Review

Sensing in Smart Cities: A Multimodal Machine Learning Perspective

Touseef Sadiq * and Christian W. Omlin

Posted Date: 28 July 2025

doi: 10.20944/preprints202507.2268.v1

Keywords: smart cities; multimodal machine learning; data fusion; deep learning architectures; intelligent urban systems; real-time decision making; privacy and ethics; scalability and interpretability

Review

Sensing in Smart Cities: A Multimodal Machine Learning Perspective

Touseef Sadiq * and Christian W. Omlin

Centre for Artificial Intelligence Research (CAIR), Department of Information and Communication Technology, University of Agder, Norway

* Correspondence: touseef.sadiq@uia.no

Highlights

What are the main findings?

- Presents a detailed framework and review of multimodal machine learning (MML) approaches utilized within smart urban environments.
- Highlights the effectiveness of MML techniques and current technical limitations in modality fusion, scalability and real-time implementation across urban domains.

What is the implication of the main finding?

- Provides essential guidance to researchers, academicians, policymakers and developers on choosing effective MML approaches for key smart city applications.
- Identifies current challenges and practical solutions to advance the deployment of multimodal machine learning in complex urban environments.

Abstract

Smart cities rely on diverse sensing infrastructures, generating vast multimodal data from IoT devices, surveillance systems, health monitors and environmental sensors. The seamless integration and interpretation of such multimodal data is key to enabling intelligent decision-making and adaptive urban services. Multimodal machine learning (MML) offers a powerful paradigm that surpasses traditional unimodal and rule-based methods and enables the effective integration of heterogeneous data from diverse applications across transportation, public safety, healthcare and environmental monitoring domains in smart cities. This review surveys the role of MML in smart city sensing, covering key data modalities, enabling technologies and state-of-the-art MML techniques such as fusion methods and deep learning-based architectures. We identify leading challenges in both MML methods including alignment, scalability and modality-specific noise, as well as in urban deployment scenarios, including infrastructure constraints, privacy concerns and ethical implications. Finally, we suggest future research directions toward the development of scalable, interpretable and ethically informed MML systems for smart cities. This survey serves as a reference for AI researchers, urban planners and policymakers seeking to understand, design and deploy multimodal learning solutions for complex urban environments.

Keywords: smart cities; multimodal machine learning; data fusion; deep learning architectures; intelligent urban systems; real-time decision making; privacy and ethics; scalability and interpretability

---

1. Introduction

With the modern growth of cities and their increasingly data-informed management, smart cities have evolved into complex ecosystems integrating traffic sensors, surveillance cameras,

 doi:10.20944/preprints202507.2268.v1

environmental monitors, GPS trackers and even social media feeds. The vast, interconnected data generated by these diverse sources not only provides insights of unprecedented scope but also poses significant challenges. Conventional data analysis methods face challenges with this blend of modalities, highlighting a clear need for more sophisticated, multimodal machine learning systems that can efficiently process and analyze multiple streams of data [1].

Traditional approaches to smart city sensing and data integration such as rule-based fusion [2], manual heuristics [3] and ontology-driven logic [4] have laid foundational groundwork but exhibit inherent limitations in scalability, adaptability and handling heterogeneous data [1]. These methods often rely on predefined rules and expert knowledge, restricting their capacity to process complex, multimodal urban data efficiently. To overcome these challenges attention has shifted towards MML, a powerful approach that addresses these complexities. MML enables systems to process and analyze information from diverse modalities, including text, audio, video, sensor data and geospatial data. This integration allows for a richer contextual understanding and significantly enhances decision making capabilities within smart city frameworks.

In smart cities, MML facilitates the integration of various data types, such as integrating video and sensor data for real-time traffic forecasting or combining CCTV feeds, emergency call transcripts and geolocation information for enhanced public safety surveillance. This paper delves into recent advancements in multimodal machine learning techniques and their applications in addressing concrete urban challenges.

Within the domain of artificial intelligence, MML represents a significant shift in how computational models handle complex environments. Comprehensive surveys highlight the increasing role of machine learning in analyzing temporal data and enhancing decision-making processes in smart city applications [5-8] . Unlike traditional methods that process only unimodal data such as text, audio or visual, MML integrates multiple data modalities simultaneously [9]. This approach includes structured and unstructured data sources like documents, images, audio, video, geographical information, biometrics and IoT sensor data. This integration is crucial for handling the complex, high-dimensional spaces of smart cities, where data from various sources must be analyzed collectively to make informed decisions.

Developing an effective MML system requires robust strategies for merging and processing heterogeneous data. Data alignment is critical in this context; ensures that outputs from different temporally or semantically misaligned modalities are mapped to a common contextual framework. For instance, aligning spoken language with text transcripts or syncing geospatial imagery with environmental sensor data ensures coherent data fusion [10].

Early fusion combines raw features from different modalities early in the processing stage, facilitating deep inter-modal interactions [11]. However, this requires extensive preprocessing to standardize data formats and scales. Late fusion analyzes each modality independently before combining their outputs, using methods like averaging or voting, suitable for modalities with significant disparities [11]. Hybrid fusion blends the strengths of early and late fusion, often incorporating mechanisms like attention to fine-tune the contribution of each modality [11].

Traditional analyses that consider unimodal data streams in isolation may lead to incomplete or misleading conclusions. For instance, sensor data may indicate rising temperatures, but without integrating it with citizen complaints and infrastructure strain reports, the implications may be missed. By integrating these diverse streams, MML provides a comprehensive framework for understanding complex urban dynamics, uncovering hidden patterns and enhancing city-wide decision-making processes.

MML can merge video feeds from traffic cameras with GPS data from vehicles and live social media updates about traffic conditions. This integration supports smart routing algorithms and dynamic signal control systems that adaptively respond to real-time traffic conditions, significantly improving urban mobility management. In public safety, emergency call audio, facial recognition data from surveillance cameras and incident reports can be fused by MML models to identify threats more swiftly and accurately. In practical applications these models have been used to detect

 doi:10.20944/preprints202507.2268.v1

gunshots, estimate their locations via GPS and cross-reference data with surveillance feeds to provide verified alerts to authorities [12]. Similarly, during natural disasters such as floods or earthquakes, MML integrates weather radar imagery, infrastructure sensor data, citizen tweets and emergency calls to paint a clearer picture of affected areas and effectively coordinate resource deployment.

Multimodal historical data informs long-term urban planning such as zoning, transportation infrastructure, energy grid expansion and sustainability efforts. By modeling scenarios that consider both structural variables and human behavior, MML enables urban planners to create more effective and resilient city systems [13].

The aim of our study is to provide a comprehensive overview of how MML methods are transforming sensing and decision-making processes in smart cities as shown in Figure 1. While numerous surveys have explored various aspects of smart cities and machine learning approaches, most tend to broadly focus on unimodal data in general or provide high-level overviews of artificial intelligence-based applications in smart cities [7,8,14]. To the best of our knowledge, there are no in-depth reviews specifically addressing the role of MML in smart city sensing. This paper fills this gap by providing a comprehensive review of MML fusion techniques, deep learning architectures, deployment challenges and real-world applications in smart cities, thus providing a distinctive focus that integrates sensing technologies and state-of-the-art multimodal data analytics.

This review is intended to serve as a resource for both AI researchers and smart city practitioners seeking to understand, develop, or apply MML-based solutions in urban environments. Specifically, the objectives of this paper are:

1. To classify and contextualize the types of multimodal data generated in smart cities, with a focus on sensing technologies and urban data sources.
2. To provide a systematic overview of multimodal machine learning techniques such as fusion strategies, cross-modal learning and attention mechanisms while addressing current challenges including alignment, scalability and data quality.
3. To review the practical applications of MML in smart city domains, including mobility, environmental monitoring, public safety, healthcare and governance.
4. To identify the current challenges related to deploying multimodal machine learning in smart city environments, including infrastructure limitations, policy constraints and ethical considerations.
5. To outline future research directions and opportunities at the intersection of MML and smart city development, aiming to inform the design of robust, ethical and scalable intelligent urban systems.

 doi:10.20944/preprints202507.2268.v1

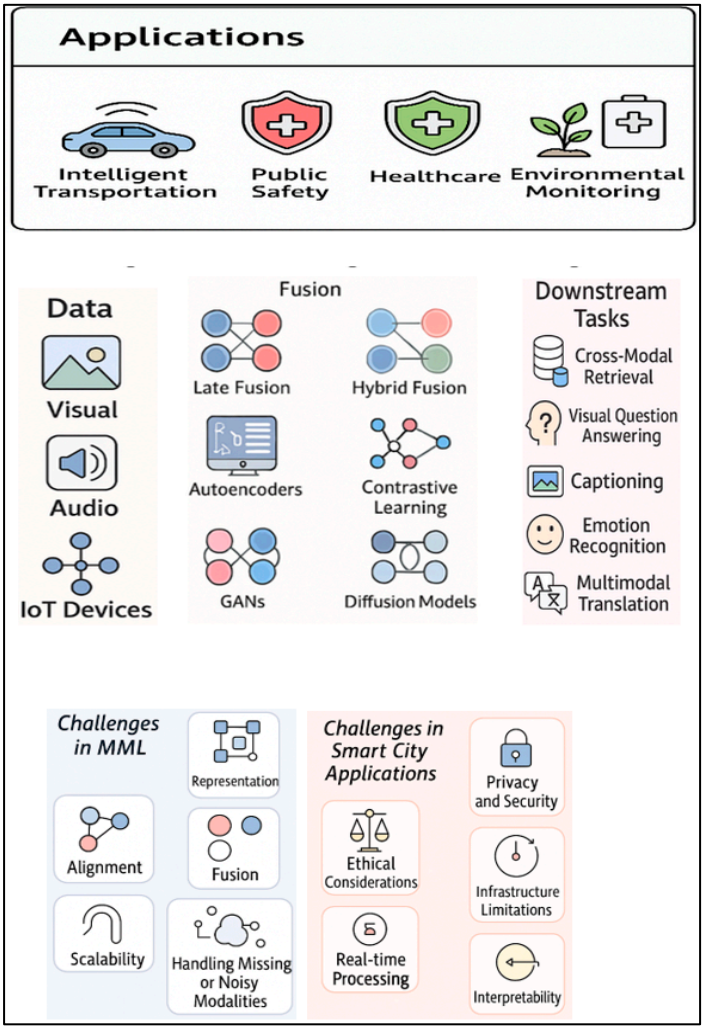

Figure 1. A comprehensive framework of multimodal deep learning for smart cities. The figure presents a high-level taxonomy starting with smart city application domains followed by the core components of multimodal learning which include data sources, fusion methods and tasks and concludes with a discussion of open challenges in both multimodal learning and its real-world deployment.

Our study is organized into seven interrelated sections, each of which contributes to a detailed analysis of MML motivated by the challenges of smart cities. In Section 2, we introduce the key concepts of smart cities and review traditional machine learning approaches for urban data analytics, while underlining the necessity of MML because of the complexity and diversity of urban data sources. Section 3 explores key MML techniques including advanced multimodal data fusion strategies based on deep learning architectures and addresses challenges such as data alignment, representation learning and handling missing or noisy data. Section 4 is dedicated to major domains of a smart city exploring practical applications of MML such as transportation, environmental monitoring, public safety, healthcare and citizen engagement. Section 5 identifies key challenges and limitations of MML deployment in smart city applications focusing on data privacy, algorithmic bias, scalability, interoperability and explainability. Section 6 discusses research gaps and future directions revolving around privacy-preserving learning frameworks, context-aware intelligent systems, explainable AI and development of standard multimodal datasets and benchmarks.

2. Background and Foundations

Smart cities are emerging as the next generation of urban environments that leverage technology and data to improve the quality of life for citizens, optimize resource management and enhance the overall functioning of urban systems. In these environments, the integration of multimodal data such

 doi:10.20944/preprints202507.2268.v1

as data from Internet of Things (IoT) sensors, smart devices and data-driven technologies enables cities to operate as adaptive real-time systems capable of efficiently addressing urban challenges [14,15].

These cities collect vast volumes of data from diverse sources including traffic sensors, surveillance cameras, environmental monitors and social media platforms. These data originate from urban infrastructures and citizen-centric platforms including sensors embedded in roads, buildings, vehicles and utility systems provide structured data on temperature, energy usage, traffic flow and air quality. Multimodal data vary in structure, frequency and reliability encompassing continuous video streams from CCTV and drones, real-time geolocation and biometric data from smartphones and wearables, unstructured social media content reflecting public sentiment and semi-structured records from transportation, emergency services and municipal databases [16]. Understanding how these diverse modalities interconnect is vital for intelligent, data-driven urban governance.

The need for smart data driven by urban management systems to monitor cities all over the world continues to grow and digitization becomes increasingly urgent. To manage, supervise and enhance the infrastructure and services, smart cities depend upon a wide range of databases [17].

Urban data can be categorized into several key modalities, each offering unique perspectives and challenges. The most prominent include:

Visual Data: Visual data, including CCTV footage, satellite imagery and drone footage, is extensively used in smart cities. Municipalities and law enforcement agencies use CCTV for real-time surveillance, traffic monitoring and crime prevention. Satellite imagery supports large-scale environmental monitoring and urban planning, while drones provide flexible, aerial views for crowd monitoring and infrastructure inspection [18].

Audio Data: Audio sensors are used to detect abnormal sounds such as vehicle collisions, public disturbances, or emergency events [19]. When combined with visual data (e.g., CCTV), audio can improve real-time event classification and response.

Textual Data: Textual data often derived from social media, citizen feedback and official reports is useful for gauging public sentiment and monitoring events. Due to its unstructured nature, text mining techniques are necessary to extract actionable insights for crisis response and public safety [20].

Sensor-Based Data: IoT sensors monitor a wide range of conditions, such as air quality, noise levels, temperature and traffic density. These provide continuous, real-time data that is essential for managing pollution, traffic congestion and health infrastructure. Smart cities leverage IoT extensively to collect diverse urban data, but face significant challenges in technology integration, data management and security [16,21].

Geospatial and Mobility Data: Geospatial data from GPS devices, public transportation systems and mobile phones is used to analyze movement patterns, optimize transit systems and reduce congestion [22].

Wearable Data: Wearables like fitness trackers and smartwatches generate personal health and activity data, which can be aggregated for public health monitoring and environmental exposure analysis [23] .

These diverse data modalities form the foundation of urban sensing in smart cities. Figure 2 provides an overview of key data sources and how they are collected from various infrastructures and citizens to support city-wide operations.

 doi:10.20944/preprints202507.2268.v1

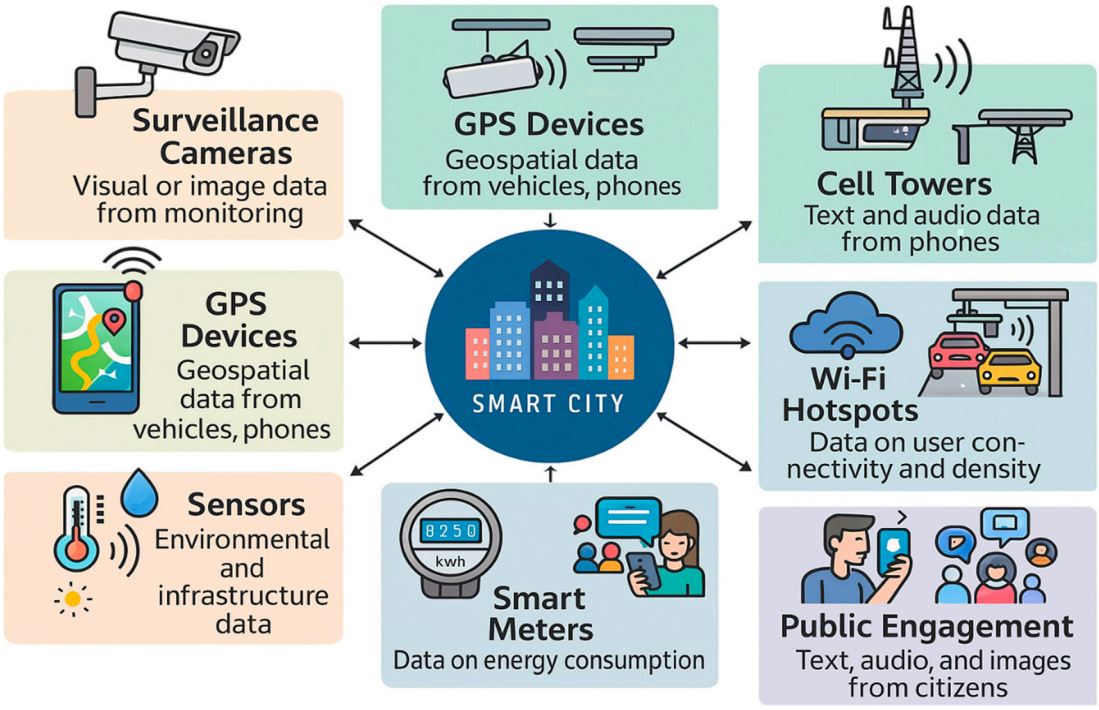

Figure 2. Overview of key data sources in a smart city environment. Diverse data streams including visual, geospatial, audio, environmental, infrastructure and citizen-generated inputs are continuously collected and integrated to support real-time urban intelligence and decision-making.

Traditional smart city sensing methods such as rule-based fusion [2] and ontology-driven logic [4], rely on explicit knowledge representations and manually crafted heuristics. Although these techniques offer interpretability and domain insight they often struggle with scalability and adapt to the complex, dynamic and multimodal nature of urban data [24]. Rule based systems [2] are inherently brittle, requiring continuous manual updates as environments change while ontology driven approaches [4] can be limited by their reliance on predefined semantic structures and difficulties handling noisy or incomplete data.

In contrast, MML leverages data-driven models capable of learning joint representations across heterogeneous modalities including visual, auditory and sensor streams enabling more flexible and robust fusion [1]. Techniques such as attention mechanisms [25] and cross-modal learning [26] facilitate effective alignment and integration of disparate data sources, overcoming key challenges faced by traditional methods [11]. This paradigm shift enhances the ability of smart city systems to process complex information and make informed decisions in real time. Recent advances in deep learning, including transformer based models [25] and graph neural networks (GNNs) [27] have made multimodal machine learning increasingly viable by enabling the learning of complex relationships between modalities and enhancing capabilities in event detection, anomaly detection and decision support.

MML systems are already being used in smart cities to improve transportation, public safety, environmental monitoring and citizen engagement [28]. By integrating multiple modalities, these systems enable more adaptive, sustainable and resilient urban solutions. A summary of major data modalities used in smart cities is provided in Table 1, it illustrates how various sources include visual, sensor-based, textual, geospatial and behavioral dataflow into centralized MML systems for holistic smart city analysis.

Table 1. Summary of the key data modalities along with their examples, characteristics and applications in smart cities.

Modality	Examples	Data Characteristics	Urban Application Areas

 doi:10.20944/preprints202507.2268.v1

Visual	CCTV footage, satellite images, drone videos	Real-time, high volume, spatial information	Traffic monitoring, public safety, event management [18,29].
Sensor-based	Air quality sensors, weather stations, noise monitors	Continuous, structured, environmental data	Pollution monitoring, climate modeling [23,30]
Textual	Tweets, public service reports, news articles	Unstructured, periodic, noisy data	Public sentiment analysis, social media monitoring [20,31].
Geospatial	GPS data, geolocation tracking, heatmaps	Spatial-temporal, dynamic	Traffic management, mobility optimization [21,22].
Behavioral	Mobile app usage, pedestrian tracking	Structured and unstructured, behavioral	Urban mobility, public health monitoring [13,32].

3. Techniques in Multimodal Machine Learning for Smart Cities

MML is revolutionizing how smart cities process and analyze the vast array of diverse data generated by urban environments. By integrating data from multiple sources such as video feeds, sensor networks, social media and geospatial data MML enables more accurate, dynamic and context-aware decision-making. This section explores the fundamental techniques that underpin MML systems focusing on fusion strategies, deep learning-based approaches and the challenges associated with processing and integrating multimodal data in smart city contexts as shown in Figure 3.

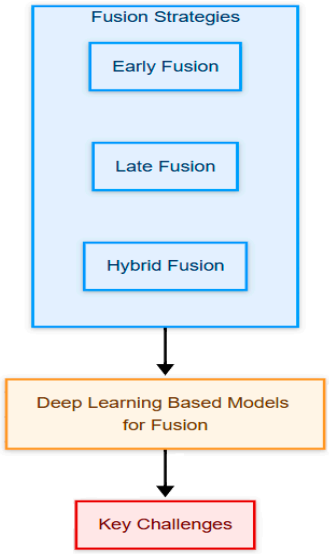

Figure 3. Hierarchical overview of fusion strategies and their connection to deep learning-based models for multimodal data integration. The three primary fusion approaches: early, late and hybrid serve as foundational strategies for designing deep learning models. These models encounter key multimodal challenges, including representation learning, modality alignment, fusion mechanisms, scalability and robustness to missing or noisy data.

3.1. Fusion Strategies in Deep Learning for Smart Cities

Handling the variety and volume of smart city data necessitates advanced fusion techniques to ensure robust and accurate insights, as summarized in recent systematic reviews [1,9,11]. Key fusion strategies in deep learning for multimodal learning include early fusion, late fusion and hybrid fusion, each offering unique ways to integrate multimodal data at various stages of the learning pipeline [33] as shown in Figure 4:

 doi:10.20944/preprints202507.2268.v1

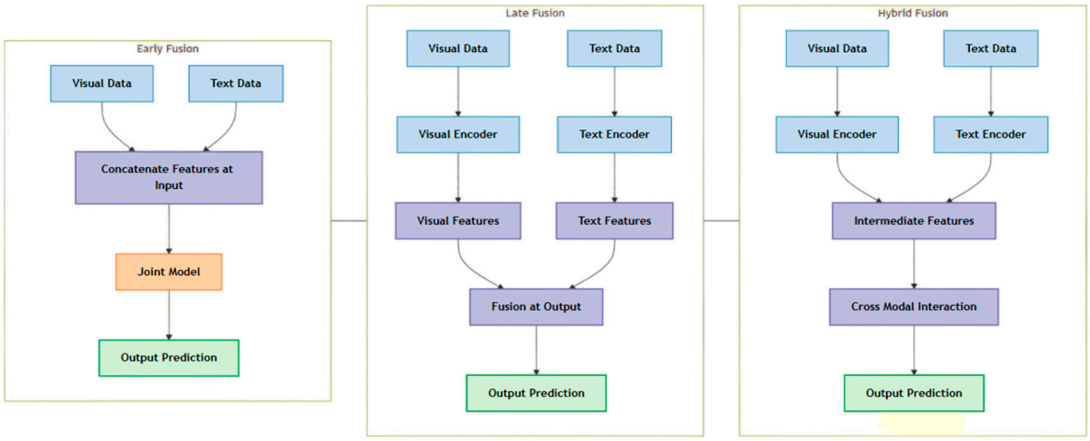

Figure 4. Illustration of multimodal fusion strategies: early fusion concatenates data at the input level; late fusion combines outputs of modality-specific encoders; hybrid fusion integrates features at intermediate layers through cross modal attention.

3.1.1. Early Fusion

Early fusion combines raw data or low-level features from different modalities before they are input into the learning model as depicted in Figure 5. This approach enables the model to learn joint representations from the outset, allowing for deeper cross-modal interaction at all layers. However, it also demands careful preprocessing and alignment of modalities to ensure compatibility in temporal and spatial dimensions [43].

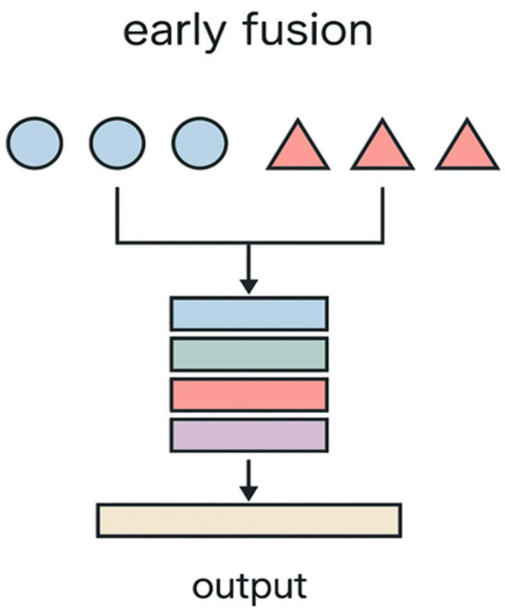

Figure 5. Early fusion approach for multimodal data integration. Features from each input modality are concatenated at the initial stage and jointly processed through subsequent layers to produce a unified output.

Transformer-based models as shown in Table 2 are the most representative architectures employing early fusion. VisualBERT [34] directly concatenates image region embeddings with tokenized text inputs and feeds the combined sequence into a standard BERT model [44], enabling early cross-modal attention [26]. Similarly, VL-BERT [35] follows a joint encoding scheme where both modalities are processed in a single transformer stream from the beginning, facilitating fine-grained token-region alignment. Unimo [45] expands this idea by unifying vision, language and even

 doi:10.20944/preprints202507.2268.v1

structured data in a common space through a shared encoder trained on generative and contrastive objectives.

Table 2. Overview of fusion strategies in multimodal machine learning, categorized into early, late and hybrid fusion. The table summarizes representative models, their fusion points and distinguishing architectural characteristics.

Fusion Type	Example Models	Fusion Point	Notes
Early Fusion	VisualBERT [34], VL-BERT [35]	Input-level	Joint transformers over concatenated modalities
Late Fusion	CLIP [36], ALIGN [37]	Output-level	Independent encoders, aligned via similarity
Hybrid Fusion	LXMERT [38], ViLBERT [39], UNITER [40], ALBEF [41], BLIP [42]	Mid-level / cross-modal layers	Modality-specific encoders + cross-attention

The strength of early fusion lies in its ability to model tight dependencies between modalities, which is particularly beneficial in tasks such as visual question answering (VQA), image captioning and multimodal sentiment analysis. However, scalability can be a challenge when modalities differ significantly in structure, dimensionality, or sampling frequency common in sensor data fusion for smart cities.

3.1.2. Late Fusion

Late fusion allows each data modality to be processed independently through specialized models tailored to the nature of each input. For instance, traffic flow data from surveillance cameras might be processed using a computer vision model, while environmental sensor readings (e.g., air quality or temperature) are handled by time-series forecasting models. The outputs from these separate streams are then combined at the decision level, typically via ensemble techniques such as majority voting, weighted averaging, or meta-learning classifiers [43]. An illustration based on late fusion is shown in Figure 6.

This approach is particularly valuable when modalities differ significantly in scale, structure, or reliability, a common scenario in smart city systems where heterogeneous sensors, social data and video analytics coexist.

 doi:10.20944/preprints202507.2268.v1

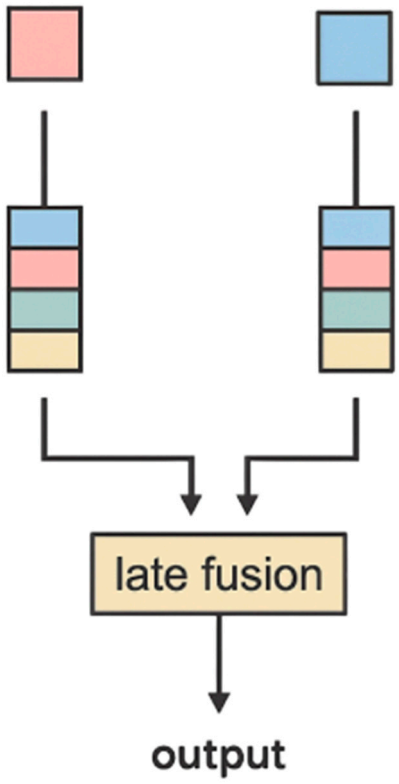

Figure 6. Late fusion approach for multimodal data integration. Each input modality is processed independently through separate feature extraction pipelines and their outputs are combined at a later stage to generate the final prediction.

Classical late fusion strategies have been widely used in multimodal affective computing, surveillance and event detection, where models for audio, video and text operate separately and contribute to a final prediction score. While deep learning approaches favor early or hybrid fusion for joint representation learning, some contrastive learning frameworks such as CLIP [36] and ALIGN [37] as shown in Table 2 retain a late fusion flavor by encoding each modality separately and aligning them only in a shared embedding space albeit not performing fusion in the strict sense. Thus, late fusion remains a practical choice when system modularity, interpretability, or robustness to noisy modalities is prioritized.

3.1.3. Hybrid Fusion

Hybrid fusion integrates elements of both early and late fusion strategies by combining modalities at multiple stages of the processing pipeline as shown in Figure 7. For instance, traffic camera data and environmental sensor readings might be processed independently to detect anomalies and pollution patterns, respectively. These intermediate outputs are then fused at a mid-level to inform a downstream decision model, allowing the system to leverage both modality-specific features and cross-modal interactions [46].

This approach is particularly valuable when complex, hierarchical relationships exist across modalities such as retrieving satellite imagery of traffic congestion based on user-submitted text reports. By learning shared intermediate representations, multimodal machine learning systems can enable more context-aware and accurate decision-making in smart city applications.

 doi:10.20944/preprints202507.2268.v1

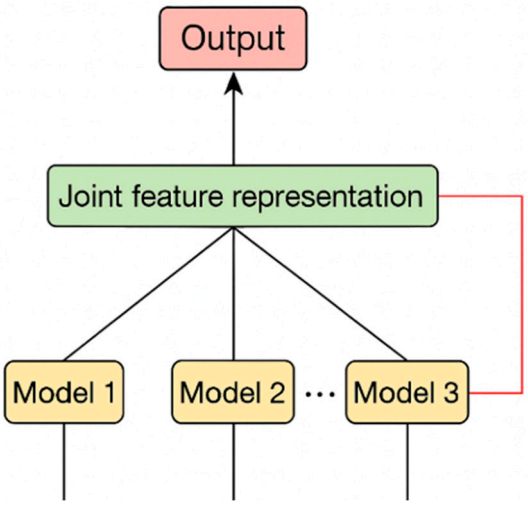

Figure 7. Hybrid fusion architecture for multimodal data integration. Each modality is first processed independently, and their features are fused at intermediate stages to form a joint feature representation. The resulting combined features are further processed and integrated at a later stage to generate the final output, leveraging both early and late fusion strategies.

Architecturally, hybrid fusion is often implemented using modality-specific encoders followed by cross-modal transformers or co-attentional layers, as discussed in recent surveys and foundational model analyses [1,11,43,46]. These approaches preserve the advantages of independent representation learning while introducing mechanisms for joint reasoning across modalities offering a strong balance between expressiveness, modularity and scalability for heterogeneous urban data streams.

These fusion strategies are often implemented through deep learning architectures, including CNNs [47], RNNs [48] and transformers [25] which are tailored to specific types of data and integration approaches. These models are essential for making sense of complex, high-dimensional data typical of smart city environments.

3.2. Deep Learning based Models for Fusion

MML leverages various deep learning architectures to process and integrate diverse data sources for smart city applications. Algorithms like CNNs [47] and LSTM [49] networks excel at processing visual and sequential data, respectively, while transformers and GNNs [27] are increasingly used to handle multimodal data integration across various types of sensors and data streams. These deep learning models, combined with specialized techniques such as reinforcement learning (RL) for adaptive decision-making, form the backbone of intelligent systems for smart cities [1,50]. Bayesian networks and fuzzy logic methods complement these approaches by incorporating uncertainty and improving model robustness, especially in real-time dynamic environments [51]. Together, these deep learning methods enable data-driven insights and context-aware decision-making, contributing to the creation of smarter, more efficient urban environments.

Table 3 summarizes the main architectures found in MML including CNN, Transformer-based and GNN. It elaborates their descriptions, special implementations among smart cities and their benefits for urban management activities like traffic forecast, public safety and environment monitoring.

Table 3. Overview of advanced deep learning architectures for MML in Smart Cities.

Architecture	Description	Applications in Smart Cities	Advantages

 doi:10.20944/preprints202507.2268.v1

Convolutional Neural Networks (CNNs)	CNNs are used for image processing but have been extended to integrate visual data with other modalities such as sensors or geospatial data.	Urban mobility prediction, environmental monitoring (analyzing traffic cameras and environmental sensor data).	Well-suited for spatial data analysis and can handle large-scale image data, which is common in traffic and environmental monitoring systems [47].
Transformer-based Models	Transformer models, initially developed for natural language processing (NLP), use self-attention mechanisms to learn relationships across data modalities.	Image-text matching, video captioning, traffic event prediction (integrating visual and textual data).	Captures long-range dependencies across multiple modalities. Self-attention mechanism allows for flexible attention to relevant features across modalities [25].
Graph Neural Networks (GNNs)	GNNs are designed to process graph-structured data, where relationships between entities are crucial for prediction tasks.	Traffic prediction, urban mobility, public safety (modeling road networks, sensors and traffic flows).	Effectively models spatial dependencies and interconnected data across multiple sources, such as roads, sensors and events [27].

CNNs are powerful models traditionally used for image and video processing tasks. In the context of multimodal fusion, CNNs are particularly effective in integrating visual data with other data types such as sensor data or geospatial information. This is crucial for applications in smart cities, where large-scale data from diverse sources such as traffic cameras, environmental sensors and satellite imagery must be processed in real-time to provide actionable insights. Figure 8 depicts how CNNs integrate visual data (e.g., traffic camera footage) for traffic monitoring in smart cities.

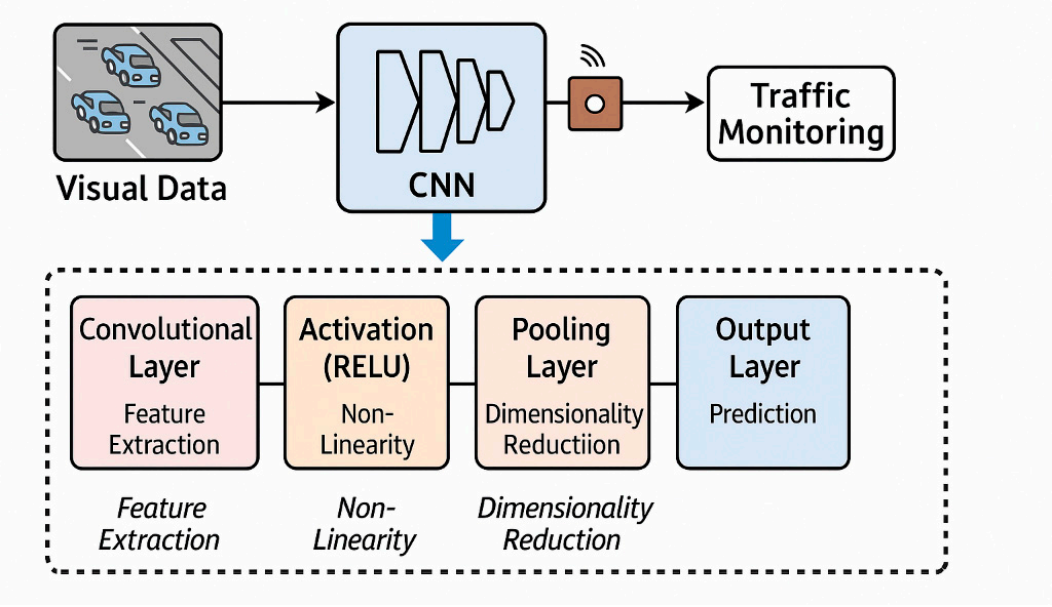

Figure 8. Illustration of CNNs fusion of visual data for traffic monitoring, showing inner workings of a CNN in the context of multimodal fusion, showing the processing flow from input data.

 doi:10.20944/preprints202507.2268.v1

CNNs identifies spatial hierarchies in visual data, starting from low-level features like edges and textures, progressing to more complex patterns such as objects or scenes. CNNs are particularly well-suited for early fusion, where data from different modalities such as video feeds from traffic cameras and sensor data from environmental monitors is combined and processed simultaneously.

A CNN typically consists of three main layers:

1. **Convolutional Layer:** This layer applies filters to the input image (or visual data) to detect low-level features, such as edges or corners. It produces feature maps that highlight important patterns in the data.
2. **Pooling Layer:** After convolution, pooling is applied to reduce the spatial dimensions of the feature maps while retaining important information. This helps the model become invariant to small translations of the input data.
3. **Fully Connected Layer**: The final layer combines the extracted features to make predictions or classifications based on the learned patterns. In multimodal fusion, these outputs are often combined with data from other sources (such as environmental sensor readings) at a later stage.

Convolutional neural networks continue to serve as the backbone for many multimodal models due to their unparalleled ability to extract hierarchical features from visual data. Despite the rise of transformer-based architecture, CNNs remain essential for tasks that involve image or video processing because of their ability to efficiently capture spatial patterns in the data [52]. In multimodal models, CNNs are typically used to process visual data, which is then combined with other modalities, such as text or sensor data, to enable more comprehensive decision-making. For instance models like VisualBERT [34] and ViLBERT [39] use features from object detectors based on CNNs and these features are later fused with textual data for tasks such as visual question answering (VQA) or image captioning. The reason for their continued use is their ability to handle large-scale image data and to learn rich, hierarchical representations that are essential in many real-time applications in smart cities, such as traffic monitoring and environmental sensing. Despite the increasing popularity of transformer models, CNNs are still favored for their computational efficiency and scalability, especially when dealing with the massive amounts of image data in urban settings [53].

CNN-based architectures, particularly in combination with transformers or other models, allow for a balance between automatic feature extraction and powerful fusion mechanisms, making them highly effective for multimodal learning. Models such as LXMERT[38] and ALBEF [41] leverage CNNs for visual data processing while the integration of textual data through cross-modality attention mechanisms provides a robust framework for applications in urban mobility, safety and environmental monitoring. As smart cities increasingly rely on real-time data from a diverse set of sources, CNNs continue to be a valuable component in these systems due to their ability to efficiently process visual data on a scale and integrate it with other sources of information for context-aware decision-making .

Transformers have emerged as the dominant architecture for many multimodal learning tasks, particularly due to their ability to handle long-range dependencies and model complex interactions between multiple data modalities [25]. Unlike traditional models such as CNNs or LSTMs, which process data sequentially or hierarchically, transformers utilize self-attention mechanisms that enable the model to attend to all parts of the input simultaneously. The Vision Transformer (ViT) architecture leverages attention mechanisms to capture high-level semantic features from image patches, improving RGB-D scene classification [54]. Figure 9 depicts that multimodal transformer model addresses the challenge of unaligned multimodal language sequences by using attention mechanisms to integrate disparate data streams effectively. This characteristic makes them exceptionally well-suited for tasks that require combining information from heterogeneous data sources, such as visual data, text and sensor data.

 doi:10.20944/preprints202507.2268.v1

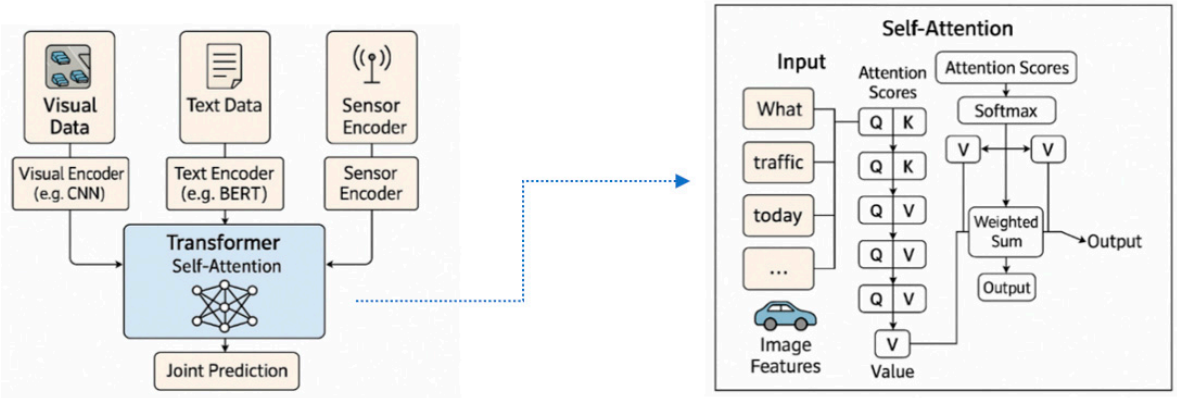

Figure 9. Diagram illustrating the self-attention mechanism in a transformer model for multimodal fusion. The diagram shows how visual data, text data and sensor data are processed by separate encoders before being integrated into the self-attention layer of the transformer for joint prediction.

Self-attention layer [55] operates within a transformer, particularly for multimodal data like images (processed via CNNs) and text.

1. **Input Sequence:** The input sequence of elements will be a mix of text tokens and image features (e.g., from CNNs for visual data).
2. **Query, Key and Value Vectors**: Each element in the input (text or visual) is transformed into three vectors: Query (Q), Key (K) and Value (V).
3. **Attention Scores:** This explains how the Query (Q) is compared to all Keys (K) to compute attention scores.
4. **Softmax and Weighted Sum:** After computing the attention scores, we'll show how they are normalized (via softmax) and used to weight the Value (V) vectors.
5. **Output:** The final output of the self-attention layer, which is a contextualized representation for each element, will be shown.

Transformers enable the integration of various modalities at different stages of processing, typically after the independent encoding of each modality. Models such as ViLBERT [39] and LXMERT [38] leverage transformers to process visual and textual data separately at first, followed by cross-modal interactions through attention layers that align these representations. This allows the model to learn how to relate and integrate visual information (e.g., images or video) with textual data (e.g., descriptions, questions) in a flexible and context-aware manner.

Transformers are also crucial in hybrid fusion strategies, where they allow parallel processing of multiple data streams while preserving the modality-specific features. The flexibility of transformers allows them to be used in a variety of configurations for multimodal tasks, from early fusion (where modalities are integrated at the input level) to late fusion (where the outputs of separate models are combined at the decision level). Their scalability and ability to learn joint representations have made transformers the go-to architecture for tasks such as image captioning, visual question answering (VQA) and image-text retrieval [40].

The key advantage of transformers in multimodal fusion lies in their ability to capture complex interactions between modalities. For instance, in smart city applications, transformers can process sensor data (such as traffic data, weather conditions, or air quality) alongside video feeds or satellite images to generate more accurate predictions for traffic management or environmental monitoring. Unlike CNNs, which are primarily focused on visual data, transformers can handle a wider variety

 doi:10.20944/preprints202507.2268.v1

of inputs and learn relationships across them, making them more flexible for complex urban environments [56].

In addition to their ability to handle diverse data types, transformers have been widely adopted due to their parallelization capabilities, which enable faster processing of large-scale multimodal datasets, which are a critical feature in real-time applications in smart cities. Contrastive learning [57] has emerged as a powerful paradigm for aligning modalities in a shared representation space. It operates by maximizing the similarity between semantically related cross-modal pairs (positive samples) while pushing apart unrelated ones (negative samples). Models like CLIP [36] and ALIGN [41] use large-scale datasets of image caption pairs to learn joint embeddings without requiring explicit alignment labels. Transformers are at the heart of models like CLIP, which learn shared representations across vision and language tasks through contrastive learning. Although primarily designed for vision-language tasks, the underlying principles of contrastive alignment are readily extensible to smart city modalities, enabling traffic management authorities to identify scenes from traffic videos using textual descriptions, such as linking camera footage with incident reports[58].

Graph neural networks [27] help show organized links like traffic patterns or sensor networks from different methods. Attention tools used alone or in bigger systems, let the model focus on important features or methods based on task-specific situations. Together, these plans help MML systems come close to how humans perceive and think about understanding speech with facial cues or text reports with related videos. In smart city settings such skills matter greatly for quick understanding and choices in areas like public safety, transport, environmental watch and governance. MML gives a strong way to analyze building wise, changing and welcoming urban systems. The initialization layer ensures that different modalities are connected and mapped on the same window of context. In order to create the significant connection that comes from different sources and ranges at various levels of time, sourced from observing mode data is synchronized.

3.3. Comparative Assessment of MML Techniques and Their Performance

The comparative analysis of various MML techniques reveals important insights into their accuracy, computational complexity and applicability to different types of multimodal data in smart cities is shown in Table 4.

Table 4. Performance metrics and suitable modalities for various multimodal machine learning techniques.

MML Technique	Accuracy (%)	Computational Complexity	Suitable Data Modalities	Application
Deep Learning	90	High	Audio, Video, Text	Urban video analytics [59,60]
Transfer Learning	85	Medium	Sensor Data, Images	Sensor-to-sensor adaptation [61]
Ensemble Methods	88	Low-Medium	Sensor Data, Social Media	Social sentiment and traffic data [62]
Graph-based Methods	82	Medium	Spatial Data, Networks	Traffic network inference [63]
Reinforcement Learning	87	High	IoT Data, Control Systems	IoT-based adaptive control [64]

Deep learning techniques, while achieving the highest accuracy of 90%, are best suited for complex big data types such as audio, video and text. However, their high computational complexity limits their use to environments where powerful hardware and resources are available. These

 doi:10.20944/preprints202507.2268.v1

techniques are highly effective in smart city applications involving large-scale and diverse datasets but come with the trade-off of requiring substantial computational resources.

Transfer learning provides a balance between accuracy and computational efficiency, achieving an 85% accuracy [65]. With medium computational complexity, this approach is particularly well-suited for processing sensor data and images, both of which are abundant in smart city applications. By leveraging pre-trained models, transfer learning reduces the need for extensive training, making it an attractive choice for real-time tasks.

Ensemble Methods, which combine multiple models to produce a final prediction, have demonstrated up to 88% accuracy on smart city datasets involving sensor fusion and social media analysis [66]. These methods are effective for integrating insights from various data sources, such as sensor data and social media content, which is frequently encountered in urban sensing and monitoring tasks.

Graph-based methods [27] provide an accuracy of 82% and are particularly useful for modeling and analyzing spatial data and networks, such as IoT sensor networks and social network data. Although they come with medium to high computational complexity, their ability to capture relationships in data makes them suitable for smart city applications that involve networked data.

Reinforcement Learning is a powerful machine learning approach that is particularly well-suited for multimodal learning in smart cities [64]. RL algorithms make decisions by interacting with their environment and receiving feedback in the form of rewards or penalties, which allows them to continuously optimize strategies over time. In traffic management, RL can adapt and optimize traffic signals and routing decisions based on both real-time and historical data effectively reducing congestion and improving traffic flow. RL is valuable in enhancing public safety by learning optimal strategies for resource allocation and incident response. By simulating various urban scenarios, RL can develop efficient safety protocols that enhance urban security and functionality. RL achieves an accuracy of 87% and excels in dynamic environments where real-time decision-making is critical [77]. Although RL involves medium to high computational complexity, it is highly effective for tasks such as adaptive traffic control and environmental optimization in smart cities [78].

Table 5 presents a comparative analysis of the pros and cons for the various MML approaches discussed in the context of smart city applications. Each of the techniques provides unique advantages for addressing the diverse challenges in urban environments, with real-world experimentation demonstrating their potential. The merits of these methods highlight their ability to optimize traffic management, environmental monitoring and other smart city tasks, making them highly effective in dynamic and complex settings. However, each approach also comes with limitations, including computational complexity, data requirements and challenges in scalability and real-time deployment. These drawbacks underscore the need for further refinement and development of multimodal learning models to overcome unexpected issues and maximize their potential for smart city application.

Table 5. Pros and cons of multimodal machine learning models in smart city applications.

Authors	Model/Approach	Pros	Cons
Ouoba et al. [67]	Multimodal Journey Planning	Addresses fragmented environments by integrating real-time data for comprehensive planning.	May require large computational resources for real-time data processing.
Botea et al. [68]	Risk-Averse Journey Advising	Accounts for uncertainties in public transport schedules, improving system reliability.	Does not fully address long-term dynamic changes in urban mobility.
Asgari et al. [69]	Multimodal Trajectories	Uses unsupervised models, effective for forecasting traffic flow without needing labeled data.	May struggle with real-time adjustments or highly dynamic urban environments.

 doi:10.20944/preprints202507.2268.v1

Alessandretti et al. [70]	Public Transportation Networks	Leverages data-driven models to analyze complex transport networks, improving planning and efficiency.	Might be less effective in highly decentralized, less connected urban settings.
Pronello et al. [71]	Travel Behavior	Provides insights into behavioral shifts using multimodal data, improving transportation planning.	Requires large amounts of data to detect subtle shifts in behavior and patterns.
Kang and Youm [72]	Extended Public Transport	Improves route optimization, enhancing public transportation efficiency.	Complexity increases with the number of variables and real-time adjustments needed.
Sokolov et al [73]	Digital Railway Infrastructure	Integrates digital frameworks to optimize railway systems and reduce urban congestion.	High computational complexity and infrastructure requirements for implementation.
Young et al. [74]	Smart-Citizen Engagement	Leverages multimodal data for interactive city management, improving citizen engagement.	May face challenges in data privacy and ethical concerns when dealing with citizen data.
Kumar et al. [75]	Crowd Monitoring	Enhances public safety through intelligent monitoring systems for crowd management.	It can be costly and difficult to scale across large urban areas without specialized infrastructure.
Zhang et al. [76]	Vehicle Tracking	Improves vehicle tracking accuracy in complex urban environments by combining multiple data sources.	Requires continuous data input and faces challenges in real-time tracking in highly dynamic settings.

Applying these techniques comes with challenges such as ensuring effective representation learning across different modalities and aligning data from diverse sources with varying temporal resolutions and semantic meanings. Overcoming these challenges is critical for building robust MML systems that can deliver actionable insights for smart city management.

3.4. Core Challenges in the Multimodal Fusion

Addressing challenges associated with processing multimodal data is critical to the successful deployment of smart city initiatives. As highlighted in Table 6, these challenges encompass issues such as multimodal representation learning, cross-modal alignment and fusion, scalability and real-time constraints and robustness to missing or noisy modalities. To overcome these obstacles, researchers and practitioners are exploring a variety of solutions including feature fusion techniques, distributed computing and explainable AI approaches [9]. These strategies aim to enhance the accuracy, efficiency and adaptability of multimodal systems in dynamic urban environments.

Table 6. Key challenges in multimodal data processing for smart city applications and their corresponding proposed solutions. The table outlines the major issues faced in processing multimodal data, such as representation learning, cross-modal alignment, scalability and robustness to noisy data, along with potential solutions that address these challenges.

Challenge	Proposed Solution(s)	References
Multimodal Representation Learning	Feature Fusion, Transfer Learning	[1, 51]
Cross-modal Alignment	Cross-modal Attention Mechanisms, Multi-modal Transformers	[52, 53]

 doi:10.20944/preprints202507.2268.v1

Scalability and Real-time Constraints	Distributed Computing, Cloud Infrastructure, Edge Computing	[54,55]
Robustness to Missing or Noisy Modalities	Data Imputation, Robust Training Methods, Noise Reduction	[56,57]
Interpretability of Models	Explainable AI Techniques, Model Visualization	[58,59]
Dataset and Benchmark Limitations	Standardized Datasets, Synchronized data, Robust Benchmarks, Composite metrics, Open Benchmarks.	[1,5]

3.4.1. Multimodal Representation Learning

The primary objective of deep learning for multimodal representations is to create an aligned representation that effectively captures the most salient features across all modalities, ensuring consistency during their integration. For example, when combining numerical data from environmental sensors (e.g., air quality or temperature) with high-dimensional video data (e.g., from traffic cameras), architectures must be designed to condense and reconcile the differences between these data types. This challenge is particularly significant in smart cities, where accurate and real-time decision-making is crucial. Despite the complexity of each modality presenting information in unique ways, successful integration can lead to a 25% improvement in system responsiveness and accuracy, making it invaluable for applications like traffic management or public safety [79].

To further illustrate the concept of representation learning in the context of smart cities, Table 7 above summarizes how this technique works in practice. Table 7 provides a concise overview of representation learning in the context of MML in smart cities. It defines representation learning as the act of mapping different types of data (such as images, text and sensor outputs) into a shared feature space which allows us to analyze them together. We aim at enabling valid comparison and integration of heterogeneous data for tasks such as cross-modal retrieval. That, in smart cities, allows systems to link information from traffic cameras, weather reports and social media to deal with congestion or emergencies. The main utility is improved context-aware decision-making through the fusion of different data sources, although there are challenges related to ensuring semantic consistency across modalities, especially when dealing with complex, spatial formats like audio and video.

Table 7. Representation learning in MML for smart cities.

Aspect	Description	Example in Smart Cities
Definition	Representation learning refers to the process of transforming raw data from different modalities into a shared latent space where they can be compared or combined.	In smart cities, this involves mapping data from traffic cameras and social media into a unified feature space to analyze traffic congestion in relation to weather [74,80].
Goal	to create a common feature space that allows for the effective comparison and integration of different data types.	By aligning traffic images and weather data, MML models create a joint representation that helps to analyze traffic congestion during different weather conditions [60].
Applications	It is widely used in cross-modal retrieval, where a system retrieves relevant data from one modality based on a query from another modality.	Cross-modal retrieval might involve retrieving relevant satellite images of an area based

 doi:10.20944/preprints202507.2268.v1

		on textual descriptions about a traffic incident [58,81].
Advantage	Learning shared representations allows for more context-aware decision-making, enabling models to integrate diverse insights more effectively.	By learning a joint representation, a smart city system can combine sensor data with social media sentiment to better respond to traffic incidents or public safety concerns.
Challenges	One challenge is ensuring that the representations learned are semantic and meaningful across different modalities, which require careful design and training.	In smart cities, aligning audio data (e.g., traffic sounds) with visual data (e.g., traffic cameras) may require sophisticated models to capture spatial and temporal context.

3.4.2. Cross-Modal Alignment

Cross-Modal alignment is crucial for preserving both the temporal and spatial integrity of multimodal datasets. For example, aligning high-frequency audio signals from traffic sounds with sporadically collected traffic data requires advanced techniques such as interpolation and resampling to ensure synchronization. Misalignments in these data streams can lead to misinterpretation by models, which may result in flawed decision-making in urban planning.

Successful alignment and fusion have been shown to significantly enhance the performance of traffic management systems, ensuring that decisions are based on coherently synchronized data, thus improving the accuracy and reliability of real-time traffic control and planning [46]. Cross-modal alignment process is shown in Figure 10. Several challenges should be addressed to achieve effective and accurate integration of multimodal data. Below are the key challenges in cross-modal alignment, along with proposed solutions:

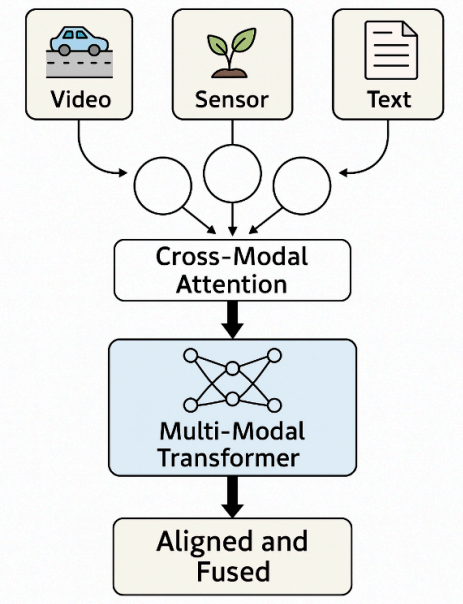

Figure 10. Illustration of the cross-modal alignment and fusion process, showing how different modalities (e.g., video, sensor data and text) are aligned in time and space and fused using advanced techniques like attention mechanisms and transformers.

Temporal Misalignment

 doi:10.20944/preprints202507.2268.v1

Different modalities often have different sampling rates or time intervals. For example, video data from traffic cameras may be captured at high frequency (e.g., 30 frames per second), whereas sensor data (such as air quality or traffic flow readings) might be recorded at a much lower frequency (e.g., once every minute). Temporal misalignment can lead to inaccurate data synchronization, where the model may struggle to combine data from these sources at the right moments in time.

Interpolation and resampling techniques can be applied to synchronize data from various sources, adjusting the frequency of the data so that it aligns temporally. Cross-modal attention mechanisms, especially in transformer-based models, can help assign weights to different data points based on their temporal relevance, ensuring that data from different time intervals are fused appropriately [82]. These mechanisms allow the model to focus on the most relevant data at each time step, improving temporal synchronization.

Spatial Misalignment

Spatial misalignment occurs when data from different modalities, such as satellite images and high-definition video, have different spatial resolutions or are captured from different perspectives. In smart city applications, sensor networks may have varying spatial resolutions, making it difficult to align them with geospatial data or video feeds from traffic cameras, which often have higher resolution.

Multi-modal transformers can help align spatial features from different modalities by transforming them into a shared representation space where their relationships can be learned and integrated effectively. Feature alignment techniques use techniques like spatial down sampling or up sampling to bring data to a common spatial scale [83]. CNNs [47] can be used to extract and align features across various modalities before fusion.

Scalability and Real-Time Constraints

Scalability and real-time constraints are critical factors in the deployment of multimodal machine learning models in smart city systems. As models need to handle large-scale data from diverse urban environments, ensuring that they can scale effectively without losing accuracy is a primary concern. A model trained in one city should generalize well to other cities with different sensor deployments and data collection protocols. This is especially important given the challenges of data heterogeneity and variability across cities. Research has shown that high model robustness can significantly reduce the need for local retraining, with some studies indicating that this can reduce localized adjustments by up to 30% [84].

Another challenge to scalability is dealing with real-time constraints. In dynamic environments such as traffic management or environmental monitoring, decisions must be made almost instantly. Models need to process and analyze data in real-time which often involves large, heterogeneous datasets from sources like traffic cameras, sensors and social media feeds. Ensuring that multimodal models can process this data with minimal latency while maintaining high accuracy remains a major hurdle.

The challenge of limited labeled training data exacerbates both scalability and real-time issues. In urban settings, data collection is often sparse and expensive particularly when it comes to labeled multimodal data. This makes it difficult to create large, high-quality training datasets. Techniques such as transfer learning [65] and semi-supervised learning [85] are valuable in this context as they allow models to leverage unlabeled data effectively. Transfer learning can help reduce the reliance on labeled data by up to 40% while maintaining and sometimes improving, model accuracy [86].

Co-learning approaches [87] where different modalities reinforce each other, can help address data disparities and enhance model performance. For example, visual data may be abundant but often noisy, while auditory data may be less frequent but of higher quality. By balancing these strengths and weaknesses in a co-learning framework, models can enhance their predictive accuracy for tasks like urban event detection, with studies showing up to a 15% improvement in accuracy when properly balanced [88,89].

 doi:10.20944/preprints202507.2268.v1

3.4.3. Robustness to Missing or Noisy Modalities

In the context of robustness to missing or noisy modalities, a significant challenge in smart city applications arises from the inherent imperfections in the data collected from diverse sources. Urban environments generate data from multiple sensors, cameras and social media feeds, which can be prone to missing values due to sensor failures or network disruptions. The data can be noisy, often due to factors such as ambient interference, poor calibration, or even environmental conditions that distort the readings. This missing or noisy data can significantly affect the accuracy and reliability of models particularly when they are tasked with real-time decision-making in areas like traffic management, environmental monitoring or public safety.

To address these challenges, several solutions have been proposed [90]. For handling missing data, data imputation techniques [91] are commonly applied, where missing values are predicted or estimated based on the relationships found within the rest of the dataset. Machine learning-based imputation techniques such as autoencoders or k-nearest neighbors can be effective in these situations, allowing the system to estimate missing data without requiring manual input [92]. When dealing with noisy data, various preprocessing techniques such as smoothing, filtering, or more advanced signal processing methods like Kalman filters can help clean the data before it is processed by the model. In addition, using robust machine learning algorithms with outlier detection or robust loss functions allows the model to be less sensitive to noise and anomalies in the data, ensuring that it is more resilient to irregularities [93].

Ensemble learning techniques [94], which combine the outputs of multiple models, can improve robustness by reducing the likelihood that a single noisy or missing modality will skew the results. By aggregating predictions from diverse sources, ensemble methods provide an inherent redundancy, making the overall system more robust and less prone to the errors introduced by individual faulty data streams. These solutions collectively enhance the reliability, accuracy and resilience of multimodal models ensuring that smart city systems can continue functioning effectively despite the challenges posed by missing or noisy data. As a result, these models improve real-time decision-making capabilities, maintaining high performance even in dynamic, data-intensive environments.

3.4.4. Interpretability of Models

Interpretability and explainability of multimodal models used in urban environments are critical for ensuring accountability and gaining the trust of stakeholders. Given that these models often drive decisions related to public services, understanding how these decisions are made is essential for ensuring transparency. Deciphering the inner workings of such complex models can be challenging due to the multifaceted nature of the data involved. Enhancing the transparency of these models can significantly improve stakeholder acceptance, with studies indicating that such improvements could boost trust by up to 20% [95,96].

To improve interpretability and explainability of machine learning models, we need to use inherently interpretable models like decision trees or linear regression which offer clear insights into how features affect predictions [97]. For complex models, techniques such as SHAP and LIME provide local explanations by highlighting the contribution of each feature to individual predictions while partial dependence plots show how changing a feature impacts the overall model output [98]. Surrogate models [99] can approximate black-box models with simpler ones for better understanding and visualizations help communicate model behavior clearly. Counterfactual explanations reveal what minimal changes in input would flip a prediction, making the decision process more transparent and actionable.

Therefore, fostering the development of tools and methodologies that explain the decision-making process of these multimodal systems is crucial for their adoption and successful deployment in smart city applications.

3.4.5. Dataset and Benchmark Limitations

 doi:10.20944/preprints202507.2268.v1

A significant obstacle to progress in MML for smart city systems is the limited availability of high-quality, task-relevant and well-annotated multimodal datasets. Most publicly available datasets are unimodal or lack the granularity and synchronization necessary for effective cross-modal learning. For instance, datasets like Cityscapes [100] focus on dense pixel-level annotations for urban scene understanding from RGB images but do not incorporate other modalities such as audio, environmental sensor data, or textual reports. Similarly, xView and xView2 [101] provide satellite imagery for infrastructure damage assessment but lack temporal data streams or complementary modalities like geolocation-based citizen input. Platforms like OpenSenseMap [102] and AQICN [103] offer IoT-based sensor data (e.g., air quality, temperature) but are often isolated from visual, behavioral, or linguistic signals critical for real-world urban decision-making.

These datasets are also inconsistent in spatial resolution, sampling frequency and annotation quality. For example, synchronizing a high-frame-rate CCTV feed (30 Hz) with low-frequency air quality measurements (0.01 Hz) poses challenges for cross-modal alignment. Geolocation data (e.g., from mobile phones or GPS tagged reports) may exhibit drift, sampling bias, or temporal sparsity. Social media-derived textual data lacks structured annotation, grounding and temporal correlation with physical sensor events. These discrepancies degrade the performance of models designed for joint representation learning and result in fragile inference pipelines under real-world conditions.

In terms of evaluation, most existing benchmarks do not sufficiently reflect the real-time, multimodal and high-stakes nature of smart city applications [104]. Tasks like visual question answering, caption generation or sentiment classification fail to capture the requirements of urban intelligence system such as multimodal event detection, incident forecasting and policy-aware decision support. Evaluation metrics like top-1 accuracy or F1-score are insufficient on their own to assess performance in dynamic, streaming environments [1,105].

To address these issues, we propose the development of domain-specific benchmarks with standardized tasks across common smart city application areas (e.g., traffic congestion prediction, multimodal public safety alerting, environmental anomaly detection) [106]. Such benchmarks should include multimodal data streams (visual, sensor, textual, geospatial, behavioral) collected under realistic urban conditions. Event-based labels and temporal markers to facilitate supervised learning and temporal segmentation and modality dropout settings to test model robustness under partial data availability [107].

The development of shared datasets and evaluation standards guided by such task-aware metrics will be essential for the reproducible benchmarking and large-scale deployment of MML systems in smart cities.

4. Applications of Multimodal Machine Learning in Smart Cities

As modern cities evolve into complex, sensor-rich environments, there is a critical need for analytical frameworks that can derive actionable insights from diverse and distributed data sources. Multimodal machine learning offers a powerful solution to this challenge by enabling joint processing and learning from heterogeneous urban data such as geospatial information, environmental sensor readings, video feeds, acoustic signals and unstructured text.

Although still an emerging field, MML has already found various applications in smart cities:

- Congestion prediction: In cities like Singapore and Barcelona, pilot projects utilize CCTV footage, traffic signal timings, vehicle GPS logs and Twitter sentiment analysis to forecast peak congestion points and proactively manage traffic flows [108,109].
- Multimodal surveillance for event monitoring: During large public events, integrated systems that analyze live video, crowd noise levels and social media activity help detect and manage instances of unrest or overcrowding [110].

 doi:10.20944/preprints202507.2268.v1

- Environmental monitoring and citizen sensing: Cities like Amsterdam and Seoul have employed MML to combine sensor readings (e.g., CO_2, temperature, noise), satellite imagery and mobile app reports from citizens to monitor pollution and urban heat islands in real time [111,112].
- Healthcare and epidemiology surveillance: In response to the COVID-19 pandemic, some regions have experimented with merging data from wearable devices, public health databases and mobility tracking to understand the spread and impact of the virus at a neighborhood level [113].

This section highlights the role of MML across eight core domains of smart city infrastructure: traffic and transportation, environmental monitoring, public safety and surveillance, urban planning and infrastructure, citizen engagement and services, IoT Platforms, cloud computing and edge computing. Each domain illustrates how MML techniques can effectively integrate multimodal data to improve situational awareness, optimize operations and support real-time decision-making in urban environments. Figure 11 provides an overview of the primary application areas in which MML contributes to smart city functionality, reflecting the integration of diverse data modalities in urban systems.

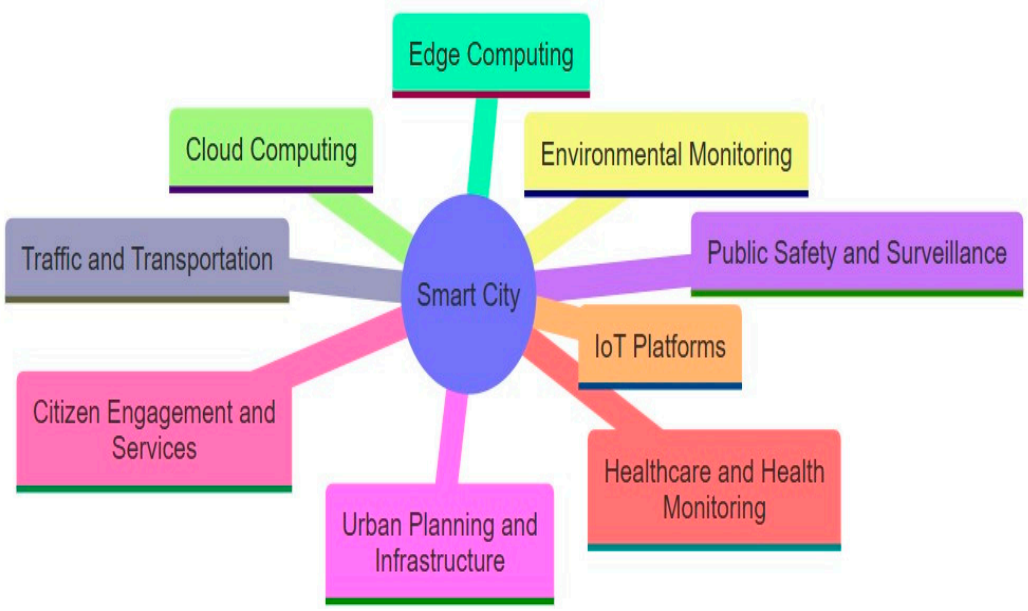

Figure 11. Framework of key application domains where MML is applied in smart cities. The central role of MML enables the integration of diverse data modalities across urban systems, supporting advanced capabilities in transportation, environment, safety, healthcare, infrastructure, citizen services and computational platforms such as IoT, cloud and edge computing.

Table 8 summarizes the significant impacts that MML has had across key aspects of smart city applications. These impacts reflect real world improvements in various sectors, including transportation, energy management, public safety, environmental monitoring and urban planning. The table provides quantitative data on the benefits realized from MML integration in smart city contexts.

Table 8. Impact of MML on key aspects of smart cities.

Aspect of Smart City	Impact of MML	Notes
Transportation	20% Reduction in Traffic Congestion [20]	Achieved through traffic flow prediction, adaptive signals and incident detection.

 doi:10.20944/preprints202507.2268.v1

Energy Management	15% Increase in Energy Efficiency [23]	Enabled by demand forecasting and optimized grid operations via MML.
Public Safety	30% Decrease in Emergency Response Times [24]	Real-time data fusion improves emergency detection and resource dispatch.
Environmental Monitoring	25% Reduction in Air Pollution-related Illnesses [15]	Sensor data integration helps in pollution forecasting and alerts.
Urban Planning	10% Improvement in Urban Infrastructure Efficiency [25]	Supports better zoning, infrastructure usage and investment decisions.

4.1. Traffic and Transportation

The infrastructure of urban mobility relies on a dense and heterogeneous network of data streams, including traffic flow sensors, road surveillance cameras, vehicular GPS traces and even public sentiment signals from social media platforms. As shown in Figure 12, MML models can effectively integrate these diverse sources to capture complex spatiotemporal correlations, thereby enabling more accurate traffic forecasting, real-time incident detection and adaptive traffic signal control even in complex environment i.e. during night [114].

For example, hybrid deep learning models that combine convolutional neural networks (CNNs) for processing visual data from traffic cameras with long short-term memory (LSTM) networks for sequential GPS data have demonstrated improved city-wide travel time prediction [48]. Integrating sensor-based traffic flow data with social media text streams enhances the detection of congestion hotspots faster than unimodal systems [115].

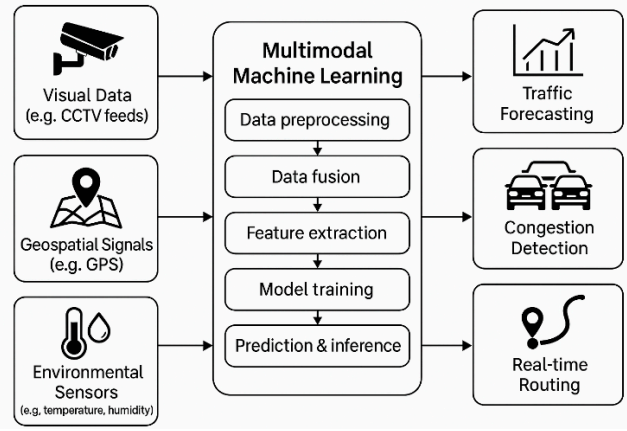

Figure 12. Integration of multimodal data for smart traffic prediction and management. The diagram illustrates how visual data (e.g., CCTV feeds), geospatial signals (e.g., GPS) and environmental sensors are fused using multimodal machine learning techniques to enable accurate traffic forecasting, congestion detection, adaptive signal control and real-time routing in smart cities.

Multimodal sensing is also crucial for autonomous mobility systems, where data from vehicle-mounted sensors must be fused with live traffic analytics to optimize navigation and reduce accident risk. MML techniques [109] support various transportation applications, including route optimization, anomaly detection and predictive maintenance of transport infrastructure.

Figure 12 visually represents the integration of diverse multimodal data sources in intelligent transportation smart city environment, processed by a central MML system. It shows how visual data (e.g., cityscapes, drone footage), data from geospatial sources (e.g., GPS, vehicle tracking) and sensor-based data (e.g., environmental sensors, fitness trackers), acts on this information, optimizing various urban functions such as traffic forecasting, congestion detection and real time routing. This

 doi:10.20944/preprints202507.2268.v1

interconnected network of data sources enables the real-time decision-making required for an intelligent, adaptive urban environment.

4.2. Environmental Monitoring

Urban ecosystems are increasingly challenged by environmental threats such as air quality degradation and noise pollution. Addressing these issues requires the effective integration of multiple data sources, a process in which MML plays a pivotal role. MML techniques focus on combining diverse data modalities, including IoT-based environmental sensors, satellite imagery and weather stations, to improve the accuracy and timeliness of environmental monitoring as shown in Figure 13.

As highlighted [15] data fusion techniques are essential in smart urban environments, where data from various sources must be integrated to offer comprehensive insights. These methods enhance environmental forecasting and enable more responsive actions by urban authorities. For example, real-time air quality index (AQI) measurements can be fused with data on wind patterns and vehicular density to predict pollution spikes, allowing for timely mitigation strategies [116].

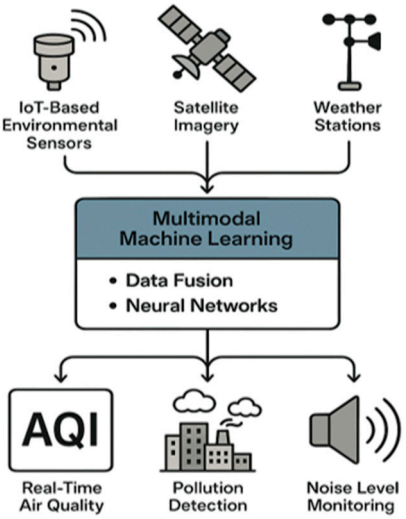

Figure 13. Environmental monitoring with the integration of IoT sensors, satellite imagery and weather data via MML to manage urban air quality, detect pollution and monitor noise levels.

The advent of the IoT solutions [15,64] has significantly advanced smart environmental management by enabling continuous monitoring of air quality, temperature and noise levels through real-time sensor data streams . In addition, visual and textual data, such as satellite images and social media posts, can complement sensor data to identify pollution hotspots and potential environmental hazards. When these diverse data types are fused using techniques such as autoencoders and graph-based neural networks, they provide high-resolution insights that allow urban planners to take proactive action [27,63]. The integration of multimodal data empowers cities to predict and mitigate environmental hazards before they escalate, supporting sustainable development and fostering healthier urban living conditions.

4.3. Public Safety and Surveillance

In modern urban environments, MML is playing an increasingly crucial role in enhancing public safety systems by monitoring urban threats and enabling real-time emergency responses. Practical applications demonstrate how big data systems can enhance public safety through real-time surveillance, anomaly detection and emergency response coordination [117]. By integrating data from various sources such as surveillance cameras, acoustic signals and emergency dispatch logs,

 doi:10.20944/preprints202507.2268.v1

MML systems are capable of accurately identifying and classifying complex events, such as fights, crowd surges, or fire outbreaks as shown in Figure 14.

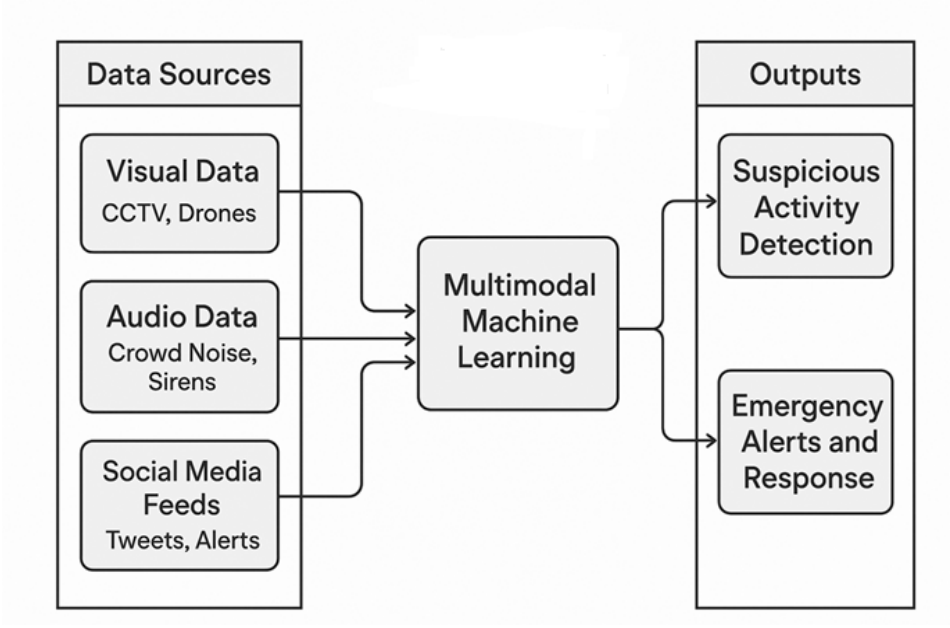

Figure 14. System flow for public safety and surveillance. Visual, audio and social media data sources are integrated using machine learning to detect suspicious activity and trigger emergency alerts and responses.

For instance, MML based systems utilize audio anomaly detection to identify suspicious sounds ranging from cracking glass to shouting and cross-verify these auditory signals against nearby CCTV video feeds [118,119]. This fusion of modalities allows for more accurate event classification and reduces false positives, which is particularly important in environments with high volumes of data and potential incidents.

MML systems are especially effective in densely populated areas, where traditional single-modality monitoring (e.g., relying solely on video surveillance) may not provide the comprehensive coverage needed to detect or assess incidents in real-time. Augmenting these systems with social media data, including tweets and posts related to events, significantly enhances the detection and analysis of public safety issues, allowing authorities to respond more effectively.

4.4. Urban Planning and Infrastructure

Urban planning increasingly relies on the integration of diverse data sources such as satellite imagery, census data and mobility data, to gain a comprehensive understanding of urban dynamics and make informed development decisions as presented in Figure 15. Satellite imagery provides high-resolution visual data that tracks land use changes, urban sprawl and infrastructure development over time. By analyzing these images, urban planners can identify areas in need of development, monitor environmental changes and assess the impact of urban policies. Census data offers valuable demographic insights, such as population density, age distribution and socio-economic characteristics, which are essential for shaping policies and development strategies [120].

 doi:10.20944/preprints202507.2268.v1

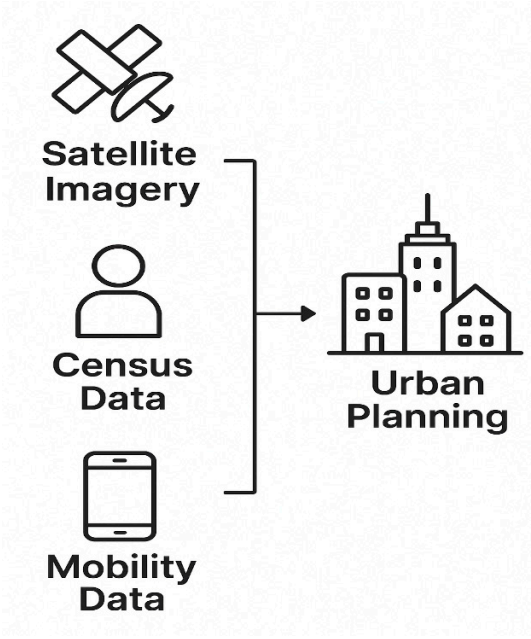

Figure 15. Simplified illustration of multimodal data integration in urban planning, combining satellite imagery, census data and mobility data to support informed decision-making.

Mobility data, gathered from IoT sensors and GPS devices, provides real-time information on transportation patterns, congestion and public transit use. This multimodal data integration allows planners to gain a more holistic view of urban development and aids in decision-making for critical areas such as housing, transportation and infrastructure development.

Recent studies demonstrate how MML techniques such as data fusion and clustering can be used to merge satellite imagery, census data and mobility information to model future urban growth scenarios, predict traffic congestion and allocate resources more effectively [121,122]. By leveraging deep learning techniques like CNNs for satellite image analysis and recurrent neural networks (RNNs) for time-series mobility data, planners can forecast urban expansion, evaluate the impact of new infrastructure projects and optimize the use of urban spaces [47,49].

This data-driven, multimodal approach to urban planning ensures that development is sustainable, efficient and responsive to the needs of growing urban populations, promoting more resilient and adaptive cities.

4.5. Citizen Engagement & Services

Modern urban governance increasingly emphasizes responsiveness and public participation. MML models play a key role in processing a variety of user-generated data, including feedback from civic applications, social media sentiment and location-based data, helping city authorities prioritize and address infrastructure or service-related issues as depicted in Figure 16. For example, complaints tagged with geolocation data, such as those about potholes, power outages, or sanitation issues, can be clustered and visualized in real-time, enabling authorities to allocate resources efficiently. By combining this data with event schedules or weather alerts, MML facilitates better decision-making and enhances the city's ability to respond to emerging problems promptly [123].

 doi:10.20944/preprints202507.2268.v1

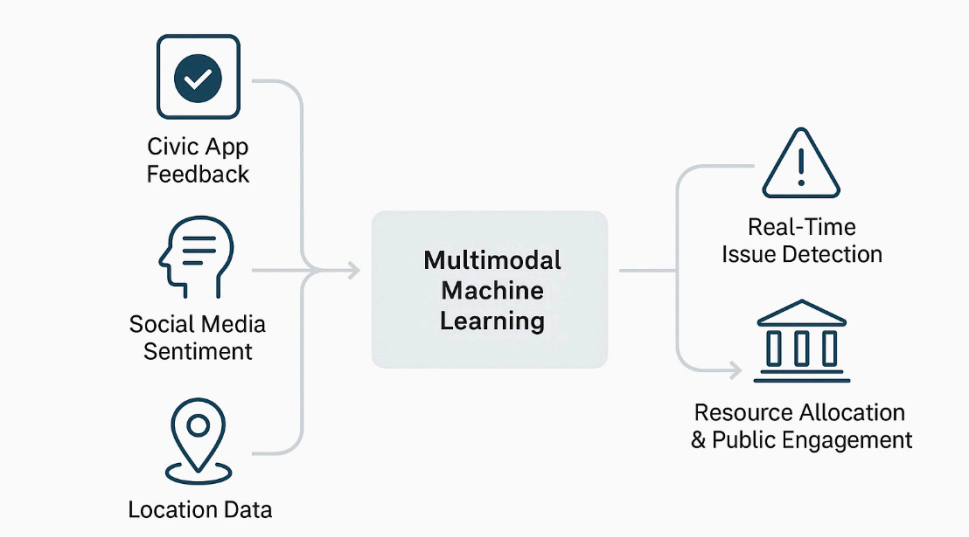

Figure 16. Conceptual diagram showing how MML integrates user-generated data such as civic app feedback, social media sentiment and location data for real-time issue detection, resource allocation and improved public engagement in urban governance.

Chatbots integrated with MML systems offer real-time interaction, allowing citizens to report issues, ask questions, or receive immediate responses regarding city services. These systems can also leverage social media as a form of social sensing, where community-driven feedback highlights issues, ideas, or complaints that emerge in real-time. Regular analysis of social media posts helps identify key themes and community sentiment, enabling more proactive governance [124].

Sentiment analysis of unstructured social media data provides valuable insights into public opinion and emotional trends over time. By tracking citizens' moods and behavioral patterns, city authorities can gain a deeper understanding of public concerns and tailor their responses accordingly. This proactive approach helps improve citizen satisfaction, build trust and foster stronger relationships between government and residents, ultimately enhancing the effectiveness of public services.

4.6. IoT Platform

The Internet of Things (IoT) refers to the network of billions of interconnected devices that communicate and exchange data across various platforms, creating a smart society as shown in Figure 17. These devices utilize standardized communication protocols to share information, thus enhancing the connectivity and efficiency of critical urban infrastructures such as mobility, safety, entertainment, agriculture and healthcare.

 doi:10.20944/preprints202507.2268.v1

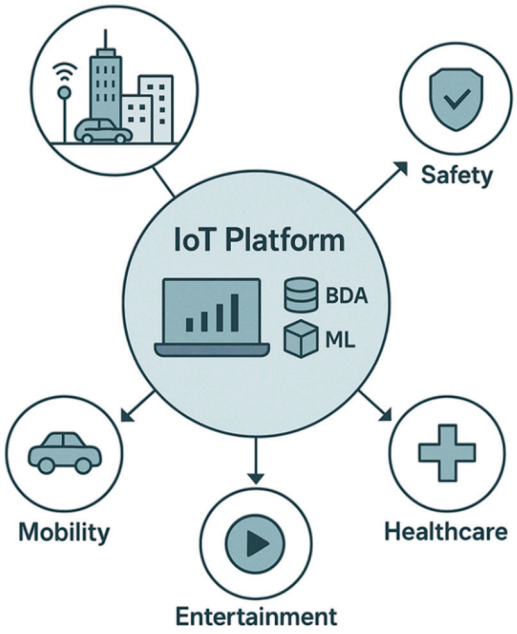

Figure 17. Conceptual illustration of an IoT platform for smart cities, showing interconnected devices and sensors collecting multi-source data from critical infrastructures (e.g., mobility, safety, healthcare). The platform utilizes big data analysis and machine learning for real-time data processing, intelligent governance and secure analysis.

For instance, recent work proposed a distributed CNN parallelism strategy, integrated with big data analysis (BDA), to process the multi-source data gathered from smart cities [125]. Their research highlights the role of digital twins in smart cities, where IoT-BDA systems, powered by machine learning, help manage the massive volumes of data generated, facilitating intelligent governance and advanced, secure data analysis. In a similar vein, work explores innovative methods for vehicle number estimation and direction of arrival (DOA) estimation for non-circular signals. By leveraging machine learning, they improve the efficiency of non-circular signal estimation, enabling better traffic monitoring and management in smart cities.

Moreover, recent work [126] demonstrates that integration of blockchain with MML approaches could enhance privacy and security in IoT-based smart city applications. Their smart blockchain architecture, utilizing a Proof of Work (PoW) mechanism and smart contracts, ensures data integrity and privacy. They introduce principal component analysis-based transformation method for data encryption, which was shown to outperform existing privacy-preserving intrusion detection approaches.

Energy management solutions are crucial for the sustainability of IoT applications in smart cities, addressing power efficiency and operational longevity of connected devices. The issue of energy efficiency in the public sector by developing ML systems for predicting energy consumption is addressed [127]. Their research proposed a smart, ML-based energy management system for public sector buildings in smart cities, using data from IoT networks and energy management information systems. By applying models such as deep neural networks (DNN) and random forest, they demonstrated the ability to accurately predict energy usage. Their architecture involves six levels, from data collection to predictive modeling and showed that the random forest model yielded the most accurate results, with a symmetric mean absolute percentage error (SMAPE) of 13.59%, a metric used to evaluate prediction accuracy by comparing forecasted and actual values as a percentage.

Finally, [128] focuses on predicting air quality in smart cities, addressing the challenges posed by overcrowding and urban development. They introduce the LSTM-SAE model, combining LSTM networks and stacked auto-encoders (SAE), to forecast air pollution levels. The LSTM model assesses

 doi:10.20944/preprints202507.2268.v1

air quality, while the SAE model extracts the intrinsic components of air pollution, helping urban planners design more sustainable and healthy urban environments.

4.7. Cloud Computing

Cloud computing plays a pivotal role in the development and operation of smart cities, providing the infrastructure to store, process and analyze vast amounts of urban data. Innovative clustering techniques using k-means are used to analyze a large dataset of household energy consumption over a decade [129]. By dividing the data into three seasonal clusters, they were able to highlight fluctuations in consumption, which are influenced by weather patterns. This methodology is crucial for smart city applications, such as energy management where understanding and predicting energy usage based on environmental factors can improve efficiency and reduce waste.

Cloud platforms enable smart cities to store and access data remotely, allowing for continuous and real-time analysis. This infrastructure is essential for urban functions such as traffic management, municipal security and government services, all of which rely on cloud-based systems to provide intelligent solutions for city management. The ubiquity of cloud computing ensures that urban decision-makers can access the data they need at any time, from any location.

The use of semantic maps composed of subject-action-object triplets derived from textual data is introduced [130]. By applying knowledge-based and deep learning algorithms, these maps are codified into formal ontologies, unifying fragmented knowledge across administrative levels. This approach enhances decision-making by offering a comprehensive view of smart city initiatives, enabling policymakers to analyze and address urban challenges at the international, national and local levels.

4.8. Edge Computing

Edge computing addresses the limitations of traditional cloud computing by processing data closer to its source, at the "edge" of the network. This approach reduces the volume of data sent to central servers, thereby alleviating bandwidth strain and reducing latency for time-sensitive applications. By decentralizing computational power, edge computing improves performance, particularly for intelligent transportation, healthcare and social services, where immediate responses are crucial.

Various challenges in smart cities highlighting the pivotal role of IoT in reducing data traffic, especially between sensors and IoT nodes are discussed [131]. They identify limitations with existing compression algorithms used in IoT systems, which struggle with the limited memory capacity of devices. Lossy compression methods are not acceptable due to data loss during transmission, while lossless compression techniques are difficult to implement on resource-constrained IoT devices, posing a significant challenge in efficiently managing large-scale data traffic in smart cities.

A work proposed [132] states a networked architecture for machine learning in smart cities, focusing on the challenges of handling big data. Their design integrates data mining techniques for data collection, storage and evaluation, specifically addressing the volume, diversity and velocity of smart city data. This system is designed to manage the complexities of decentralized data in smart city environments by storing it in distributed clusters, facilitating scalable and efficient processing.

In the realm of real-time person recognition, deep learning approaches integrated with edge and cloud computing facilitate identification in crowded environments by processing high-resolution video streams. This method achieved a 95% detection accuracy, outperforming conventional algorithms and demonstrates the significant potential of edge computing for real-time image processing and video surveillance in smart cities.

4.9. Healthcare and Health Monitoring

Healthcare in smart cities is evolving to include continuous monitoring through wearable devices, electronic health records (EHRs) and environmental exposure data. MML frameworks

 doi:10.20944/preprints202507.2268.v1

enable the integration of temporal health metrics, contextual data and symptom self-reports to create unified patient models [23]. This multimodal approach supports predictive modeling for health events such as breathing distress or stress due to heat and facilitates remote diagnostics by cross-correlating biometric signals, speech and postural data. Such systems are particularly valuable for the aging population, enabling proactive healthcare management and real-time interventions.

Recent studies [133] have formulated predictive multimodal deep learning healthcare analytics frameworks integrating data from EHR, wearable devices and environmental sensors to predict health outcomes and enable timely interventions. These frameworks have shown significant improvements in predictive accuracy and patient outcomes. MML enhances healthcare in smart cities through smart health monitoring systems that analyze data from wearable devices, electronic health records and environmental sensors to predict health events and automate emergency responses as shown in Figure 18. For example, these systems can predict asthma attacks based on air quality data and individual health records, allowing for preemptive medical interventions. Integrating multimodal data from wearable sensors, medical records and environmental factors enhances predictive analytics and proactive health management in smart cities.

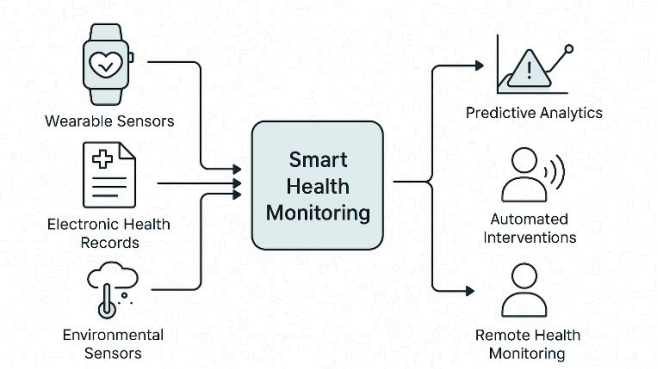

Figure 18. Conceptual illustration of a smart healthcare monitoring system in a smart city. Data from EHR, wearable sensors and environmental sensors are integrated using multimodal machine learning (MML) for predictive analytics, early detection and automated interventions. The system enables city-scale remote health monitoring and supports proactive, decentralized healthcare management.

In the context of smart healthcare, multimodal data often include physiological signals (e.g., heart rate, EEG), environmental conditions (e.g., air quality), imaging data (e.g., CT, MRI) and textual or tabular data from EHRs [17,23,133]. MML techniques facilitate the fusion of these modalities to improve diagnostic accuracy, early disease detection and intervention strategies. Feature level and decision level fusion strategies can integrate wearable sensor data with patient history to predict adverse events in real time [134]. Healthcare in smart cities also emphasizes remote and ambient health monitoring using non-invasive technologies. With the support of edge computing and AI, city-scale deployments can monitor chronic conditions and support elderly populations in home settings. This is particularly crucial in urban areas with growing aging populations and increased demand for decentralized healthcare services.

Information fusion methods are central to this transformation. As noted in systematic reviews [135], integrating multimodal streams from smart environments enhances the system's context-awareness and decision-making capacity. For instance, combining audio-visual cues with clinical data using deep learning enables early detection of neurological disorders and cognitive decline, which is challenging using unimodal systems.

Despite these advances, several challenges remain. Data alignment across modalities, especially those collected at different temporal resolutions or sampling rates, presents significant hurdles. Data quality, privacy and ethical concerns related to continuous health surveillance must be addressed through robust governance frameworks [17,135]. The potential of MML in urban healthcare is also influenced by broader smart city infrastructure and policy. For instance, intelligent transportation

 doi:10.20944/preprints202507.2268.v1

systems can provide mobility data for patients with disabilities, while environmental monitoring can inform public health interventions during air pollution episodes or pandemics.

The integration of large multimodal foundation models into healthcare applications may unlock new capabilities in transfer learning, cross-modal representation learning and generalization across urban contexts. However, future research must focus on making these systems explainable, equitable and interoperable with existing health information systems [136].

In conclusion, IoT platforms, cloud computing and edge computing are enabling technologies that drive the functionality and optimization of smart cities. These technologies enable seamless data collection, processing and analysis across various domains such as transportation, environmental monitoring, public safety, healthcare and citizen engagement. By facilitating real-time decision-making and efficient resource management, they play a crucial role in improving urban living conditions. As these technologies continue to evolve, they will further enhance the sustainability, efficiency and responsiveness of smart cities, paving the way for smarter, more connected urban environments.

5. Challenges and Limitations of MML Deployment in Smart Cities

Although MML offers tremendous opportunities in revolutionizing smart cities, there are several challenges and limitations to consider and overcome to enable successful deployment. Challenges in data privacy, algorithmic bias, scalability, interoperability among others are abundant.

5.1. Privacy and Security Concerns

Smart cities produce massive amounts of sensitive data that can include people's personal information, vehicles and public interactions. A key challenge is data privacy and security because one must be careful to prevent access to the data by anyone except the system. MML models should be secured by using differential privacy [137] and data anonymization techniques [138] to protect citizen data. Privacy concerns in smart cities are critical due to the extensive data collection from sensors, cameras and IoT devices. While these technologies enhance urban services and resource management, they also pose significant risks to individual privacy. The aggregation of data can lead to the creation of detailed profiles, potentially exposing sensitive information about citizens' behaviors and movements without their consent. Researchers emphasize the need for robust data governance frameworks that prioritize transparency, informed consent and stringent data protection measures [139,140] . Additionally, engaging communities in discussions about privacy can help build trust and ensure that the deployment of smart technologies aligns with public expectations and ethical standards.

Privacy-preserving techniques [141] are of utmost importance in MML for smart cities, allowing for full data utility while preserving individual privacy. Differential privacy inserts noise into the data at a scale that is proportionate to sensitivity and, thus, effectively protects it from identification [137]. For example, with the use of differential privacy, the risks of re-identification can be drastically reduced without compromising that utility for a traffic management application.

It is important to ensure that such laws are complied with through robust data governance frameworks, ensuring regulatory compliance with very strict laws such as GDPR [139] for data processing and user consent. This is important because non-compliance with the same may result in penalties equaling up to 4% of the company's annual global turnover or 20 million euros, whichever is higher. Governing frameworks clearly defines the exact role and responsibility because managing multimodal data emanating from multiple sources, releasing many IoT devices, public surveillance systems and whatnot, remains the order of the day.

Security concerns in smart cities are critical due to the interconnected nature of IoT devices and urban management systems. These technologies create vulnerabilities that can be exploited by cybercriminals, leading to potential attacks on critical infrastructure such as transportation, energy, healthcare and public safety. Such breaches can result in service disruptions, data theft and even physical harm. The lack of standardized security protocols across different devices further

 doi:10.20944/preprints202507.2268.v1

complicates protection efforts. Researchers emphasize the need for robust security frameworks, including encryption [142], access controls [143] and regular security assessments, to safeguard urban environments [144]. Collaborative efforts among technology developers, city planners and law enforcement are essential to establish resilient smart city ecosystems that prioritize security and build public trust.

The rapid proliferation of IoT devices in smart cities has significantly increased their vulnerability to cyber-attacks. A report [145] indicates that such attacks have surged by 300% in the last two years. This growing threat landscape highlights the urgent need for comprehensive cybersecurity frameworks and regular security audits. Implementing such measures can potentially reduce breaches by up to 60%, thereby safeguarding critical infrastructure and sensitive data. In parallel, secure data transmission and storage practices play a vital role in defending against cyber threats. Techniques such as AES-256, encryption and controlled access mechanisms have proven effective in minimizing risks [146]. Adopting advanced encryption and secure storage protocols has led to a 40% reduction in unauthorized data access incidents within smart city networks [147]. Together, these cybersecurity strategies form the foundation of a resilient and secure smart urban ecosystem.

5.2. Ethical Considerations

Ethical considerations in smart cities are increasingly important as they navigate the complexities of technology integration and data usage. The collection and analysis of vast amounts of data from citizens raise significant ethical questions regarding consent, ownership and the potential for surveillance. There is a need to revolutionize transparent policies that inform citizens about how their data is used and the implications for their privacy.

Ethical frameworks must ensure that technologies are implemented equitably, avoiding biases that could compromise fairness and inclusion. There is also a critical need to address the ethical implications of automated decision-making systems, which may lack accountability and transparency. Engaging with diverse stakeholders, including community members, ethicists and policymakers, is essential to develop inclusive guidelines that prioritize human rights and social equity. Addressing these ethical considerations is vital for fostering trust and ensuring that smart city initiatives enhance the quality of life for all residents.

This bias in such large data often engaged in multimodal models, might lead to unfairness and discrimination [148]. Based on limited data MML systems can unintentionally reproduce bias in decision-making processes, meaning that if the original data contains such bias, the societal inequalities will be reproduced [43]. Bias in these smart city applications use cases can be particularly harmful to marginalized communities. Algorithmic bias is tackled through carefully curated and representative data & frequent audits of the model to maintain fairness [28,149]. Two key follow-up steps that should be taken to mitigate this risk are the diversification of training sets and the use of fairness-aware algorithms in those models [150]. Diversified training sets can reduce bias in urban surveillance algorithms by up to 30%, hence making outcomes fairer for law enforcement and social services.

It is important for public trust that transparency and accountability be observed. As noted in [151], public trust in smart city applications could be raised to 25%, when transparent decision-making processes and audit trails are implemented. This requires that decision rationale be accessible and comprehensible to stakeholders and that clear mechanisms exist for redress and correction when errors occur.

Ethical frameworks must mandate continuous monitoring, bias detection and mitigation strategies to uphold equity. Defining clear accountability structures is necessary to assign responsibility for decisions or harms caused by MML systems. This includes legal and ethical responsibilities of developers, operators and governing bodies, particularly in sensitive smart city domains such as public safety and healthcare.

 doi:10.20944/preprints202507.2268.v1

6. Research Gaps and Future Directions in Multimodal Sensing for Smart City Applications

While the challenges related to MML and its deployment in smart cities have been extensively discussed in previous sections, several research gaps as shown in Table 9 remain unaddressed. These gaps highlight the need for more comprehensive, scalable and context-aware solutions beyond state-of-the-art limitations. In this section, we identify the key areas where further investigation is essential to advance the field and suggest future research.

Table 9. Prospective research directions in multimodal machine learning for smart city applications, highlighting key areas for future development, including handling big data, improving model interpretability, addressing privacy concerns, enhancing robustness and exploring novel data modalities to advance smart city technologies.

Research Direction	Description
Handling Big Data [152]	Developing scalable MML algorithms for large datasets
Improving Models Interpretability [153]	Exploring techniques for explaining complex MML models
Addressing Privacy and Ethical Concerns [142]	Investigating methods for preserving privacy in MML models
Enhancing Robustness [154]	Researching strategies for improving the robustness of MML
Exploring Novel Data Modalities [155]	Investigating the use of emerging data modalities in MML

Table 9 presents several promising research directions that can drive the development of MML for smart city applications. These directions are pivotal in overcoming the challenges faced by MML models when applied to urban environments.

Handling Big Data: One of the primary challenges in smart cities is managing the vast amounts of data generated by various sensors, cameras and IoT devices. Current models often struggle to scale effectively to the volume, velocity and variety of urban data streams, limiting their practical deployment [152]. There is no universal "one-size-fits-all" MML architecture capable of handling diverse data modalities and application requirements across different smart city domains [33]. Real-time processing demands further exacerbate these challenges, as latency-sensitive applications require efficient and rapid data fusion. Resource constraints on edge devices complicate the deployment of large-scale models outside centralized cloud environments. Future research should focus on developing modular, adaptive and scalable MML frameworks that leverage distributed and parallel computing [156], incorporate data reduction technique and support incremental learning [157] .

Artificial Neural Networks (ANNs) including deep learning models are widely used to analyze smart city data. While traditional ANNs struggle with large complex datasets, deep learning offers greater capacity but requires substantial computational resources, limiting real-time deployment in resource-constrained urban environments. Shallow machine learning methods remain in use but fall short in efficiently processing large-scale, multimodal data [158]. This underscores a critical research gap for scalable, lightweight and adaptable multimodal machine learning models capable of handling heterogeneous urban data effectively.

The deployment of large language models (LLMs) at the edge of networks is opening new opportunities for real-time decision-making. Integrating LLMs on edge devices can facilitate complex multimodal data processing directly within the city infrastructure, improving scalability and responsiveness. The combination of deep learning with IoT big data analytics provides powerful tools for tackling urban development challenges [125]. There is still needed to bridge the performance gap between deep models and lightweight alternatives while ensuring scalability, interpretability and robustness in complex urban environments.

 doi:10.20944/preprints202507.2268.v1

One promising area of research is online hyperparameter estimation, which adjusts certain hyperparameters dynamically based on real-time data [159]. This approach could allow systems to adapt to changing environmental conditions. However, the unpredictability and complexity of real-world environments make it a challenging task to ensure optimal performance continuously. This gap in capability may offer opportunities to address existing limitations and inspire future developments in multimodal systems.

Improving Models Interpretability: As MML models become more complex, their interpretability becomes essential. Many MML models especially those based on deep learning function as "black boxes," providing high accuracy but little insight into how inputs from multiple modalities combine to produce outputs. This lack of transparency limits trust and acceptance among stakeholders such as city planners, policymakers and citizens. It's important for stakeholders such as city planners, policymakers and citizens to understand how these models make decisions.

Improving interpretability involves developing methods that explain model predictions in an understandable and actionable way [153]. Techniques include attention mechanisms that highlight important input features, model-agnostic explanation tools like SHAP or LIME and designing inherently interpretable architectures [98]. Enhancing model interpretability will enable better validation, debugging and ethical oversight, facilitating more responsible and trustworthy deployment of MML systems in smart city environments. Research into techniques that improve model transparency and explainability will foster trust in these systems, especially in critical applications like public safety and healthcare.

Addressing Privacy and Ethical Concerns: Urban data often involves sensitive information, including personal data from individuals and public safety data. The pervasive collection and processing of such data can lead to inadvertent invasions of individual privacy and potential misuse. Ethical frameworks must prioritize data minimization, restrict unnecessary surveillance and ensure compliance with legal regulations such as the General Data Protection Regulation (GDPR) and other jurisdiction-specific privacy laws [139]. Transparent policies on data retention and access control are vital to prevent abuse. Developing privacy-preserving methodologies for MML models, such as federated learning and differential privacy, will help safeguard citizens' privacy while enabling the benefits of smart city technologies[137].

Ensuring informed and voluntary citizen consent for data collection is a fundamental ethical obligation often under-addressed in smart city deployments. Consent mechanisms should be clear, accessible and revocable, empowering individuals with control over their personal information. The opaque nature of many MML models challenges trust, as decision-making processes may be difficult to interpret or contest. Algorithmic transparency and explainability are therefore critical for accountability, enabling stakeholders and citizens to understand how data inputs lead to particular outputs or actions [149]. Participatory design involving communities in setting ethical guidelines and oversight mechanisms can help align technological advancements with societal values.

Enhancing Robustness: Given the dynamic and sometimes unpredictable nature of urban environments, it is vital to develop MML models that are robust and capable of adapting to changes in the data they process. In smart city applications, multimodal data often suffer from noise, missing values, sensor failures and dynamic environmental changes, which can degrade model effectiveness.

Current MML models may be sensitive to such imperfections, leading to unreliable predictions that could compromise critical urban services like traffic management or emergency response. Improving robustness involves designing models that can handle noisy, incomplete, or corrupted data gracefully, ensuring consistent and trustworthy outputs. Research should focus on strategies for improving the resilience of MML systems, especially when faced with noisy, incomplete, or misleading data. This should investigate techniques such as robust fusion strategies, noise-resistant architecture, data augmentation and uncertainty quantification [160].

Exploring Novel Data Modalities: Smart city sensing traditionally relies on modalities such as cameras, GPS and environmental sensors. To meet the growing complexity and diversity of urban challenges, there is a need to explore novel data modalities that offer richer, complementary or more

 doi:10.20944/preprints202507.2268.v1

resilient information. Emerging modalities, such as acoustic sensors, LiDAR, thermal imaging, wearable devices and crowd-sourced social media data provide new dimensions of insight into city dynamics, human behavior and environmental conditions [155]. Investigating these novel data modalities can open new possibilities for enhancing urban intelligence and improving the accuracy and reliability of smart city applications. For example, acoustic sensing can improve the monitoring of urban noise pollution, traffic density or emergency situations in ways visual sensors cannot effectively capture. LiDAR and radar modalities provide high-accuracy 3D spatial information indispensable for autonomous transportation and infrastructure mapping. Thermal imaging allows energy efficiency monitoring and public safety surveillance in situations where visible light sensors do not function. Wearable sensors and biometric sensors introduce opportunities for personalized health and mobility information, supporting more intelligent public health management.

The combination of crowd-sourced information from mobile apps and social media delivers real-time, human-centered information, revealing events, sentiments or anomalies that static sensors might miss. New biochemical sensors, including those for air pollutants or environmental DNA (eDNA), yield essential information for urban environmental health and biodiversity monitoring [161].

Despite these encouraging prospects, several of these modalities are still underexploited or encounter data heterogeneity, privacy and deployment expense challenges. Thus, investigating and incorporating these new modalities within multi-modal learning frameworks is a crucial research gap and a thrilling future direction in the evolution of smart city sensing systems.

Collectively, these research directions will shape the future of MML in smart cities, ensuring that these systems are not only effective but also ethical, secure and adaptable to the complexities of urban environments.

7. Conclusions

Multimodal machine learning is rapidly transforming how smart cities perceive, understand and respond to complex urban phenomena. By integrating diverse data sources from visual, audio, geospatial, physiological and IoT modalities, MML provides a powerful foundation for intelligent sensing, contextual decision-making and adaptive urban services. The challenge of effective data fusion remains a central and unresolved problem within MML. Unlike traditional single-modality approaches, there is no universal fusion strategy that fits all scenarios; the optimal method depends critically on the specific data modalities, task requirements and data quality. Current solutions tend to be a diverse set of specialized techniques rather than a unified framework. This ongoing complexity underscores the need for continued research into adaptive, scalable and context-aware fusion mechanisms.

Our analysis also identifies critical challenges in deploying MML systems in real-world smart city environments, including data alignment, scalability, privacy, infrastructure constraints and ethical considerations. Despite these challenges, the synergy between MML and smart city systems offers significant opportunities. Emerging developments in large multimodal foundation models, edge AI and privacy-preserving learning present promising pathways to build robust, adaptive and human-centric urban intelligence systems. We conclude that the integration of MML into smart cities is not merely a technical enhancement but foundational to developing sustainable, resilient and responsive urban ecosystems. Future research should prioritize the development of scalable and generalizable fusion architectures, improve interpretability and ensure that these systems remain inclusive, transparent and ethically aligned with the needs of urban populations.

Author Contributions: Conceptualization, T.S. and C.W.O.; literature review and analysis, T.S.; writing—original draft preparation, T.S.; writing—review and editing, T.S. and C.W.O.; visualization, T.S.; supervision, C.W.O.; project administration, T.S. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202507.2268.v1

Funding: This research was funded by the Research Council of Norway under project number 310105, as part of the NORCICS - Norwegian Center for Cybersecurity in Critical Sectors initiative (2020–2028).

Data Availability Statement: Data are available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **2018**, *41*, 423–443.
2. Valet, L.; Mauris, G.; Bolon, P.; Keskes, N. A fuzzy rule-based interactive fusion system for seismic data analysis. *Information Fusion* **2003**, *4*, 123–133.
3. Florescu, D.; Koller, D. Using probabilistic information in data integration. In Proceedings of the In Proc. of the Int. Conf. on Very Large Data Bases (VLDB, 1997.
4. Buccella, A.; Cechich, A.; Rodríguez Brisaboa, N. An ontology approach to data integration. *Journal of Computer Science & Technology* **2003**, *3*.
5. Sharma, H.; Haque, A.; Blaabjerg, F. Machine learning in wireless sensor networks for smart cities: a survey. *Electronics* **2021**, *10*, 1012.
6. Anwar, M.R.; Sakti, L.D. Integrating artificial intelligence and environmental science for sustainable urban planning. *IAIC Transactions on Sustainable Digital Innovation (ITSDI)* **2024**, *5*, 179–191.
7. Ortega-Fernández, A.; Martín-Rojas, R.; García-Morales, V.J. Artificial intelligence in the urban environment: Smart cities as models for developing innovation and sustainability. *Sustainability* **2020**, *12*, 7860.
8. Ullah, Z.; Al-Turjman, F.; Mostarda, L.; Gagliardi, R. Applications of artificial intelligence and machine learning in smart cities. *Computer communications* **2020**, *154*, 313–323.
9. Pawłowski, M.; Wróblewska, A.; Sysko-Romańczuk, S. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors* **2023**, *23*, 2381.
10. Huang, X.; Wang, S.; Yang, D.; Hu, T.; Chen, M.; Zhang, M.; Zhang, G.; Biljecki, F.; Lu, T.; Zou, L. Crowdsourcing geospatial data for earth and human observations: A review. *Journal of Remote Sensing* **2024**, *4*, 0105.
11. Lahat, D.; Adali, T.; Jutten, C. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE* **2015**, *103*, 1449–1477.
12. Kang, H.-W.; Kang, H.-B. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one* **2017**, *12*, e0176244.
13. Srivastava, S.; Vargas-Munoz, J.E.; Tuia, D. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote sensing of environment* **2019**, *228*, 129–143.
14. Prawiyogi, A.G.; Purnama, S.; Meria, L. Smart cities using machine learning and intelligent applications. *International Transactions on Artificial Intelligence* **2022**, *1*, 102–116.
15. Alam, F.; Mehmood, R.; Katib, I.; Albogami, N.N.; Albeshri, A. Data fusion and IoT for smart ubiquitous environments: A survey. *Ieee Access* **2017**, *5*, 9533–9554.
16. Lifelo, Z.; Ding, J.; Ning, H.; Dhelim, S. Artificial intelligence-enabled metaverse for sustainable smart cities: Technologies, applications, challenges, and future directions. *Electronics* **2024**, *13*, 4874.
17. Nasr, M.; Islam, M.M.; Shehata, S.; Karray, F.; Quintana, Y.J.I.a. Smart healthcare in the age of AI: recent advances, challenges, and future prospects. **2021**, *9*, 145248–145270.
18. Myagmar-Ochir, Y.; Kim, W. A survey of video surveillance systems in smart city. *Electronics* **2023**, *12*, 3567.
19. Bello, J.P.; Mydlarz, C.; Salamon, J. Sound analysis in smart cities. In *Computational analysis of sound scenes and events*; Springer: 2017; pp. 373–397.
20. Lim, C.; Cho, G.-H.; Kim, J. Understanding the linkages of smart-city technologies and applications: Key lessons from a text mining approach and a call for future research. *Technological Forecasting and Social Change* **2021**, *170*, 120893.

 doi:10.20944/preprints202507.2268.v1

21. Musa, A.A.; Malami, S.I.; Alanazi, F.; Ounaies, W.; Alshammari, M.; Haruna, S.I. Sustainable traffic management for smart cities using internet-of-things-oriented intelligent transportation systems (ITS): challenges and recommendations. *Sustainability* **2023**, *15*, 9859.
22. Mete, M.O. Geospatial big data analytics for sustainable smart cities. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2023**, *48*, 141–146.
23. Panahi, O. Wearable sensors and personalized sustainability: Monitoring health and environmental exposures in real-time. *European Journal of Innovative Studies and Sustainability* **2025**, *1*, 11–19.
24. Shahzad, S.K.; Ahmed, D.; Naqvi, M.R.; Mushtaq, M.T.; Iqbal, M.W.; Munir, F. Ontology driven smart health service integration. *Computer Methods and Programs in Biomedicine* **2021**, *207*, 106146.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
26. Driver, J.; Spence, C. Cross–modal links in spatial attention. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **1998**, *353*, 1319–1331.
27. Ektefaie, Y.; Dasoulas, G.; Noori, A.; Farhat, M.; Zitnik, M. Multimodal learning with graphs. *Nature Machine Intelligence* **2023**, *5*, 340–350.
28. Wolniak, R.; Stecuła, K. Artificial intelligence in smart cities—applications, barriers, and future directions: a review. *Smart cities* **2024**, *7*, 1346–1389.
29. Sadiq, T.; Omlin, C.W. NLP-based Traffic Scene Retrieval via Representation Learning.
30. Concas, F.; Mineraud, J.; Lagerspetz, E.; Varjonen, S.; Liu, X.; Puolamäki, K.; Nurmi, P.; Tarkoma, S. Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Transactions on Sensor Networks (TOSN)* **2021**, *17*, 1–44.
31. Rodríguez-Ibánez, M.; Casánez-Ventura, A.; Castejón-Mateos, F.; Cuenca-Jiménez, P.-M. A review on sentiment analysis from social media platforms. *Expert Systems with Applications* **2023**, *223*, 119862.
32. Luca, M.; Barlacchi, G.; Lepri, B.; Pappalardo, L. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)* **2021**, *55*, 1–44.
33. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM computing surveys* **2024**, *56*, 1–36.
34. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* **2019**.
35. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* **2019**.
36. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning, 2021; pp. 8748–8763.
37. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International conference on machine learning, 2021; pp. 4904–4916.
38. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* **2019**.
39. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **2019**, *32*.
40. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the European conference on computer vision, 2020; pp. 104–120.
41. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* **2021**, *34*, 9694–9705.
42. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International conference on machine learning, 2022; pp. 12888–12900.
43. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* **2023**, *91*, 424–444.

 doi:10.20944/preprints202507.2268.v1

44. Koroteev, M.V. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943* **2021**.
45. Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409* **2020**.
46. Xu, X.; Wang, Y.; He, Y.; Yang, Y.; Hanjalic, A.; Shen, H.T. Cross-modal hybrid feature fusion for image-sentence matching. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **2021**, *17*, 1–23.
47. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In Proceedings of the 2020 IEEE 23rd international conference on information fusion (FUSION), 2020; pp. 1–6.
48. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural computation* **2020**, *32*, 829–864.
49. Saleh, K.; Hossny, M.; Nahavandi, S. Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks. In Proceedings of the 2017 IEEE 20th international conference on intelligent transportation systems (ITSC), 2017; pp. 1–6.
50. Rudovic, O.; Zhang, M.; Schuller, B.; Picard, R. Multi-modal active learning from human data: A deep reinforcement learning approach. In Proceedings of the 2019 international conference on multimodal interaction, 2019; pp. 6–15.
51. Wang, X.; Lyu, J.; Kim, B.-G.; Parameshachari, B.; Li, K.; Li, Q. Exploring multimodal multiscale features for sentiment analysis using fuzzy-deep neural network learning. *IEEE Transactions on Fuzzy Systems* **2024**, *33*, 28–42.
52. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International conference on machine learning, 2021; pp. 10347–10357.
53. Takahashi, S.; Sakaguchi, Y.; Kouno, N.; Takasawa, K.; Ishizu, K.; Akagi, Y.; Aoyama, R.; Teraya, N.; Bolatkan, A.; Shinkai, N. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems* **2024**, *48*, 84.
54. Ahmed, M.W.; Sadiq, T.; Rahman, H.; Alateyah, S.A.; Alnusayri, M.; Alatiyyah, M.; AlHammadi, D.A. MAPE-ViT: multimodal scene understanding with novel wavelet-augmented Vision Transformer. *PeerJ Computer Science* **2025**, *11*, e2796.
55. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* **2018**.
56. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
57. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? *Advances in neural information processing systems* **2020**, *33*, 6827–6839.
58. Sadiq, T.; Omlin, C.W. Scene Retrieval in Traffic Videos with Contrastive Multimodal Learning. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), 2023; pp. 1020–1025.
59. Wang, L.; Sng, D. Deep learning algorithms with applications to video analytics for a smart city: A survey. *arXiv preprint arXiv:1512.03131* **2015**.
60. Xiao, H.; Zhao, Y.; Zhang, H. Predict vessel traffic with weather conditions based on multimodal deep learning. *Journal of Marine Science and Engineering* **2022**, *11*, 39.
61. Liu, Y.; Yang, C.; Liu, K.; Chen, B.; Yao, Y. Domain adaptation transfer learning soft sensor for product quality prediction. *Chemometrics and Intelligent Laboratory Systems* **2019**, *192*, 103813.
62. Soni, U. Integration of traffic data from social media and physical sensors for near real time road traffic analysis. University of Twente, 2019.
63. Luan, S.; Ke, R.; Huang, Z.; Ma, X. Traffic congestion propagation inference using dynamic Bayesian graph convolution network. *Transportation research part C: emerging technologies* **2022**, *135*, 103526.

 doi:10.20944/preprints202507.2268.v1

64. Zhuang, D.; Gan, V.J.; Tekler, Z.D.; Chong, A.; Tian, S.; Shi, X. Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning. *Applied Energy* **2023**, *338*, 120936.
65. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. In Proceedings of the International conference on artificial neural networks, 2018; pp. 270–279.
66. Saini, K.; Sharma, S. Smart Road Traffic Monitoring: Unveiling the Synergy of IoT and AI for Enhanced Urban Mobility. *ACM Computing Surveys* **2025**, *57*, 1–45.
67. Ouoba, J.; Lahti, J.; Ahola, J. Connecting digital cities: Return of experience on the development of a data platform for multimodal journey planning. In *International Summit, Smart City 360°*; Springer: 2015; pp. 91–103.
68. Botea, A.; Berlingerio, M.; Braghin, S.; Bouillet, E.; Calabrese, F.; Chen, B.; Gkoufas, Y.; Nair, R.; Nonner, T.; Laumanns, M. Docit: An integrated system for risk-averse multimodal journey advising. In *Smart Cities and Homes*; Elsevier: 2016; pp. 345–359.
69. Asgari, F. Inferring user multimodal trajectories from cellular network metadata in metropolitan areas. Institut National des Télécommunications, 2016.
70. Alessandretti, L.; Karsai, M.; Gauvin, L. User-based representation of time-resolved multimodal public transportation networks. *Royal Society open science* **2016**, *3*, 160156.
71. Pronello, C.; Gaborieau, J.-B. Engaging in pro-environment travel behaviour research from a psycho-social perspective: A review of behavioural variables and theories. *Sustainability* **2018**, *10*, 2412.
72. Kang, Y.; Youm, S. Multimedia application to an extended public transportation network in South Korea: optimal path search in a multimodal transit network. *Multimedia Tools and Applications* **2017**, *76*, 19945–19957.
73. Sokolov, I.; Kupriyanovsky, V.; Dunaev, O.; Sinyagov, S.; Kurenkov, P.; Namiot, D.; Dobrynin, A.; Kolesnikov, A.; Gonik, M. On breakthrough innovative technologies for infrastructures. The Eurasian digital railway as a basis of the logistic corridor of the new Silk Road. *International Journal of Open Information Technologies* **2017**, *5*, 102–118.
74. Young, G.W.; Naji, J.; Charlton, M.; Brunsdon, C.; Kitchin, R. Future cities and multimodalities: how multimodal technologies can improve smart-citizen engagement with city dashboards. **2017**.
75. Kumar, S.; Datta, D.; Singh, S.K.; Sangaiah, A.K. An intelligent decision computing paradigm for crowd monitoring in the smart city. *Journal of Parallel and Distributed Computing* **2018**, *118*, 344–358.
76. Zhang, J.; Xiao, W.; Coifman, B.; Mills, J.P. Vehicle tracking and speed estimation from roadside lidar. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2020**, *13*, 5597–5608.
77. Jordan, S.; Chandak, Y.; Cohen, D.; Zhang, M.; Thomas, P. Evaluating the performance of reinforcement learning algorithms. In Proceedings of the International Conference on Machine Learning, 2020; pp. 4962–4973.
78. Maadi, S.; Stein, S.; Hong, J.; Murray-Smith, R. Real-time adaptive traffic signal control in a connected and automated vehicle environment: optimisation of signal planning with reinforcement learning under vehicle speed guidance. *Sensors* **2022**, *22*, 7501.
79. Nigam, N.; Singh, D.P.; Choudhary, J. A review of different components of the intelligent traffic management system (ITMS). *Symmetry* **2023**, *15*, 583.
80. Wu, P.; Zhang, Z.; Peng, X.; Wang, R. Deep learning solutions for smart city challenges in urban development. *Scientific Reports* **2024**, *14*, 5176.
81. Yu, W.; Wu, G.; Han, J. Deep Multimodal-Interactive Document Summarization Network and Its Cross-Modal Text–Image Retrieval Application for Future Smart City Information Management Systems. *Smart Cities* **2025**, *8*, 96.
82. Wu, C.; Wang, T.; Ge, Y.; Lu, Z.; Zhou, R.; Shan, Y.; Luo, P. $\pi $-Tuning: Transferring Multimodal Foundation Models with Optimal Multi-task Interpolation. In Proceedings of the International Conference on Machine Learning, 2023; pp. 37713–37727.
83. Bian, L. Multiscale nature of spatial data in scaling up environmental models. In *Scale in remote sensing and GIS*; Routledge: 2023; pp. 13–26.

 doi:10.20944/preprints202507.2268.v1

84. Pang, T.; Lin, M.; Yang, X.; Zhu, J.; Yan, S. Robustness and accuracy could be reconcilable by (proper) definition. In Proceedings of the International conference on machine learning, 2022; pp. 17258–17277.
85. Yang, X.; Song, Z.; King, I.; Xu, Z. A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering* **2022**, *35*, 8934–8954.
86. Alzubaidi, L.; Al-Amidie, M.; Al-Asadi, A.; Humaidi, A.J.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Santamaría, J.; Duan, Y. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* **2021**, *13*, 1590.
87. Barua, A.; Ahmed, M.U.; Begum, S. A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions. *Ieee access* **2023**, *11*, 14804–14831.
88. Kieu, N.; Nguyen, K.; Nazib, A.; Fernando, T.; Fookes, C.; Sridharan, S. Multimodal colearning meets remote sensing: Taxonomy, state of the art, and future works. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *17*, 7386–7409.
89. Lopes, P.P. PONDIÔNSTRACKER: A FRAMEWORK BASED ON GTFS-RT TO IDENTIFY DELAYS AND ESTIMATE ARRIVALS DYNAMICALLY IN PUBLIC TRANSPORTATION NETWORK.
90. Wu, R.; Wang, H.; Chen, H.-T.; Carneiro, G. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825* **2024**.
91. Seu, K.; Kang, M.-S.; Lee, H. An intelligent missing data imputation techniques: A review. *JOIV: International Journal on Informatics Visualization* **2022**, *6*, 278–283.
92. Psychogyios, K.; Ilias, L.; Ntanos, C.; Askounis, D. Missing value imputation methods for electronic health records. *IEEE Access* **2023**, *11*, 21562–21574.
93. Çetin, V.; Yıldız, O. A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* **2022**, *28*, 299–312.
94. Younis, E.M.; Zaki, S.M.; Kanjo, E.; Houssein, E.H. Evaluating ensemble learning methods for multi-modal emotion recognition using sensor data fusion. *Sensors* **2022**, *22*, 5611.
95. Ferrario, A.; Loi, M. How explainability contributes to trust in AI. In Proceedings of the Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022; pp. 1457–1466.
96. Bell, A.; Solano-Kamaiko, I.; Nov, O.; Stoyanovich, J. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In Proceedings of the Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022; pp. 248–266.
97. Mahbooba, B.; Timilsina, M.; Sahal, R.; Serrano, M. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity* **2021**, *2021*, 6634811.
98. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems* **2025**, *7*, 2400304.
99. Zhu, X.; Wang, D.; Pedrycz, W.; Li, Z. Fuzzy rule-based local surrogate models for black-box model explanation. *IEEE Transactions on Fuzzy Systems* **2022**, *31*, 2056–2064.
100. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp. 3213–3223.
101. Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; McCord, B. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856* **2018**.
102. Pfeil, M.; Bartoschek, T.; Wirwahn, J.A. Opensensemap-a citizen science platform for publishing and exploring sensor data as open data. **2018**.
103. Bui, L. Breathing smarter: A critical look at representations of air quality sensing data across platforms and publics. In Proceedings of the 2015 IEEE First International Smart Cities Conference (ISC2), 2015; pp. 1–5.
104. Sahoh, B.; Choksuriwong, A. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review. *Journal of Ambient Intelligence and Humanized Computing* **2023**, *14*, 7827–7843.
105. Naidu, G.; Zuva, T.; Sibanda, E.M. A review of evaluation metrics in machine learning algorithms. In Proceedings of the Computer science on-line conference, 2023; pp. 15–25.

 doi:10.20944/preprints202507.2268.v1

106. Anjuma, K.; Arshad, M.A.; Hayawi, K.; Polyzos, E.; Tariq, A.; Serhani, M.A.; Batool, L.; Lund, B.; Mannuru, N.R.; Bevara, R.V.K. Domain Specific Benchmarks for Evaluating Multimodal Large Language Models. *arXiv preprint arXiv:2506.12958* **2025**.
107. Zhou, Y.; Gallego, G.; Lu, X.; Liu, S.; Shen, S. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE transactions on neural networks and learning systems* **2021**, *34*, 4868–4880.
108. Peng, W.; Bai, X.; Yang, D.; Yuen, K.F.; Wu, J. A deep learning approach for port congestion estimation and prediction. *Maritime Policy & Management* **2023**, *50*, 835–860.
109. Liu, J.; Ong, G.P. Prediction of Next-Time Traffic Congestion with Consideration of Congestion Propagation Patterns and Co-occurrence. *IEEE Transactions on Vehicular Technology* **2025**.
110. Wattacheril, C.Y.; Hemalakshmi, G.; Murugan, A.; Abhiram, P.; George, A.M. Machine Learning-Based Threat Detection in Crowded Environments. In Proceedings of the 2024 International Conference on Smart Technologies for Sustainable Development Goals (ICSTSDG), 2024; pp. 1–7.
111. Jiang, Q.; Kresin, F.; Bregt, A.K.; Kooistra, L.; Pareschi, E.; Van Putten, E.; Volten, H.; Wesseling, J. Citizen sensing for improved urban environmental monitoring. *Journal of Sensors* **2016**, *2016*, 5656245.
112. Lim, C.C.; Kim, H.; Vilcassim, M.R.; Thurston, G.D.; Gordon, T.; Chen, L.-C.; Lee, K.; Heimbinder, M.; Kim, S.-Y. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environment international* **2019**, *131*, 105022.
113. Hu, T.; Wang, S.; She, B.; Zhang, M.; Huang, X.; Cui, Y.; Khuri, J.; Hu, Y.; Fu, X.; Wang, X. Human mobility data in the COVID-19 pandemic: characteristics, applications, and challenges. *International Journal of Digital Earth* **2021**, *14*, 1126–1147.
114. Almujally, N.A.; Qureshi, A.M.; Alazeb, A.; Rahman, H.; Sadiq, T.; Alonazi, M.; Algarni, A.; Jalal, A. A novel framework for vehicle detection and tracking in night ware surveillance systems. *Ieee Access* **2024**, *12*, 88075–88085.
115. Son, H.; Jang, J.; Park, J.; Balog, A.; Ballantyne, P.; Kwon, H.R.; Singleton, A.; Hwang, J. Leveraging advanced technologies for (smart) transportation planning: A systematic review. *Sustainability* **2025**, *17*, 2245.
116. Zaib, S.; Lu, J.; Bilal, M. Spatio-temporal characteristics of air quality index (AQI) over Northwest China. *Atmosphere* **2022**, *13*, 375.
117. Xu, Z.; Mei, L.; Lv, Z.; Hu, C.; Luo, X.; Zhang, H.; Liu, Y. Multi-modal description of public safety events using surveillance and social media. *IEEE Transactions on Big Data* **2017**, *5*, 529–539.
118. Alrashdi, I.; Alqazzaz, A.; Aloufi, E.; Alharthi, R.; Zohdy, M.; Ming, H. Ad-iot: Anomaly detection of iot cyberattacks in smart city using machine learning. In Proceedings of the 2019 IEEE 9th annual computing and communication workshop and conference (CCWC), 2019; pp. 0305–0310.
119. Islam, M.; Dukyil, A.S.; Alyahya, S.; Habib, S. An IoT enable anomaly detection system for smart city surveillance. *Sensors* **2023**, *23*, 2358.
120. Zhong, C.; Guo, H.; Swan, I.; Gao, P.; Yao, Q.; Li, H. Evaluating trends, profits, and risks of global cities in recent urban expansion for advancing sustainable development. *Habitat International* **2023**, *138*, 102869.
121. Jadhav, S.; Durairaj, M.; Reenadevi, R.; Subbulakshmi, R.; Gupta, V.; Ramesh, J.V.N. Spatiotemporal data fusion and deep learning for remote sensing-based sustainable urban planning. *International Journal of System Assurance Engineering and Management* **2024**, 1–9.
122. Qiu, J.; Zhao, Y. Traffic Prediction with Data Fusion and Machine Learning. *Analytics* **2025**, *4*, 12.
123. Karagiannopoulou, A.; Tsertou, A.; Tsimiklis, G.; Amditis, A. Data fusion in earth observation and the role of citizen as a sensor: A scoping review of applications, methods and future trends. *Remote Sensing* **2022**, *14*, 1263.
124. Hsu, I.-C.; Chang, C.-C. Integrating machine learning and open data into social Chatbot for filtering information rumor. *Journal of Ambient Intelligence and Humanized Computing* **2021**, *12*, 1023–1037.
125. Li, X.; Liu, H.; Wang, W.; Zheng, Y.; Lv, H.; Lv, Z. Big data analysis of the internet of things in the digital twins of smart city based on deep learning. *Future Generation Computer Systems* **2022**, *128*, 167–177.
126. Liu, Q.; Huang, Y.; Jin, C.; Zhou, X.; Mao, Y.; Catal, C.; Cheng, L. Privacy and integrity protection for IoT multimodal data using machine learning and blockchain. *ACM Transactions on Multimedia Computing, Communications and Applications* **2024**, *20*, 1–18.

 doi:10.20944/preprints202507.2268.v1

127. Zekić-Sušac, M.; Mitrović, S.; Has, A. Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International journal of information management* **2021**, *58*, 102074.
128. Liu, H.; Cui, W.; Zhang, M. Exploring the causal relationship between urbanization and air pollution: Evidence from China. *Sustainable Cities and Society* **2022**, *80*, 103783.
129. Malatesta, T.; Breadsell, J.K. Identifying home system of practices for energy use with k-means clustering techniques. *Sustainability* **2022**, *14*, 9017.
130. Kilicay-Ergin, N.; Barb, A.S. Semantic fusion with deep learning and formal ontologies for evaluation of policies and initiatives in the smart city domain. *Applied Sciences* **2021**, *11*, 10037.
131. Naoui, M.A.; Lejdel, B.; Ayad, M.; Amamra, A.; kazar, O. Using a distributed deep learning algorithm for analyzing big data in smart cities. *Smart and Sustainable Built Environment* **2021**, *10*, 90–105.
132. Atitallah, S.B.; Driss, M.; Boulila, W.; Ghézala, H.B. Computer Science Review. **2020**.
133. Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; Luo, Y. Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine* **2022**, *5*, 171.
134. Dautov, R.; Distefano, S.; Buyya, R. Hierarchical data fusion for smart healthcare. *Journal of Big Data* **2019**, *6*, 1–23.
135. Nazari, E.; Chang, H.-C.H.; Deldar, K.; Pour, R.; Avan, A.; Tara, M.; Mehrabian, A.; Tabesh, H. A comprehensive overview of decision fusion technique in healthcare: A systematic scoping review. *Iranian Red Crescent Medical Journal* **2020**, *22*, 1–17.
136. Haltaufderheide, J.; Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ digital medicine* **2024**, *7*, 183.
137. Demelius, L.; Kern, R.; Trügler, A. Recent advances of differential privacy in centralized deep learning: A systematic survey. *ACM Computing Surveys* **2025**, *57*, 1–28.
138. Sampaio, S.; Sousa, P.R.; Martins, C.; Ferreira, A.; Antunes, L.; Cruz-Correia, R. Collecting, processing and secondary using personal and (pseudo) anonymized data in smart cities. *Applied Sciences* **2023**, *13*, 3830.
139. Labadie, C.; Legner, C. Building data management capabilities to address data protection regulations: Learnings from EU-GDPR. *Journal of Information Technology* **2023**, *38*, 16–44.
140. Oladosu, S.A.; Ike, C.C.; Adepoju, P.A.; Afolabi, A.I.; Ige, A.B.; Amoo, O.O. Frameworks for ethical data governance in machine learning: Privacy, fairness, and business optimization. *Magna Sci Adv Res Rev* **2024**.
141. Qu, Y.; Nosouhi, M.R.; Cui, L.; Yu, S. Privacy preservation in smart cities. In *Smart cities cybersecurity and privacy*; Elsevier: 2019; pp. 75–88.
142. Rao, P.M.; Deebak, B.D. Security and privacy issues in smart cities/industries: technologies, applications, and challenges. *Journal of Ambient Intelligence and Humanized Computing* **2023**, *14*, 10517–10553.
143. Daoudagh, S.; Marchetti, E.; Savarino, V.; Bernabe, J.B.; García-Rodríguez, J.; Moreno, R.T.; Martinez, J.A.; Skarmeta, A.F. Data protection by design in the context of smart cities: A consent and access control proposal. *Sensors* **2021**, *21*, 7154.
144. Al-Turjman, F.; Zahmatkesh, H.; Shahroze, R. An overview of security and privacy in smart cities' IoT communications. *Transactions on Emerging Telecommunications Technologies* **2022**, *33*, e3677.
145. Rusinova, V.; Martynova, E. Fighting cyber attacks with sanctions: Digital threats, economic responses. *Israel Law Review* **2024**, *57*, 135–174.
146. Narasimha Rao, K.P.; Chinnaiyan, S. Blockchain-Powered Patient-Centric Access Control with MIDC AES-256 Encryption for Enhanced Healthcare Data Security. *Acta Informatica Pragensia* **2024**, *13*, 374–394.
147. Ahmed, S.; Ahmed, I.; Kamruzzaman, M.; Saha, R. Cybersecurity Challenges in IT Infrastructure and Data Management: A Comprehensive Review of Threats, Mitigation Strategies, and Future Trend. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology* **2022**, *1*, 36–61.
148. Balayn, A.; Lofi, C.; Houben, G.-J. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal* **2021**, *30*, 739–768.
149. De Falco, C.C.; Romeo, E. Algorithms and geo-discrimination risk: What hazards for smart cities' development? In *Smart Cities*; Routledge: 2025; pp. 104–117.

 doi:10.20944/preprints202507.2268.v1

150. Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; Ntoutsi, E. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2022**, *12*, e1452.
151. Herdiansyah, H. Smart city based on community empowerment, social capital, and public trust in urban areas. *Glob. J. Environ. Sci. Manag* **2023**, *9*, 113–128.
152. Sarker, I.H. Smart City Data Science: Towards data-driven smart cities with open research issues. *Internet of Things* **2022**, *19*, 100528.
153. Gao, L.; Guan, L. Interpretability of machine learning: Recent advances and future prospects. *IEEE MultiMedia* **2023**, *30*, 105–118.
154. Rashid, M.M.; Kamruzzaman, J.; Hassan, M.M.; Imam, T.; Wibowo, S.; Gordon, S.; Fortino, G. Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications. *Computers & Security* **2022**, *120*, 102783.
155. Dutta, H.; Minerva, R.; Alvi, M.; Crespi, N. Data-driven Modality Fusion: An AI-enabled Framework for Large-Scale Sensor Network Management. *arXiv preprint arXiv:2502.04937* **2025**.
156. Huang, J.; Zhang, Z.; Zheng, S.; Qin, F.; Wang, Y. {DISTMM}: Accelerating distributed multimodal model training. In Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), 2024; pp. 1157–1171.
157. Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; Liu, Z. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**.
158. Jafari, F.; Moradi, K.; Shafiee, Q. Shallow learning vs. Deep learning in engineering applications. In *Shallow Learning vs. Deep Learning: A Practical Guide for Machine Learning Solutions*; Springer: 2024; pp. 29–76.
159. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.L. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2023**, *13*, e1484.
160. Schmitt, M. Securing the digital world: Protecting smart infrastructures and digital industries with artificial intelligence (AI)-enabled malware and intrusion detection. *Journal of Industrial Information Integration* **2023**, *36*, 100520.
161. Pearson, M. Pioneering Urban Biodiversity: Using AI-sensors, eDNA and traditional methods to create a novel biodiversity monitoring toolkit and assessment framework. 2024.