Explainable Spatiotemporal Multi-Sensor Fusion for Urban Environmental Sensing

Ozlem Kilickaya[1,2], *Basant Kumar*[3]
[1]Lincoln University College, Petaling Jaya, Selangor, Malaysia.
[2] University of the People-595 E Colorado Blvd Suite 623, Pasadena, CA 91101, USA.
[3] Modern College of Business and Science, Oman.
ozlemgulsumkilickaya.writer@gmail.com ; basant@mcbs.edu.om

Abstract: Urban environmental sensing is a critical component of smart city infrastructure, yet current deep learning solutions remain fragmented. A primary motivation for this study is the observation that temporal modeling, multi-sensor fusion, and explainability are often developed independently, which limits predictive reliability and restricts real-world deployment. To address this lack of integration, a novel conceptual framework named ST-MFXAI (SpatioTemporal Multi-modal Fusion with Explainable AI) is proposed. This architecture unifies graph-based spatial modeling, temporal learning, and attention-driven fusion with embedded interpretability. Significant findings from the analysis highlight that current limitations stem from the absence of unified frameworks capable of jointly addressing heterogeneous data dynamics and model transparency. By bridging the gap between complex modeling and interpretative clarity, this study contributes a comprehensive synthesis of the literature and introduces a deployment-oriented perspective. The proposed framework and review find direct applications in real-time air quality monitoring, urban resource management, and the development of resilient, transparent sensing systems for future smart city environments.

Keywords: Urban Sensing; Sensor Fusion; Spatiotemporal Modeling; Explainable AI; Smart Cities

---

Introduction

As cities grow at an unprecedented pace, they face a complex web of challenges—from rapid industrialization to the urgent need for environmental sustainability. To meet these demands, smart city initiatives have emerged, using digital technology and data-driven insights to improve urban life. At the heart of this shift is the Internet of Things (IoT), which allows for the large-scale deployment of interconnected sensors. These devices provide a constant pulse on the city, monitoring everything from air quality and temperature to shifting traffic patterns [1], [5], [6].

Environmental sensing, in particular, has become a priority due to its profound impact on public health. The rise of affordable, high-resolution sensing technology has transformed how cities track pollution and climate shifts, providing a level of detail that was previously impossible [2]. This progress is further

bolstered by IoT-enabled frameworks that allow for real-time data analysis in the most unpredictable urban settings [7], [8].

To keep up with the sheer variety of data being generated, researchers have increasingly turned to advanced machine learning. Multimodal learning has proven effective at merging data from different sources into a single, cohesive picture of the urban landscape [3]. Similarly, newer temporal models, such as transformer-based architectures, have made it much easier to spot long-term trends and improve the accuracy of time-series forecasting [4].

Despite these strides, a critical disconnect remains. Most current research treats time-series modeling, sensor fusion, and explainability as separate problems rather than parts of a single, integrated system. This fragmented approach creates a "silo" effect: a model might be highly accurate on paper, but it often lacks the transparency and resilience needed for real-world use in a smart city. Furthermore, when explainability is treated as an afterthought—added only after a model is built—it rarely builds the deep trust required for high-stakes decision-making.

Closing these gaps requires a new generation of integrated frameworks. There is a clear need for systems that can simultaneously navigate spatial and temporal dependencies, fuse diverse sensor data, and offer built-in interpretability. This study explores the current state of urban sensing, identifies where existing models fall short, and proposes a unified conceptual path—the ST-MFXAI framework—to create more reliable and deployment-ready intelligent sensing systems.

Related work

The landscape of urban environmental sensing has evolved through a series of distinct technological shifts, transitioning from a primary focus on physical connectivity toward the current demand for integrated intelligence. Early research was largely preoccupied with the fundamental engineering of smart city infrastructures, prioritizing the deployment of large-scale sensor networks and the communication protocols necessary to sustain data acquisition across sprawling urban environments. These foundational studies established that scalable architectures and system interoperability are the essential precursors to any functional smart city ecosystem [1]. Building upon these structural basics, subsequent work transitioned toward more sophisticated IoT-based frameworks, demonstrating that interconnected sensing systems could move beyond simple data collection to facilitate real-time monitoring and informed decision-making in high-complexity settings [5], [6].

As infrastructure matured and data became more accessible, the research community naturally turned its attention toward the nuances of data acquisition strategies. The rise of low-cost sensing solutions, in particular, democratized high-resolution monitoring, enabling cities to track air pollution and localized environmental shifts with unprecedented spatial granularity [2]. However, this expansion introduced a fresh set of hurdles: ensuring data quality, maintaining sensor calibration, and establishing long-term reliability in the field. To mitigate these issues, more recent methodologies have begun integrating multiple sensing modalities within unified IoT platforms. While these advancements significantly enhance monitoring depth, they have also notably increased the technical complexity of the underlying analytical systems [7], [8].

To navigate the difficulties of managing such heterogeneous data, multimodal machine learning has emerged as a vital tool. By integrating diverse streams—including environmental, spatial, and contextual

information—these methods offer a more holistic digital representation of urban systems than single-source models [3]. Yet, a critical gap remains; these fusion-centric approaches often prioritize cross-modal integration at the expense of capturing the fluid temporal dependencies and dynamic variations inherent in city environments.

In a parallel development, temporal modeling techniques have seen significant breakthroughs, particularly in the realm of time-series forecasting. Transformer-based architectures have demonstrated a powerful capacity for capturing long-term dependencies and sharpening predictive accuracy [4]. While their integrated attention mechanisms offer a degree of interpretability, their utility is generally restricted to the temporal domain. Consequently, these models frequently fail to incorporate the complex spatial relationships or multimodal inputs required for a truly unified urban sensing framework. To clarify these existing limitations and the resulting need for a cohesive architecture, a comparative analysis of seminal studies is synthesized in Table 1.

Table 1. Comparison of existing approaches and proposed framework

Study	IoT Infrastructure	Temporal Modeling	Sensor Fusion	Explainability	Key Limitation
[1], [5], [6]	✓	✗	✗	✗	Focus on architecture, lacks analytical integration
[2]	✓	✗	✗	✗	Data quality issues, no modeling integration
[7], [8]	✓	✗	Partial	✗	Increased complexity, weak integration of analytics
[3]	✗	✗	✓	✗	Lacks temporal modeling and deployment focus
[4]	✗	✓	✗	Partial	Limited to temporal analysis, no multimodal fusion
This Study	✓	✓	✓	✓ (embedded)	Unified, deployment-oriented framework

As synthesized in Table 1, existing research remains largely partitioned into specialized silos, with studies typically isolating individual components such as IoT infrastructure, temporal modeling, or sensor fusion. While these narrow approaches may achieve high performance in controlled, isolated tasks, they frequently lack the cohesive integration required for systemic robustness and practical utility. This fragmentation creates a significant barrier to the real-world deployment of intelligent systems, as they often struggle to provide the interpretability necessary for high-stakes urban governance.

To resolve these systemic limitations, this study introduces a novel conceptual framework: ST-MFXAI (SpatioTemporal Multi-modal Fusion with Explainable AI). Moving away from traditional, disjointed

methodologies, the proposed framework unifies spatial modeling, temporal learning, multimodal fusion, and explainability within a single, streamlined architecture.

The structural integrity of ST-MFXAI is built upon three primary pillars designed to mirror the complexity of the urban environment. First, it leverages graph-based representations to capture the intricate spatial dependencies of urban topography. Second, it utilizes transformer-based mechanisms to model the nuanced temporal dynamics inherent in city life. Finally, it employs attention-driven fusion techniques to seamlessly merge heterogeneous data sources into a meaningful, holistic view. Crucially, explainability is not treated as an external, post-hoc process; instead, it is embedded directly into the model's core. This intrinsic transparency is designed to foster the trust and clarity essential for the successful deployment of autonomous sensing systems in modern smart cities.

Conclusions

This study has evaluated the evolving landscape of urban environmental sensing, leading to several key conclusions regarding the transition from fragmented data collection to integrated intelligent systems. While urban sensing is a cornerstone of smart city infrastructure, current research remains siloed. The primary motivation for this work stems from the observation that temporal modeling, multi-sensor fusion, and explainability are typically developed as independent modules. This fragmentation results in sensing systems that, despite high technical performance, lack the robustness and transparency required for high-stakes real-world deployment. To address these gaps, a structured review of IoT-enabled sensing, multimodal fusion, and spatiotemporal modeling was conducted. Based on this synthesis, a novel conceptual framework—ST-MFXAI (SpatioTemporal Multi-modal Fusion with Explainable AI)—was developed. This architecture moves beyond disjointed models by unifying graph-based spatial analysis, transformer-based temporal learning, and attention-driven fusion into a single, cohesive system. The analysis reveals that while current IoT frameworks excel at data acquisition and transformers offer superior predictive accuracy, a critical integration gap persists. Notably, explainability is frequently relegated to a post-hoc process, which diminishes stakeholder trust. The study finds that embedding interpretability directly into the fusion architecture is essential for creating systems that are not only accurate but also actionable for urban governance. As a conceptual review and framework proposal, this study is currently limited by the absence of empirical validation across diverse urban topologies. Consequently, future research will focus on the technical implementation of the ST-MFXAI architecture. This will involve rigorous testing against real-world datasets to evaluate its scalability, computational efficiency, and the practical utility of its embedded explainability features in live smart city applications.

References

1. A. Zanella, N. Bui, A. Castellani, L. Vangelista and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014. https://doi.org/10.1109/JIOT.2014.2306328
2. P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford and R. Britter, "The rise of low-cost sensing for managing air pollution in cities," *Environment*

International, vol. 75, pp. 199–205, 2015. https://doi.org/10.1016/j.envint.2014.11.019

3. T. Baltrušaitis, C. Ahuja and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019. https://doi.org/10.1109/TPAMI.2018.2798607
4. B. Lim, S. Ö. Arık, N. Loeff and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021. https://doi.org/10.1016/j.ijforecast.2021.03.012
5. M.-D. González-Zamar, E. Abad-Segura, E. Vázquez-Cano and E. López-Meneses, "IoT Technology Applications-Based Smart Cities: Research Analysis," *Electronics*, vol. 9, no. 8, p. 1246, 2020. https://doi.org/10.3390/electronics9081246
6. P. Bellini, P. Nesi and G. Pantaleo, "IoT-Enabled Smart Cities: A Review of Concepts, Frameworks and Key Technologies," *Applied Sciences*, vol. 12, no. 3, p. 1607, 2022. https://doi.org/10.3390/app12031607
7. A. Baranwal, "IoT-based environmental sensing solutions for smart city monitoring," *Smart City Insights*, vol. 2, no. 1, pp. 1–16, 2025. https://doi.org/10.22105/sci.v2i1.28
8. A. Waqar, T. A. H. Barakat, H. R. Almujibah, M. A. Alnowibet, M. A. Alqarni, A. M. Alshahrani and M. A. Alharthi, "Analytical approach to smart and sustainable city development with IoT," *Scientific Reports*, vol. 15, p. 23617, 2025. https://doi.org/10.1038/s41598-025-08861-y