Green AI power impact on development of smart cities

Ioannis APOSTOLIDIS,
University of Bucharest, Bucharest, Romania
ioannis.apostolidis@s.unibuc.ro

Isabelle COICIU,
University of Bucharest, Bucharest, Romania
isabelle.coiciu@s.unibuc.ro

Ioan Florin VOICU,
ING Tech, Bucharest, Romania
ioan-florin.voicu@ing.com

Daniel Constantin DIACONU,
Dr., Interdisciplinary Center for Advanced Studies (CISA-ICUB), University of Bucharest,Romania
daniel.diaconu@unibuc.ro

Abstract

Smart cities are increasingly integrating Internet of Things (IoT) and Artificial Intelligence (AI) technologies to collect real-time data and respond to citizens' needs. The objective of this study is to examine the impact of Green Artificial Intelligence (Green AI) on the sustainable development of smart cities, focusing on how energy-efficient algorithms and environmentally conscious machine learning systems can optimize urban operations. The research aims to demonstrate how Green AI supports the creation of climate-resilient, resource-efficient, and human-centered urban ecosystems. Building upon prior work in sustainable innovation and AI for environmental management, this study extends the discourse toward the intersection of Green AI and smart city transformation. Previous research in sustainable urban design and AI-enabled infrastructure has shown the potential of intelligent technologies to reduce emissions and improve resource utilization, yet few studies explicitly address the sustainability of the AI systems themselves. The present paper contributes to this emerging field by proposing an integrative conceptual framework grounded in theoretical review. The approach synthesizes existing literature and case evidence to analyze the role of Green AI in energy management, mobility optimization, and digital governance. The analysis suggests that Green AI has the potential to enhance predictive analytics for urban planning, support smart grid efficiency, enable adaptive traffic and waste management systems, and strengthen environmental monitoring through low-carbon computation. The implications of this study are significant for academics, policymakers, and urban developers, offering insights into how sustainable AI deployment can balance technological advancement with ecological preservation. The value of the paper lies in its holistic perspective on Green AI as both a catalyst and a constraint in smart city evolution, providing an original theoretical foundation for future empirical research and policy development in sustainable digital urbanization.

Keywords: sustainable innovation, urban resilience, digital governance, AI efficiency, carbon reduction.

1. Introduction

The rapid proliferation of data-driven technologies has transformed the epistemic foundations of urban governance [1]. As smart cities increasingly depend on interconnected sensor networks and algorithmic systems, the question of how knowledge is produced, processed, and operationalized becomes central to understanding smart sustainable urban development [2]. Contemporary urban infrastructures rely not only on physical assets but also on computational mechanisms that interpret real-time data and guide decision-making [3]. Within this evolving technological landscape, Artificial Intelligence (AI) and the Internet of Things (IoT) function as epistemic agents that shape the very ways in which cities perceive environmental conditions, allocate resources, and respond to societal needs [4]. Yet, the sustainability of these knowledge-generating systems has received comparatively limited scholarly scrutiny. A recent theoretical review by [5] emphasizes that artificial intelligence significantly influences sustainable innovation across domains such as green products, green energy, and green farming, further underscoring the need to examine not only AI for sustainability but also the sustainability of AI systems themselves.

While the literature on smart cities extensively documents the benefits of AI-enabled systems for optimizing energy use, mobility flows, and public services, an epistemic gap persists regarding the environmental costs of the AI models themselves [6]. Much of the existing research assumes the neutrality of algorithmic intelligence, often overlooking the materiality of computation: model training, data processing, and inference all consume energy, generate emissions, and depend on hardware ecosystems whose ecological footprints remain substantial [7] This disconnect raises a fundamental epistemic tension. Smart cities seek to become more sustainable through AI, yet the very knowledge infrastructures that enable this transformation may undermine ecological goals if not designed with sustainability in mind.

Green Artificial Intelligence (Green AI) emerges as a response to this tension by advocating for transparent, energy-efficient, and environmentally conscious approaches to machine learning. Green AI extends beyond the development of efficient algorithms to encompass reductions in energy consumption, minimization of carbon emissions, and the promotion of responsible AI practices throughout the entire lifecycle of machine-learning systems [8].

This philosophy is embodied in the MERIDA system, a system that combines IoT sensors, Big Data tools like Hadoop and Spark, and machine learning predictions to handle energy use in public buildings at both individual (micro) and city-wide (macro) scales [9]. It works through six key steps: gathering data, cleaning it up, running predictions with methods like deep neural networks and Random Forest, creating easy-to-read visuals, offering decision advice, and tracking results [9]. This setup connects smoothly to smart cities using smart IoT networks, helping public officials focus spending on upgrades that lower costs, cut CO_2 emissions, and improve services.

In this study we address the epistemic gap by conceptualizing Green AI not merely as a technical improvement but as a transformative lens for re-evaluating the role of algorithmic intelligence in sustainable urban development. Building upon both the theoretical review and the empirical findings of this research, the paper integrates insights from sustainable innovation, environmental informatics, and AI ethics with real-time power-consumption measurements derived from controlled experiments using different AI model sizes (Gemma 4B and Gemma 12B). Through this combined approach, the paper proposes a framework that articulates how Green AI reconfigures the production, validation, and application of knowledge in smart cities by linking conceptual principles of energy efficient computation with measurable environmental impacts. This epistemic framing highlights Green AI as both a catalyst that enhances predictive modelling, resource efficiency, and environmental monitoring, and a constraint that challenges prevailing assumptions about scalability, accuracy, and the pursuit of increasing computational capacity, demonstrated empirically through the significant energy differences observed between GPU-fit and CPU-offloaded model executions.

2. Methodology

This study employs a mixed-method research design that integrates conceptual analysis with empirical experimentation to examine the role of Green Artificial Intelligence in sustainable smart-city development. To remain consistent with methodological standards in smart-city research, the present work also acknowledges the need to account for geographical, cultural, and infrastructural differences when analysing urban technologies, as highlighted in comparative studies such as that of Ismagilova et al. (2019) [10]. Given that the environmental impacts of AI computation remain an emerging area of inquiry, the methodology combines a structured theoretical review with a controlled hardware-based case study that measures real-time energy consumption during AI model execution.

The research proceeds in three interconnected stages. The first stage consists of a systematic review of scholarly literature on Green AI, energy-efficient machine learning, and sustainable urban informatics. Academic databases including Scopus, Web of Science, IEEE Xplore, and Google Scholar were queried using targeted keywords related to sustainable AI, smart cities, computational energy use, and low-carbon urban technologies. The selected works were screened for conceptual relevance and their contribution to understanding the ecological footprint of algorithmic systems. This stage establishes the theoretical foundation by identifying knowledge gaps related to the sustainability of AI infrastructures deployed in urban environments.

The second stage involves the design and execution of a controlled computational experiment aimed at empirically evaluating the energy footprint of different AI model sizes under identical workload conditions. A dedicated test environment was constructed using an Intel 13th-generation CPU, 128 GB RAM, and an NVIDIA RTX 3050 GPU with 8 GB VRAM. Two AI models, Google Gemma 4B and Gemma 12B, were executed through the Ollama framework across repeated benchmark runs. Real-time power consumption was monitored using a Shelly WiFi power-measurement plug connected to a separate Home Assistant server running InfluxDB and Grafana. This setup allowed high-resolution logging of system wattage, execution time, idle baseline consumption, and the effect of GPU memory fitting versus CPU offloading. The empirical component directly measures the environmental cost of model execution, offering practical insights into hardware-aware “right-sizing” of AI for smart-city infrastructures.

The third stage synthesizes findings from both the literature review and the experimental results into an integrative Green AI framework for smart cities. This synthesis uses abductive reasoning, allowing the study to iteratively relate theoretical principles, such as energy transparency, computational efficiency, and resource-aware design, to the empirical patterns observed in the benchmark tests. The framework interprets how model choice, hardware constraints, and workload characteristics influence energy use in urban AI applications such as traffic prediction, environmental sensing, digital governance, and

grid optimization. The empirical results, demonstrating significant differences in total energy consumption between models that fit GPU memory and those that do not, are contextualized as practical guidance for sustainable AI deployment in urban ecosystems.
This methodological strategy enables a holistic understanding of Green AI as both a theoretical concept and a measurable computational phenomenon. By grounding the study in existing scholarly evidence while incorporating real-world power-consumption data, the research provides a robust foundation for future empirical investigations and supports evidence-based policy development in sustainable digital urbanization.

2.1. Case study hardware setup

The empirical component of this study was conducted on a dedicated workstation designed to reflect a realistic, energy-constrained computing environment such as those increasingly used in smart-city infrastructures. The hardware selection aimed to balance computational capability with accessibility and energy efficiency, ensuring that the findings would be applicable to practical urban AI deployments rather than relying on large-scale data-center hardware.

The system was built around an Intel 13th-generation CPU, representing a modern hybrid architecture that combines performance cores and efficiency cores. This type of processor is widely available in commercial and professional edge-computing contexts and is well-suited for mixed workloads involving both conventional data processing and machine-learning inference. The CPU plays a central role in the experiment, especially when running models that exceed the capacity of the available GPU memory and therefore require partial or full CPU-based computation. Understanding the power implications of CPU-dominant workloads is essential for evaluating the real-world efficiency of AI systems intended for smart-city operations such as traffic modelling, environmental monitoring, and automated municipal services.

To support the execution of memory-intensive models and avoid bottlenecks, the workstation was equipped with 128 GB of RAM. This configuration ensures that both the smaller Gemma 4B and the larger Gemma 12B models can operate under consistent conditions without triggering disk swapping or memory exhaustion, both of which would distort power-usage measurements. In smart-city scenarios, adequate system memory is particularly important for AI applications that process large sensor datasets, maintain real-time analytics pipelines, or run multiple inference tasks in parallel.

The primary accelerator used in the case study was an NVIDIA RTX 3050 GPU with 8 GB of VRAM. Although categorized as an entry-level or mid-range GPU, this model is highly relevant to Green AI research because it reflects hardware commonly found in local government IT systems, university labs, and small enterprises, contexts where energy constraints are real and widespread adoption of large data-center GPUs is neither practical nor cost-effective. The 8 GB VRAM limit also plays a defining role in the experiment: the smaller Gemma 4B model fits entirely within the GPU's memory, enabling rapid inference with a higher but short-duration power draw. In contrast, the larger Gemma 12B model exceeds the available VRAM, forcing the system to rely on CPU processing and memory paging. This difference in execution strategy directly influences energy consumption and highlights the importance of “right-sizing” AI models for the available hardware.

To accurately capture the power behavior of the system throughout the experiment, the workstation was connected to a Shelly WiFi power-monitoring plug (Fig. 1). This device provides real-time measurement of instantaneous wattage and cumulative energy consumption (kWh), enabling precise tracking of idle, load, and post-load states. The Shelly plug communicated with a separate monitoring server running Home Assistant OS, which integrated data into InfluxDB for time-series storage and Grafana for visualization. This monitoring architecture ensures high-resolution, reliable data suitable for analyzing short and long inference runs. It also mirrors the type of energy-aware sensing that is increasingly incorporated into smart buildings and smart-city IoT infrastructures.

Fig. 1. Shelly WiFi power-monitoring plug
Source: www.shelly.com/products/shelly-plug-s-gen3

Overall, this hardware setup provides a controlled yet realistic platform for evaluating the environmental impact of AI workloads. The combination of a hybrid CPU, mid-range GPU, substantial system memory, and high-precision power monitoring allows for meaningful comparison of the energy performance of differently sized models. This setup supports the study’s broader aim of demonstrating how hardware-conscious model selection can significantly reduce the computational footprint of smart-city AI systems, thereby contributing to more energy-efficient and sustainable urban digital infrastructures.

2.2. Case study software setup
The software environment was designed to support controlled, repeatable evaluation of AI model energy consumption while ensuring compatibility with the monitoring infrastructure used in the experiment. All AI workloads were executed using Ollama, a lightweight, containerized runtime optimized for efficient local deployment of large language models. Ollama was selected due to its low overhead, reproducible model

execution pipeline, and native support for hardware-aware inference, making it suitable for evaluating differences between GPU-fit and CPU-dominant workloads.

Two models from Google's Gemma family were used in the benchmark tests: Gemma 4B and Gemma 12B. These models were chosen to illustrate the energy implications of running AI workloads that either fully utilize GPU memory (4B) or exceed it (12B) (Fig. 2 and 3). Each model was executed ten times under identical conditions to reduce variability and ensure statistically reliable comparisons (Tab. 1). The use of repeated benchmarks also reflects realistic smart-city scenarios, where AI systems often run continuous or scheduled inference cycles rather than single isolated tasks.

Table 1. Core parameters of Google's Gemma family models.

Parameters	**12B**	**4B**
(Embedding Size) d_model	3840	2560
Layers	48	34
Feedforward hidden dims	30720	20480
Num heads	16	8
Num Key Value heads	8	4
Query Key Value heads	256	256
Vocab size	262144	262144
Context window	128k	128k

Source:developers.googleblog.com/en/gemma-explained-whats-new-in-gemma-3

To measure real-time power consumption during inference, the workstation was integrated with an external monitoring system. A separate server running Home Assistant OS served as the central hub for energy data collection. The server connected directly to the Shelly power plug API, enabling high-frequency sampling of instantaneous wattage. Power readings were stored in InfluxDB, a time-series database optimized for rapid write operations and precise temporal alignment. Visualization and exploratory analysis were performed using Grafana, which provided detailed wattage curves and allowed comparison of load profiles across different model sizes and execution cycles.

Together, this software configuration ensured accurate, repeatable model execution and reliable power-measurement workflows, creating a robust foundation for evaluating the environmental efficiency of AI workloads under realistic hardware constraints.

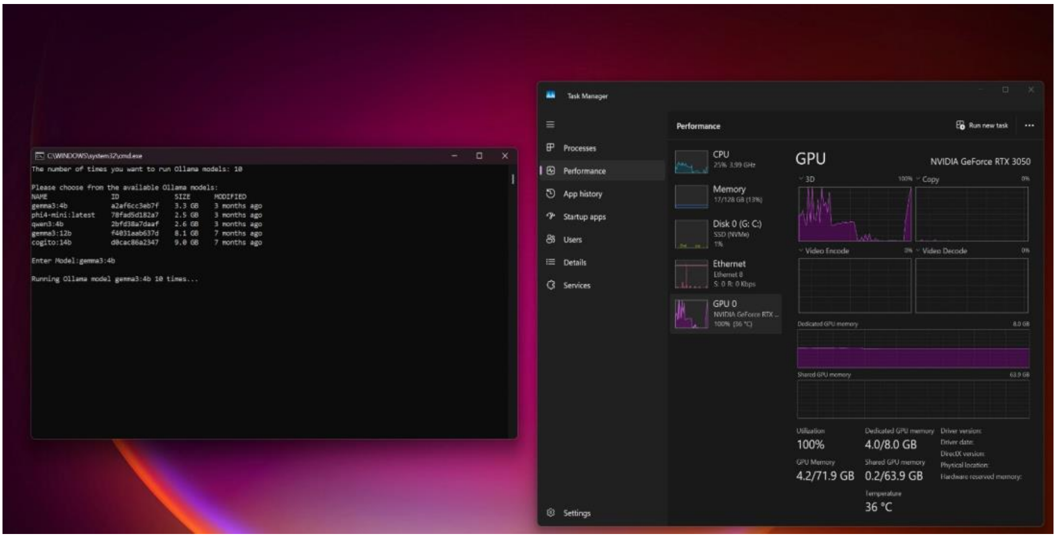

Fig. 2. Side-by-side view of an Ollama benchmark in a command-line window and Windows Task Manager, highlighting GPU utilization and memory usage while running the gemma3:4b model.
Source: author's testing

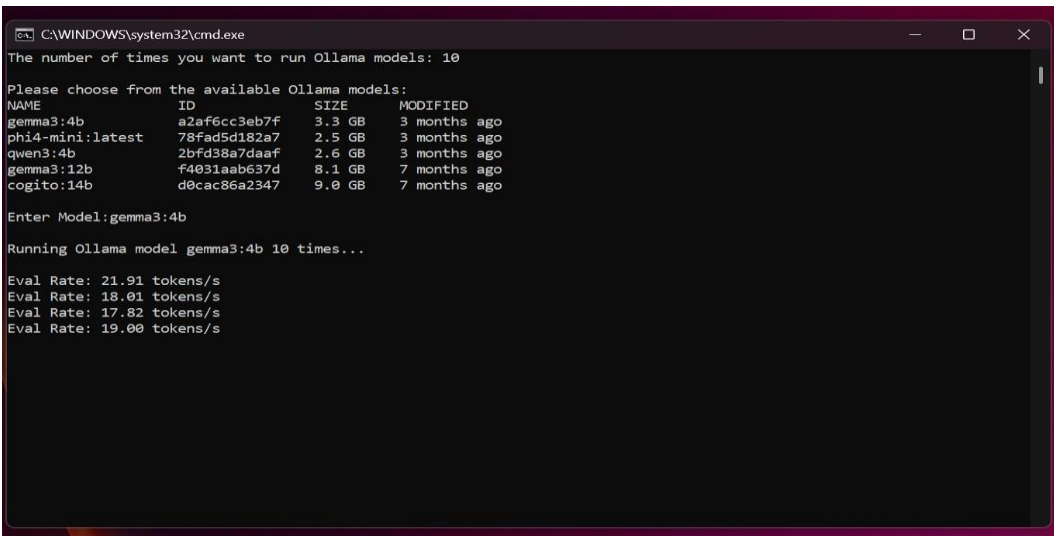

Fig. 3. Command-line window showing an Ollama benchmark run for the gemma3:4b model on Windows, listing available models and displaying evaluation speed results in tokens per second.
Source: author's testing

3. Results

3.1. Server consumption monitoring results

Power-consumption data were captured using the Shelly plug at 1–2 second resolution. Baseline measurements indicated that the server operated at 150–170 W under idle conditions, with no AI workload applied. During the execution of the Gemma 4B model, the system displayed a pronounced but short power spike, sustaining approximately 300 W for a duration of around 2 minutes. Because the 4B model fits entirely within the GPU's

8 GB VRAM, the workload was processed efficiently, resulting in a brief high-load interval and relatively low total energy use.

In contrast, running the Gemma 12B model produced a markedly different consumption profile. The system maintained an elevated load of approximately 250 W for about 20 minutes. As the 12B model exceeds the available GPU memory, execution shifted largely to CPU processing and memory paging, extending the computation time despite the lower instantaneous wattage. This prolonged execution resulted in significantly higher overall energy consumption.

Calculated energy usage for the two workloads demonstrates this disparity clearly:
Gemma 4B: ~0.012 kWh (2 minutes at 300 W) (Fig. 4)
Gemma 12B: ~0.10 kWh (20 minutes at 250 W) (Fig. 5)

These results indicate that the smaller, GPU-fit model required roughly 10 times less energy to complete the same benchmark task, assuming both models meet the performance requirements of the application (Fig. 6).

The findings underscore a central principle of Green AI: right-sizing the model to the available hardware is critical for minimizing environmental impact. For organizations or municipalities selecting sustainable IT solutions, understanding hardware power limits and aligning AI model size accordingly can substantially reduce energy consumption while maintaining functional performance.

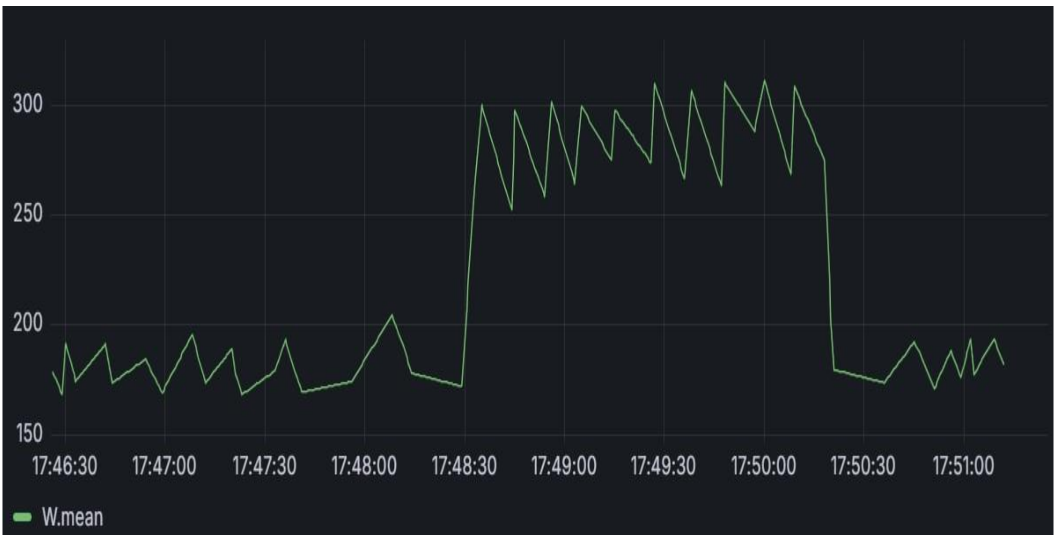

Fig. 4. Power consumption during Gemma 4B inference showing brief spike to 300 W over 2-minute execution.
Source: author's testing

Fig. 5. Power consumption during Gemma 12B inference showing sustained elevation to 250 W over 20-minute execution.
Source: author's testing

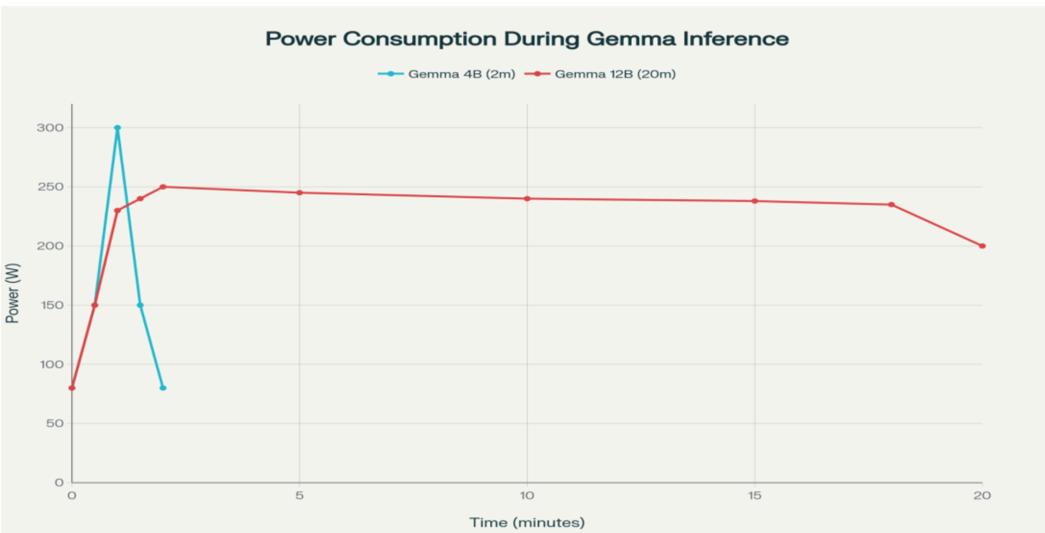

Fig. 6. Power consumption profiles for Gemma 4B (short, high-intensity run) and Gemma 12B (long, sustained run)
Source: author's testing

4. Discussions

The empirical findings reveal that Green AI principles, particularly model right-sizing to hardware constraints, significantly reduce energy consumption in smart city applications. The Gemma 4B model's efficient GPU fit resulted in 0.012 kWh usage over 2 minutes at 300 W, compared to the Gemma 12B's 0.10 kWh over 20 minutes at 250 W due to CPU offloading. This tenfold difference underscores how sustainable AI deployment can optimize urban operations like traffic prediction and environmental monitoring without sacrificing performance. These results align with the proposed framework, emphasizing energy transparency and resource-aware design for climate-resilient cities.

4.1 Applicability
The case study's insights apply broadly to resource-constrained smart city infrastructures, such as local government IT systems using mid-range GPUs like the NVIDIA RTX 3050. Municipalities can adopt smaller, GPU-fit models for real-time tasks like smart grid management and waste optimization, achieving substantial carbon reductions while maintaining predictive accuracy. This hardware-conscious approach extends to edge computing in IoT networks, enabling scalable deployment across diverse urban settings from Bucharest to other mid-sized cities. Policymakers benefit by prioritizing low-carbon AI in digital governance, fostering human-centered ecosystems that balance innovation with ecological limits.

4.2. Limitations of case study
The experiment used a single workstation with Intel 13th-gen CPU, 128 GB RAM, and RTX 3050 GPU, limiting generalizability to high-end data centers or mobile edge devices. Benchmarks focused on Ollama-executed Gemma models under controlled conditions, potentially overlooking variability in real-world workloads like continuous IoT data streams. Power measurements via Shelly plug captured aggregate consumption but did not isolate component-level efficiency, such as GPU versus CPU contributions. Future studies should incorporate diverse hardware, dynamic urban datasets, and longitudinal emissions tracking for comprehensive validation.

5. Conclusions
Green AI emerges as a vital catalyst for sustainable smart-city development, as demonstrated by empirical evidence showing that energy-efficient model selection can reduce computational footprints by up to an order of magnitude. AI may also constrain the capacity of regions to develop specialization in green technologies, underscoring the need for careful governance and context-sensitive deployment [11]. The framework developed in this work integrates theoretical principles with practical benchmarks, offering urban planners guidance on selecting right-sized AI systems for energy management, mobility, and urban governance. By addressing key epistemic gaps in AI sustainability, this research provides policymakers and developers with actionable strategies for building climate-resilient digital infrastructures. Future research should extend these findings through multi-model comparisons and city-scale pilot projects to fully realize pathways toward low-carbon digital urbanization.

The importance of citizen participation in smart cities and also the understanding of both the discursive and material dimensions of smart citizenship is of major significance [12]. The way smart cities frame citizenship through language, policy, and governance, together with the tangible technological infrastructures that shape everyday experiences, reveals how deeply intertwined these perspectives are in practice [12]. This interplay between rhetoric and reality underscores the need for aligning official narratives with actual citizen engagement, ensuring that smart city development remains inclusive, responsible, and centered on the rights and duties of its residents.

References

[1] S. E. Bibri, „The core academic and scientific disciplines underlying data-driven smart sustainable urbanism: an interdisciplinary and transdisciplinary framework,” *Computational Urban Science,* vol. 1, nr. 1, p. 1, 2021.

[2] S. Cohen, „Interconnected sensor networks and digital urban governance in data-driven smart sustainable cities,” *Geopolitics, History, and International Relations,* vol. 13, nr. 1, pp. 97-107, 2021.

[3] B. N. Silva, M. Khan, C. Jung, J. Seo, D. Muhammad, J. Han și K. Han, „Urban planning and smart city decision management empowered by real-time data processing using big data analytics,” *Sensors,* vol. 18, nr. 9, p. 2994, 2018.

[4] S. E. Bibri, Artificial Intelligence of Things for Smarter Eco-Cities: Pioneering the Environmental Synergies of Urban Brain, Digital Twin, Metabolic Circularity, and Platform, CRC Press, 2024.

[5] B. Artha, U. T. Sari, B. Bahri, A. S. Hadi, N. P. Sari, C. P. Asri și B. Satriadhi, „Artificial Intelligence and Sustainable Innovation: A Theoretical Review,” *Journal of Sustainable Innovation Engineering,* vol. 1, nr. 2, pp. 39-43, 2025.

[6] A. Zhuk, „Artificial intelligence impact on the environment: Hidden ecological costs and ethical-legal issues,” *Journal of Digital Technologies and Law,* vol. 1, nr. 4, pp. 932-954, 2023.

[7] U. Felt, „Environmental Intelligence?! On the Consequences of Sidelining the Materiality of AI,” *Harvard Data Science Review,* vol. 7, nr. 4, 2025.

[8] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela și A. Alonso-Betanzos, „A review of green artificial intelligence: Towards a more sustainable future,” *Neurocomputing,* vol. 599, p. 128096, 2024.

[9] M. Zekić-Sušac, S. Mitrović și A. Has, „Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities,” *International journal of information management,* vol. 58, p. 102074, 2021.

[10] E. Ismagilova, L. Hughes, Y. K. Dwivedi și K. R. Raman, „Smart cities: Advances in research—An information systems perspective,” *International journal of information management,* vol. 47, pp. 88-100, 2019.

[11] G. Cicerone, A. Faggian, S. Montresor și F. Rentocchini, „Regional artificial intelligence and the geography of environmental technologies: does local AI knowledge help regional green-tech specialization,” *Regional Studies,* vol. 57, nr. 2, pp. 330-343, 2023.

[12] D. Kumar, „Actual practices of citizen participation in smart cities,” *Orașe inteligente și dezvoltare regională,* vol. 8, nr. 2, pp. 19-30, 2024.