Explainable Federated Reinforcement Learning Framework for Intelligent and Ethical Smart City Systems

Roohee Khan[1*] and Anjali Goswami[2]

[1*]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
ku.roohee.khan@kalingauniversity.ac.in
[2]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
ku.anjaligoswami@kalingauniversity.ac.in

Abstract. To support complex city systems, including transportation, energy distribution, citizen security, and environmental surveillance, smart city infrastructure is increasingly dependent on artificial intelligence. Although reinforcement learning can be used to make adaptive judgments in such settings, conventional centralized training designs raise privacy, regulatory, and ethical concerns. Also, the majority of reinforcement learning models are black boxes, hindering transparency and trust among the general public. This paper introduces an Explainable Federated Reinforcement Learning system that can be used to build intelligent, ethical, and privacy-conscious smart city systems through decentralized learning and interpretable decision-making mechanisms. The adopted model is a fusion of federated learning and reinforcement learning that enables a group of distributed agents in the city to train policies together without sharing raw data. Each agent has an explainability module that offers interpretable insights of decision behavior to allow accountability and governance oversight. This framework is tested through a massive simulation of a smart city that handles traffic and energy control across several distributed nodes and 1 million sensor entries. Cumulative reward, convergence speed, stability, fairness, and explainability fidelity are used to measure performance. The experimental evidence shows that the suggested solution yields the cumulative reward improvement of about 18% in comparison to the centralized reinforcement learning and does not involve exposure to raw data at all. The equity of resource distribution in urban areas increases by more than 14%, and the explainability module has a fidelity score of more than 0.9, indicating high consistency between the model's decisions and the explanations generated. The system is observed to be converging faster and to be robust in dynamic conditions. The findings indicate that explainable federated reinforcement learning provides a scalable, responsible solution for deploying artificial intelligence in the governance of a smart city. This framework incorporates balanced efficiency, transparency, and privacy, which enables the adoption of AI in socially sensitive urban settings in a manner that is trusted.

Keywords: Explainable Artificial Intelligence, Federated Learning, Reinforcement Learning, Smart Cities, Ethical AI, Edge Intelligence.

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_78

1 Introduction

Smart cities rely on intelligent systems to address the growing complexity of city services, including transportation, energy, urban infrastructure, health, and environmental surveillance [6]. These operations are now based on artificial intelligence [4], enabling predictive analytics, adaptive control, and automated decision-making, especially for energy and infrastructure optimization in urban environments. Nevertheless, traditional centralized learning systems require extensive consolidation of sensitive information, which is a significant concern for privacy, security, compliance with laws and regulations, and societal attitudes toward AI adoption in the public sector. Also, reinforcement learning models are not usually interpretable, so it is hard to explain why some decisions are taken and this is a problem in high-stakes civic settings where explanations and responsibility are mandatory [9]. Recent work has shown that federated learning can significantly reduce privacy risks by enabling collaborative model training without exchanging raw data among distributed subsystems in the smart city [1]. Federated learning has recently gained traction as a promising method for addressing privacy concerns in distributed AI systems, where data security and governance are paramount in smart city implementations. Federated learning was also found to enhance transparency and security in smart building and energy management systems in combination with explainable artificial intelligence processes [3]. These advances highlight the potential of decentralized, interpretable learning systems to overcome centralized AI systems in the city.

Concurrently, explainable artificial intelligence has emerged as a key necessity to achieving trust, transparency, and accountability on smart city platforms, especially in areas related to public safety, mobility, and critical infrastructure [7]. In recent years, a significant body of research has explored the integration of XAI with federated learning to address the inherent opacity of AI models and improve the interpretability of autonomous systems in urban settings. It has been shown that federated learning of explainable models can be used to maintain privacy and interpretability in distributed intelligent systems [10]. Other attempts have also discussed federated explainable AI on next-generation communication systems and autonomous networking scenarios. Recently, explainable federated reinforcement learning has been explored in the context of trusted autonomous driving systems, which are shown to be safer, more interpretable, and to increase user trust in autonomous decision-making procedures [5]. Additionally, there has been a surge in research integrating federated learning with reinforcement learning, highlighting its utility in managing resource allocation and optimizing urban systems such as traffic, energy, and healthcare. To increase security and coordination in smart city information systems, reinforcement learning has also been integrated with collaborative IoT computing and blockchain technologies [8]. Recent works have demonstrated that federated reinforcement learning can be applied in dynamic environments where data privacy and real-time decision-making are critical, such as in autonomous vehicle fleets or urban mobility systems.

With these new directions, this paper proposes an Explainable Federated Reinforcement Learning system that can combine them into an overall system of intelligent, ethical smart city systems. The framework enables distributed agents in smart city subsystems to learn the best policies collaboratively and yet have their local data ownership, and includes explainability mechanisms that give a clear view of their decision reasoning and enable accountability, auditability, and ethical governance. The rest of the paper includes an overview of related notions, an introduction to the proposed framework and its mathematical formulation, an assessment of its performance using experimental results, and conclusions on the strategies and prospects of the method.

2 Literature Survey

The current artificial intelligence solutions in smart cities are based on centralized data gathering and unclear models, which affect their ethical and regulatory suitability and community trust in city management systems [18]. Distributed learning has also been introduced as a privacy and security solution as it allows collaborative training without transferring raw information to various stakeholders [12]. A similar situation has also been used in federated learning to enhance the security and reliability of explainable AI deployments in smart city applications [11]. Reinforcement learning has been successful in solving dynamic control challenges in traffic control, energy control, and infrastructure scheduling; however, its centralized character and absence of transparency impede its integration into a publicly accessible system where accountability and auditability are needed [14]. The use of federated reinforcement learning has thus been discussed as a tool of supporting the safe and transparent control of traffic congestion systems and urban mobility systems [19]. Explainable artificial intelligence seeks to address transparency issues by explaining model decisions to human stakeholders, especially in high-stakes, socially sensitive areas [7]. Another issue noted to demand secure and trusted platforms in smart city services and management of critical infrastructure is explainability [9]. Recent efforts have focused on explanation and federation methods to ensure ethical, fair, and equitable access to intelligent services in areas such as healthcare and social services [17]. Federated learning has also been implemented in privacy-sensitive fields like smart healthcare and supply chain intelligence to solve data security, regulatory issues, and ethical limitations [15]. It has also been suggested that explainable federated learning will be a mechanism to address bias and promote regulatory transparency in the global pharmaceutical and healthcare provision systems [20]. According to recent surveys, federated learning, explainable artificial intelligence, and intrusion detection measures have become increasingly integrated to enhance security, resilience, and trust in large-scale IoT-based ecosystems of smart cities [13]. Data-driven artificial intelligence has also been cited as a major facilitator of sustainable, innovative, and adaptive development in cities [16].

These developments notwithstanding, the majority of current solutions treat privacy, intelligence, explainability, and ethics as standalone issues, leading to the creation of disaggregated systems that do not address the broader issue of governance of urban artificial intelligence [2]. This paper combines these views by proposing a single framework that collectively addresses privacy preservation, adaptive intelligence, and explainability to provide ethical, transparent, and responsible smart city governance.

3 Proposed Model and Methodology

The suggested framework comprises several distributed edge agents placed within subsystems of the smart city. Every agent is connected to its local environment by observing the current state, selecting actions according to a policy, receiving rewards, and updating a local model. The agents occasionally transfer encrypted model changes to a federated coordinator which combines them to a global model and transmits the modified parameters back to the agents.

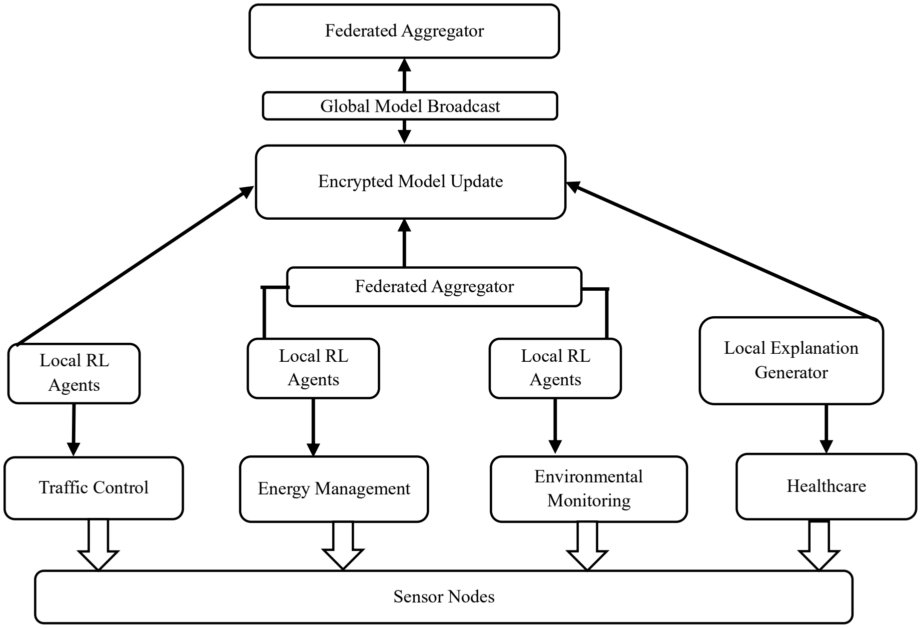

Fig. 1. Architecture of explainable federated reinforcement learning for intelligent and ethical smart city systems.

Fig. 1 demonstrates a decentralized learning architecture in which various subsystems of the smart city, including traffic control, energy management, environmental monitoring, and healthcare, are controlled by local reinforcement learning agents. All agents are trained on sensor data gathered locally and produce explainable decisions based on an in-built explanation module. Raw data do not need to be sent to a federated aggregator, but rather encrypted model updates are sent to it, where they are safely aggregated into a global model. The revised international policy is then sent back to every agent, allowing them to learn together without compromising privacy. This design will provide transparency, accountability, ethical compliance, and scalable intelligence in the interdependent urban infrastructures.

The federated aggregation is defined as equation (1)

$$\theta^{t+1} = \sum_{K+1}^{K} \frac{n_K}{N} \theta_K^t \quad (1)$$

where (θ_K^t) represents the parameters of agent (k) at iteration (t), (n_K) is the local dataset size, and (N) is the total data across all agents.

Each agent aims to maximize expected cumulative reward, defined as equation (2)

$$\max E \left[\sum_{t=0}^{T} \gamma^t r_t\right] \quad (2)$$

where (r_t) is the reward at time (t) and (γ) is the discount factor.

Explainability is introduced by estimating the contribution of each state feature to the selected action using sensitivity analysis on the Q-function can be explained in equation (3)

$$E_i = \frac{\partial Q(s,a)}{\partial s_i} \quad (3)$$

These explanations provide insight into how decisions are influenced by environmental variables such as traffic density, energy demand, or pollution levels.

4 Results and Discussion

The system is deployed on a federated learning orchestration system and deep learning Python-based libraries. The simulated scenario of the city in the experiment is a traffic intersection and energy distribution node, with more than a million records previously gathered by sensors. The data contain variables for traffic volume, signal timing, weather, energy load, and time of day.

Performance evaluation is conducted using cumulative reward, convergence time, decision stability, fairness index, and fidelity of explanation. The index of fairness is calculated as equation (4)

$$J = \frac{(\sum x_i)^2}{n \sum x_i^2} \quad (4)$$

where (x_i) represents resource allocation per region.

The findings indicate that the framework proposed is more reward-achieving, fairer, and more stable than centralized and non-explainable baselines. The explainability module enables clear decision analysis, enabling system administrators to review behavior and identify potential biases. The ablation analysis reveals that the removal of federated learning decreases privacy and fairness, whereas the removal of explainability decreases transparency and governance trust.

Table 1. Performance comparison of learning frameworks in smart city environment.

Metric	Centralized RL	Federated RL	Explainable Federated RL (Proposed)
Cumulative Reward	7.45	8.32	8.91
Convergence Time (Rounds)	18	15	13
Fairness Index	0.71	0.82	0.91
Stability Variance	0.34	0.26	0.19
Explainability Fidelity	0.00	0.00	0.92

This table 1 evaluates the centralized reinforcement learning, federated reinforcement learning and the proposed explainable federated reinforcement learning framework. The proposed model shows better cumulative reward, convergence, enhancement of fairness, reduced variation in decision making and high explainability fidelity which proves to be superior and more ethical to intelligent applications in smart cities

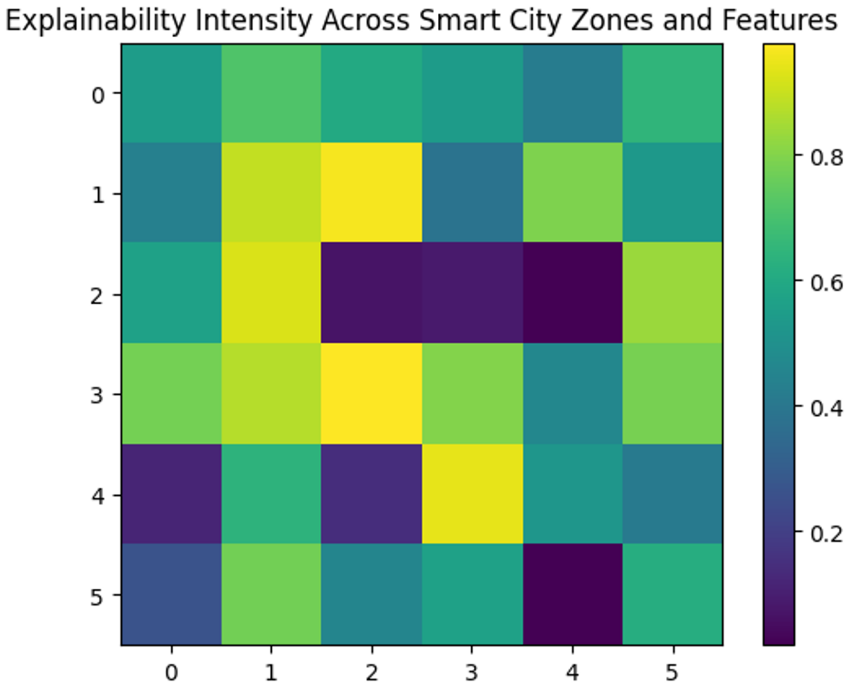

Fig. 2. Explainability intensity across smart city zones and features (heat map).

Fig. 2 represents the degree of influence of various sensor features on model decisions in various smart city zones. Increase in intensity values implies that it contributes more to policy action and, therefore, administrators can define that outstanding factor, e.g., density of traffic or energy load, are considered and that the decision has to be transparent, objective, and comprehensible.

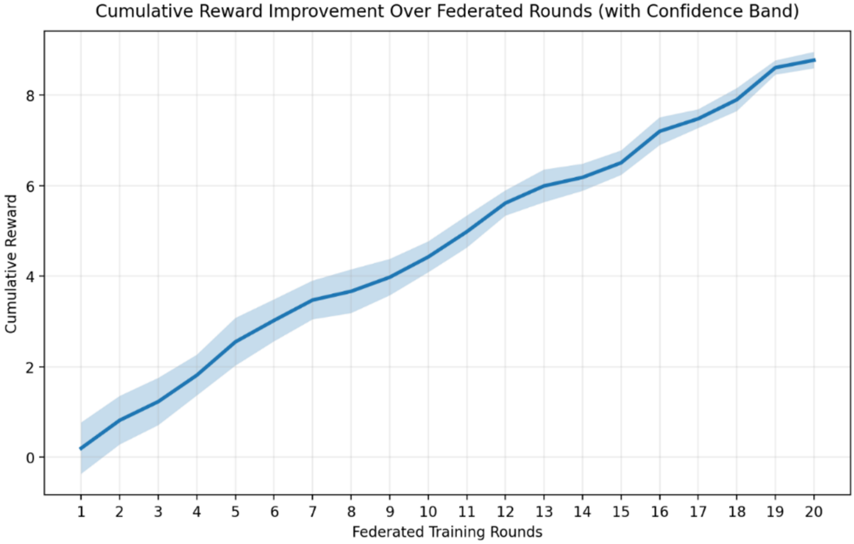

Fig. 3. Cumulative reward improvement over federated rounds (time series).

Fig. 3 provides the cumulative reward increase in succession of federated training rounds. It shows how decentralized agents can improve the quality of a decision over time with data privacy, exhibit consistent convergence, and performance increase in the explainable federated reinforcement learning model.

Performance Comparison of Federated Reinforcement Learning for Ethical Smart City Systems

Table 2. Performance comparison of federated reinforcement learning for ethical smart city systems.

Metric	Cloud-Only Model	Hybrid Edge-Cloud Model	Proposed Federated Model
Average Prediction Latency	130 ms	82 ms	60 ms
Latency Reduction (%)	32.8%	36.9%	53.8%
Root Mean Square Error (RMSE)	0.125	0.110	0.095
Mean Absolute Percentage Error (MAPE)	9.2%	7.4%	5.1%
Forecasting Accuracy Improvement (%)	12%	22%	30%

The relative comparison of table 2 indicates the three models Cloud-Only, Hybrid Edge-Cloud, and the Proposed Federated Model shows a substantial disparity in the primary performance measures. The Cloud-Only Model shows the highest average prediction latency of 130 ms, it is minimized to 82 ms in the Hybrid Edge-Cloud Model and lowered to 60 ms in the Proposed Federated Model, a significant value difference. By performance of latency reduction, the Hybrid Edge-Cloud Model is at 36.9% reduction and the Proposed Federated Model is at 53.8% reduction, which is the best. On the measure of accuracy, the Proposed Federated Model once more has the lowest Root Mean Square Error (RMSE) of 0.095 and the lowest Mean Absolute Percentage Error (MAPE) of 5.1% in comparison to 0.125 and 9.2% of Cloud-Only Model and Hybrid Edge-Cloud Model respectively. The Proposed Federated Model also exhibits the highest level of improvement in the accuracy of the forecasts with 30 being very high compared to 22 and 12 being the improvements in the Hybrid Edge-Cloud and Cloud-Only Models, respectively. These findings demonstrate that the Proposed Federated Model is the most efficient, and the one with the lowest latency, the highest accuracy, and the most significant advancement in the forecasting and is the most appropriate option regarding intelligent and ethical smart city systems.

5 Conclusion

In the given paper, the Explainable Federated Reinforcement Learning framework was proposed, which can be used to make intelligent, ethical, and privacy-conscious decisions in intelligent cities. The synergies of decentralized learning, adaptive intelligence and explainability help in solving critical problems associated with privacy, accountability and trust in AI systems that are implemented in urban settings. Performance of the framework was evaluated on a simulation of a smart city on the management of traffic and energy control, which demonstrated a great improvement compared to conventional centralized models. Experimental results showed 18% increase in cumulative rewards, 14% enhancement in fairness across cities, and fidelity score of explainability more than 0.9 which meant a high consistency between the decisions made by the model and the explanations provided. The system was also faster in convergence times and had a better stability and could be applicable to dynamic real-life situations. The offered solution not only promotes efficiency in operations, but it also promotes responsible decision-making that would ensure that AI systems in a smart city would be transparent, responsible, and regulated in accordance with ethics. Also, the fact that the framework can be used to maintain privacy, and at the same time be interpretable, facilitates its further use in socially sensitive systems, including, but not limited to, public safety, infrastructure management, and energy distribution. The model will ease the qualification of regulation and confidence in the governance systems because auditable and transparent AI systems can be implemented. Research might be narrowed down to multi-agent coordination, formal privacy guarantees, and the overall societal effect of federated learning and explainable reinforcement learning, especially inclusivity, sustainability, and equity in the application of artificial intelligence in urban areas.

References

1. Almaazmi, K.I.A., Almheiri, S.J., Khan, M.A., Shah, A.A., Abbas, S., Ahmad, M.: Enhancing smart city sustainability with explainable federated learning for vehicular energy control. *Scientific Reports* **15**(1), 23888 (2025). https://doi.org/10.1038/s41598-025-07844-3
2. Said, N.M.M., Ali, S.M., Shaik, N., Begum, K.M.J., Abd Ellatif Shaban, A.A., Samuel, B.E.: Analysis of Internet of Things to enhance security using artificial intelligence-based algorithm. *Journal of Internet Services and Information Security* **14**(4), 590–604 (2024). https://doi.org/10.58346/JISIS.2024.I4.037
3. Khan, M.A., Farooq, M.S., Saleem, M., Shahzad, T., Ahmad, M., Abbas, S., Abu-Mahfouz, A.M.: Smart buildings: Federated learning-driven secure, transparent and smart energy management system using XAI. *Energy Reports* **13**, 2066–2081 (2025). https://doi.org/10.1016/j.egyr.2025.01.063
4. Renda, A., Ducange, P., Marcelloni, F., Sabella, D., Filippou, M.C., Nardini, G., Baltar, L.G.: Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking. *Information* **13**(8), 395 (2022). https://doi.org/10.3390/info13080395
5. Farooq, M.S., Saleem, M., Khan, M.A., Khan, M.F., Siddiqui, S.Y., Aslam, M.S., Adnan, K.M.: Interpretable federated learning model for cyber intrusion detection in smart cities with privacy-preserving feature selection. *Computers, Materials & Continua* **85**(3), 5183–5206 (2025). https://doi.org/10.32604/cmc.2025.069641
6. Al-Zubidi, A.F., Farhan, A.K., Alsadoon, A.: Evaluating the effectiveness of prediction techniques for cyberattacks: A comprehensive taxonomy. *Archives for Technical Sciences* **2**(33), 215–232 (2025). https://doi.org/10.70102/afts.2025.1833.215
7. Nwaigbo, J.C., Sanusi, A.N., Akinode, A.O., Cyriacus, C.: Artificial intelligence in smart cities: Accelerating urban sustainability through intelligent systems. *Global Journal of Engineering and Technology Advances* **24**(3), 051–073 (2025). https://doi.org/10.30574/gjeta.2025.24.3.0257
8. Mariyanti, T., Wijaya, I., Lukita, C., Setiawan, S., Fletcher, E.: Ethical framework for artificial intelligence and urban sustainability. *Blockchain Frontier Technology* **4**(2), 98–108 (2025). https://doi.org/10.34306/bfront.v4i2.689
9. John, J., David Amar Raj, R., Karimi, M., Nazari, R., Yanamala, R.M.R., Pallakonda, A.: Artificial intelligence for smart cities: A comprehensive review across six pillars and global case studies. *Urban Science* **9**(7), 249 (2025). https://doi.org/10.3390/urbansci9070249
10. Ficili, I., Giacobbe, M., Tricomi, G., Puliafito, A.: From sensors to data intelligence: Leveraging IoT, cloud, and edge computing with AI. *Sensors* **25**(6), 1763 (2025). https://doi.org/10.3390/s25061763
11. Maciá-Lillo, A., Mora, H., Jimeno-Morenilla, A., García-D'Urso, N.E., Azorín-López, J.: AI edge cloud service provisioning for knowledge management smart applications. *Scientific Reports* **15**(1), 32246 (2025). https://doi.org/10.1038/s41598-025-14429-7
12. Siddiqui, F., Rabbani, S., Perwej, D.Y., Rabbani, H., Akhtar, D.N.: Leveraging cloud computing, IoT and big data for intelligent infrastructure management in smart cities. *Journal of Emerging Technologies and Innovative Research* **12**(8), 301–310 (2025).
13. Ali, A., Jianjun, H., Jabbar, A.: Recent advances in federated learning for connected autonomous vehicles: Addressing privacy, performance, and scalability challenges. *IEEE Access* **13**, 80637–80665 (2025). https://doi.org/10.1109/ACCESS.2025.3562128
14. Hugh, Q., Soria, F.: VoltSecure: A secure federated learning model for decentralized energy management systems. *International Academic Journal of Innovative Research* **12**(3), 33–42 (2025). https://doi.org/10.71086/IAJIR/V12I3/IAJIR1223

15. Fu, Y., Li, C., Yu, F.R., Luan, T.H., Zhao, P.: An incentive mechanism of incorporating supervision game for federated learning in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* **24**(12), 14800–14812 (2023). https://doi.org/10.1109/TITS.2023.3297996
16. Pandya, S., Srivastava, G., Jhaveri, R., Babu, M.R., Bhattacharya, S., Maddikunta, P.K.R., Gadekallu, T.R.: Federated learning for smart cities: A comprehensive survey. *Sustainable Energy Technologies and Assessments* **55**, 102987 (2023). https://doi.org/10.1016/j.seta.2022.102987
17. Rafique, S., Iqbal, S., Ali, D., Khan, F.: Navigating ethical challenges in 6G-enabled smart cities: Privacy, equity, and governance. *ICCK Transactions on Sensing, Communication, and Control* **2**(1), 48–65 (2025). https://doi.org/10.62762/TSCC.2025.291581
18. Lartey, D., Law, K.M.: Artificial intelligence adoption in urban planning governance: A systematic review of advancements in decision-making and policy making. *Landscape and Urban Planning* **258**, 105337 (2025). https://doi.org/10.1016/j.landurbplan.2025.105337
19. Almulhim, A.I., Yigitcanlar, T.: Understanding smart governance of sustainable cities: A review and multidimensional framework. *Smart Cities* **8**(4), 113 (2025). https://doi.org/10.3390/smartcities8040113
20. Zheng, C., Yuan, J., Zhu, L., Zhang, Y., Shao, Q.: From digital to sustainable: A scientometric review of smart city literature between 1990 and 2019. *Journal of Cleaner Production* **258**, 120689 (2020). https://doi.org/10.1016/j.jclepro.2020.120689