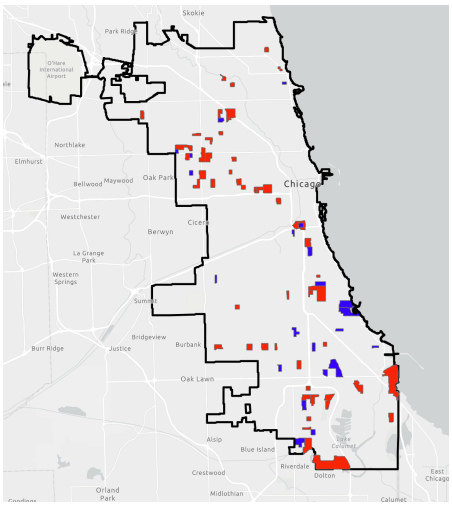
Performance Across Contamination Thresholds

We evaluate model robustness across different contamination thresholds to assess operational flexibility:

	5% Threshold		**10% Threshold**	
Method	**BA**	**Recall**	**BA**	**Recall**
CPH	0.486	0.354	0.520	0.354
SVM	0.636	0.840	0.658	0.697
GAT-only	0.682	0.832	0.679	0.706
Tabular XGBoost	0.797	0.876	0.724	0.812
GAT–XGBoost	**0.848**	0.916	**0.741**	**0.861**

Table 2: Model performance at different contamination thresholds.

The framework maintains balanced accuracy above 0.74 across both thresholds, including the more stringent 10% threshold, indicating adaptability to varying regulatory standards. Improvements over all baselines at both thresholds suggest that the hybrid approach is robust across different evaluation settings Table 2.



(a) Model predictions with test data

Figure 2: GAT-XGBoost predictions and ground truth: red dots mark unsafe predictions ($\geq$ 10 ppb), and blue dots mark safe predictions ($<$ 10 ppb).

Ablation Study

To understand the contribution of each component, we systematically evaluate variants of the proposed framework:

Model Variant	BA	Δ from Full
GAT-XGBoost (Full Model)	0.848	–
w/o spatial embeddings	0.797	-5.1 pp
w/o demographic features	0.682	-16.6 pp

Table 3: Ablation study showing the impact of removing spatial embeddings and demographic features on model performance ($\tau = 0.05$).

Table 3 indicates that demographic features contribute most significantly to performance (16.6%), while spatial embeddings provide an additional 5.1% point improvement. These results indicate that both types of information play meaningful roles in model. The model achieves optimal performance at $\rho = 1.0$ km, capturing local contamination patterns while avoiding noise from distant neighborhoods. This aligns with environmental health literature suggesting contamination sources primarily affect immediate surroundings.

Discussion

The 44.4% reduction in false positives compared to GAT-only (15 vs. 27) and 57.1% reduction compared to SVM (15 vs. 35) demonstrate substantial gains in operational efficiency; for a typical quarterly inspection cycle targeting 100 block groups, this equates to avoiding approximately 12 unnecessary inspections relative to GAT-only and 20 relative to SVM. These improvements stem from the model's integration of spatial relationships, which proves especially valuable in differentiating superficially similar neighborhoods, with feature importance analysis showing that GAT-derived embeddings rank among the top predictors by capturing contamination spillover effects that traditional methods such as SVM cannot model effectively.

Validation Against Domain Knowledge

To ensure model predictions align with established environmental health principles, we examine the relationship between predicted risk and known contamination correlates:

Variable	Safe	Unsafe	d
Per Capita Income ($)	74,588	42,684	0.87
Median Home Value ($)	383,145	302,982	0.34

Table 4: Socioeconomic differences between predicted safe and unsafe groups.

The effect size for per capita income (Cohen's $d = 0.87$) confirm that model predictions correlate strongly with established risk factors(U.S. Department of Housing Urban Development 2022), enhancing interpretability and stakeholder trust table 4. In contrast, the smaller effect size for median home value ($d = 0.34$) suggests that property value alone is a less direct proxy for lead contamination risk. These associations improve the interpretability of predictions and may help inform targeted intervention strategies in resource-limited settings.

The performance hierarchy (GAT-XGBoost $>$ XGBoost $>$ GAT $>$ SVM $>$ CPH) reveals important insights into the nature of lead contamination patterns. Traditional machine learning approaches such as SVM, despite using sophisticated kernels, achieve only 0.636 balanced accuracy, suggesting that lead risk cannot be adequately captured through standard feature transformations alone. The performance of GAT-only model (0.682) demonstrates that spatial relationships provide a valuable signal, but substantial gain of the full hybrid model (0.848) confirms that both spatial and demographic information are essential for accurate prediction and in help reduce lead crises.

Conclusion and Future Work

The block group–level aggregation required for privacy can mask variation within neighborhoods, potentially misdirecting inspections. The model's reliance on static 2016–2023 data also ignores Chicago's ongoing lead service line replacements, leading to possible overestimation of risk in remediated areas. GAT retraining is required when boundaries change, limiting real-time updates, and the exclusive use of second-draw samples may underestimate exposure in some homes. Future work will explore hierarchical models incorporating property-level features, incremental learning to update risk with remediation data, and external validation in cities with differing infrastructure and water chemistry. Additional notes on deployment constraints and operational trade-offs are provided in Appendix under (Deployment Considerations) for deployment details.

Appendix A

Equation A1: Feature importance quantification by mean decrease in Gini impurity:

$$I_j^{(r)} = \frac{1}{n_{\text{trees}}} \sum_{t=1}^{n_{\text{trees}}} \Delta\text{Gini}_{j,t}^{(r)} \tag{1}$$

Equation A2: Recursive feature elimination iteration:

$$\mathcal{F}^{(r+1)} = \mathcal{F}^{(r)} \setminus \left\{ \arg\min_{j \in \mathcal{F}^{(r)}} I_j^{(r)} \right\}, \quad \mathcal{F}^{(0)} = \{1, 2, \ldots, p\} \tag{2}$$

Equation A3: Edge weighting by inverse distance with offset:

$$w_{ij} = \frac{1}{d(i,j) + 0.1} \tag{3}$$

where $d(i,j)$ is the Haversine distance (in kilometers) between block group centroids, and the offset prevents division by zero for colocated nodes.

Equation A4: GAT node update at layer l:

$$\mathbf{h}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)} \right) \tag{4}$$

Equation A5: Attention logits and normalized coefficients:

$$e_{ij}^{(l)} = \text{LeakyReLU}\Big(\mathbf{a}^{(l)\top} \big[\mathbf{W}^{(l)} \mathbf{h}_i^{(l-1)} \,\|\, \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)} \big] \Big),$$
$$\alpha_{ij}^{(l)} = \frac{\exp\big(e_{ij}^{(l)}\big)}{\sum_{k \in \mathcal{N}(i)} \exp\big(e_{ik}^{(l)}\big)}. \tag{5}$$

Equation A6: GAT spatial embedding generation:

$$\mathbf{z}_i = \text{GAT}_\theta(\mathbf{x}_i, \mathcal{N}(i)) \in \mathbb{R}^{32} \tag{6}$$

Equation A7: XGBoost final classification:

$$\hat{y}_i = \text{XGBoost}_\phi([\mathbf{x}_i \| \mathbf{z}_i]) \tag{7}$$

Equation A8: GAT cross-entropy training loss:

$$\mathcal{L}_{\text{GAT}} = -\frac{1}{|V_{\text{train}}|} \sum_{i \in V_{\text{train}}} \left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right] \tag{8}$$

Equation A9: Final feature vector composition:

$$\mathbf{f}_i = [\mathbf{x}_i^{(1:64)} \| \mathbf{z}_i^{(1:32)} \| \text{deg\_cent}_i \| \text{clust\_coeff}_i] \in \mathbb{R}^{98} \tag{9}$$

Equation A10: Class imbalance scale weight:

$$\text{scale\_pos\_weight} = \frac{|V_{\text{safe}}|}{|V_{\text{unsafe}}|} \approx 1.92 \tag{10}$$

Equation A11: Computational complexity of the hybrid model:

$$\mathcal{O}_{\text{GAT}} = |E| \cdot d \cdot h, \quad \mathcal{O}_{\text{XGB}} = n \cdot m \cdot \text{depth} \cdot \text{trees}$$

where $|E|$ is the number of edges, d the feature dimension, h the hidden size, n the number of samples, and m the number of features.

Component	Parameter	Search Space	Selected Value
GAT	Hidden dimensions	[128, 64], [32, 16]	[64, 32]
	Dropout	0.2, 0.3, 0.4, 0.5	0.4
	Learning rate	0.001, 0.003, 0.005, 0.01	0.005
	Weight decay	$10^{-5}, 5 \times 10^{-4}, 10^{-3}$	5×10^{-4}
XGBoost	n_estimators	300, 450, 600, 1000	600
	max_depth	3, 4, 5, 6	4
	learning_rate	0.01, 0.05, 0.1, 0.3	0.05
	subsample	0.7, 0.8, 0.9, 1.0	0.9
	reg_lambda	0.1, 0.5, 1.0, 2.0	1.0
Graph	Radius (km)	0.5, 1.0, 1.5, 2.0	1.0
	Min degree k	5, 8, 10	8
Feature Sel.	RFE features	32, 48, 64, 96, 128	64

Table A1: Table: Hyperparameter selection for GAT and XGBoost components.

Appendix B: Deployment Considerations

Technical Feasibility. The model is computationally practical for municipal IT environments. Training completes in approximately one hour on a single NVIDIA V100 GPU, and inference for all 800 block groups completes in under 15 minutes. The model requires less than 500MB memory for deployment and exports to ONNX format for platform-agnostic integration.

Integration with Existing Systems. Model outputs (risk scores per block group) are formatted as GeoJSON files compatible with municipal GIS systems including ArcGIS and QGIS, enabling seamless integration with existing inspection scheduling workflows.

Acknowledgements

This work was supported by the National Science Foundation under Award No. 2230180. The authors also thank their project partners, including MassDEP; the Drinking Water Treatment Plants and Water and Sewer Departments of Andover, Dracut, Lawrence, and Lowell; YWCA Lowell; the Merrimack River Watershed Council; and the teachers and students of K-12 schools in Massachusetts

References

Abernethy, J.; Chojnacki, A.; Farahi, A.; Schwartz, E. M.; and Webb, J. 2018. Active Remediation: The Search for Lead Pipes in Flint, Michigan. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 5–14. London, United Kingdom: ACM.

Agency, U. E. P. 2024. Lead and Copper Rule: Proposed Improvements. Technical report, U.S. EPA.

Bradham, K. D.; Nelson, C. M.; Sowers, T. D.; Blackmon, M. D.; Kovalcik, K.; et al. 2023. A national survey of lead and other metal(loids) in residential drinking water in the United States. *Journal of Exposure Science & Environmental Epidemiology*, 33: 160–167.

Centers for Disease Control and Prevention. 2016. Community Assessment for Public Health Emergency Response (CASPER) After the Flint Water Crisis: May 17–19, 2016, Flint Michigan. Technical report, CDC.

Chakraborty, J. 2011. Disproportionate Proximity to Environmental Health Hazards: Methods, Models, and Measurement. *American Journal of Public Health*, 101(Suppl 1): S27–S36.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.

City of Chicago Department of Water Management. 2016–2023. Water Quality Test Results Database. https://chicagowaterquality.org/. Accessed: 2025-07-30.

City of Chicago Department of Water Management. 2025. Operations Distribution. https://www.chicago.gov/city/en/depts/water/provdrs/ops.html. Accessed: 2025-08-18.

City of Chicago Office of Inspector General. 2018. Department of Buildings Complaint-Based Inspections Audit. Technical report.

Darst, B. F.; Malecki, K. C.; and Engelman, C. D. 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genomic Data*, 19(1): 55.

Gayton, L. D. 1950. Chicago Area Water Supply. Technical report, Illinois State Water Survey.

Hajiseyedjavadi, S.; Blackhurst, M.; and Karimi, H. A. 2020. A Machine Learning Approach to Identify Houses with High Lead Tap Water Concentrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13300–13305.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2020. Graph Representation Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3): 1–159.

Illinois Department of Public Health. 2025. Lead Testing High-Risk ZIP Codes Notice. https://dph.illinois.gov/resource-center/news/2025/july/release-20250701.html. Accessed: 2025-07-30.

Jin, G.; Liang, Y.; Fang, Y.; Shao, Z.; Huang, J.; Zhang, J.; and Zheng, Y. 2024. Spatio-Temporal Graph Neural Networks for Predictive Learning in Urban Computing: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(10): 5388–5408.

Kaplowitz, S. A.; Perlstadt, H.; and Post, L. A. 2010. Comparing Lead Poisoning Risk Assessment Methods: Census Block Group Characteristics vs. Zip Codes as Predictors. *Public Health Reports*, 125(2): 234–245.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*.

Moody, H. A.; Darden, J. T.; and Pigozzi, B. W. 2016. Relationship of Neighborhood Socioeconomic Differences and Racial Residential Segregation to Childhood Blood Lead Levels in Metropolitan Detroit. *Journal of Urban Health*, 93(5): 802–817.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Potash, E.; Ghani, R.; Walsh, J.; Jorgensen, E.; Lohff, C.; Prachand, N.; and Mansour, R. 2020. Validation of a Machine Learning Model to Predict Childhood Lead Poisoning. *JAMA Network Open*, 3(9): e2012734.

Present, E.; Keene, D.; Brown, M. J.; et al. 2022. Detecting New Sources of Childhood Environmental Lead Exposure Using a Statistical Surveillance System, 2015–2019. *American Journal of Public Health*, 112(S7): S715–S722.

Suthaharan, S. 2016. Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207–235. Springer.

Triantafyllidou, S.; Burkhardt, J.; Tully, J.; Cahalan, K.; DeSantis, M.; Lytle, D.; and Schock, M. 2021. Variability and sampling of lead (Pb) in drinking water: Assessing potential human exposure depends on the sampling protocol. *Environment International*, 146: 106259.

U.S. Department of Housing Urban Development. 2022. American Housing Survey (AHS): Overview and Methodology. https://www.census.gov/programs-surveys/ahs.html. Accessed: 2025-08-17.

U.S. Department of Housing and Urban Development, Office of Inspector General. 2022. Risk Indicators of Lead-Based Paint Hazards in Public Housing Agencies. Technical report, HUD OIG. Accessed: 2025-08-17.

U.S. Environmental Protection Agency. 2024. Proposed Lead and Copper Rule Improvements (LCRI). Technical report, U.S. EPA. URL: https://www.epa.gov/ground-water-and-drinking-water/proposed-lead-and-copper-rule-improvements. Accessed: 2025-07-30.

U.S. Environmental Protection Agency. 2025. Lead and Copper Rule 90th Percentile Calculator. https://www.epa.gov/region8-waterops/lead-and-copper-rule-90th-percentile-calculator. Accessed: 2025-08-14.

Van Ness, M.; Bosschieter, T. M.; Halpin-Gregorio, R.; and Udell, M. 2023. The Missing Indicator Method: From Low to High Dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5004–5015. New York, NY, USA: Association for Computing Machinery.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.

Wang, Q.; Phillips, N. E.; Small, M. L.; and Sampson, R. J. 2017. Multi-Contextual Segregation and Environmental Justice Research: Toward Fine-Scale Spatiotemporal Approaches. *International Journal of Environmental Research and Public Health*, 14(10): 1205.

Yaseen, Z. M.; and Alhalimi, F. L. 2025. Heavy metal adsorption efficiency prediction using biochar properties: a comparative analysis for ensemble machine learning models. *Scientific Reports*, 15(1): 13434.

Zeighami, S.; Ahuja, R.; Ghinita, G.; and Shahabi, C. 2021. A Neural Database for Differentially Private Spatial Range Queries. In *Proceedings of the 2021 ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 528–531.

Zhang, P.; Yang, M.; Wang, Y.; Yang, T.; Yu, H.; and Yan, X. 2024. Integrating Metro Passenger Flow Data to Improve the Classification of Urban Functional Regions Using a Heterogeneous Graph Neural Network. *International Journal of Digital Earth*, 17(1): 2443468.