AI-Driven Performance Evaluation of Network Digital Twins for 6G: A Unified Taxonomy and Polymetric Twinning Index

Elif Ak, *Member, IEEE,* Trung Q. Duong, *Fellow, IEEE,* Berk Canberk, *Senior Member, IEEE.*

***Abstract*—This paper addresses the challenge of evaluating Network Digital Twins (NDTs) and discusses how to evaluate the performance of NDTs in AI-driven 6G systems. The paper makes three main contributions. First, it provides a layer-by-layer analysis of enabling technologies supporting NDTs, examining how tools like edge computing, blockchain, federated learning, and various data collection methods fulfill specific NDT requirements across key characteristics of NDTs. Building on this, it proposes a comprehensive taxonomy of performance evaluation metrics. This taxonomy systematically classifies both digital twin-specific metrics and general performance metrics. Finally, it introduces the polymetric twinning index, a simple, unified metric that aggregates multiple performance indicators into a single score. The approach is validated through five case studies spanning IoT and 5G networks, demonstrating how different metric subsets can be selected based on specific NDT focus areas. Results show that the taxonomy and polymetric twinning index together form an extensible framework for assessing NDTs across domains, which enables fair baseline comparisons when appropriate subsets are chosen.**

***Index Terms*—network digital twin, digital twin for networks, digital twin, performance metrics, enabling technologies.**

I. Introduction

The digital twin (DT) concept, representing a virtual counterpart of a product, process, or physical system, provides a platform to validate configurations, simulate changes, and enhance management efficiency. Originating from the NASA project, later presented by Grieves in 2002, the application of DT has expanded to include manufacturing processes and even human anatomical modeling [1]. It also plays a crucial role in addressing the complex challenges inherent in managing and operating communication networks. Due to its conceptual nature rather than being a definitive methodology or technology, the definitions, objectives, and technologies associated with DT vary widely both in academic literature and among companies. The concept of DT extends beyond a single technology or methodology, since it has a wider range representing a system of systems. Furthermore, as might be expected, there is no single standardized architecture for designing and developing DTs. This variability makes it essential to question *how we can evaluate the performance of a designed DT and ensure it functions as intended.*

Answering that question is not straightforward, especially given the broad scope of the DT concept. However, this ambiguity is more manageable when we narrow the focus, such as a DT of a manufactured product or a DT of a networking system, as is the case in this paper. Despite the pivotal role of DTs in enhancing network management and operations, the lack of standardization and ambiguity not only hinders the consistent assessment of the network DTs (NDTs) but also impedes potential improvements and innovations within this research domain. This paper aims to bridge this gap by introducing a unified approach for performance evaluation metrics specifically designed for NDTs. By synthesizing existing evaluation methodologies and aligning them with the specific needs of different networks, reviewing enabling technologies and guidelines from Standard Development Organizations (SDOs) [2]–[4], this paper strives to offer a comprehensive perspective that can be universally used across diverse NDT scenarios, including but not limited to smart cities, Internet of Things applications, 5G/6G developments, and core&access networks. While existing surveys comprehensively review NDT architectures [1], enabling technologies [5], security threats [6], and application scenarios [7], [8], none provide a systematic framework for evaluating NDT performance. The main contributions of the paper are summarized below:

- We provide a layer-by-layer analysis of technologies and enablers supporting NDTs, showing how each maps to key DT characteristics, which later guides the selection of appropriate evaluation metrics.
- We propose a broader taxonomy of performance evaluation metrics for NDTs, a novel addition to the literature that systematically classifies metrics based on the key characteristics defined in ITU-T Y.3090 [2] and IETF standards [3].
- We introduce a new metric, the *polymetric twinning index*, specifically designed to assess the performance of NDTs comprehensively and uniformly. This paper also includes a thorough statistical analysis to validate the efficacy and accuracy of our proposed metric.

Section II provides a foundational overview of NDTs.

E. Ak and T. Q. Duong are with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada. (e-mail: {elif.ak, tduong}@mun.ca).

B.Canberk is with the School of Engineering and Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT, UK, (e-mail: B.Canberk@napier.ac.uk).

The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109, in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program RGPIN-2025-04941. The work of B. Canberk is supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK) Frontier R&D Laboratories Support Program for BTS Advanced AI Hub: BTS Autonomous Networks and Data Innovation Lab Project 5239903.

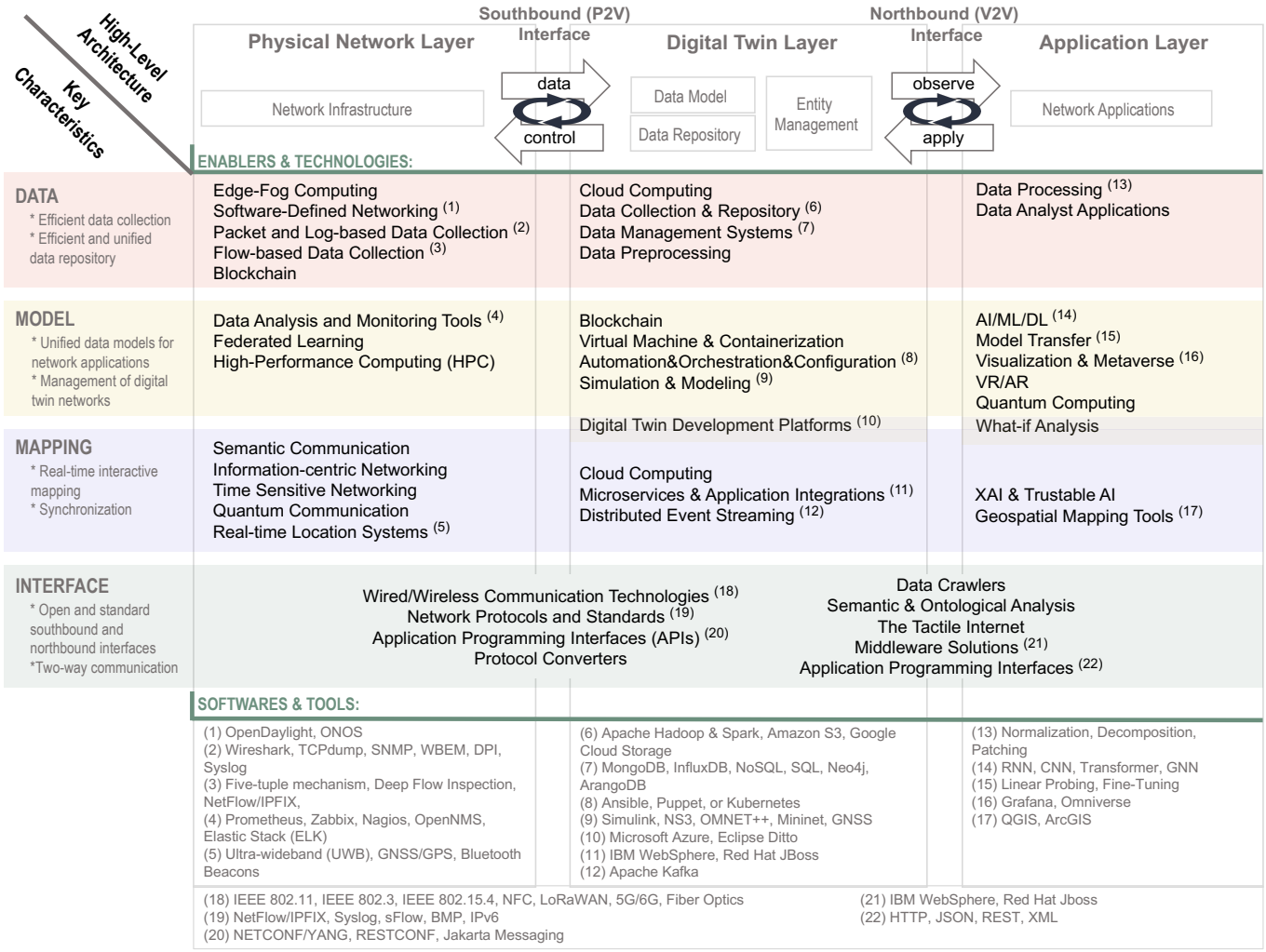


Fig. 1: Key Characteristics of NDTs and its Relation with Enablers, Technologies and Tools of the Networking Systems

Section III explores the devised taxonomy of the existing metrics for NDTs. Section IV presents and evaluates our newly proposed metric, the *polymetric twinning index*, and discuss the open research issues in Section V.

II. An Insight into Network Digital Twins

Given the broad and flexible nature of the DT concept, as previously discussed, it is not surprising that the terminology used to describe NDTs is equally varied. Terms such as "network digital twin" [9], "digital twin networks" [1], and "digital twin for/of [sometimes specifying a type of network] networks" [7] are frequently used interchangeably, each referring to the application of digital twins in networking environments. While this paper does not delve into which term is most appropriate, we adopt the term "network digital twin" in alignment with the terminology used by IEEE Network Magazine, which featured a Special Issue titled "network digital twin" [9], and it is observed that the term NDT has been more widely adopted and recognized in both academic and industry contexts[1].

Continuing our discussion on the varied terminology, it is crucial to recognize that DTs are conceptualized in two distinct aspects in the current literature: i) *networks for DTs* and ii) *DTs for networks* [7]. The former views communication networks as the main enablers for various twinning operations, treating DTs' communication requirements as challenges to be addressed/satisfied. Conversely, the latter sees DTs as solutions, emphasizing the development of efficient communication methods enabled by DTs. Our study explores both dimensions. Accordingly, we will first focus on *DTs for networks*, particularly the NDT concept, to assess the key characteristics of NDTs in Section II-A. Subsequently, we will explore *networks for DTs* as enabling technologies in Section III.

[1]The term "Digital Twin Networks" might be interpreted as networks of a digital twin rather than a digital twin of networks [1]. Since our focus is on the digital representation of a single network rather than on multiple interconnected digital twins, we will consistently use the term "network digital twin" throughout this paper.

A. Key Characteristics of Network Digital Twins

The literature presents several high-level architectural models for DTs, including 2-layer [5], 3-layer [1]–[4], [7], 4-layer [6], and 5-layer [8] architectures. While these frameworks essentially aim to achieve similar objectives and hold equivalent characteristics, they differ in how they distribute functionalities across various layers, resulting in diverse structural approaches. Consequently, it is not appropriate to deem any specific architecture as inherently superior or inferior. However, for the purposes of this paper, we adhere to the ITU-T recommendation Y.3090 [2], a standard specifically developed for Digital Twins within network infrastructures. This recommendation guides our use of terminology, the architectural framework we adopt, and the definitions and requirements we reference throughout the paper. The architecture of the NDT is defined in the ITU-T Y.3090 recommendation [2]. As shown in Fig. 1, it consists of three primary layers. The bottom layer (the leftmost), the physical network layer, contains all the tangible network elements and infrastructure. Above this lies the digital twin layer, where digital replicas of physical assets are managed. The topmost layer is the application layer, where network applications leverage insights generated from the digital twin to enhance network management and operational decision-making. According to both ITU-T Y.3090 [2] and IETF's Internet-Draft 07 [3], the defining key characteristics of the NDT include *Data, Model, Mapping*, and *Interface*, shown in Fig. 1 and explained below.

- **Data:** Data serves as the foundational element in NDTs, collected from network infrastructures and stored in a *unified data repository* to support crucial functions like analysis and decision-making.
- **Model:** NDTs utilize sophisticated *data models* to enable data-driven modeling that supports a wide range of network applications and to enhance the agility and programmability of network services through model instances that must undergo thorough emulation and verification. These models are categorized into (i) *basic models*, which provide real-time, accurate representations and verifications of the network's physical state and configurations, and (ii) *functional models*, which utilize extensive (sometimes historical) data to support complex operations like network analysis, diagnosis, and optimization across various network domains and functionalities.
- **Mapping:** The *real-time interactive* mapping between the physical network and its digital twin distinguishes NDTs from traditional network simulation systems. This feature ensures that changes and conditions in the physical network are immediately reflected in the digital twin in a synchronized way, facilitating up-to-date simulations and analytics.
- **Interface:** Standardized interfaces, including both *southbound interfaces*, i.e. physical-to-virtual (P2V), that connect with the physical network and *northbound interfaces*, i.e. virtual-to-virtual (V2V), that communicate with *network applications*, are crucial. These interfaces ensure that the digital twin is both scalable and interoperable with various network components and applications, adapting to specific use case requirements such as latency sensitivity and data volume.

These four key characteristics provide a clear framework of the functionalities essential for effectively realizing NDTs. To expand on these foundations, Fig. 1 offers a comprehensive outlook, the first of its kind in the literature, that aligns these characteristics with the architectural framework and requirements specified in ITU-T Y.3090. This alignment shows how each characteristic can be realized through specific enablers, technologies, and tools. It is important to note that while the list of enablers shown in Fig. 1 is extensive and based on a comprehensive review of literature, the examples of software and tools are illustrative and not exhaustive for each enabler. In subsequent sections, we will explore how this detailed mapping deepens our understanding of evaluating the specific features of NDTs in light of functional and operational needs, ensuring their effective deployment.

III. From Enablers to Evaluation: Understanding Performance Metrics in NDTs

As previously discussed, DTs are considered systems of systems, adapting to the specific context or domain in which they are utilized. Thus, evaluating the performance of the NDT thus necessitates an understanding of the underlying technologies that influence key metrics like accuracy, responsiveness, and reliability used in the creation of a virtual model. For instance, while a communication medium of the southbound interface is essential, the choice of a specific communication protocol depends on the unique requirements of the NDT application.

Consequently, the proposed taxonomy is systematically constructed by following the key characteristics defined in ITU-T Y.3090 and IETF's Internet-Draft on the NDT. For each characteristic (data, model, mapping, interface), we first identified the associated enablers and technologies as shown in Section II-A, then derived corresponding evaluation metrics based on the functional requirements specified in these standards. Instead of adhering to a high-level, layered architectural approach, we categorize enablers, thus the proposed performance evaluation taxonomy of NDTs from the perspective of key characteristics. This approach also aligns with current research trends in DT studies, where studies typically focus on specific elements of DTs based on their individual expertise and domain knowledge instead of adopting a layer-by-layer analysis or constructing fully operational, ready-to-use systems.

A. Data Handling and Mapping

In this section, we explore enablers and performance evaluation criteria of data and mapping elements of NDTs together. Starting with item "1.1.Data Collection", as delineated in our taxonomy in Fig. 2, the data collection process is critical for completeness, diversity, and efficiency of resource usage. The selection of tools and methods for gathering data from physical networks is tailored to the specific network infrastructure, topology, and the needs of targeted network applications, as detailed in the referenced survey [10]. For instance, simpler techniques such as Simple Network Management Protocol

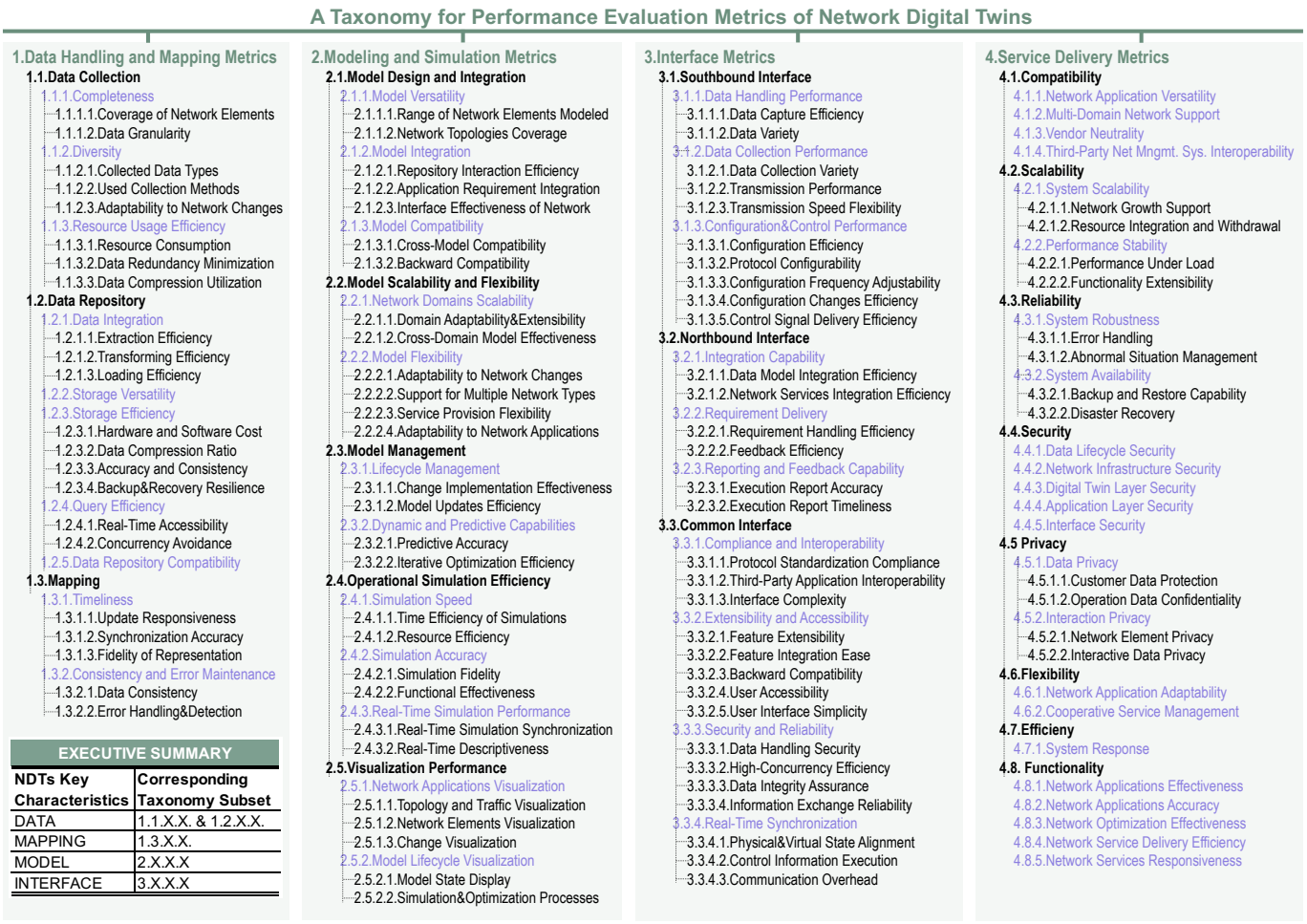


Fig. 2: A Taxonomy for Digital Twin Network Evaluation Metrics

(SNMP), TCPdump, and log-based methods like Syslog are apt for smaller networks requiring minimal data. In contrast, more intricate methods like Deep Packet Inspection (DPI) and NetFlow/IPFIX enable the collection of detailed data but might be limited by available resources. Additionally, advanced technologies such as the YANG data model with NETCONF telemetry and in-band network telemetry (INT) facilitate comprehensive data collection but can significantly impact resource consumption, including CPU, memory, and bandwidth. Real-time data collection, essential for applications such as anomaly detection, benefits from integrating multiple methods, enabling concurrent processing of diverse data streams. Enhancements in scalability and efficiency are achieved through Edge-Fog Computing and Software-Defined Networking (SDN), which support the collaborative use of various collection methods in a distributed approach. Furthermore, the adoption of blockchain technology enhances data consistency and trustworthiness, ensuring reliable data collection across diverse network environments.

Moving further with "1.2.Data Repository", the ITU-T recommendation also underscores the necessity for a robust data repository capable of managing the extensive data demands of NDTs. Tools such as Apache Hadoop and Apache Spark support high-volume data handling and real-time processing, while cloud solutions like Amazon S3 and Google Cloud Storage offer scalable data management. These platforms integrate, store and manage the data within the DT Layer. Databases such as MongoDB, Neo4j, and InfluxDB are typically employed for various data structuring and transaction speed, each bringing its own set of advantages and challenges to the NDTs' data repository landscape. Choosing the right technologies for the data repository involves systems that support parallel processing and can handle high levels of concurrency with minimal conflicts. Metrics for evaluating storage efficiency, cost-effectiveness, concurrency management, data integrity and security are essential when selecting a data repository for NDTs. Moreover, employing data compression techniques can also enhance storage efficiency. Additional functionalities such as fast search engines, data federation, batch data services, and capabilities for snapshot and rollback of historical data also play a significant role in NDTs.

Moving with "1.3.Mapping", often called *twinning* in NDTs, leverages several advanced technologies to meet its requirements for low latency and minimal overhead in data transmission. For instance, semantic communication improves data synchronization between physical networks and digital twins by transmitting only critical, meaningful information. Additionally, techniques such as information-centric network-

ing and time-sensitive networking enhance the timeliness of updates across network entities. These methods are supported by real-time location systems like ultra-wideband (UWB) and GNSS/GPS, which deliver precise geospatial data, particularly vital in mobile scenarios. On the software side, platforms like IBM WebSphere and Red Hat JBoss streamline the integration of complex network data into operational models. Apache Kafka facilitates robust event streaming, crucial for maintaining high fidelity in mirroring the physical network. Moreover, the incorporation of Explainable AI (XAI) and Trustable AI into the digital twins promotes transparency and reliability in decision-making processes.

B. Modeling and Simulation

Under "2.1.Model Design and Integration", NDTs develop unified data models to represent various network elements like base stations and routers, as well as network topologies. The effectiveness of these models depends on integration with a unified data repository, facilitating fast search, data federation, and batch processing. Tools such as Prometheus, Zabbix, and the Elastic Stack ensure that collected data is integrated into these models timely and accurate. Virtual machine and containerization technologies enhance model integration by providing scalable environments that adapt to network application requirements and efficiently report simulation results.

Moving to "2.2.Model Scalability and Flexibility," a robust automation and orchestration approach is critical for constructing data models that may focus on a single network domain, such as mobile access, transmission, core, or bearer networks or span multiple domains [11]. Federated learning plays a key role in enhancing model responsiveness by enabling local training at various network points, such as base stations or routers [7]. This reduces latency in decision-making and increases scalability by allowing for simultaneous model training across multiple nodes. High-performance computing (HPC) is essential for processing extensive datasets and making iterative refinements to models, thereby improving the accuracy and utility of predictions and impacting key performance metrics like system responsiveness and delay reduction.

Virtual machines, containerization, automation, and orchestration technologies collectively underpin "2.3.Model Management," facilitating the automated management of data, model instances, topologies, and security [12]. Blockchain technology ensures that these models are stored in an immutable and transparent manner, safeguarding them for future use [5]. Advanced AI and ML algorithms, particularly RNNs and GNNs, are utilized to refine these models for specific network scenarios, with techniques like fine-tuning to enhance their precision.

For "2.4.Operational Simulation Efficiency," platforms specifically designed for digital twins, such as Microsoft Azure, along with simulation tools like OMNET++ and NS3, support the development of predictive models for hypothetical scenarios. These tools enhance predictive accuracy and operational effectiveness by enabling emulation and iterative optimization of network applications.

Lastly, under "2.5.Visualization Performance," tools like Grafana provide powerful interfaces that transform complex model outputs into actionable insights, aiding effective decision-making. The user interfaces of simulation platforms are also crucial and should be evaluated to ensure they meet the needs of network management effectively.

C. Interface

The "3.1 Southbound Interface Performance" criterion assesses data acquisition from the physical network using a blend of wired and wireless technologies, including IEEE standards and 5G/6G. This criterion is divided into two distinct aspects based on the IETF's description of data collection technologies [3]: *data handling* and *data collection*. Data handling focuses on what types and formats of data can be captured from network elements, evaluating the interface's capability to acquire diverse network data such as configuration, operational state, topology, and telemetry data. Data collection, on the other hand, concerns how data is gathered and transmitted, including the selection of methods (passive, active, subscription-based, and on-demand) and protocols like NetFlow, Syslog, and NETCONF/YANG. The selection of these protocols and methods is crucial as they impact performance, efficiency, and reliability of data exchanges. It also measures the interface's ability to handle varying data update frequencies and deliver control signaling and configuration changes to network elements effectively.

In contrast, "3.2 Northbound Interface Performance" focuses on the bidirectional data exchange between the digital twin and network applications. It uses advanced data crawlers and semantic analysis for intent-based communication. Middleware solutions like IBM WebSphere, augmented with API technologies such as REST and JSON, ensure accurate and relevant communication and facilitate report execution from network digital twin models. The integration of Tactile Internet and ontological frameworks enhances interface responsiveness and intuitiveness.

The combined functionalities of these interfaces are grouped under "3.3 Common Interface Metric" in the taxonomy, designed to manage large data volumes and high concurrency within a seamless loop. Evaluation includes compliance, interoperability, and the interface's capacity to support digital copies or common data models for third-party applications. The performance evaluation also covers high concurrency management, massive data handling, and secure, reliable data transfers, maintaining open and standardized protocols with high extensibility and backward compatibility.

D. Service Delivery

This part of the performance evaluation taxonomy addresses the service requirements of NDTs from criteria "4.1 to 4.7" and their functionality in delivering intended services as listed in "4.8." Specifically, compatibility assesses support for diverse network applications, multi-domain infrastructures, and devices, including interoperability with services from different vendors. The NDT design is also observed for scalability, evaluating both system capacity and performance scalability. Reliability testing focuses on system robustness, particularly error

Taxonomy Number	Metric Name	Description	Range	Unit	Related NDT Layers	Selected References	Similar Metrics
Digital Twin-Specific Adapted Metrics							
1.3.1.1.	twinning rate	the ratio of the update period (time between updates) to the sum of the update period and synchronization delay, showing how much of the sync cycle is spent on actual updates rather than waiting. it can also be expressed as the proportion of successful synchronizations within a given time window.	[0,1]	N/A	Physical Network Layer, P2V Interface, Digital Twin Layer	[1], [8], [14]	update latency, data freshness index, real-time alignment factor
1.1.1.X	twin fidelity	the extent to which the digital twin collects data from all relevant network elements and devices	[0-100]	%	Physical Network Layer, P2V Interface, Digital Twin Layer	[10], [8], [14]	data accuracy index, granularity fidelity, model update rate, age-of-information
1.3.2.1	twin reliability	the probability that the digital twin operates without failure over a specified period, measured by uptime percentage, mean time between failures (MTBF), or error rate in state synchronization	[0,1]	N/A	P2V Interface, Digital Twin Layer	[1], [6], [8], [10]	twin uptime, error handling capacity, failover effectiveness, freshness of data, correlation-based measure
1.1.X.X.	twin geo-coverage	the proportion of the targeted area that is covered by the digital twin compared to the total area of intended network	[0-100]	%	Physical Network Layer, Digital Twin Layer	[4], [9]	spatial resolution, mobility rate, virtualization quality
2.4.2.1	twin behavioral correlation	the correlation coefficient (such as Pearson or Spearman correlation) between the time-series data of the digital twin's outputs and the real system's outputs, during what-if simulation observation period	[0,1]	N/A	Physical Network Layer, Digital Twin Layer	[6], [10], [13]	decision impact accuracy, simulation iteration accuracy, status consistency, temporal alignment of data
General Performance Metrics							
1.2.4.1.	query response time	measures the time taken for the data repository to respond to a query	[0, ∞)	second or ms	Digital Twin Layer, Application Layer	[6], [12]	data retrieval latency, database load time, transaction processing time, robustness to failure
2.2.2.3.	service scaling latency	measures the time taken for scaling services up or down in response to changing demand	[0, ∞)	second or ms	Digital Twin Layer	[4], [11], [12]	resource utilization efficiency, pod startup success rate, migration cost, transferability, flexibility
2.4.1.2.	resource utilization	measures the percentage of CPU and memory resources used during modeling operations	[0-100]	%	Digital Twin Layer, Application Layer	[1], [6]	system load average, peak memory usage, resource saturation level
2.4.3.X.	time-to-process	the total time required for processing and completing simulation tasks within a Network Digital Twin system	[0, ∞)	second or ms or hour	Digital Twin Layer, Application Layer	[9], [12]	event handling rate, jitter, concurrency capacity, event throughput
3.2.1.1.	integration success rate	quantifies the percentage of successful data model integrations into the digital twin system	[0-100]	%	Digital Twin Layer, V2V Interface	[1], [11], [12]	integration error rate, configuration success rate, data fusion quality
3.1.2.2.	throughput impact ratio	measures the relative change in network throughput before and after modifications to the southbound interface	[0-100]	%	Physical Network Layer, P2V Interface	[10], [12]	bandwidth utilization, packet delivery rate, link capacity utilization
4.8.1.X.	round trip time	measures the time taken for a signal to travel from the source to the destination and back	[0, ∞)	ms	ALL	[10], [12]	ping time, network latency, propagation delay, traffic rate
4.8.2.X	mean absolute error	calculates the average of the absolute differences between predicted values and actual values	[0, ∞)	N/A	ALL	[1], [4], [5]	Root Mean Square Error (RMSE), prediction accuracy rate, precision, recall, F1 score

Fig. 3: Example Subset of Metrics from Each Top-Level Category in the Proposed Taxonomy

management and handling unusual scenarios. Also, system reliability is measured by operational uptime, aiming for 99.99% reliability, which translates to a maximum yearly downtime of 52 minutes [13]. This criterion also encompasses evaluations of backup, restoration, and disaster recovery procedures [7]. Security and privacy are evaluated across all service layers and functionalities, ensuring robust defenses against various types of security threats [6] . Flexibility is assessed based on how effectively they manage data collection, storage, and processing. This evaluation also considers the system's end-to-end adaptability across different scenarios outlined in the intended case study, ensuring that the NDT meets specific flexibility criteria tailored to its objectives. Finally, "4.8.Functionality" measures how well the NDT enhances the performance of the underlying physical network. This includes evaluating the NDT's effectiveness post-implementation to determine improvements in network-specific performance metrics, such as packet delivery ratios in backbone networks or coverage range in wireless networks. The evaluation also examines the robustness of the digital twin and data models in accurately representing the physical network, predicting performance trends, identifying potential faults, and optimizing network operations using AI/ML algorithms [13].

IV. A New NDT Metric: The Polymetric Twinning Index

We reviewed the enablers, their roles, and their relationships with NDT design to clearly illustrate the proposed taxonomy presented in Fig. 2. Overall, we leave the decision of selecting subsets of evaluation criteria from our taxonomy, as well as the choice of enablers and technologies, to the discretion of professionals, recognizing that evaluating NDTs from all perspectives simultaneously is both challenging and impractical.

This section starts with an overview of the proposed *polymetric twinning index*. It is followed by an examination of five explanatory case studies that apply this taxonomy to assess NDT performance. We then evaluate and compare their performances using both the proposed *polymetric twinning index* and conventional performance metrics, all of which are detailed in Fig. 3. These metrics exemplify selections from each top-level category of the taxonomy, demonstrating representative metrics for each NDT characteristic, such as data, model, mapping, and interface. We have categorized the subset metrics presented in Fig. 3 into two groups: *digital twin-specific adapted metrics* and *general performance metrics*, which are frequently referenced in the literature. It is important to note that digital twin-specific metrics are seldom discussed in existing studies, making our selection some of the few explored in this area.

UC#	Intended Network	Data	Model	Mapping	Interface	Enablers	Best Practice Subset
UC1	IoT	●	○	●	◐	NETCONF, Azure, WiFi, ML	1.1. & 1.3. & 3.1.
UC2	IoT	●	○	●	◐	Neo4j, IPv6, Protocol Converters, DL	1.1. & 1.3. & 3.1.
UC3	IoT	◐	●	○	●	REST APIs, Cloud Computing, NS3, ETL, Elasticsearch	1.2. & 2.2. & 2.4. & 3.2.
UC4	5G	◐	●	○	●	REST APIs, Cloud Computing, NS3, ETL, Elasticsearch	1.2. & 2.2. & 2.4. & 3.2.
UC5	5G	◐	●	○	◐	Kubernetes, Microservices, NS3, Caching, InfluxDB, ML	1.2. & 2.2. & 2.4.

●: fully applied ○: not applied ◐: partially applied

Fig. 4: Example Use Cases and Their Most Suitable Performance Metric Subsets Based on Their Contributions to NDTs

A. Overview of the Proposed Polymetric Twinning Index

The proposed *polymetric twinning index* is a straightforward yet novel evaluation method designed to assess the performance of NDT across multiple dimensions quantitatively. This index is visually represented as the area of a polygon on a radar (or kiviat) chart, where each vertex represents a distinct performance metric relevant to NDT functionalities, as presented in Fig. 5. The benefits, drawbacks, and best practices of this approach are discussed in Sec. IV-D, following the presentation of experimental results. Here is a brief overview of how the index is utilized:

- A subset of metrics from the taxonomy is selected based on the focus and enablers of the NDT development.
- Concise metrics are then chosen, and as depicted in Fig. 3, both general and digital twin-specific metrics are considered for adaptation.
- Each selected metric is normalized to a uniform scale (e.g., from 0 to 1) using an appropriate normalization technique, such as min/max normalization, z-standardization, or softmax normalization with a logistic sigmoid.
- Metrics that inherently decrease in value to indicate improvement are inverted to ensure that all values represent positive increments.
- Metrics are plotted on the radar chart axes following their taxonomical order, either clockwise or counterclockwise. While the specific ordering affects the absolute area value (since adjacent metrics contribute to triangular segments in the area calculation), it does not affect fair comparison across different NDT implementations as long as the same order is consistently applied to all cases.
- Finally, the area enclosed by the axes connecting points is computed to derive the *polymetric twinning index*.

B. Case Studies

Drawing inspiration from existing studies, we developed five distinct case studies as summarized in Fig. 4 to demonstrate the application of our proposed taxonomy and evaluate the performance using the *polymetric twinning index*. Each case study, numbered UC1 to UC5, adapts enablers and methods from the literature to suit NDTs.

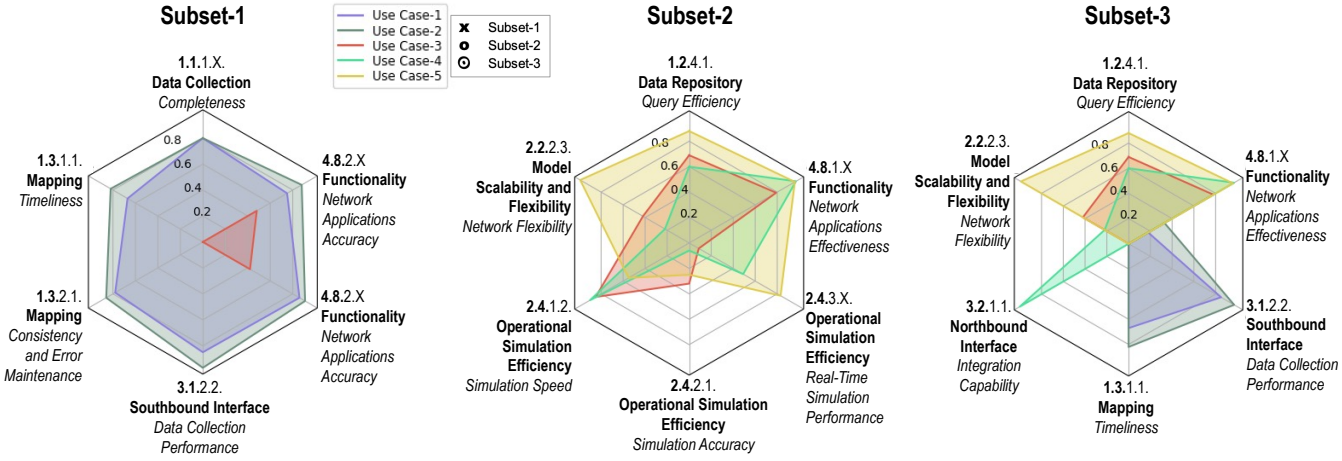
For diversity, we utilized two different physical networks: IoT and 5G. In the IoT scenario, various sensors generate data for smart city net-zero goals, as discussed in [14]. The aim is to predict waste amounts using sensor data and external information (e.g., weather, location) to optimize garbage collection and recycling routes autonomously. The primary focus is on *data* (particularly collection), *mapping*, and the *southbound interface*. In UC1, the WiFi protocol, supported by the NETCONF protocol, facilitates data collection in the southbound interface and mapping, while Microsoft Azure hosts sensor data for real-time processing and mapping. A multi-layer perception (MLP) model predicts truck path planning based on the dataset [14]. UC2 replicates UC1's structure but uses IPv6 over Ethernet for southbound interface and a protocol converter with Digital Twins Definition Language (DTDL) for mapping, as described in [15]. Neo4j replaces Azure to store graph-structured data in the digital twins, employing a gradient boosting (GB) model for the network application.

Skipping UC3 for now, UC4 and UC5 explore 5G networks, emphasizing mobile edge computing (MEC) by leveraging task offloading and caching to reduce latency, crucial for metaverse applications [4]. UC4 processes 5G network logs through an Extract-Transform-Load (ETL) pipeline into Elasticsearch on Amazon AWS, supporting NS3 simulations that inform network decision-making. A simplified optimization formula is used to minimize latency. UC5, while retaining the same problem definition, switches the data repository to InfluxDB and runs multiple caching and NS3 simulation instances orchestrated by Kubernetes. Microservice APIs manage the northbound interface communications, partly facilitating NS3 modeling. As a network application, UC5 employs deep reinforcement learning (DRL) instead of an optimization formula, optimizing offloading latency [4].

Finally, UC3 serves as a transitional case study, utilizing the problem definition from the IoT network outlined in UC1 and UC2 but employing enablers described in UC4. Instead of gathering data through IoT sensors, this study directly uses CSV files to transfer data to the cloud for modeling and storage, similar to the methods detailed in UC4 and UC5. The objective is to observe the outcomes when applying enablers initially designed for a different physical network, such as the 5G network in UC4, to another network type, specifically IoT. After developing these case studies, we assigned specific subsets from our taxonomy to each, analyzing their contributions to NDT aspects such as *data, model, mapping*, and *interface*. In this manner, we evaluate various strategies within the same network, and, as demonstrated by UC3, we also assess how enablers designed for one network type perform in another. This approach provides valuable insights into the generalization capabilities of our taxonomy and its applicability across various network environments.

C. Performance Evaluation

We utilized three subsets from the taxonomy for the evaluation: Subset-1, Subset-2, and Subset-3. Subset-1 emphasizes data (specifically the data collection aspect), mapping, and the southbound interface. In contrast, Subset-2 focuses on data (particularly the data repository aspect) and the model. Subset-3 aims to incorporate all essential elements of NDT to provide



	1.1.1.X. twin fidelity	1.2.4.1. query response time	1.3.1.1. twinning rate	1.3.2.1. twin reliability	2.2.2.3. service scaling latency	2.4.1.2. resource utilization	2.4.2.1. twin behavioral correlation	2.4.3.X. time-to-process	3.1.2.2. throughput impact ratio	3.2.1.1. integration success rate	4.8.1.X. round trip time	4.8.2.X. mean absolute error	4.8.2.X. prediction accuracy	Polymetric Twinning Index of Subset-1	Polymetric Twinning Index of Subset-2	Polymetric Twinning Index of Subset-3
Goal	↑	↓	↑	↑	↓	↓	↑	↓	↓	↑	↓	↓	↑	N/A		
UC1	80	-	0.67	0.78	-	-	-	-	15%	-	825ms	112	0.75	1,60	-	0,31
UC2	80	-	0.82	0.86	-	-	-	-	3%	-	680ms	89	0.88	1,98	-	0,48
UC3	-	338ms	-	-	289sec	15%	0.32	1830ms	-	-	250ms	308	0.48	0,09	0,64	0,36
UC4	-	427ms	-	-	387sec	10%	0.06	1043ms	-	100	73ms	143	0.69	-	0,68	0,40
UC5	-	159ms	-	-	0.14sec	45%	0.25	368ms	-	0	68ms	68	0.85	-	1,47	0,74
	×	o ⊙	× ⊙	×	o ⊙	o	o	o	× ⊙	⊙	o ⊙	×	×			

Fig. 5: Evaluation of *polymetric twinning index* throughout the proposed NDT taxonomy

TABLE I: Reproducibility Summary for Use Cases (UC1–UC5)

UC#	Data & Platform	Toolchain & Configuration	Setup Details
UC1	Smart city waste sensors dataset [14] Microsoft Azure IoT Hub IEEE 802.11ac + NETCONF/YANG	MLP 16-8-4-8 (Scikit-learn 1.0) NS-3 3.36, Python 3.9 CBOR and XML data mapping	Sensors: 100, Update: 30-sec Validation: 10-fold CV Duration: 800 s simulation
UC2	Same as UC1 Neo4j 4.4 (graph database) IPv6/Ethernet + DTDL [15]	Gradient Boosting (Scikit-learn 1.0) NS-3 3.36, Python 3.9	Sensors: 100, Neo4j graph nodes: ∼150 Validation: 10-fold CV Duration: 800 s simulation
UC3	Synthetic CSV file generated from UC1 AWS EC2 t3.large Elasticsearch 8.x (3-node cluster)	ETL pipeline (Pandas 1.5) NS-3 3.36, REST API ingestion Time-series sharding	Records: 50k, Batch: 1,000 Logging: ES query latency Duration: 800 s simulation
UC4	5G MEC task offloading AWS EC2 + Elasticsearch 8.x REST APIs (HTTP/JSON)	Latency optimization formula NS-3 3.36, ETL pipeline Metaverse application scenario	MEC servers: 5, Users: 100 Task rate: 10 req/s Logging: ES + NS-3 tracing
UC5	Same as UC4 Kubernetes 1.26 + InfluxDB 2.x Microservice APIs (gRPC/REST)	DRL with PyTorch 2.0 Multi-instance NS-3 3.36 Caching-enabled offloading	K8s pods: 8, Episodes: 1,000 Learning rate: 0.001 (Adam) Logging: InfluxDB time-series

a comprehensive evaluation. Given that none of the case studies were designed as fully operational NDTs, some aspects are expected to show suboptimal performance. Specifically, UC1, UC2, and UC3 were evaluated under Subset-1; UC4, UC5, and UC3 under Subset-2; and all case studies under Subset-3.

The red-shaded cells in the table placed below of Fig. 5 indicate the poorest results within each subset, while the green cells depict the highest *polymetric twinning index* results across all subsets. Additionally, we employed min-max normalization for the metrics with units presented in the table of Fig. 5. Moreover, the UC3 color code is also deliberately chosen as a red-shaded area to clearly illustrate its suboptimal performance across all subsets, serving as a critical analysis point for evaluating both generalization capabilities and baseline suitability of the proposed taxonomy, the importance of the subset and outcome of the *polymetric twinning index*.

D. Outcomes

The main outcomes related to the *polymetric twinning index* and the way of its usage are discussed below:

- The outcomes from Subset-1 and Subset-2 demonstrate that both the radar chart and the *polymetric twinning index*, with the proposed taxonomy, offer a robust framework for fair baseline comparisons within the selected subsets.
- The proposed taxonomy for selecting evaluation criteria enables the design of varied evaluation metrics, such as the *polymetric twinning index*. This facilitates a fair comparison by aiming to develop a use case-specific unified metric.
- Even without employing the *polymetric twinning index*, using a radar chart is highly recommended for comparing different NDT approaches within the same subset, thanks to its visual representation capabilities among various

evaluation criteria.

- The selection of subsets and the number of metrics from each top-level category are crucial and should be chosen based on existing studies to maintain a baseline for fair comparison.
- The number of metrics included is critical; too many can complicate evaluation, while too few (such as three or four) may not adequately represent performance.
- The order of metrics on the radar chart affects the absolute area calculation, as adjacent metrics form triangular segments that contribute to the total area. Therefore, it is advisable to maintain a fixed ordering, preferably following the taxonomy number sequence, across all evaluated cases. Grouping similar metrics together can also reduce artificial inflation of areas caused by unrelated high-performing metrics being adjacent.
- Given the varying importance of different metrics, considering a metric contribution effect may be beneficial. Weighting them or employing techniques like Principal Component Analysis can provide a more balanced evaluation.
- The normalization required in the *polymetric twinning index* can result in the loss of some informative values among metrics. However, as most metrics are naturally normalized on different scales, the index remains effective, useful, and informative.
- The choice of normalization technique significantly impacts the outcomes and should be clearly emphasized, as it can drastically change the results.

V. Conclusion

By aligning the performance metrics with specific network requirements, this research contributes significantly to the literature by offering a comprehensive and systematic approach that can be universally applied across various NDT scenarios. A detailed analysis of technologies and enablers, a novel taxonomy of performance metrics, and the introduction of the polymetric twinning index facilitate targeted assessments and highlight the importance of metric selection. The findings suggest that a detailed, structured approach to metric selection is essential for accurate and fair performance analysis, guiding future implementations and enhancements in network digital twin technologies.

Acknowledgements

The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109, in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program RGPIN-2025-04941. The work of B. Canberk is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) 1515 Frontier R&D Laboratories Support Program for BTS Advanced AI Hub: BTS Autonomous Networks and Data Innovation Lab, Project 5239903.

References

[1] Y. Wu, K. Zhang, and Y. Zhang, "Digital twin networks: A survey," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13 789–13 804, 2021.

[2] "Digital twin network - Requirements and architecture," ITU Recommendation ITU-T Y.3090, document, 2022. [Online]. Available: https://handle.itu.int/11.1002/1000/14852

[3] "Digital Twin Network: Concepts and Reference Architecture," IETF Draft, Rev. 07, document, 2022. [Online]. Available: https://datatracker.ietf.org/doc/draft-zhou-nmrg-digitaltwin-network-concepts/

[4] N. P. Kuruvatti, M. A. Habibi, S. Partani, B. Han, A. Fellan, and H. D. Schotten, "Empowering 6g communication systems with digital twin technology: A comprehensive survey," *IEEE Access*, vol. 10, pp. 112 158–112 186, 2022.

[5] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," *IEEE Access*, vol. 8, pp. 108 952–108 971, 2020.

[6] C. Alcaraz and J. Lopez, "Digital twin: A comprehensive survey of security threats," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 3, pp. 1475–1503, 2022.

[7] L. U. Khan, Z. Han, W. Saad, E. Hossain, M. Guizani, and C. S. Hong, "Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2230–2254, 2022.

[8] K. Duran, L. Verda Cakir, Y. Yigit, K. Huseynov, S. Ram Kusu, M. Ali Ertürk, and B. Canberk, "Toward digital twin-as-a-service (dtaas) platforms: A survey on architecture, design requirements, and performance metrics," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 1845–1878, 2026.

[9] Y. Cui, J. Liu, M. Yu, J. Jiang, L. Zhang, and L. Lu, "Network digital twin," *IEEE Network*, vol. 38, no. 1, pp. 5–6, 2024.

[10] Z. Wang, D. Jiang, and S. Mumtaz, "Network-wide data collection based on in-band network telemetry for digital twin networks," *IEEE Transactions on Mobile Computing*, vol. 24, no. 1, pp. 86–101, 2025.

[11] L. Hui, M. Wang, L. Zhang, L. Lu, and Y. Cui, "Digital twin for networking: A data-driven performance modeling perspective," *IEEE Network*, vol. 37, no. 3, pp. 202–209, 2023.

[12] D. González-Sánchez, L. Bellido, I. D. Martinez-Casanueva, D. Martínez-García, D. Fernández, and D. R. Lopez, "Towards building a digital twin for network operation and management," *IEEE Open Journal of the Communications Society*, pp. 1–1, 2025.

[13] R. Poorzare, D. N. Kanellopoulos, V. K. Sharma, P. Dalapati, and O. P. Waldhorst, "Network digital twin toward networking, telecommunications, and traffic engineering: A survey," *IEEE Access*, vol. 13, pp. 16 489–16 538, 2025.

[14] E. Ak, K. Duran, O. A. Dobre, T. Q. Duong, and B. Canberk, "T6conf: Digital twin networking framework for ipv6-enabled net-zero smart cities," *IEEE Communications Magazine*, vol. 61, no. 3, pp. 36–42, 2023.

[15] T. Bilen, E. Ak, B. Bal, and B. Canberk, "A proof of concept on digital twin-controlled wifi core network selection for in-flight connectivity," *IEEE Communications Standards Magazine*, vol. 6, no. 3, pp. 60–68, 2022.