Geospatial Deep Learning for Predicting Public Health Risks Through Multimodal Analysis of Social Media Sentiment and Urban Environmental Imagery

Jasmina Weimer
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
weimer1981@unr.edu

Klaus Dieter Dietz
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
dietz835@ucf.edu

Renzo Grisoni
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
hellorenzo@uc.edu

Prisha Oak
School of Computing, Clemson University, Clemson, SC, USA.
prishamail@clemson.edu

Abstract

Public health risk varies sharply across small urban areas, but conventional surveillance is often too sparse, delayed, or aggregated to support neighborhood-scale screening. Social media sentiment and street-level imagery provide complementary observations: text streams capture expressed distress, lifestyle change, and service complaints, while environmental images encode visible greenness, built density, disorder, and exposure conditions. This manuscript develops a geospatial deep learning framework that fuses geotagged social media sentiment with urban environmental imagery and spatial covariates to estimate area-time public health risk. The contribution is a complete methodological architecture rather than a reported field trial. The framework specifies reliability-weighted sentiment aggregation, pretrained image encoding, gated multimodal fusion, graph-based spatial regularization, calibrated probabilistic risk estimation, and uncertainty-aware decision thresholds. It treats spatial support mismatch, geotag imprecision, platform selection bias, temporal misalignment between posts and images, and visually confounded urban form as first-class modeling problems. The central argument is that multimodal gains depend less on adding modalities than on representing when each modality is trustworthy at a given location and time. Spatial blocking, calibration, subgroup assessment, and decision-cost analysis are therefore integrated into the evaluation logic. The resulting framework supports public health screening and urban planning as a triage aid, not as a substitute for clinical or epidemiological diagnosis. It also provides an auditable basis for assessing where social media signals enrich environmental imagery and where imagery corrects sentiment-only blind spots.

Keywords

multimodal deep learning; geospatial public health; social media sentiment; street-level imagery; risk calibration; graph neural networks

1. Introduction

Fine-grained public health risk assessment is constrained by the spatial and temporal resolution of conventional surveillance. Administrative counts, survey estimates, and clinical records are often reported at coarse intervals, while risk-relevant exposures and behaviors vary at the block or neighborhood scale. Large digital data streams appear to fill this gap, but they also create well-known inferential hazards. The widely discussed mismatch between Google Flu Trends and later epidemiological evidence illustrates how unmodeled drift, selection effects, and feedback can undermine big-data health claims [1]. A credible geospatial model must therefore specify not only predictive architecture but also the conditions under which digital observations represent health-relevant population states.

Social media have a longer record in public health informatics. Linguistic signals from posts have been used to detect illness trends and to study health behavior [2]. Geotagged posts can predict county-level health outcomes and provide neighborhood-scale context that surveys cannot resolve [3]. This line of work connects with the broader volunteered geographic information paradigm, in which citizens act as distributed sensors of urban conditions [4]. Yet text alone often conflates complaint, humor, reporting, and actual exposure. A distress-heavy post about housing may be relevant to mental-health risk but not to respiratory risk; a complaint about trash may indicate environmental nuisance without individual exposure.

Street-level imagery provides a complementary view. Deep learning applied to Google Street View and similar sources can estimate neighborhood demographic composition and physical form at scale [5]. Imagery captures visible greenness, sidewalk condition, facade quality, traffic infrastructure, and visible disorder, which text streams rarely describe consistently. Joint modeling of geotagged social media and street-view imagery with deep neural networks has been demonstrated for urban streetscape representation, indicating that the two data types can support a shared spatial learning problem [6]. What remains undertheorized is how such a representation should be converted into calibrated public health risk, how modality reliability should vary across space and time, and how evaluation should avoid spatial leakage.

The present work addresses that gap by specifying a geospatial multimodal risk inference framework for area-time public health screening. Its central contribution is an integrated architecture in which sentiment, imagery, and geospatial structure are fused through reliability-aware gating, regularized by spatial graph propagation and spatial autocorrelation, and converted into calibrated risk probabilities with uncertainty-aware decision thresholds. The framework is designed for settings in which a risk label or proxy is available from surveillance, survey, or administrative sources for some areas and times, and the model is intended to interpolate or forecast risk where direct labels are unavailable. The manuscript develops the data semantics, mathematical formulation, architectural components, loss structure, calibration logic, and validity threats. It does not report an empirical deployment; rather, it provides a technically substantive basis for reproducible implementation and evaluation.

2. Problem Formulation and Data Semantics

The unit of analysis is an area a at time t, where a may be a census tract, a regular grid cell, or a locally defined health planning zone. The temporal unit may be a week or month, chosen to balance label availability against the dynamics of social media and environmental conditions. The target Y_{a,t} is a latent or proxy risk state, such as elevated community mental-health burden, environment-sensitive symptom risk, or a composite vulnerability index derived from surveillance and survey sources. Because public health labels are often counts of events, the risk state may be modeled as a probability of exceeding a threshold, a rate, or a continuous score after suitable transformation.

The conceptual risk function is written as

$$Y_\{a,t\} = f(X\^\{text\}_\{a,t\}, X\^\{img\}_\{a,t\}, X\^\{geo\}_\{a\}, X\^\{time\}_\{t\}) + \varepsilon_\{a,t\}$$

where X^{text}_{a,t} denotes aggregated social media signals in area a and time t, X^{img}_{a,t} denotes environmental image representations assigned to that area, X^{geo}_{a} contains static or slowly changing spatial covariates, and X^{time}_{t} captures seasonality, day-of-week effects, and secular trends. The term ε_{a,t} represents unobserved risk determinants and measurement error. The function f is not estimated in this manuscript; it defines the object that the proposed architecture approximates.

Three data semantics issues are decisive. First, spatial support differs across modalities. A geotagged post has a point location, an image has a capture point and viewing direction, and a health label is aggregated over an administrative polygon. Assigning point observations to polygons discards within-area variation; interpolating labels to points can create false precision. The framework treats each observation as a noisy sample of the area-time state and preserves its location, timestamp, and confidence metadata. Second, missingness is informative. Areas with few posts may be low-population, digitally disengaged, or affected by platform moderation rather than genuinely low risk. Third, temporal alignment is imperfect. Street-view images may predate or postdate the target period, and sentiment can change rapidly around local events. The model therefore needs explicit reliability variables rather than treating all records as equally informative.

3. Multimodal Geospatial Architecture

3.1 Textual Sentiment and Temporal Signals

The text branch converts each post into a sentiment valence s_i and a reliability weight m_i. A transformer encoder provides contextual token representations and supports sentiment classification or regression, following the attention-based sequence modeling paradigm [7]. The post-level reliability m_i combines geotag precision, language confidence, duplicate or bot likelihood, and topical relevance. Posts with coarse geotags or ambiguous location references receive lower weight. The area-time sentiment signal is then aggregated as

$$S_{a,t} = \sum_{i \in P_{a,t}} w_i m_i \frac{s_i}{\sum_{i \in P_{a,t}} w_i m_i}$$

where P_{a,t} is the set of posts assigned to area a and time t, s_i is sentiment valence, m_i is post reliability, and w_i is a spatial kernel weight. A Gaussian kernel w_i = exp(−d_i^2/(2σ^2)) assigns lower influence to posts farther from the area centroid or from a representative population-weighted point. The denominator prevents areas with many low-quality posts from dominating. This aggregation yields an interpretable sentiment feature, but

it is not treated as a direct measure of population distress. The model also preserves post volume, lexical diversity, and topic composition as auxiliary inputs because a high volume of negative posts may indicate either widespread distress or a localized service disruption.

Temporal modeling uses lagged sentiment, rolling means, and event indicators. A sequence encoder can capture persistence and abrupt changes without requiring the sentiment branch to model spatial dependence directly. This separation is useful because sentiment dynamics and spatial structure have different failure modes: sentiment models are sensitive to platform drift and sarcasm, while spatial models are sensitive to boundary specification and spillover.

3.2 Urban Environmental Image Encoding

The image branch summarizes street-level environmental conditions. Images are sampled at or near area locations and encoded by a convolutional or transformer-based vision model. A residual network provides a robust baseline for visual representation learning [8]. For each area, the image representation is

$$E_a \;=\; (1/|I_a|)\,\sum_\{j \in I_a\}\,\varphi(I_j;\;\theta_img)$$

where I_a is the set of images assigned to area a, φ is the image encoder, I_j is an image, and θ_img denotes encoder parameters. The encoder may be pretrained on scene or place recognition tasks and fine-tuned only when sufficient labels exist. In addition to the pooled embedding, the framework retains interpretable visual indicators derived from semantic segmentation or object detection: green view index, sky view, building density, road surface, sidewalk presence, visible waste, and facade condition. These indicators connect to environmental health mechanisms such as restorative green space, heat exposure, walkability, and physical disorder [12].

Image reliability is explicitly modeled. An image receives lower weight if its capture date is far from t, if its location is uncertain, or if the view is dominated by a construction site or temporary event. This matters because a single visible renovation can make a block appear disordered for years in an image archive. The image branch should also record viewing direction and sampling density, since a block with one downward-facing image is poorly represented relative to one with panoramic coverage.

3.3 Spatially Aware Fusion and Graph Regularization

The text, image, and geospatial branches produce representations h_{S,a,t}, h_{E,a,t}, and h_{G,a}, respectively. Simple concatenation assumes that every modality is equally reliable at every location and time, an assumption that is rarely defensible. The framework therefore uses gated attention to produce a fused representation:

$$z_\{a,t\} \;=\; \sum_\{m \in \{S,E,G\}\}\, g_\{m,a,t\}\; h_\{m,a,t\}$$

where m indexes the sentiment, environmental image, and geospatial covariate branches, h_{m,a,t} is the branch representation, and g_{m,a,t} is its gate weight. The gate is computed as

$$g_\{m,a,t\} \;=\; softmax_m(q\text{^}T\; tanh(W\; h_\{m,a,t\} \;+\; b))$$

where q is a learned query vector, W and b are learned parameters, and softmax_m normalizes the weights across modalities. This formulation allows the model to down-weight sentiment in areas with sparse posting and to down-weight imagery when temporal alignment

is poor. The learned gates are not treated as causal attributions; they are reliability indicators that should be audited against data quality metadata.

Spatial dependence is modeled with a graph over areas. Nodes represent areas, and edges connect adjacent areas or areas with strong commuting or service linkage. Graph propagation updates area representations using neighborhood information:

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

where H^{(l)} contains node representations at layer l, A is the adjacency matrix, D is its degree matrix, W^{(l)} is a layer-specific weight matrix, and σ is a nonlinear activation. This graph convolution follows the spectral-inspired formulation introduced for semi-supervised learning [9]. The graph is not assumed to be complete or homogeneous: edge weights can reflect distance, administrative adjacency, or estimated mobility. Multimodal fusion is consistent with the view that different data sources encode partially independent information and that fusion should account for modality-specific noise and missingness [10].

4. Risk Estimation, Calibration, and Decision Use

The fused representation z_{a,t} feeds a risk head that estimates the probability of an elevated public health risk state. The choice of likelihood depends on the label: a Bernoulli likelihood for thresholded risk events, a Poisson or negative-binomial likelihood for counts, or a regression likelihood for continuous indices. Social media signals have been used to predict depression-related states and mental-health risk, but those models are sensitive to self-selection and linguistic variation [11]. The risk head should therefore include area-level intercepts, temporal smoothness, and spatial autocorrelation terms rather than relying solely on neural features.

The training objective combines predictive loss, cross-modal alignment, spatial regularization, and parameter shrinkage:

$$L = L_{risk} + \lambda_1 L_{align} + \lambda_2 \sum_{a,b} A_{ab} ||r_a - r_b||^2 + \lambda_3 ||\theta||_2^2$$

where L_{risk} is the negative log-likelihood or binary cross-entropy, L_{align} encourages matched text-image-area triplets to have similar representations, r_a is the risk logit for area a, A_{ab} is the spatial adjacency weight, θ denotes model parameters, and λ_1, λ_2, and λ_3 control the relative contributions. The spatial penalty encodes the expectation that neighboring areas often share unobserved risk determinants, a property formalized in spatial association analysis [13]. It is not a substitute for substantive covariates; it regularizes otherwise unstable estimates in sparse areas.

Uncertainty is represented at two levels. Aleatoric uncertainty captures irreducible noise in labels and observations, while epistemic uncertainty reflects limited evidence for a given area-time combination. Bayesian approximations or ensemble variance can provide both [14]. Calibration is assessed separately from discrimination because a model can rank areas well while assigning probabilities that are systematically too high or low. Modern neural networks are often miscalibrated, and post hoc calibration methods such as temperature scaling or isotonic regression may be required [15]. A calibrated risk probability supports decision thresholds that reflect intervention costs:

$$Flag_{a,t} = 1[P(Y_{a,t} = 1 \mid X) \geq \tau_{a,t}]$$

where Flag_{a,t} indicates whether an area-time unit is flagged for review and τ_{a,t} is a threshold chosen from decision costs, capacity constraints, or public health policy. The

threshold should be evaluated with a decision-curve or cost-benefit analysis rather than a default value of 0.5. A screening model that over-flags low-capacity areas can waste resources and stigmatize neighborhoods; one that under-flags high-risk areas can delay outreach.

5. Analytical Assessment and Validity Threats

The framework requires an evaluation design that respects spatial and temporal dependence. Random cross-validation leaks information because neighboring areas and adjacent time periods are correlated. Spatial blocking, temporal holdout, and clustered cross-validation are therefore necessary [16]. Evaluation should compare at least five baselines: text-only sentiment, imagery-only environmental features, static geospatial covariates, early concatenation fusion, and a non-spatial multimodal model. These comparisons clarify whether gains come from multimodal content, spatial regularization, or simply increased model capacity. External validation in a held-out city or health jurisdiction is more informative than in-sample improvement, but it also requires checking platform coverage and image recency.

Several failure mechanisms are foreseeable. Platform drift can change the meaning and volume of sentiment independently of health risk. Geotag sparsity can make wealthy or digitally active areas appear more variable than they are. Image archives can contain outdated views, seasonal mismatch, and construction artifacts. Label leakage can occur if surveillance data are generated from the same complaints or service records that also shape social media discussion. Confounding is substantial: an area may have high distress and poor visual conditions because of poverty, disinvestment, or chronic stress rather than because either modality causes the other. The model should therefore be interpreted as a risk screening association, not as a causal mechanism.

Fairness and privacy require explicit treatment. Algorithms used in health management can encode racial and socioeconomic bias through label choice and access patterns [17]. Subgroup calibration, error-rate parity, and spatial error mapping should be reported across demographic and socioeconomic strata. Social media posts and street-level images can be re-identifying when combined at fine resolution. The framework should use aggregation, geotag coarsening, text de-identification, and restricted access to image locations. These safeguards are part of the model's validity, not an external ethical add-on.

6. Technical Discussion and Implications

The framework's main technical claim is that multimodal geospatial risk prediction is a reliability-weighted inference problem, not a simple feature-concatenation problem. Sentiment and imagery observe different aspects of urban life: sentiment captures expressed experience and attention, while imagery captures visible environmental conditions. Their combination is most useful when the model can learn where each source is trustworthy. The gating mechanism provides that capability at the area-time level, while graph propagation shares information across neighboring areas and reduces variance in sparse locations. Calibration and uncertainty then translate model outputs into decision-relevant probabilities.

This design has practical implications for public health screening. A city health department could use the model to prioritize outreach, inspection, or service referral in areas with elevated predicted risk and high uncertainty. An urban planning unit could examine image-derived environmental indicators that co-occur with persistent distress signals. The model should not be used to diagnose individuals, allocate clinical care, or justify punitive

enforcement. Its appropriate role is triage and hypothesis generation, followed by epidemiological verification and community engagement.

The architecture also suggests a research agenda centered on robustness rather than marginal accuracy gains. Future work should examine platform shift, cross-city transfer, image recency weighting, and the sensitivity of gates to missing modalities. Such work requires prospective validation and should report calibration, spatial error, and subgroup performance alongside aggregate metrics. The framework offered here provides a coherent structure for those evaluations because it separates representation learning, spatial regularization, risk estimation, calibration, and decision rules.

7. Conclusion

This manuscript has specified a geospatial deep learning framework for predicting area-time public health risk from social media sentiment, urban environmental imagery, and spatial covariates. The framework treats sentiment and imagery as complementary but imperfect observations, aggregates them with reliability weights, fuses them through gated attention, and regularizes risk estimates over a spatial graph. It then converts model outputs into calibrated probabilities and cost-sensitive screening decisions. The central methodological insight is that multimodal prediction is only as credible as its representation of spatial support, temporal alignment, modality reliability, and calibration. The framework does not replace epidemiological surveillance; it provides a technically explicit and auditable approach for using digital environmental and social signals as screening evidence. Its value lies in making assumptions visible, failure modes testable, and decision thresholds accountable.

References

1. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. Science, 343(6176), 1203–1205.

2. Paul, M. J., & Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), 265–272.

3. Nguyen, Q. C., Li, D., Meng, H.-W., Kath, S., Nsoesie, E., Li, F., & Wen, M. (2016). Geotagged US tweets as predictors of county-level health outcomes, 2015. American Journal of Public Health, 106(12), 2241–2247.

4. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211–221.

5. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences, 114(50), 13108–13113.

6. Aman, J., Matisziw, T. C., Kim, J. B., & Luo, D. (2022, April). Sensing the city: Leveraging geotagged social media posts and street view imagery to model urban streetscapes using deep neural networks. In Proceedings of the International Conference on Computer-Aided Architectural Design Research in Asia (pp. 595-604). Association for Computer-Aided Architectural Design Research in Asia.

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

9. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.

10. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

11. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1), 128–137.

12. Hartig, T., Mitchell, R., de Vries, S., & Frumkin, H. (2014). Nature and health. Annual Review of Public Health, 35, 207–228.

13. Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93–115.

14. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

15. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

16. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929.

17. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.