A Trustworthy Artificial Intelligence Approach for Satellite Systems Digital Twins

Umberto Francesco Carolini*, Saverio Ieva*, Ivano Bilenchi*, Filippo Gramegna†, Agnese Pinto*, Corrado Fasciano*, Floriano Scioscia*, Michele Ruta*

* Polytechnic University of Bari – Via E. Orabona 4, Bari (I-70125), Italy – {name.surname}@poliba.it

† LUM "Giuseppe Degennaro" University – Casamassima BA (I-70010), Italy – {gramegna}@lum.it

***Abstract*—Mission-critical environments require decision support systems that are both accurate and trustworthy. Traditional black-box Artificial Intelligence (AI) models are unsuitable because anomaly detection without explainability or data traceability poses unacceptable operational risks. This paper presents a Knowledge-based Digital Twin framework that ingests, processes, and interprets telemetry data through a hybrid inference engine. Beyond the space domain, its general architecture is also applicable to complex Smart Cities domains like infrastructure monitoring and mobility management. By integrating eXplainable Artificial Intelligence (XAI) metrics with semantic data provenance analysis based on Provenance Ontology (PROV-O), the framework ensures end-to-end auditability, making automated predictions explainable, traceable to the original sensor, and operationally reliable.**

Index Terms*—Semantic Digital Twin, eXplainable AI, data provenance, anomaly detection, telemetry analysis, hybrid inference engine

I. Introduction

The increasing complexity of cyber-physical systems is pushing Digital Twins (DTs) beyond their traditional role of virtual replicas toward advanced simulation and decision-support environments. High-fidelity DTs can synchronize with physical assets through sensor streams, reproduce component and subsystem states, and support diagnostics and predictive analytics [1]. These capabilities are especially relevant in mission-critical domains, where decisions must be timely, reliable, and understandable by human operators.

Space systems provide a demanding reference scenario. Satellites operate in harsh environments with constrained resources, complex dynamics, and potentially catastrophic faults. Telemetry streams from on-board components can support anomaly detection, degradation assessment, and predictive maintenance [2]. Similar requirements arise in Smart City ecosystems, which rely on distributed and heterogeneous sensing infrastructures to monitor mobility, environmental conditions, and critical assets. In all these contexts, trustworthy simulation-based decision support must combine data-driven accuracy with interoperability, traceability, and explainability.

A major limitation of current Artificial Intelligence (AI)-enabled DT platforms is their reliance on opaque black-box models, which limit transparency in decision-making [3]. In mission-critical contexts, alerts lacking clear justification or data traceability pose unacceptable risks. Moreover, space systems require seamless communication among components and subsystems from different domains and manufacturers, creating semantic inconsistencies and integration bottlenecks.

To address these issues, this paper outlines a semantics-enhanced simulation framework for explainable and traceable DTs. Semantic models enrich raw telemetry and sensing streams with machine-interpretable metadata, reducing ambiguities across heterogeneous subsystems and enabling knowledge-based integration [4]. The annotated data feed an inference layer that supports anomaly detection and what-if analysis by combining data-driven models, knowledge-based representations, and physics-aware constraints.

The remainder of the paper is organized as follows. Section II introduces the reference architecture, while Section III presents the envisioned validation scenarios and expected benefits before the conclusion.

II. Proposed Framework

The proposed architecture, shown in Figure 1, is a modular end-to-end pipeline that transforms raw telemetry streams into actionable and traceable knowledge. The *Inbound layer* collects observation data through low-latency messaging protocols and forwards them to a DataService for orchestration and a DataStorage component for persistent archival and integrity preservation. The ingested data then enter the *Decision Support System (DSS)*, where a semantic filter enriches telemetry with metadata, contextual information, and ontology-based standard definitions, improving interoperability across heterogeneous systems [4]. The annotated data are processed by the Inference Engine, which integrates Physics-Informed Machine Learning (PIML) by embedding governing physical equations into neural-network loss functions, thus supporting consistent and generalizable predictions [5].

The engine also addresses transparency through two complementary mechanisms: eXplainable Artificial Intelligence (XAI) and Semantic Provenance. XAI applies post-hoc quantitative techniques, including Grad-CAM, [6] Integrated Gradients, [7] SmoothGrad, [8] SHAP, [9] LIME, [10] Feature Importance, [11] and Saliency Maps, [12] to identify the telemetry variables driving anomaly predictions and convert detection outputs into interpretable diagnostics [3]. However, XAI alone cannot reconstruct model versions, data origins, or preprocessing steps. To ensure end-to-end trustworthiness, the framework adopts Provenance Ontology (PROV-O) [13] to represent the data lifecycle through: (i) *entities*, such as

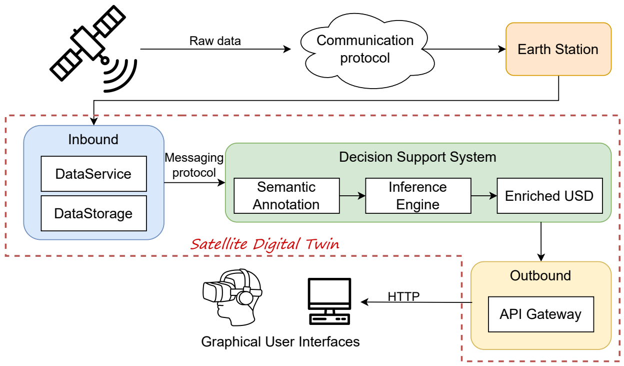

Fig. 1. Framework architecture

raw packets, processed datasets, and Machine Learning (ML) checkpoints; (ii) *activities*, namely algorithms and processing operations; (iii) *agents*, including people, organizations, and software components.

When an anomaly is detected, XAI identifies the immediate predictors, while provenance reconstructs the complete processing chain. The result is encoded as a knowledge graph fragment, integrated into a Universal Scene Description (USD) file [14], and delivered to the *Outbound layer*, which exposes a RESTful Application Programming Interface (API) for interactive 3D monitoring through multimodal interfaces.

III. Preliminary Validation Scenarios

The proposed framework serves as a reference architecture for semantics-enhanced simulation and DT platforms. Accordingly, the work focuses on functional requirements, integration principles, and the expected benefits of the approach.

A first validation scenario will address space systems, where telemetry streams from satellite subsystems can support anomaly detection, predictive maintenance, and what-if analysis. In this context, the framework is expected to make alerts explainable through XAI techniques and traceable through semantic provenance metadata, linking the original telemetry packet to the final decision-support output.

Assessment will rely on theoretical, [15] structured, [16], [17] and ontology-based [18] frameworks for user-centered XAI. The evaluation will consider the semantic harmonization of heterogeneous data streams, the quality and usefulness of generated explanations, and the completeness of provenance traces in reconstructing the decision-making process.

IV. Conclusion

This paper has introduced a Knowledge-based DT framework that improves explainability and auditability compared with black-box AI approaches. By combining PIML, post-hoc XAI metrics, and end-to-end traceability through the PROV-O ontology, the framework supports justification and verification of automated predictions and decisions. Ongoing case studies are assessing its compliance with the strict reliability requirements of satellite (sub)systems and missions. The methodology also offers a reusable foundation for intelligent DT platforms in other mission-critical Smart City domains.

Acknowledgments

This work has been supported by the *Space It Up!* project, funded by the Italian Space Agency (ASI) and the Ministry of University and Research (MUR), under contract n. 2024-5-E.0 - CUP n. I53D24000060005.

References

[1] E. Mikołajewska, D. Mikołajewski, T. Mikołajczyk, and T. Paczkowski, “Generative AI in AI-based digital twins for fault diagnosis for predictive maintenance in Industry 4.0/5.0,” *Applied Sciences*, vol. 15, no. 6, p. 3166, 2025.

[2] S. Cuéllar, M. Santos, F. Alonso, E. Fabregas, and G. Farias, “Explainable anomaly detection in spacecraft telemetry,” *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108083, 2024.

[3] H. A. Tahir, M. R. Jabbarpour, B. Q. Vo, R. Kowalczyk, J. Barr, and T. Bessell, “Toward Explainable AI in Spacecraft Health Monitoring: Comparative Benchmarking of Anomaly Detection Models and Open-Source Datasets,” *IEEE Access*, vol. 13, pp. 209 368–209 398, 2025.

[4] M. Elhajj, “Semantic foundations for digital twins: the contribution of ontological analysis,” *Frontiers in Computer Science*, vol. 8, 2026.

[5] H. Farhat and A. Altarawneh, “Physics-Informed Machine Learning for Intelligent Gas Turbine Digital Twins: A Review,” *Energies*, vol. 18, no. 20, 2025.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. Jour. of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[7] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Int. Conf. on Machine Learning*, 2017, pp. 3319–3328.

[8] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” *arXiv preprint arXiv:1706.03825*, 2017.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in NeurIPS*, vol. 30, 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, “”Why should i trust you?” Explaining the predictions of any classifier,” in *22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[11] H. Mandler and B. Weigand, “A review and benchmark of feature importance methods for neural networks,” *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–30, 2024.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, “Visualising image classification models and saliency maps,” *Deep Inside Convolutional Networks*, vol. 2, no. 2, 2014.

[13] T. Lebo, S. Sahoo, and D. McGuinness, “PROV-O: The PROV Ontology,” https://www.w3.org/TR/prov-o/, 2013, W3C Recommendation.

[14] S. Ieva, D. Loconte, F. De Feudis, V. Di Ceglie, and F. Scioscia, “Annotating 3D Scenes with Knowledge Graphs for Smart Infrastructure Digital Twins,” in *Int. Conf. on Web Engineering*. Springer, 2025, pp. 98–109.

[15] M. N. Keas, “Systematizing the theoretical virtues,” *Synthese*, vol. 195, no. 6, pp. 2761–2793, 2018.

[16] S. Naveed, G. Stevens, and D. Robin-Kern, “An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI,” *Applied Sciences*, vol. 14, no. 23, p. 11288, 2024.

[17] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert, “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI,” *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–42, 2023.

[18] S. Chari, O. Seneviratne, M. Ghalwash, S. Shirai, D. M. Gruen, P. Meyer, P. Chakraborty, and D. L. McGuinness, “Explanation Ontology: A general-purpose, semantic representation for supporting user-centered explanations,” *Semantic Web*, vol. 15, no. 4, pp. 959–989, 2024.