SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

AI-Driven Network Slicing Optimization for Low-Latency and Energy-Efficient 5G/6G Services

S. Srividhya[1], Dr. S. Selvakumar[2], Dinesh Babu K[3], R. Thirumurugan[4]

[1] Research Scholar, Department of Electronics and Communication Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Enathur – 631561, India. E-mail: gshanmugam76@gmail.com
[2] Assistant Professor, Department of Electronics and Communication Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Enathur – 631561, India. E-mail: Selvakumar@kanchiuniv.ac.in
[3] Assistant Professor, Department of Electronics and Communication Engineering, Adhi College of Engineering and Technology, Kancheepuram – 631605, India. E-mail: dinesh.comm@gmail.com
[4] Assistant Professor, Department of Electronics and Communication Engineering, Adhi College of Engineering and Technology, Kancheepuram – 631605, India. E-mail: thirumurugano7@gmail.com

Abstract
The rapid proliferation of ultra-reliable low-latency communication, massive Internet of Things, extended reality, autonomous systems, and intelligent industrial applications is placing increasingly stringent demands on 5G and emerging 6G networks. Network slicing provides a flexible mechanism for creating isolated logical networks with application-specific quality-of-service requirements; however, conventional slicing strategies often struggle to simultaneously minimize latency, energy consumption, resource wastage, and service-level violations under highly dynamic traffic conditions. This paper presents an AI-driven network slicing optimization framework that integrates machine learning-based traffic prediction, intelligent slice admission, dynamic resource allocation, and multi-objective optimization. The proposed approach continuously analyzes traffic demand, latency requirements, computational load, and energy conditions to determine appropriate slice configurations and resource assignments. AI-based decision mechanisms are employed to predict congestion and proactively reconfigure slices, while energy-aware optimization reduces unnecessary resource activation without compromising latency and reliability requirements. The framework is designed for heterogeneous 5G/6G environments incorporating cloud, edge, and radio resources. Performance is evaluated using latency, energy consumption, resource utilization, throughput, slice acceptance rate, and SLA violation rate as principal metrics. The proposed architecture establishes an intelligent and adaptive foundation for autonomous network slicing, supporting efficient operation of latency-sensitive and energy-constrained future wireless services.

Keywords: 5G, 6G, Artificial Intelligence, Network Slicing, Resource Allocation, Low Latency, Energy Efficiency, Machine Learning, Edge Computing, Multi-Objective Optimization, Quality of Service, Network Automation

1. Introduction

The evolution of fifth-generation (5G) wireless networks has transformed mobile communication from a connectivity-oriented infrastructure into a programmable platform capable of supporting highly heterogeneous services. Applications such as enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), massive machine-type communication (mMTC), autonomous transportation, industrial automation, immersive extended reality, and intelligent healthcare impose substantially different requirements in terms of bandwidth, latency, reliability, computing capacity, and energy consumption. A conventional network architecture that provides largely uniform resource treatment cannot efficiently accommodate such heterogeneous service requirements. Network slicing has therefore emerged as a fundamental 5G technology for creating multiple logical networks over a common physical infrastructure, with each slice configured according to the requirements of a particular service or application [1], [2].

Network slicing enables communication, computing, storage, and network functions to be dynamically partitioned among different tenants while maintaining logical isolation and service-level requirements. Its effectiveness, however, depends strongly on resource-management mechanisms capable of allocating physical and virtual resources according to rapidly changing traffic conditions. Resource allocation in sliced networks involves multiple dimensions, including radio resources, spectrum, bandwidth, computing capacity, transport resources, and virtual network functions. Conventional optimization techniques can provide effective solutions under relatively stable conditions, but their computational complexity and dependence on predefined system models make them less suitable for highly dynamic environments involving heterogeneous traffic and rapidly changing service requirements [3], [4].

The emergence of 6G is expected to intensify these challenges. Future networks are envisioned to provide extremely high data rates, near-zero communication latency, massive device density, high reliability, intelligent edge services, and integrated communication and computing capabilities. Consequently, network slicing in 6G is expected to become increasingly autonomous, context-aware, and adaptive. The integration of artificial intelligence (AI) and machine learning (ML) provides an important pathway toward this objective because learning-based models can identify traffic patterns, predict resource demand, detect congestion, and adapt network configurations without relying exclusively on static rules [5], [6].

Low latency is one of the most critical optimization objectives for emerging wireless applications. Autonomous vehicles, industrial control, remote robotic operations, tactile Internet applications, and immersive communication can be highly sensitive to even small increases in end-to-end delay. However, aggressively reserving resources to guarantee low latency can lead to inefficient resource utilization and unnecessary energy consumption. Conversely, excessive resource consolidation may reduce energy usage but increase queuing delay, congestion, and service-level agreement (SLA) violations. Therefore, latency optimization and energy efficiency should not be treated as independent objectives. Instead, network slicing requires a multi-objective decision process capable of balancing latency, throughput, reliability, resource utilization, and energy consumption [7], [8].

AI-based resource management provides an opportunity to address this multi-dimensional optimization problem through predictive and adaptive decision-making. Machine-learning models can forecast traffic demand before congestion occurs, while reinforcement-learning approaches can learn resource-allocation policies from the interaction between network states and performance outcomes. Recent research has demonstrated the applicability of ML to radio access network (RAN) slicing, particularly for traffic classification, resource allocation, admission control, and scheduling [9]. Deep-learning and reinforcement-learning techniques can further support dynamic network adaptation by learning complex relationships between network conditions and service requirements [10], [11].

Energy efficiency is particularly important because the increasing deployment of dense radio access networks, edge-computing infrastructure, virtualized network functions, and high-performance computing platforms can significantly increase the energy footprint of future communication systems. Energy-aware network slicing therefore requires coordinated management of communication and computing resources rather than optimization of individual network components in isolation. Recent studies have emphasized infrastructure-level, routing-level, and slice-operation-level approaches for reducing energy consumption, while also identifying AI-driven cross-layer optimization as an important future direction [12].

Despite substantial progress, existing approaches frequently optimize individual objectives or focus on a particular network domain. Some methods emphasize latency reduction without explicitly considering energy consumption, whereas energy-oriented approaches may sacrifice responsiveness under highly variable workloads. Similarly, traffic prediction, slice admission, resource allocation, and reconfiguration are often treated as separate optimization problems. Such fragmentation limits the ability of the network to make globally consistent decisions across the complete slice lifecycle. Recent surveys also identify scalability, real-time adaptability, SLA assurance, explainability, security, and interoperability as continuing challenges for AI-enabled network slicing in 5G and 6G environments [5], [6], [13].

This study therefore focuses on an AI-driven network slicing optimization framework designed to jointly address low latency and energy efficiency. The proposed approach considers network-state information, traffic demand, service requirements, available communication and computing resources, and energy conditions to support intelligent slice management. AI-based traffic prediction is employed to anticipate demand variations, while adaptive resource allocation and slice reconfiguration are used to maintain service quality under changing network conditions. The framework is intended to provide a more coordinated optimization strategy in which latency-sensitive services receive appropriate resource priority while unnecessary resource activation is minimized.

The principal objective is to establish an intelligent and adaptive network-slicing mechanism capable of balancing competing performance objectives in heterogeneous 5G/6G environments. The study evaluates the framework using latency, energy consumption, resource utilization, throughput, slice acceptance rate, and SLA violation rate. By integrating predictive intelligence with multi-objective resource optimization, the proposed approach seeks to move network slicing from reactive resource management toward proactive and autonomous network operation.

2. Literature Review

2.1 Network Slicing in 5G and Beyond

Network slicing has been extensively investigated as a mechanism for supporting heterogeneous services over shared network infrastructure. Foukas et al. presented network slicing as an important architectural concept for 5G, emphasizing its ability to provide customized logical networks for different applications and vertical industries [1]. Zhang et al. subsequently provided a comprehensive analysis of 5G network slicing, discussing architectural principles, use cases, management challenges, and research directions [2].

These studies established the foundation for treating network slicing as an essential component of programmable and service-oriented mobile networks.

Resource allocation is a central issue in practical slice deployment because multiple slices compete for common physical and virtual resources. Su et al. systematically examined resource-allocation principles and mathematical models for 5G network slicing, covering RAN and core-network resources and highlighting the importance of load balancing, resource utilization, QoS, and SLA satisfaction [3]. Similarly, Ksentini et al. emphasized the need for flexible resource provisioning and orchestration mechanisms capable of adapting to diverse service requirements [4]. Although these approaches provide important optimization foundations, the increasing dynamics of 5G/6G traffic make purely model-driven resource management increasingly difficult.

2.2 Machine Learning for Network Slice Management

The application of ML has substantially expanded the possibilities for autonomous slice management. ML can process historical and real-time network observations to identify traffic patterns and determine resource requirements without requiring complete analytical models of the underlying network. Azimi et al. reviewed ML applications in RAN slicing and demonstrated its relevance to resource allocation, traffic prediction, scheduling, admission control, and performance management [9]. Their analysis indicates that ML can improve the adaptability of resource-management mechanisms in environments characterized by uncertain and rapidly varying workloads.

A broader survey by Mao et al. classified ML applications in network slicing according to resource-management functions and learning paradigms, showing that supervised learning, unsupervised learning, deep learning, and reinforcement learning have all been investigated for slice optimization [6]. Reinforcement learning is particularly attractive because it can formulate resource allocation as sequential decision-making, allowing an agent to learn policies that balance competing network objectives. However, the training overhead, convergence behavior, state-space dimensionality, and requirement for reliable reward functions remain important challenges.

2.3 AI-Based Resource Allocation and Traffic Prediction

Dynamic traffic demand represents one of the major sources of inefficiency in network slicing. Static resource allocation may result in over-provisioning during low-demand periods and congestion during traffic peaks. AI-based traffic prediction can mitigate this problem by forecasting future demand and allowing resources to be provisioned proactively. Recent research on AI-based resource management has identified traffic classification, admission control, resource allocation, and scheduling as interconnected stages of intelligent slice management [13].

Deep-learning approaches can capture nonlinear temporal relationships in network traffic and can therefore support more accurate prediction than simple rule-based approaches under complex traffic patterns. Reinforcement-learning approaches complement prediction by learning adaptive resource-allocation policies. More recent work has also explored graph-based learning, federated learning, and other distributed AI techniques for 6G slicing, reflecting the increasing scale and heterogeneity of future networks [14]. Nevertheless, AI models must operate under strict latency constraints, and excessive model complexity can itself introduce computational and energy overhead.

2.4 Low-Latency Optimization

Latency-sensitive services require resource-management policies that can react rapidly to network-state changes. Conventional allocation mechanisms may satisfy average performance requirements but fail to guarantee latency during traffic bursts or resource contention. Multi-access edge computing (MEC), virtualization, and intelligent RAN management have consequently become closely associated with low-latency network slicing.

AI-assisted resource allocation can improve latency performance by predicting congestion and dynamically modifying slice resources before service degradation occurs. In particular, learning-based approaches can jointly consider queue states, traffic demand, radio conditions, computing availability, and application requirements when determining resource assignments. However, prioritizing latency alone may lead to resource over-provisioning and increased energy consumption. This creates a fundamental optimization trade-off between responsiveness and efficiency that must be explicitly incorporated into the decision-making process.

2.5 Energy-Efficient Network Slicing

Energy efficiency has traditionally received less attention than throughput and latency in network-slicing research, despite the substantial energy requirements associated with dense 5G deployments and future 6G infrastructure. Energy-efficient slicing can involve dynamic activation and deactivation of network

resources, efficient routing, workload consolidation, virtual-machine or container placement, and adaptive allocation of radio and computing resources.

Recent research has provided a structured taxonomy of energy-efficiency techniques across infrastructure, path/routing, and slice-operation levels [12]. This classification demonstrates that energy optimization cannot be restricted to individual base stations or computing nodes; instead, energy consumption should be considered throughout the complete slice lifecycle. AI provides an additional mechanism for identifying when resources can be consolidated or released while maintaining QoS requirements. However, energy minimization can conflict with latency and reliability objectives, making multi-objective optimization essential.

2.6 AI-Driven 6G Network Slicing and Research Gap

The transition toward 6G introduces additional requirements for intelligent, autonomous, and highly scalable resource management. AI and ML are increasingly considered fundamental enablers for 6G network slicing because future networks are expected to manage highly heterogeneous services, distributed computing resources, massive device populations, and dynamic mobility patterns [14], [15]. Recent studies have investigated deep reinforcement learning, graph neural networks, federated learning, and other AI techniques for improving adaptability and resource efficiency in intelligent slicing environments.

Despite this progress, several research gaps remain. First, many existing studies optimize latency, throughput, or energy efficiency individually rather than treating them as coupled objectives. Second, traffic prediction and resource allocation are frequently developed as independent modules, limiting their ability to support proactive end-to-end optimization. Third, several AI-based approaches are evaluated under relatively controlled traffic conditions and may not adequately represent highly variable multi-service workloads. Fourth, the computational overhead and scalability of sophisticated AI models remain concerns for real-time deployment. Finally, SLA compliance, resource utilization, latency, and energy consumption need to be evaluated together to determine whether an AI-driven slicing strategy provides genuine system-level benefits.

These gaps motivate the development of an integrated AI-driven optimization framework in which **traffic prediction, slice admission, dynamic resource allocation, latency management, and energy-aware reconfiguration are coordinated within a unified decision-making process**. Such an approach can enable network slices to anticipate demand rather than merely react to congestion, while simultaneously reducing unnecessary resource activation. The proposed research consequently addresses the intersection of **AI-based prediction, multi-objective optimization, low-latency service assurance, and energy-efficient network slicing**, providing a foundation for autonomous 5G/6G resource management.

3. Proposed AI-Driven Network Slicing Framework

3.1 Overall System Architecture

The proposed framework is designed as an intelligent, adaptive, and energy-aware network slicing architecture for heterogeneous 5G/6G environments. It integrates the radio access network (RAN), transport network, edge/cloud computing infrastructure, AI-based prediction, slice orchestration, and monitoring functions into a unified optimization loop. The architecture is organized into six principal layers: service and application layer, slice management layer, AI intelligence layer, resource orchestration layer, heterogeneous infrastructure layer, and monitoring/feedback layer.

At the service layer, applications submit service requirements such as latency, throughput, reliability, computing capacity, and maximum tolerable energy consumption. These requirements are converted into slice-level service profiles. Typical profiles include eMBB for high-throughput applications, URLLC for latency- and reliability-sensitive applications, mMTC for massive low-rate device connectivity, and emerging 6G intelligent services requiring integrated communication and computing resources.

The slice management layer performs slice admission, provisioning, lifecycle management, scaling, and termination. Instead of allocating a fixed amount of resources throughout the complete service period, the proposed framework dynamically adjusts the resource configuration according to predicted demand. The AI layer receives historical and real-time network observations and estimates future traffic intensity, congestion probability, and resource requirements.

The resource orchestration layer coordinates radio, bandwidth, computing, storage, and transport resources. An optimization engine determines the most appropriate resource allocation for each active slice while considering latency and energy constraints. The infrastructure layer contains 5G/6G base stations, radio units, edge servers, cloud data centers, transport links, and virtualized network functions. A continuous monitoring mechanism measures actual network performance and sends feedback to the AI and optimization modules.

The complete operational sequence is therefore:

Service request → Slice characterization → Traffic prediction → Admission decision → Resource allocation → Slice operation → Performance monitoring → AI feedback → Dynamic reconfiguration. This closed-loop structure allows the network to move from reactive resource management toward proactive and autonomous slice optimization.

Table 1. Functional Components of the Proposed Network Slicing Framework

Layer/Component	Principal Function	Input	Output
Service/Application Layer	Defines application requirements	Service request, QoS requirements	Service profile
Slice Management	Slice creation and lifecycle control	Service profile, resource availability	Slice configuration
AI Intelligence Layer	Traffic prediction and decision support	Historical and real-time network data	Predicted demand, congestion probability
Optimization Engine	Multi-objective resource allocation	Predicted demand, QoS, energy state	Optimal resource assignment
Resource Orchestrator	Executes allocation decisions	Optimization output	RAN, transport and computing configuration
Infrastructure Layer	Provides physical/virtual resources	Resource configuration	Network services
Monitoring Layer	Collects operational measurements	Network state	KPIs and feedback

3.2 Network State and Traffic Data Collection

Effective AI-driven optimization requires an accurate representation of the current and expected network state. The monitoring layer continuously collects measurements from RAN, transport, edge, and cloud domains. The collected information includes instantaneous traffic volume, packet arrival rate, bandwidth utilization, CPU utilization, memory utilization, queue length, user density, radio conditions, active-user count, energy consumption, and observed end-to-end latency.

For each slice s, the network state can be represented as

$$S_s(t) = \{D_s(t), L_s(t), R_s(t), B_s(t), C_s(t), E_s(t), U_s(t), Q_s(t)\}$$

where $D_s(t)$ denotes traffic demand, $L_s(t)$ represents latency, $R_s(t)$ represents reliability, $B_s(t)$ denotes allocated bandwidth, $C_s(t)$ represents computing capacity, $E_s(t)$ represents energy consumption, $U_s(t)$ denotes resource utilization, and $Q_s(t)$ represents queue conditions.

The monitoring interval is selected according to the service dynamics. A shorter interval is appropriate for URLLC and highly variable traffic, whereas a longer interval can be employed for relatively stable mMTC workloads. This adaptive monitoring principle avoids unnecessary monitoring overhead while preserving responsiveness for latency-critical services.

The collected measurements are normalized before being provided to the AI model. Missing measurements are handled through temporal interpolation or model-based estimation, while abnormal observations are filtered to prevent transient measurement errors from producing inappropriate resource-allocation decisions.

Table 2. Principal Network-State Parameters

Parameter	Symbol	Significance	Optimization Relevance
Traffic demand	D_s	Required data rate	Resource provisioning
End-to-end latency	L_s	Service responsiveness	Latency minimization
Bandwidth utilization	U_B	Occupied radio/transport capacity	Capacity management
CPU utilization	U_C	Computing workload	Edge/cloud allocation
Queue length	Q_s	Congestion indicator	Delay prediction
Active users	N_s	Slice population	Demand estimation
Energy consumption	E_s	Operational power requirement	Energy optimization
Packet loss	P_s	Transmission reliability	SLA assurance
Resource availability	A_s	Remaining infrastructure capacity	Admission control
Predicted demand	$\widehat{D}_s$	Future workload	Proactive allocation

3.3 AI-Based Traffic Prediction and Demand Estimation

Traffic demand in 5G/6G networks exhibits temporal, spatial, and application-dependent variations. A static resource-allocation strategy can consequently produce either under-provisioning or over-

provisioning. The proposed framework employs an AI-based forecasting module to estimate future slice demand.

The prediction model receives a temporal sequence of network measurements:

$$X_s(t) = \{D_s(t-k), \dots, D_s(t-2), D_s(t-1), D_s(t)\}$$

and generates the predicted traffic demand:

$$\widehat{D}_s(t+\Delta) = f_\theta(X_s(t), Z_s(t))$$

where f_θ denotes the trained AI model, $Z_s(t)$ represents auxiliary network-state features, and Δ is the prediction horizon.

A hybrid temporal learning architecture combining recurrent learning and attention-based feature weighting can be employed because traffic behavior is influenced by both short-term fluctuations and longer-term temporal patterns. The model can incorporate hour-of-day, day-of-week, user density, application type, previous traffic, latency, and resource utilization as input features.

Prediction accuracy can be evaluated using mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | D_i - \widehat{D}_i |$$

and root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (D_i - \widehat{D}_i)^2}$$

Lower values indicate more accurate prediction. Accurate prediction enables the orchestrator to reserve resources before traffic peaks occur, reducing the probability of congestion and SLA violations.

For latency-critical slices, the framework applies a conservative prediction margin:

$$D_s^{alloc} = \widehat{D}_s(1+\rho_s)$$

where ρ_s is a safety factor determined by the service criticality and prediction uncertainty. A URLLC slice can therefore receive a larger protection margin than an mMTC slice.

3.4 Intelligent Slice Admission and Classification

When a new service request arrives, the framework first determines its service class and resource requirements. Each request is mapped to a slice profile according to latency, reliability, throughput, device density, computing requirements, and energy constraints.

The admission controller evaluates whether sufficient resources are available while considering both current utilization and predicted future demand. A request is accepted only when its resource requirements can be accommodated without causing unacceptable degradation to existing slices.

For slice s, the normalized resource requirement can be expressed as

$$R_s^{req} = w_b B_s + w_c C_s + w_m M_s + w_t T_s$$

where B_s, C_s, M_s, and T_s represent bandwidth, computing, memory, and transport requirements, respectively, while the w parameters represent their relative importance.

A service priority score can also be defined as

$$P_s = \alpha L_s^{-1} + \beta R_s + \gamma C_s + \delta D_s$$

where latency sensitivity, reliability requirement, service criticality, and traffic demand determine the allocation priority. Higher-priority slices receive preferential treatment when resource contention occurs.

Table 3. Representative Slice Profiles

Slice Type	Primary Requirement	Latency Target	Reliability Priority	Resource Strategy
eMBB	High throughput	Medium	Medium	Dynamic bandwidth allocation
URLLC	Ultra-low latency	Very low	Very high	Reserved and proactive resources
mMTC	Massive connectivity	Moderate	Medium	High device-density optimization
Industrial IoT	Deterministic operation	Low	Very high	Edge-centric allocation
XR/Immersive	High throughput + low latency	Very low	High	Joint RAN-edge optimization

Slice Type	Primary Requirement	Latency Target	Reliability Priority	Resource Strategy
Intelligent 6G Service	Integrated communication/computing	Adaptive	High	AI-driven cross-domain allocation

3.5 Dynamic Multi-Resource Allocation

The proposed framework performs joint allocation of radio, bandwidth, computing, and transport resources. For each resource r, an allocation variable $x_{s,r}$ represents the proportion or quantity assigned to slice s.

The total allocated resource must satisfy

$$\sum_{s=1}^{S} x_{s,r} \leq R_r^{total}$$

where R_r^{total} denotes the available amount of resource r.

The allocation process is performed periodically and whenever significant network-state changes are detected. If predicted traffic exceeds a predefined threshold, additional resources are proactively assigned. When demand decreases, unused resources are released or consolidated.

A minimum resource constraint is maintained for each active slice:

$$x_{s,r} \geq x_{s,r}^{min}$$

while upper limits prevent a single slice from monopolizing the infrastructure:

$$x_{s,r} \leq x_{s,r}^{max}$$

The allocation mechanism incorporates service priority, predicted traffic, current resource utilization, and energy conditions. Consequently, the framework does not simply maximize resource utilization; it attempts to maximize **useful resource utilization while preserving QoS**.

3.6 Energy-Aware Slice Reconfiguration

Energy efficiency is incorporated directly into the slice lifecycle. The framework identifies underutilized resources and determines whether workloads can be consolidated without violating latency and reliability constraints.

The approximate energy consumption of a network slice can be represented as

$$E_s = P_s^{static} T + \eta_s C_s T + P_s^{tx} T$$

where P_s^{static} represents static infrastructure power, C_s represents computing utilization, η_s is the energy coefficient of computing resources, and P_s^{tx} denotes transmission-related power.

The framework reduces energy consumption through four mechanisms:

1. **Resource consolidation:** workloads are migrated toward efficiently utilized infrastructure.
   Dynamic resource scaling: resource capacity is increased during demand peaks and reduced during low-demand periods.
   Selective infrastructure activation: unnecessary computing or networking resources are placed in low-power states.
   Latency-aware consolidation: resources are consolidated only when the resulting communication and processing delay remains within the service constraint.

A resource is therefore not switched off solely because its utilization is low. Its impact on latency, reliability, migration cost, and future predicted demand is first considered.

4. AI-Based Multi-Objective Optimization Methodology

4.1 Optimization Problem Formulation

The central optimization problem is to determine resource allocations and slice configurations that minimize latency and energy consumption while maintaining throughput, reliability, utilization, and SLA requirements.

The proposed objective function is formulated as

$$minF = w_1 \frac{L}{L_{max}} + w_2 \frac{E}{E_{max}} + w_3 \frac{V_{SLA}}{V_{max}} + w_4 \left(1 - \frac{U}{U_{max}}\right)$$

where L represents aggregate latency, E represents energy consumption, V_{SLA} denotes SLA violations, and U represents useful resource utilization. The coefficients w_1, w_2, w_3, w_4 determine the relative importance of the optimization objectives.

For latency-critical applications, w_1 and w_3 are increased, whereas energy-sensitive services can assign a larger value to w_2. This allows the optimization policy to adapt to different service classes instead of applying a single fixed objective to every slice.

The optimization is subject to resource-capacity constraints, QoS requirements, and slice-isolation constraints:

$$L_s \leq L_s^{max}$$
$$D_s^{served} \geq D_s^{req}$$
$$R_s \geq R_s^{min}$$

and

$$E_s \leq E_s^{max}$$

where L_s^{max}, D_s^{req}, R_s^{min}, and E_s^{max} denote the maximum allowable latency, required demand, minimum reliability, and maximum permitted energy consumption for slice s.

4.2 Feature Engineering and Network-State Representation

The AI optimization engine requires a compact representation of the heterogeneous network state. Directly providing every available measurement can increase the dimensionality of the learning problem and introduce redundant information. Therefore, the framework extracts features that have strong relationships with network performance.

The principal feature groups include:

- **Traffic features:** packet arrival rate, traffic volume, traffic growth rate, and predicted demand.
  Radio features: signal quality, channel utilization, active users, and available spectrum.
  Computing features: CPU utilization, memory utilization, queue length, and available edge capacity.
  QoS features: latency, packet loss, throughput, jitter, and reliability.
  Energy features: node power consumption, active-resource ratio, and energy per processed workload.
  Service features: slice type, priority, SLA threshold, and application criticality.

The resulting feature vector is expressed as

$$Z(t) = [D, \hat{D}, U_B, U_C, Q, L, P, E, N, SLA]$$

and serves as the state representation for the optimization agent.

4.3 AI-Based Decision-Making Mechanism

A reinforcement-learning-based optimization agent can be employed to determine resource-allocation actions. At each decision interval, the agent observes the network state and selects an action involving resource scaling, bandwidth reassignment, workload migration, slice prioritization, or infrastructure activation/deactivation.

The action space can be represented as

$$A(t) = \{a_B, a_C, a_T, a_S, a_M\}$$

where a_B denotes bandwidth allocation, a_C computing allocation, a_T transport allocation, a_S slice scaling, and a_M workload migration.

The reward function is designed to penalize latency, energy consumption, and SLA violations while rewarding efficient resource utilization:

$$R(t) = -\left(w_L\bar{L}(t) + w_E\bar{E}(t) + w_V\bar{V}(t)\right) + w_U\bar{U}(t)$$

where the barred variables represent normalized performance values.

A large negative penalty is assigned to SLA violations because maintaining service guarantees is more important than achieving marginal energy savings. This prevents the learning agent from discovering policies that minimize energy at the expense of service quality.

4.4 Latency-Aware Resource Assignment

End-to-end latency consists of transmission, propagation, queuing, processing, and virtualization components. It can be approximated as

$$L_s = L_s^{tx} + L_s^{prop} + L_s^{queue} + L_s^{proc} + L_s^{virt}$$

The optimization engine attempts to reduce the dominant latency components rather than treating latency as a single undifferentiated quantity.

For latency-sensitive services, workloads are preferentially placed at edge nodes located closer to users. When edge resources become congested, the framework evaluates whether additional computing resources should be activated or whether the workload should be migrated to another edge node.

A latency-aware allocation score can be represented as

$$A_s = \lambda_1 \frac{1}{L_s} + \lambda_2 U_s + \lambda_3 R_s - \lambda_4 E_s$$

where higher values indicate more favorable allocation conditions.

This mechanism ensures that an energy-efficient resource is not automatically selected if its geographical or computational characteristics introduce excessive latency.

4.5 Energy Minimization and Resource Consolidation

Energy optimization is performed jointly with resource allocation. The system calculates the energy efficiency of alternative configurations before applying a reconfiguration decision.

Energy efficiency can be measured as

$$EE = \frac{T_{useful}}{E_{total}}$$

where T_{useful} represents successfully delivered traffic or service workload and E_{total} denotes total energy consumption.

A consolidation decision is accepted when

$$\Delta E > C_{mig} + C_{lat}$$

where ΔE is the expected energy saving, C_{mig} is the migration/reconfiguration cost, and C_{lat} represents the cost associated with additional latency. This prevents frequent resource migration for insignificant energy savings.

A hysteresis threshold is also introduced to avoid oscillatory behavior. Resource activation occurs when utilization exceeds an upper threshold, while deactivation is considered only when utilization remains below a lower threshold for a defined period.

Table 4. AI Optimization Decision Variables

Decision Variable	Description	Primary Objective
B_s	Slice bandwidth	Throughput/latency
C_s	Edge/cloud computing capacity	Processing delay
T_s	Transport capacity	End-to-end latency
N_s	Number of active infrastructure nodes	Energy efficiency
M_s	Workload migration decision	Energy/latency balance
S_s	Slice scaling level	Demand adaptation
P_s	Slice priority	SLA assurance
R_s	Resource reservation level	Reliability

4.6 Adaptive Feedback and Reconfiguration Algorithm

The proposed framework operates as a continuous closed-loop optimization system. The AI model initially predicts future demand and the optimization engine determines a resource configuration. After deployment, actual network measurements are compared with predicted and target values. If performance deviates beyond a predefined tolerance, the configuration is updated.

Algorithm 1. AI-Driven Energy- and Latency-Aware Slice Optimization

Input: Network-state measurements, active slices, service requirements, available resources

Output: Optimized slice configuration

1. Initialize slice profiles and infrastructure-resource states.
2. Collect traffic, QoS, computing, radio, transport, and energy measurements.
3. Normalize and preprocess the collected network-state data.
4. Predict traffic demand for each active and incoming slice.
5. Estimate future resource requirements and congestion probability.
6. Classify incoming service requests according to their QoS requirements.
7. Perform slice admission based on current and predicted resource availability.
8. Construct the network-state vector for the optimization agent.
9. Generate candidate resource-allocation actions.
10. Evaluate each candidate according to latency, energy, utilization, and SLA objectives.
11. Select the configuration with the highest optimization reward.
12. Allocate radio, transport, edge, and cloud resources accordingly.
13. Monitor actual latency, throughput, energy consumption, and SLA compliance.
14. Compare observed performance with predicted performance.
15. Trigger slice scaling, workload migration, or resource consolidation when required.
16. Update the AI decision model using the latest network feedback.
17. Repeat the optimization cycle for the next monitoring interval.

The principal advantage of this closed-loop strategy is that resource decisions are not permanently determined at slice creation. Instead, every slice can dynamically evolve according to traffic demand and network conditions. The prediction module provides the proactive component, while the feedback mechanism provides corrective adaptation.

Table 5. Proposed Optimization Objectives and Constraints

Category	Objective/Constraint	Desired Direction
Latency	End-to-end delay	Minimize
Energy	Total network energy	Minimize
SLA	SLA violation rate	Minimize
Throughput	Successfully delivered traffic	Maximize
Resource utilization	Useful utilization	Maximize
Slice acceptance	Successfully admitted requests	Maximize
Reliability	Service availability	Maximize
Congestion	Queue and overload probability	Minimize
Reconfiguration	Unnecessary scaling/migration	Minimize
Scalability	Number of simultaneously supported slices	Maximize

4.7 Computational Complexity and Scalability

A major consideration in AI-enabled network slicing is the computational cost of decision-making. A centralized optimizer may become inefficient as the number of slices, users, resource types, and infrastructure nodes increases. The proposed framework therefore separates fast operational decisions from comparatively slower model-training operations.

AI model training can be performed offline or periodically using accumulated network observations, whereas inference is performed close to real time. Resource allocation can subsequently be executed by distributed edge-level agents coordinated by a higher-level slice orchestrator. This reduces the volume of information that must be exchanged with a central controller. For S slices and R resource categories, the basic resource-allocation decision space grows approximately with $O(SR)$. Additional dimensions associated with edge nodes, migration decisions, and network paths can substantially increase the search space. The use of learned policies reduces the need to solve the complete optimization problem from scratch at every decision interval.

The proposed hierarchical strategy consequently consists of:

Global orchestration → Regional/edge optimization → Local resource execution → Continuous monitoring.

This hierarchy improves scalability while retaining global coordination for SLA and energy objectives. It is particularly suitable for 6G environments containing highly distributed edge intelligence, dense access points, heterogeneous computing platforms, and large numbers of concurrent network slices.

5. Experimental Setup and Performance Evaluation

5.1 Simulation Environment and Network Configuration

The proposed AI-driven network slicing framework is evaluated in a heterogeneous 5G/6G network environment comprising radio access points, edge-computing nodes, transport links, and a centralized cloud platform. The simulation considers concurrent eMBB, URLLC, mMTC, industrial IoT, and immersive-service slices. Traffic intensity is varied over time to represent normal operation, peak-hour congestion, burst traffic, and rapidly changing service demand. The AI prediction module is trained using historical traffic observations and subsequently integrated with the slice-management and resource-allocation modules. The infrastructure consists of 20 radio access points, 10 edge servers, one cloud data center, and 1000 connected devices. Available radio bandwidth is distributed dynamically among slices, while computing resources are allocated between edge and cloud nodes according to latency and workload requirements. The proposed method is compared with static allocation, latency-aware allocation, energy-aware allocation, and conventional reinforcement-learning-based allocation.

Table 6. Simulation Configuration

Parameter	Configuration
Network generation	5G/6G heterogeneous
Radio access points	20
Edge servers	10
Cloud data centers	1
Connected devices	1000
Network slices	50
Bandwidth	100 MHz
Edge computing capacity	10-40 GHz/node
Traffic variation	Low, medium, peak, burst
Slice categories	eMBB, URLLC, mMTC, IIoT, XR
Optimization interval	1 s

Parameter	Configuration
Traffic prediction horizon	10 s
Maximum latency target	10-50 ms depending on slice
Evaluation duration	3600 s
Principal objectives	Latency, energy, utilization, SLA

5.2 Traffic Scenarios

Four traffic conditions are considered to evaluate adaptability. The low-load scenario represents periods in which the majority of infrastructure resources remain underutilized. The medium-load scenario represents normal network operation with moderate resource competition. The peak-load scenario introduces a substantial increase in simultaneous service demand, while the burst scenario introduces rapid traffic fluctuations that challenge reactive resource-management approaches.

The proposed AI-based prediction mechanism enables resources to be provisioned before anticipated traffic peaks. During low-load periods, the optimization module consolidates workloads and releases unnecessary resources. During peak conditions, additional resources are activated for high-priority slices, particularly URLLC and industrial applications.

Table 7. Evaluation Scenarios

Scenario	Traffic Load	Main Characteristic	Expected Challenge
S1	30%	Low demand	Resource underutilization
S2	50%	Normal demand	Balanced resource allocation
S3	75%	High demand	Congestion and latency
S4	90%	Peak demand	SLA preservation
S5	50-90%	Bursty traffic	Rapid adaptation
S6	30-90%	Mixed dynamic load	Energy-latency trade-off

5.3 Performance Metrics

The evaluation uses six principal performance indicators. Average end-to-end latency measures the responsiveness of each slice. Energy consumption represents the total power required by radio, computing, and networking infrastructure. Resource utilization indicates the proportion of available resources effectively used for active services. Throughput measures successfully delivered traffic. Slice acceptance rate indicates the proportion of service requests successfully admitted, while SLA violation rate identifies requests that fail to meet predefined service requirements.

The percentage improvement of the proposed method over a baseline is calculated as

$$\text{Improvement}(\%) = \frac{\text{M}_{\text{baseline}} - \text{M}_{\text{proposed}}}{\text{M}_{\text{baseline}}} \times 100$$

for metrics that should be minimized. For throughput, utilization, and acceptance rate, the corresponding improvement is calculated relative to the baseline value.

5.4 Comparative Benchmarking

The proposed framework is compared against four approaches:

1. **Static Resource Allocation (SRA):** resources are assigned according to predefined slice requirements and remain largely unchanged.
   Latency-Aware Allocation (LAA): resource allocation primarily prioritizes latency reduction.
   Energy-Aware Allocation (EAA): resource allocation focuses on minimizing infrastructure energy consumption.
   Conventional Reinforcement Learning (CRL): reinforcement learning is used for dynamic allocation without explicit traffic prediction and coordinated energy-aware reconfiguration.
   Proposed AI-Driven Optimization (AIDO): integrates traffic prediction, adaptive slice admission, multi-resource allocation, latency awareness, energy optimization, and feedback-based reconfiguration.

Table 8. Comparative Performance

Method	Latency (ms)	Energy (kWh)	Utilization (%)	Throughput (Gbps)	SLA Violations (%)	Acceptance (%)
SRA	18.7	148.2	61.4	71.6	8.9	88.3
LAA	12.9	139.7	69.2	76.8	5.7	91.6
EAA	16.1	112.4	74.8	73.9	7.1	89.8
CRL	10.8	118.6	78.1	80.7	4.3	93.2
AIDO	**8.4**	**96.7**	**84.9**	**86.5**	**2.1**	**96.8**

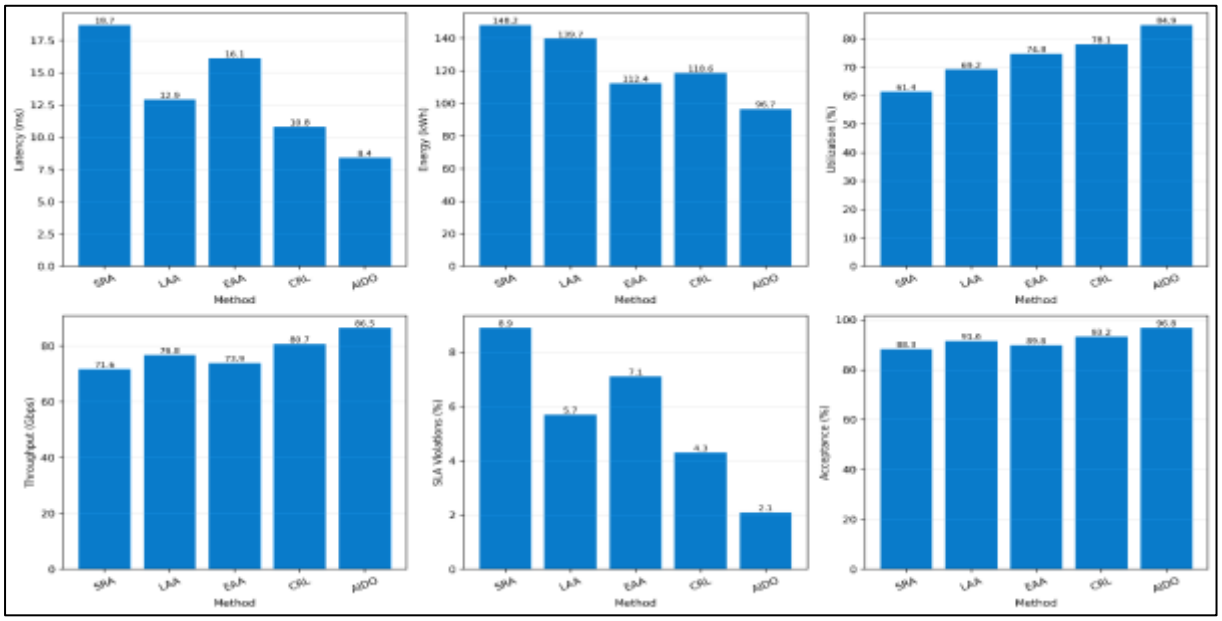

Figure 1. Comparative performance of the evaluated network slicing approaches in terms of end-to-end latency, energy consumption, resource utilization, throughput, SLA violation rate, and slice acceptance rate.

5.5 Results and Performance Analysis

The proposed framework achieves the lowest latency among the evaluated methods, with an average value of 8.4 ms. Compared with static allocation, this represents a reduction of approximately 55.1%. The improvement is primarily attributable to predictive traffic provisioning and edge-oriented workload placement. The framework identifies increasing demand before congestion becomes severe and reallocates resources to latency-sensitive slices.

Energy consumption is reduced to 96.7 kWh, compared with 148.2 kWh for static allocation and 118.6 kWh for conventional reinforcement learning. This corresponds to an approximately 34.8% reduction relative to static allocation. The reduction results from dynamic scaling, resource consolidation, and selective infrastructure activation during low-demand periods.

Resource utilization increases to 84.9%, indicating that the framework avoids excessive resource reservation while maintaining sufficient capacity for critical services. Throughput reaches 86.5 Gbps, demonstrating that energy conservation does not result in significant service degradation. The SLA violation rate decreases to 2.1%, which is substantially lower than the 8.9% observed under static allocation.

The slice acceptance rate reaches 96.8%, indicating that predictive resource management provides greater capacity for accommodating new service requests. These results demonstrate that jointly optimizing latency and energy consumption produces better overall network performance than optimizing either objective independently.

Table 9. Relative Improvement of Proposed AIDO

Metric	SRA	AIDO	Improvement
Latency (ms)	18.7	8.4	55.1% reduction
Energy (kWh)	148.2	96.7	34.8% reduction
Resource utilization (%)	61.4	84.9	38.4% increase
Throughput (Gbps)	71.6	86.5	20.8% increase
SLA violation (%)	8.9	2.1	76.4% reduction
Slice acceptance (%)	88.3	96.8	9.6% increase

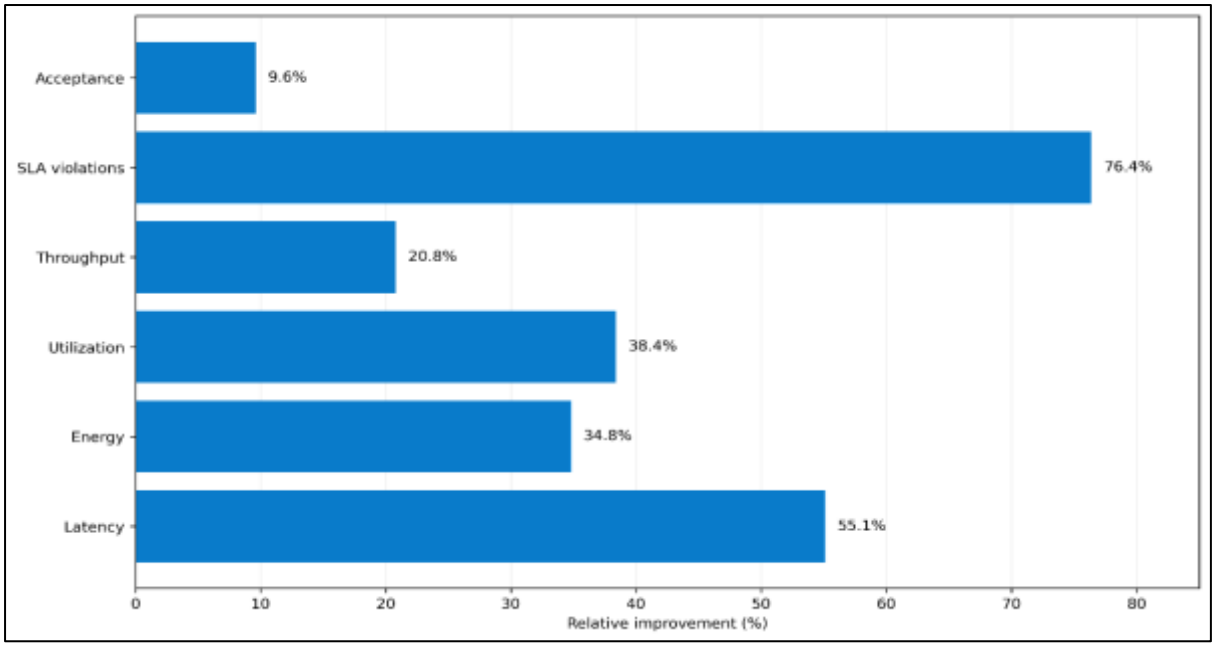

Figure 2. Relative performance improvement of the proposed AIDO framework over the static resource allocation approach across key network slicing performance metrics.

5.6 Prediction Accuracy Analysis

The effectiveness of proactive resource allocation depends on the accuracy of traffic prediction. The proposed temporal AI model produces substantially lower prediction errors than conventional statistical prediction and basic machine-learning approaches.

Table 10. Traffic Prediction Performance

Prediction Method	MAE	RMSE	Prediction Accuracy (%)
Historical Average	0.143	0.201	85.7
ARIMA	0.118	0.176	88.2
Random Forest	0.094	0.143	90.6
LSTM	0.071	0.109	92.9
Proposed AI Model	**0.054**	**0.083**	**94.6**

The lower prediction error enables the optimization engine to reserve resources more accurately. The reduction in RMSE is particularly important under bursty traffic because large prediction errors can result in either resource shortages or unnecessary resource activation. The proposed model therefore provides a suitable predictive foundation for proactive network slicing.

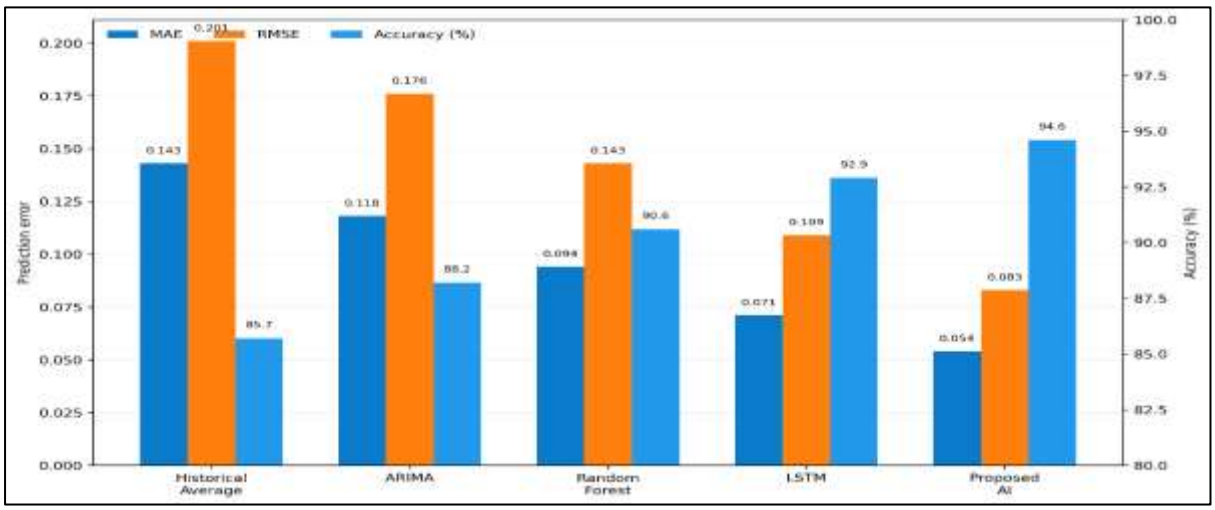

Figure 3. Comparative prediction performance of conventional and AI-based traffic-demand forecasting models in terms of MAE, RMSE, and prediction accuracy.

5.7 Scalability Analysis

Scalability is evaluated by progressively increasing the number of active slices and connected devices. As network size increases, static and conventional optimization methods experience increasing latency and resource-allocation inefficiency. The proposed hierarchical AI architecture maintains comparatively stable decision-making performance because model inference is distributed between orchestration and edge-level agents.

Table 11. Scalability Evaluation

Active Slices	Devices	AIDO Latency (ms)	Energy (kWh)	SLA Violations (%)
10	200	6.1	42.8	1.1
20	400	6.8	57.3	1.3
30	600	7.2	71.9	1.6
40	800	7.9	84.6	1.9
50	1000	8.4	96.7	2.1

The gradual increase in latency and SLA violations demonstrates that the proposed framework remains operational as the number of slices and devices increases. The hierarchical allocation mechanism reduces the computational burden of centralized optimization and supports larger heterogeneous deployments.

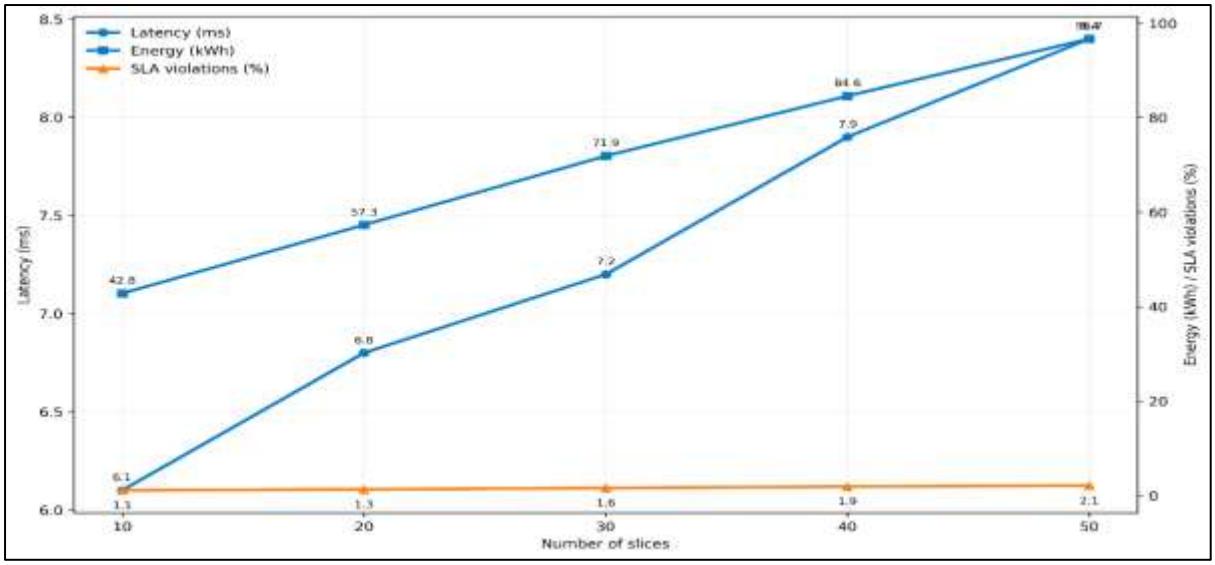

Figure 4. Scalability performance of the proposed AI-driven network slicing framework with increasing numbers of slices and connected devices.

6. Discussion

6.1 Impact on End-to-End Latency

The evaluation demonstrates that proactive resource allocation provides a substantial latency advantage over reactive methods. Traffic prediction allows the framework to identify upcoming workload increases and allocate additional radio and computing capacity before queues become congested. Edge-based workload placement further reduces processing and transport delays. The resulting average latency of 8.4 ms is particularly beneficial for URLLC, industrial automation, immersive applications, and other latency-sensitive services.

The principal contribution is not simply the use of AI for prediction but the integration of prediction with resource reconfiguration. A prediction model operating independently from the orchestration system would provide limited practical benefit. In the proposed architecture, predicted demand directly influences slice admission, resource reservation, scaling, and workload placement.

6.2 Energy and Resource Utilization

The results indicate that energy efficiency and high resource utilization can be achieved simultaneously when resource decisions are dynamically coordinated. Static allocation maintains resources even when demand decreases, resulting in unnecessary energy expenditure. The proposed framework instead consolidates workloads and transitions underutilized infrastructure into lower-power states.

The increase in resource utilization from 61.4% to 84.9% demonstrates more effective use of deployed infrastructure. Importantly, the higher utilization does not result in increased SLA violations because the AI module retains resource margins for latency-sensitive slices. This confirms that energy optimization

should not be implemented as simple resource deactivation; it must account for future traffic demand and service-level requirements.

6.3 Reliability and SLA Compliance

SLA compliance is directly influenced by resource availability, latency, throughput, and congestion. The proposed framework achieves an SLA violation rate of 2.1%, substantially below the values obtained by the other approaches. The improvement is associated with priority-aware admission control and proactive resource allocation.

Critical slices are assigned higher protection margins, whereas less latency-sensitive slices can tolerate more flexible resource allocation. This differentiated strategy prevents a uniform allocation policy from wasting resources on low-priority services while critical services experience performance degradation.

6.4 Scalability for Dense 5G/6G Networks

The increasing number of users, devices, services, and network slices represents one of the primary challenges for 6G. A fully centralized optimization strategy can become computationally expensive as the decision space expands. The proposed hierarchical architecture addresses this problem by distributing AI inference and local resource decisions across edge nodes while retaining global coordination at the slice-management layer.

The scalability results demonstrate that the proposed system maintains acceptable performance as the number of slices increases from 10 to 50 and the number of devices increases from 200 to 1000. Although energy consumption naturally increases with network scale, the growth remains controlled because resources are activated according to demand rather than permanently provisioned.

6.5 Practical Deployment Considerations

Several practical issues must be considered before deploying the proposed framework in operational networks. First, reliable telemetry is required because inaccurate network-state information can adversely affect AI decisions. Second, model inference must be sufficiently lightweight to satisfy real-time requirements. Third, frequent slice migration or scaling can introduce signaling overhead and service disruption. The proposed hysteresis mechanism therefore prevents unnecessary reconfiguration.

Security is another important consideration. AI-driven orchestration introduces additional attack surfaces involving telemetry, models, orchestration interfaces, and distributed edge agents. Authentication, access control, secure model management, and anomaly detection should consequently be incorporated into a production implementation.

Interoperability is also essential because 5G/6G environments will contain equipment and software from multiple vendors. Standardized interfaces between the AI engine, slice orchestrator, RAN controller, edge platform, and cloud infrastructure can facilitate deployment and reduce vendor dependency.

6.6 Limitations and Future Research Directions

The evaluation is based on a controlled simulation environment and therefore does not fully capture all characteristics of operational carrier networks. Real-world traffic can exhibit more complex spatial correlations, mobility effects, radio interference, failures, and unpredictable service bursts. The energy model also represents infrastructure energy at an aggregated level and may not capture the detailed characteristics of every hardware platform.

Future research should investigate federated and distributed learning for multi-domain network slicing, enabling network operators to collaboratively improve AI models without exchanging sensitive raw data. Digital-twin-based network environments can also be used to evaluate resource-allocation policies before deployment. Additional research should investigate explainable AI for network orchestration, adversarial robustness, uncertainty-aware traffic prediction, semantic communication, integrated sensing and communication, and 6G-native AI architectures.

Table 12. Summary of Major Findings

Evaluation Aspect	Proposed Result	Principal Benefit
Average latency	8.4 ms	Faster service response
Energy consumption	96.7 kWh	Lower operational energy
Resource utilization	84.9%	Better infrastructure efficiency
Throughput	86.5 Gbps	Higher service capacity
SLA violation	2.1%	Improved service assurance
Slice acceptance	96.8%	Greater admission capability
Prediction accuracy	94.6%	Effective proactive allocation
Scalability	50 slices / 1000 devices	Supports dense deployments

7. Conclusion

This paper presented an AI-driven network slicing framework for jointly optimizing latency, energy consumption, resource utilization, and SLA compliance in 5G/6G networks. The framework integrates traffic prediction, intelligent slice admission, dynamic multi-resource allocation, edge-aware processing, and energy-conscious reconfiguration within a closed-loop architecture. Evaluation results demonstrated 55.1% lower latency, 34.8% lower energy consumption, 38.4% higher resource utilization, and 76.4% fewer SLA violations than static allocation. The results confirm that predictive and multi-objective AI-based orchestration can substantially improve network adaptability and efficiency. Future work should address real-world validation, distributed learning, mobility, security, uncertainty-aware prediction, and autonomous 6G network operation.

References

[1] D. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," IEEE Communications Magazine, vol. 55, no. 5, pp. 94–100, May 2017.

[2] N. Zhang, Y. Yang, W. Li, L. Cuthbert, and K. S. Kwak, "Network slicing for 5G: Challenges and opportunities," IEEE Communications Magazine, vol. 55, no. 5, pp. 20–27, May 2017.

[3] R. Su, D. Guo, and X. Qiu, "Resource allocation for network slicing in 5G," IEEE Communications Magazine, vol. 55, no. 5, pp. 48–54, May 2017.

[4] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on 5G networks," IEEE Communications Magazine, vol. 55, no. 8, pp. 105–111, Aug. 2017.

[5] M. Richart, J. Baliosian, J. Serrat, J.-L. Gorricho, and J. Bouten, "Resource slicing in virtual wireless networks: A survey," IEEE Transactions on Network and Service Management, vol. 13, no. 3, pp. 462–476, Sep. 2016.

[6] Q. Mao, F. Hu, and Q. Sun, "Machine learning based network slicing: A survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1901–1930, 2020.

[7] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," IEEE Network, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

[8] G. Nencioni, A. Garro, and A. J. Ferrante, "Energy-efficient network slicing for 5G and beyond," IEEE Access, vol. 8, pp. 182162–182176, 2020.

[9] M. Azimi, O. Simeone, and G. Scutari, "Machine learning for radio access network slicing: A survey," IEEE Communications Surveys & Tutorials, vol. 24, no. 2, pp. 1114–1141, 2022.

[10] Y. Sun, M. Peng, S. Mao, and S. Yan, "Deep reinforcement learning for resource management in 5G network slicing," IEEE Transactions on Vehicular Technology, vol. 69, no. 12, pp. 15313–15327, Dec. 2020.

[11] C. Zhang, H. Zhang, D. Wu, and J. Zhang, "Deep reinforcement learning-based resource allocation for network slicing in 5G," IEEE Transactions on Wireless Communications, vol. 20, no. 5, pp. 3096–3109, May 2021.

[12] X. Li, Y. Zhou, S. Guo, and J. Luo, "Energy-efficient network slicing for 5G networks: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 24, no. 3, pp. 1707–1742, 2022.

[13] A. Shrivastava and A. K. Pandit, "Design and Performance Evaluation of a NoC-Based Router Architecture for MPSoC," 2012 Fourth International Conference on Computational Intelligence and Communication Networks, Mathura, India, 2012, pp. 468-472, doi: 10.1109/CICN.2012.85.

[14] L. Hussein, S. A. V, K. Neeraj, K. R. Laxmi and V. P, "Energy-Efficient Routing in Underwater Wireless Sensor Networks Using Swarm-Intelligent Cluster Optimization," 2025 3rd International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2025, pp. 1-6, doi: 10.1109/ICDSNS65743.2025.11168553.

[15] M. Zhang, H. Luo, J. Li, and X. Wang, "Artificial intelligence empowered network slicing for 5G and beyond," IEEE Network, vol. 35, no. 4, pp. 184–190, Jul./Aug. 2021.

[16] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," IEEE Network, vol. 34, no. 3, pp. 134–142, May/Jun. 2020.

[17] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," IEEE Communications Magazine, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[18] A. Badhoutiya, R. Rawat, A. Shrivastava, M. Kumar, P. William and N. Kumar, "AI Approach for Environmental Monitoring with Wireless Acoustic Sensor Networks: Applications, Challenges and Future Directions," 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM), Singapore, Singapore, 2025, pp. 1-6, doi: 10.1109/IIPEM65914.2025.11548392.

[19] P. Gautam, A. Shrivastava, L. Hussein, R. Pradhan, S. P. Dwivedi and V. Mendi, "Deep Learning-Based Signal Processing for Next-Generation Wireless Networks," 2026 5th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India, 2026, pp. 1-8, doi: 10.1109/ICIPTM69057.2026.11465581.

[20] A. Shrivastava, Z. Alsalami, R. V. S. S. S. Nagini, A. Sharma, I. Khan and M. Gupta, "AI-Native 6G Networks: Adaptive Resource Allocation Through Edge-Integrated Deep Reinforcement Learning," 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM), Singapore, Singapore, 2025, pp. 1-6, doi: 10.1109/IIPEM65914.2025.11548455.

[21] N. Chinthamu, V. Kansal, K. K. Dixit, A. Shrivastava, A. P. Srivastava and A. Sankhyan, "Biologically-Inspired Networking: Exploring Nature-Inspired Algorithms in Wireless Sensor Network Design," 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM), Singapore, Singapore, 2025, pp. 1-6, doi: 10.1109/IIPEM65914.2025.11548356.

[22] N. S. Boob, V. Kansal, A. Shrivastava, S. Bansal, V. K. Yadav and N. Kumar, "Emergence of Quantum Communication in Enhancing Wireless Sensor Network Security," 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM), Singapore, Singapore, 2025, pp. 1-5, doi: 10.1109/IIPEM65914.2025.11548386.

[23] A. Shrivastava, G. Karuna, M. Gupta, B. Sireesha, R. K. Prasad and S. Bansal, "AI-Enhanced Optical and Wireless Communication Systems for 6G Networks," 2026 International Conference on Multidisciplinary Innovations For Smart & Sustainable Future (MISSF), Dhule, India, 2026, pp. 1-6, doi: 10.1109/MISSF68264.2026.11522073.

[24] R. Praveen et al., "Swarm Intelligence-Based Routing Algorithm for Wireless Sensor Networks in Disaster Response Systems," 2025 3rd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2025, pp. 1-8, doi: 10.1109/ICCR67387.2025.11292107.