International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 721-731
https://doi.org/10.70917/ijcisim-2026-2215
Article

Digital Twin Technology for Hospital Infrastructure: Integrating 5G Private Network Design, Coverage Analysis, and Lifecycle Automation in Large Healthcare Facilities

Mohit Upadhyay

Tech Mahindra (Americas) Inc., USA
ORCID: 0009-0007-8072-0380

Abstract: The convergence of Digital Twin (DT) technology and fifth-generation (5G) private network architecture presents a transformative paradigm for the design, deployment, and operational management of wireless network infrastructure in large healthcare facilities. Hospital campuses represent uniquely challenging deployment environments: electromagnetically hostile, geometrically complex, operationally continuous, and subject to stringent regulatory constraints governing both patient safety and data privacy. Traditional wireless network planning methods, manual site surveys, empirical path loss models, and iterative post-installation rework are systematically inadequate for these environments, producing persistent coverage failures and extended commissioning timelines that concentrate their most damaging consequences in the highest-acuity zones of the hospital. This article presents a unified architectural framework integrating multi-modal digital twin construction, machine learning-enhanced RF simulation, microservices-based deployment architecture, and operational sustainability strategies specifically designed for hospital-scale 5G private network deployment. The framework encompasses five interdependent technical components: Building Information Modeling (BIM) and LiDAR-based DT construction; ML-corrected RF accuracy optimization deploying Gaussian Process Regression, convolutional neural networks, reinforcement learning, Graph Neural Networks, and federated learning; a seven-service microservices platform with Kubernetes orchestration; caching and computation efficiency strategies reducing end-to-end simulation latency; and carbon-aware energy management aligned with sustainable cloud computing frameworks. Author's primary research projections indicate that DT-enabled planning reduces total cost of ownership by USD 3-6 million on a standard 600-bed hospital deployment and reduces commissioning timelines by 15-30% relative to traditional methods.
Keywords: Digital Twin, 5G Private Network, Hospital Infrastructure, RF Simulation, Machine Learning, Network Slicing, URLLC, IoT Healthcare, Lifecycle Automation, Microservices Architecture

1. Introduction: The Business Case for Digital Twin-Enabled 5G Network Design

1.1 The Healthcare Connectivity Challenge

A 600-bed academic medical center encompasses more than 50,000 square meters of built space, operates a device population growing toward tens of thousands of connected endpoints, and runs clinical workflows where wireless latency is not a quality-of-service metric but a patient safety variable [9, 10]. The wireless network serving these operations must sustain sub-millisecond latency for robotic surgery, multi-gigabit throughput for medical imaging, and five-nines reliability across a physical environment engineered, at every level, to resist the radio frequency propagation that the network requires.

Reinforced concrete slabs attenuate 5G signals to a degree that empirical planning tools routinely underestimate. Lead-lined radiology suites, purpose-built to protect patients from ionizing radiation scatter, create RF isolation so severe that DAS-independent coverage analysis is required. Operating theater equipment generates electromagnetic interference profiles that conflict directly with the frequencies allocated to private 5G operations. Manual site survey methodologies, developed for environments far less demanding than a major hospital, are not failing incrementally in this context; they are failing categorically, producing coverage maps that do not reflect the actual electromagnetic environment and commissioning timelines that consistently exceed contractual targets by weeks [9, 10].

1.2 Digital Twin Technology: Principles and Relevance to Hospital Infrastructure

At its core, a digital twin is a computational system that maintains real-time, bi-directional synchronization with a physical counterpart, enabling prediction, simulation, and optimization without physical trial and error [1, 2]. The concept traces its formal theoretical development through Grieves and Vickers' foundational work on emergent behavior mitigation in complex systems and has since been operationalized across infrastructure-intensive sectors from aerospace to smart urban development [3]. The National Institute of Standards and Technology's 2025 definition anchors the concept in a lifecycle perspective: a set of computational models that continuously represent the state and behavior of a unique physical asset from construction through decommission [7].

Three capabilities differentiate a hospital digital twin from a static building model with RF overlays. The twin serves as a geometric substrate, a simulation-ready three-dimensional representation of every structural and environmental element influencing radio wave propagation. It provides predictive simulation capacity, allowing antenna configurations, network element placements, and service area assignments to be evaluated and rejected computationally before any physical installation occurs. And it maintains a continuous feedback loop: a synchronization mechanism that keeps the virtual model accurate as the physical hospital evolves through renovation cycles, equipment additions, and occupancy changes [4].

1.3 The Quantified Business Case

Author's primary research projections place the digital twin's cost impact in the range of a USD 3-6 million reduction on a standard USD 18-25 million private 5G network deployment for a 600-bed hospital campus. This reduction is driven by three mechanisms operating in parallel: elimination of rework cycles generated by inaccurate survey-derived coverage predictions; reduction in materials over-provisioning caused by uncertainty about actual attenuation characteristics; and a 15-30% compression of end-to-end commissioning timelines through parallel simulation-based validation replacing sequential physical survey and adjustment cycles. Deployed hospital private 5G programs have validated the operational direction of these impact categories [9, 10].

Beyond the capital program itself, the digital twin generates compounding operational returns: a continuously current analytical platform for proactive network optimization before service degradation manifests clinically; structured documentation supporting the HIPAA Security Rule compliance audits that HIPAA-regulated environments require [8]; and capacity planning capability as the hospital's device population expands through new clinical technology acquisition. Table 1 compares traditional and DT-enabled planning approaches across the key performance dimensions established by the author's primary research framework.

Table 1. Comparative Performance Metrics: Traditional vs. Digital Twin Network Planning

Metric	Traditional Planning	Digital Twin-Enabled
Site survey duration (600-bed hospital)	8-14 weeks	2-4 weeks
Post-installation coverage gap rate	18-35% of zones	<5% of zones
Commissioning timeline reduction	Baseline	15-30% reduction
Cost overrun rate	25-40% of projects	<10% of projects
Year-1 post-deployment rework events	3-5 events	0-1 events

Note: All Digital Twin-Enabled figures represent the author's primary research projections.

1.4 Hospital Zone Classification and Network Element Mapping

Not all hospital square footage presents equivalent 5G design challenges. An operating theater running robotic surgery requires URLLC-class connectivity with sub-1 ms latency and five-nines reliability [5]. A radiology suite handling DICOM image transfer at clinical throughput imposes lead-lined walls that attenuate RF signal to levels most planning tools underestimate by four to eight decibels. Patient wards require reliable handover support for devices moving continuously with nursing staff across multiple floors. Underground car parks and mechanical sub-basements present attenuation levels so severe they constitute effectively separate propagation environments requiring independent DAS design.

The digital twin models each zone's requirements independently, simulates antenna placement against each zone's physical constraints and performance targets, and validates the complete multi-zone configuration before any hardware procurement occurs. This zone-resolved simulation approach is the primary mechanism through which DT-enabled planning eliminates the rework cycles that account for the largest share of cost overruns in traditional deployments. Table 2 summarizes the zone-to-slice mapping driving the simulation analysis.

Table 2. Hospital Zone Network Element and Slice Requirements

Zone Type	5G Slice	Latency Target	Primary Use Case	Key Design Challenge
Operating Theater	URLLC	<1 ms	Robotic surgery, monitoring	RF shielding from medical equipment
Radiology Suite	eMBB	<10 ms	Medical imaging transfer	Lead shielding attenuation
ICU/CCU	URLLC	<1 ms	Continuous patient monitoring	Dense device interference
Patient Wards	eMBB/mMTC	<20 ms	Mobile nursing, patient IoT	Multi-floor handover continuity
Public Lobbies	eMBB	<20 ms	Staff and visitor devices	High device density management
Underground or Sub-basement	DAS extension	<50 ms	Maintenance, emergency access	Severe RF attenuation

2. Business Case Expansion: Peer Industries and Multi-Modal Digital Twin Construction

2.1 Applicability Across Critical Infrastructure Sectors

The hospital campus presents, in concentrated form, a set of RF design challenges that recur across multiple critical infrastructure sectors. Advanced automotive manufacturing facilities operating Industry 4.0 processes deploy robotic assembly and automated guided vehicle systems requiring URLLC-class connectivity in reinforced concrete structures with noise floor characteristics comparable to surgical environments. Smart airport terminals, vast atrium-style spaces with irregular geometry and operational systems ranging from boarding infrastructure to ground handling coordination, demand five-nines reliability at passenger-scale device densities. Power generation facilities and offshore energy platforms require private 5G for SCADA connectivity and autonomous inspection operations in electromagnetic environments that defeat empirical planning methods entirely.

Research university campuses, multi-building complexes with underground pedestrian tunnel networks, laboratory environments generating continuous RF interference, and device populations driven by research instrumentation rather than standard enterprise endpoints present design challenges matching hospital campus complexity across nearly every technical dimension. The common thread is the gap between what traditional planning delivers and what mission-critical operations require. The hospital digital twin framework addresses this gap with an architectural pattern that transfers across sectors through parametric reconfiguration rather than ground-up redesign, extending the capital and engineering investment in the hospital program to adjacent infrastructure contexts.

2.2 Multi-Modal Input Methods for Digital Twin Construction

The accuracy of RF predictions generated by a hospital digital twin is bounded above by the accuracy of the building geometry model from which those predictions are derived. A twin constructed from inaccurate structural data will produce inaccurate propagation simulations regardless of the sophistication of the simulation engine applied to it. Multi-modal data acquisition addresses this constraint by selecting the highest-accuracy input method available for each zone and building vintage within the hospital estate [20].

BIM models in IFC format provide the highest-fidelity starting point for facilities where this documentation exists: complete three-dimensional structural geometry, material specifications, and spatial organization in a format directly consumable by simulation pipelines. For the substantial portion of the existing hospital estate designed before BIM adoption, terrestrial LiDAR scanning produces as-built geometry at millimeter accuracy from point cloud data processed through automated reconstruction pipelines, capturing installed structural reality rather than design intent, which in older facilities diverges in ways that matter for RF propagation analysis [20]. Drone photogrammetry captures external campus areas and rooftops. The Building Management System feed injects real-time operational state, including HVAC modes and equipment activation patterns, that modify the ambient electromagnetic environment in ways a static geometry model cannot represent. Material property libraries provide the electromagnetic characterization data without which ray-tracing simulation accuracy degrades to levels insufficient for clinical-zone design validation. Table 3 details each modality's contribution to twin construction and RF simulation accuracy.

Table 3. Digital Twin Input Modalities and Impact on RF Simulation Accuracy

Input Modality	Data Type	Accuracy Contribution	Applicable When
BIM/IFC models	3D geometry, materials	High (full structural geometry)	New construction or BIM-documented facilities
Terrestrial LiDAR scanning	Point cloud, mm accuracy	High (as-built interior geometry)	Existing buildings without BIM documentation
Drone photogrammetry	Georeferenced 3D model	Medium (external areas)	Campus exterior, rooftops, parking structures
BMS/IoT sensor feeds	Real-time occupancy, equipment state	Medium (dynamic interference modeling)	Post-commissioning feedback and optimization
Material property libraries	Electromagnetic parameters	Critical for ray-tracing accuracy	All simulation scenarios

3. Machine Learning Enhancement of Digital Twin RF Modeling Accuracy

3.1 The Accuracy Limitation of Deterministic RF Simulation

Deterministic ray-tracing propagation models, applied to hospital environments, carry systematic error sources of eight to twelve decibels that cannot be resolved by refinement of the geometric model or extension of the reflection chain depth [12, 13]. Material property uncertainties, which are the difference between the electromagnetic parameters assigned to a concrete slab in the simulation library and the actual electromagnetic behavior of that specific slab, depend on aggregate composition, rebar density, and hydration state, and introduce irreducible prediction variance. Medical equipment and furnishings generate scattering contributions that are impractical to model at the individualobject level across a 50,000 square meter facility. The computational cost of modeling full multi-path environments to the reflection depth required for clinical-zone accuracy exceeds practical simulation budgets.

Machine learning addresses these limitations not by eliminating their physical causes but by learning their statistical consequences from measurement data and applying empirical corrections to simulation outputs. The result is a hybrid prediction pipeline: deterministic ray tracing provides the physics-grounded spatial estimate; ML models correct the systematic biases; and the combined output achieves the three-to-five decibel residual error target required for hospital deployment planning [12]. Five distinct ML techniques, each addressing a different aspect of the accuracy optimization problem, form the complete framework.

3.2 ML Techniques and Deployment Architecture

3.2.1 Gaussian Process Regression for Coverage Prediction Correction

GPR treats RF signal strength at any spatial location as a random variable with a structured covariance function learned from calibration measurements. The trained model produces not only a corrected point estimate but an explicit uncertainty quantification, a property with direct operational value in hospital environments where uncertainty about coverage performance in a surgical theater carries different risk implications than equivalent uncertainty about a public lobby. When applied to path loss prediction in mobile communication bands, ML-augmented approaches that include GPR corrections have shown consistent improvements over purely deterministic baselines, achieving accuracy levels that are sufficient for clinical-zone design validation [19]. In this framework, GPR serves as the post-ray-tracing correction layer, with a residual error target of three to five decibels against an eight-to-twelve-decibel deterministic baseline (author's primary research projection).

3.2.2 Convolutional Neural Networks for Floor Plan Feature Extraction

For hospitals without three-dimensional BIM documentation, CNN-based floor plan analysis provides a computationally efficient path to RF-relevant building characterization. CNNs trained on corpora of building images paired with RF measurement campaigns learn which architectural features, doorway positions, corridor geometries, partition configurations, and stairwell openings are most predictive of propagation behavior and extract them from two-dimensional floor plan images [12, 13]. This approach makes preliminary network design computationally tractable in buildings for which full three-dimensional geometry construction would otherwise require extended LiDAR scanning and post-processing campaigns, enabling earlier design iteration and faster program schedule execution.

3.2.3 Deep Reinforcement Learning for Network Element Placement Optimization

The antenna placement problem in a hospital, selecting a K-element configuration from N candidate mounting locations to satisfy simultaneously the URLLC requirements of operating theaters, the eMBB requirements of radiology, and the coverage continuity requirements of patient wards, is a combinatorial optimization problem of significant practical difficulty for any facility at hospital campus scale. Deep reinforcement learning methods address this problem by training an agent to interact with the digital twin simulation environment, receiving coverage quality metrics as reward signals and converging on near-optimal configurations across the joint constraint space [15, 18]. The DQN framework is particularly well-suited to this application because the digital twin simulation constitutes a high-fidelity environment model against which the agent can train without any physical experimentation.

3.2.4 Graph Neural Networks for Inter-Cell Interference Prediction

Inter-cell interference in a dense multi-floor hospital deployment is modeled most naturally as a graph structure: nodes represent network elements, edges represent interference coupling relationships, and the prediction task is inferring per-node SINR degradation across the deployment graph. GNNs trained on this representation scale efficiently to the deployment densities that hospital environments require, dozens to hundreds of small cells across multiple floors, and their permutation-invariant architecture generalizes across floor configurations without retraining [22]. This generalization property becomes operationally significant during the frequent renovation cycles that characterize active hospital facilities, where floor configurations change on timescales requiring rapid model recalibration.

3.2.5 Federated Learning for Cross-Campus Model Improvement

Post-deployment, federated learning across hospital campus deployments continuously improves the digital twin's RF prediction accuracy. Each site trains local model updates on its live measurement data and transmits only the parameter updates, not the underlying building schematics, network configuration data, or patient-area location records, to a federation aggregator [16, 17]. The FedAvg algorithm combines per-site updates into an improved global model redistributed across all participants. Over time, this process produces a shared accuracy capability that no single hospital campus could develop from its measurement data alone, while preserving the data sovereignty requirements that healthcare regulatory environments impose. Table 4 summarizes the complete ML technique deployment architecture.

Table 4. ML Techniques for Digital Twin RF Accuracy Optimization [12, 13, 15, 16, 17, 18, 19, 22]

ML Technique	Application in DT Pipeline	Primary Benefit
Gaussian Process Regression	Coverage prediction correction post-ray tracing	Uncertainty quantification; 3-5 dB error target
Convolutional Neural Networks	Floor plan feature extraction for path loss modeling	Computationally efficient 2D-to-RF feature mapping
Deep Reinforcement Learning	Network element placement optimization	Near-optimal multi-zone placement without exhaustive search
Graph Neural Networks	Inter-cell interference prediction in dense deployments	Scalable to multi-floor deployment densities
Federated Learning	Continuous post-deployment model improvement across campuses	Privacy-preserving cross-site accuracy gains

4. Architectural Design and Operational Sustainability

4.1 Seven-Service Microservices Platform Architecture

Seven microservices, each bounded to a single functional context, comprise the digital twin platform. The BIM Ingestion Service normalizes IFC, DWG, and point cloud inputs and stores validated geometry in a spatial database with format-agnostic APIs. The RF Simulation Engine executes deterministic ray tracing as a stateless REST API, enabling horizontal scaling under load without session state management overhead. The ML Inference Service hosts trained GPR, CNN, and GNN models, returning corrected coverage predictions with uncertainty bounds to downstream design tools. The Equipment Catalog Service maintains the RF parameter library, effective isotropic radiated power values, antenna radiation patterns, frequency band configurations, for the complete hardware inventory under management. The Workflow Orchestration Service manages the end-to-end design-to-construction lifecycle, tracking simulation jobs, approval gates, and construction milestones against the program schedule. The Telemetry Service ingests live KPI streams from deployed radio units via the O-RAN O1 interface [6]. The 3D Visualization Service streams WebGL renderings with coverage overlays to engineering workstations, supporting distributed design review without file transfer.

This decomposition matches resource allocation to actual computational demand patterns: the RF simulation service and ML inference service scale independently under concentrated simulation load, while the lightweight catalog and visualization services maintain responsiveness at the request rates their consumer patterns generate. Bounded service contexts also enable independent development, deployment, and failure isolation, reducing the blast radius of service-level incidents during active program execution.

4.2 Kubernetes Orchestration and Security Architecture

Kubernetes manages service orchestration deployed on private cloud infrastructure within the hospital network perimeter to satisfy data sovereignty requirements of the HIPAA Security Rule [8]. Horizontal Pod Autoscaling governs the RF simulation service and ML inference service; long-duration ray-tracing jobs are submitted as batch jobs to a GPU node pool with priority queuing that preserves interactive simulation responsiveness during peak design review periods. Mutual TLS secures all inter-service communication. The O-RAN O1 interface integration connects the Telemetry Service to deployed 5G radio units [6].

The security architecture accounts for the threat landscape that digital twin deployments introduce at the intersection of network management and clinical operations. Model poisoning attacks, introducing adversarially modified measurement data to corrupt coverage prediction accuracy, require data provenance verification at the Telemetry Service ingestion boundary. Adversarial network state manipulation requires integrity validation of the O-RAN management plane communication. 5G control plane exploitation, including evil twin and SIM-based attacks relevant to the 5G/6G transition period, requires continuous monitoring architecture informed by current threat characterizations [11, 14].

4.3 Image Processing Subsystems

Three image processing subsystems operate within the platform's data ingestion and visualization layer. The floor plan vectorization pipeline uses OpenCV-based preprocessing, deskewing, binarization, and noise removal, followed by a custom CNN performing architectural element segmentation, enabling simulation-ready geometry extraction from scanned PDF floor plans at the quality levels typical of older hospital documentation. The coverage heat map generation subsystem uses GDAL-based raster processing to reproject simulation outputs into Web Mercator overlays served from CDN edge caches to distributed engineering teams, eliminating the coordination overhead of static file-based design review. Computer vision as-built verification processes contractor photographs during construction phases, detecting network element installation compliance against the digital twin's predicted configuration and flagging deviations before they are covered by subsequent construction work.

4.4 Continuous Physical-Virtual Synchronization Mechanism

Three data streams drive the digital twin's continuous synchronization with the physical hospital. Deployed radio units stream per-cell KPIs via the O1 interface; the Telemetry Service compares these against DT-predicted values and flags deviations above tolerance for engineering review. Network-connected devices, staff tablets, mobile nursing workstations, and connected medical equipment generate passive measurement observations supplementing the infrastructure KPI stream with user-plane data distributed across facility zones. Renovation events initiate structured BIM update workflows: new LiDAR scans of the affected areas are ingested, the relevant floors are re-simulated, and coverage gaps introduced by the physical change are identified before they manifest as clinical service complaints [3, 4]. The synchronization mechanism is what distinguishes a live digital twin from a static planning model; it is the mechanism through which the computational investment in the initial simulation maintains its accuracy and operational value over the five-to-ten-year lifecycle of the deployment.

5. Infrastructure Optimization for Scalable Simulation Performance

5.1 Containerization and Compute Resource Architecture

GPU node pools host the RF simulation and computationally intensive ML training containers; memory-mapped Bounding Volume Hierarchy (BVH) geometry data structures enable concurrent simulation requests to share scene data without duplication, reducing memory footprint proportionally to concurrent request concurrency. Inference containers run on CPU-optimized nodes; ONNX Runtime and TensorRT INT8 quantization deliver a fourfold reduction in model size and a two-to-threefold improvement in inference throughput with less than one percent accuracy degradation relative to full-precision baselines (author's primary research projection). WebAssembly compilation of the visualization service's geometry processing logic offloads mesh simplification to client browsers, reducing server-side rendering cost for interactive multi-user design sessions at a level proportional to session concurrency.

5.2 Intelligent Workload Management and Scheduling

Predictive autoscaling acts on historical simulation request patterns; peaks coincide with business hours and pre-construction design review milestones to provision compute capacity ahead of demand growth rather than reacting to queue depth accumulation. This distinction matters operationally: queue-reactive autoscaling introduces latency proportional to node provisioning time, which for GPU instances is measured in minutes; pattern-predictive provisioning eliminates that latency by maintaining availability in advance of anticipated load. Spot instance scheduling assigns nightly drive test correlation runs, weekly accuracy audit jobs, and monthly full-campus re-simulation to lower-cost, preemptible compute with checkpoint-based restart tolerance for individual job segments. Data locality optimization co-locates simulation jobs with nodes holding the relevant BIM geometry cache, eliminating per-job network transfer overhead for scene files ranging from 500 megabytes to five gigabytes per hospital floor.

6. Performance Optimization and Carbon-Aware Sustainability

6.1 Multi-Tier Caching Architecture

Multi-tier caching is the primary mechanism converting the platform's inherent computational intensity from a user-experience liability into a throughput advantage. Redis hosts BVH geometry representations with a 24-hour time-to-live: initial construction from IFC source geometry requires 40-90 seconds; cache retrieval completes in three to five milliseconds (author's primary research projection). This four-order-of-magnitude latency reduction is the enabling condition for interactive simulation response times in design review sessions where engineers iteratively adjust antenna placements and immediately re-query coverage predictions for the modified configuration.

Simulation results are cached by configuration hash, eliminating redundant ray-tracing computation across the minor placement iterations that dominate the design refinement workflow, iterations where the dominant cost is traversal of the same geometry rather than exploration of meaningfully different configurations. GPR prediction outputs are held in Memcached with a one-hour TTL, serving the common pattern of concurrent engineers querying the same building area during review sessions. Coverage heat map tiles are CDN-delivered with standard cache-control headers, enabling geographically dispersed design teams to interact with the same visualization at full responsiveness without server-side tile regeneration.

6.2 Adaptive Ray-Tracing Depth and GPU Acceleration

Adaptive ray-tracing depth allocation matches simulation fidelity to workflow stage requirements. Two-bounce modeling at approximately 47 seconds per floor serves initial exploration and rough placement decisions; six-bounce processing at eight to twelve minutes per floor serves design validation review; eight-bounce modeling at 30-50 minutes per floor serves final engineering sign-off documentation (author's primary research projections). The multi-bounce computational cost scales nonlinearly with depth; allocating full-depth simulation only to the final validation pass maintains interactive design responsiveness without sacrificing the accuracy required at sign-off.

GPU-accelerated Monte Carlo sampling handles geometrically irregular zones, radiology suites with non-rectangular lead-lined geometries, and multi-level atria with complex diffraction environments, with a 10 to 40-fold speed advantage over deterministic methods for these specific geometry classes (author's primary research projection). TensorRT model quantization reduces the inference service's GPU memory footprint by a factor of four, enabling higher request concurrency per node at the accuracy level required for clinical-zone design validation.

6.3 Carbon-Aware Energy Management

Kubernetes' Vertical Pod Autoscaler continuously right-sizes pod memory and CPU allocations based on observed resource consumption patterns, improving cluster utilization and reducing idle power draw across the platform. Carbon-aware workload scheduling routes non-time-sensitive batch jobs, overnight simulation audits, monthly accuracy recalibration runs, and federated learning aggregation cycles to execution windows with lower grid carbon intensity, consistent with the energy efficiency and sustainability frameworks emerging as operational standards in cloud infrastructure management [21]. For small-area simulation tasks covering individual rooms or minor clinical areas, CPU-based ray tracing is more energy-efficient than GPU offload due to GPU initialization and context overhead; the workload manager applies a geometry complexity threshold to route jobs to the appropriate compute class. Dynamic voltage and frequency scaling applied to inference workloads during off-peak periods reduces GPU power consumption by 20-40% while maintaining acceptable throughput for non-interactive batch prediction jobs (author's primary research projection).

7. Conclusion and Forward Trajectory

7.1 Scalability Analysis

The framework's scalability operates across three independent axes. Horizontal scaling allows individual microservices to scale out under increased simulation and inference load without platform-wide redeployment, matching resource consumption to actual demand patterns rather than provisioning for peak capacity across the entire platform. Vertical scaling through GPU node pool upgrades benefits the simulation engine and ML inference service directly, with performance improvements translating immediately to reduced simulation latency and increased concurrent session capacity. Multi-campus federated architectures allow health systems to maintain per-campus data sovereignty, keeping building geometry, network configuration, and patient-area telemetry behind individual facility perimeters, while contributing to shared federated learning model improvement pools that enable accuracy gains exceeding what any single campus could achieve from its own measurement data alone [16, 17].

7.2 Emerging Technology Integration

Sub-terahertz 6G frequencies, projected for commercial standardization by 2030, will amplify the hospital environment's planning complexity by an order of magnitude: their extreme sensitivity to millimeter-scale surface roughness and geometric irregularity makes digital twin-based planning not a competitive advantage but an engineering necessity for any program targeting clinical-grade coverage. 3GPP Release 18's native AI integration [5] will shift the digital twin's ML layer from design-time analysis to runtime network intelligence, enabling continuous, automated adjustment of antenna beam patterns and power allocations in response to live clinical occupancy and device distribution changes, without requiring manual engineering intervention.

Emerging interoperability frameworks, the Digital Twin Consortium's modeling language specification and the ITU-T Focus Group on Digital Twins specification will enable the hospital network twin to connect with building management, medical device, and facility operations twins within a unified facility intelligence layer. This integration trajectory positions the 5G network digital twin as one component in a broader hospital facility awareness system, where insights from the network model inform building management decisions and building state changes are reflected immediately in network simulation, a closed-loop intelligence infrastructure for clinical operations.

7.3 A Clinical Infrastructure Commitment

A hospital's private 5G network is a clinical utility. It underpins robotic surgery, continuous patient monitoring, medication administration automation, and radiological imaging systems that have no analog fallback in facilities designed around wireless-native clinical workflows. Coverage failures in these environments are not service interruptions in the commercial sense; they are clinical risk events with direct patient safety implications.

The question facing healthcare infrastructure engineering teams is not whether to deploy private 5G but whether to deploy it with an engineering foundation equal to the clinical consequences of failure. Digital twin-enabled planning, integrated with the ML accuracy optimization pipeline and microservices architecture described in this framework, provides that foundation. It converts the probabilistic uncertainty of empirical network planning, surveys that cannot fully characterize electromagnetically complex environments before they are built, into the engineering determinism that hospital operations require. In doing so, it shifts 5G private network deployment in healthcare from a technical infrastructure project into a defensible clinical infrastructure commitment: one that can be validated computationally before the first antenna is mounted and continuously monitored against its design specification throughout its operational life.

REFERENCES

1. Michael W. Grieves, "Completing the cycle: Using PLM information in the sales and service functions," in SME Management Forum, vol. 19, p. 3, SME Forum, 2002. Available: https://www.researchgate.net/publication/356192963
2. Michael Grieves and John Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in Transdisciplinary Perspectives on Complex Systems, Springer, 2016, pp. 85-113. Available: https://link.springer.com/chapter/10.1007/978-3-319-38756-7_4
3. Adil Rasheed, Omer San, and Trond Kvamsdal, "Digital twin: Values, challenges and enablers from a modeling perspective," IEEE Access, vol. 8, pp. 21980-22012, 2020. Available: https://ieeexplore.ieee.org/abstract/document/8972429
4. Seppo Horsmanheimo et al., "5G goes underground: A proof-of-concept using digital twin for real-time control and monitoring," in Proc. EuCNC/6G Summit, pp. 961-966, IEEE, 2024. Available: https://ieeexplore.ieee.org/abstract/document/10597103
5. 3rd Generation Partnership Project, Technical Specification 23.501, 2017. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144
6. O-RAN Alliance, "O-RAN Architecture Overview," 2026. Available: https://docs.o-ran-sc.org/en/latest/architecture/architecture.html
7. National Institute of Standards and Technology, "NIST IR 8356: Security and Trust Considerations for Digital Twin Technology," 2025. Available: https://csrc.nist.gov/pubs/ir/8356/final
8. U.S. Department of Health and Human Services, "Security Rule Guidance Material," 2026. Available: https://www.hhs.gov/hipaa/for-professionals/security/guidance/index.html
9. Andrea Moglia et al., "5G in healthcare: From COVID-19 to future challenges," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8, pp. 4187-4196, Aug. 2022. Available: https://ieeexplore.ieee.org/abstract/document/9792172
10. Baozhan Chen et al., "Construction and application of a private 5G standalone medical network in a smart health environment," Journal of Medical Internet Research, vol. 26, p. e52404, 2024. Available: https://www.jmir.org/2024/1/e52404/
11. Cristina Alcaraz and Javier Lopez, "Digital twin: A comprehensive survey of security threats," IEEE Communications Surveys & Tutorials, vol. 24, no. 3, pp. 1475-1503, 2022. Available: https://ieeexplore.ieee.org/abstract/document/9765576
12. Shengjie Ma, Hong Cheng, and Hyukjoon Lee, "A practical approach to indoor path loss modeling based on deep learning," Journal of Computing Science and Engineering, vol. 15, no. 2, pp. 84-95, 2021. Available: https://jcse.kiise.org/files/V15N2-04.pdf
13. Aristeidis Seretis and Costas D. Sarris, "Toward physics-based generalizable convolutional neural network models for indoor propagation," IEEE Transactions on Antennas and Propagation, vol. 70, no. 6, pp. 4112-4126, Jun. 2022. Available: https://ieeexplore.ieee.org/abstract/document/9670666
14. Aris Cahyadi Risdianto et al., "Validating evil twin attacks in 5G/6G networks based on MITRE FiGHT," in Proc. IEEE INFOCOM 2025 Workshops, pp. 1-6, 2025. Available: https://ieeexplore.ieee.org/abstract/document/11153022
15. Nguyen Cong Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3133-3174, 2019. Available: https://ieeexplore.ieee.org/abstract/document/8714026
16. Brendan McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in Proc. AISTATS, pp. 1273-1282, PMLR, 2017. Available: https://proceedings.mlr.press/v54/mcmahan17a
17. Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, May 2020. Available: https://ieeexplore.ieee.org/abstract/document/9084352
18. Volodymyr Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529-533, 2015. Available: https://www.nature.com/articles/nature14236
19. Jakob Thrane, Darko Zibar, and Henrik Lehrmann Christiansen, "Model-aided deep learning method for path loss prediction in mobile communication systems at 2.6 GHz," IEEE Access, vol. 8, pp. 7925-7936, 2020. Available: https://ieeexplore.ieee.org/abstract/document/8950164
20. Jin Huang, Jantien Stoter, Ravi Peters, and Liangliang Nan, "City3D: Large-scale building reconstruction from airborne LiDAR point clouds," Remote Sensing, vol. 14, no. 9, p. 2254, 2022. Available: https://www.mdpi.com/2072-4292/14/9/2254
21. Rajkumar Buyya, Shashikant Ilager, and Patricia Arroba, "Energy-efficiency and sustainability in new generation cloud computing," Software: Practice and Experience, vol. 54, no. 1, pp. 24-38, 2024. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/spe.3248

22. Yifei Shen, Yuanming Shi, Jun Zhang, and Khaled B. Letaief, "Graph neural networks for scalable radio resource management," IEEE Journal on Selected Areas in Communications, vol. 39, no. 1, pp. 101-115, Jan. 2021. Available: https://ieeexplore.ieee.org/abstract/document/9252917