

An Integrated LLM-GNN-RL Framework for Network Traffic Prediction and Optimization

Ikani Lucy Hassana, Olumide Owolabi, Benjamin Okike, Abdullahi Fatimah Binta
Department of Computer Science, University of Abuja, Gwagwalada, Nigeria

Abstract - The rapid growth of Internet traffic driven by cloud services, IoT, multimedia applications, and emerging 5G and 6G infrastructures has made network traffic prediction and optimization increasingly difficult. Conventional statistical and machine learning methods often perform poorly in highly dynamic and topologically complex network environments because they do not jointly model semantic context, structural dependencies, and adaptive control. This paper presents a framework that integrates Large Language Models (LLMs), Graph Neural Networks (GNNs), and Reinforcement Learning (RL) for intelligent network traffic prediction and optimization. In the proposed framework, LLMs are used to extract semantic representations from heterogeneous traffic logs, GNNs learn the relational dependencies among nodes and links, and RL agents optimize routing and resource allocation in real time. The system was evaluated using real and simulated traffic environments derived from Cooperative Association for Internet Data Analysis (CAIDA), Measurement and Analysis on the WIDE Internet (MAWI), Network Simulator 3 (NS-3), and Mininet. Findings from the study show that the integrated model reduced prediction error by up to 17% compared with an LSTM-based baseline and improved key performance indicators such as accuracy, latency, throughput, packet delivery ratio, and jitter. In the integrated evaluation, the proposed model achieved 98.7% accuracy, 14.7ms latency, 1,045 Mbps throughput, 98.2% packet delivery ratio, and 7.1ms jitter. These results suggest that combining semantic modeling, graph learning, and adaptive control provides a practical path toward more reliable and scalable AI-driven network management.

Keywords: Network Traffic Prediction, Network Optimization, Large Language Models, Graph Neural Networks, Reinforcement Learning, Intelligent Networking

1. INTRODUCTION

Modern communication networks are now expected to support large volumes of heterogeneous traffic generated by cloud platforms, IoT ecosystems, streaming services, edge computing, and intelligent applications. As a result, network traffic has become more bursty, high dimensional, and context dependent, making accurate prediction and efficient optimization increasingly difficult. Traditional network management techniques based on fixed rules, heuristics, and classical statistical models often fail to adapt quickly to changes in topology, workload, and service demand. This leads to congestion, packet loss, latency increase, throughput degradation, and reduced quality of service. More also, modern network management requires models that can simultaneously understand semantic traffic context, capture structural network relationships, and learn adaptive optimization policies.

In recent years, graph learning and reinforcement learning (RL) have become prominent tools for intelligent networking. Graph Neural Networks are particularly suitable for networking problems because network infrastructures are naturally graph-structured, with routers, switches, and hosts represented as nodes and the communication links represented as edges. Prior work such as RouteNet demonstrated that GNNs can predict delay, jitter, and loss across arbitrary topologies by learning relationships among routing, topology, and offered traffic. Later developments such as RouteNet-Fermi further improved network modeling under more realistic traffic assumptions (Verma et al., 2024; Ferriol-Galmés et al., 2023).

At the same time, RL has shown strong potential for dynamic traffic engineering and adaptive routing because it can learn policies directly from network state and reward feedback. Recent studies on RL-driven traffic engineering in software-defined networks show that deep reinforcement learning can react to changing traffic demands and improve congestion management (Kurtuluş, 2025; Ibronke, 2025). Though challenges remain in sample efficiency, realism of training environments, and scalability.

Large Language Models are also beginning to influence the networking domain. While they were originally developed for natural language processing, recent studies show that LLMs can assist with contextual representation learning, log understanding, domain knowledge extraction, and traffic-related reasoning. Emerging work on LLM-enhanced traffic forecasting and specialized traffic foundation models suggests that large pretrained models can improve representation quality when used as semantic encoders rather than standalone predictors (Tu et al., 2026; Chen et al., 2025).

Despite these advances, a major research gap remains. GNNs are effective for relational learning but are not sufficient on their own for semantic interpretation of heterogeneous logs (Yang et al., 2022; Sankar et al., 2019). RL is effective for optimization but depends heavily on the quality of network state representation (He et al., 2023; Solozabal et al., 2019). LLMs can provide rich contextual embeddings but do not inherently model network topology or control decisions (Chen et al., 2024). This study therefore proposes an integrated LLM-GNN-RL framework in which the strengths of the three paradigms are combined in a unified architecture for traffic prediction and network optimization. The framework processes traffic logs with an LLM-based encoder, converts network behavior into graph representations for GNN-based prediction, and applies RL for real-time adaptive optimization of routing and bandwidth decisions.

The significance of this study lies in its attempt to move beyond isolated machine learning solutions toward a coordinated AI architecture for network management. The use of official traffic repositories such as Cooperative Association for Internet Data Analysis (CAIDA) and Measurement and Analysis on the WIDE Internet (MAWI), together with simulation environments such as Network Simulator 3 (NS-3) and Mininet, provides a useful basis for testing both predictive performance and optimization behavior. CAIDA provides anonymized passive traces collected from high-speed backbone monitors (Favale et al., 2021), while the MAWI archive offers long-running traffic traces from operational backbone links, making both suitable for traffic analysis research (Wei et al., 2026; Stoenescu et al., 2015).

This paper is guided by the following objectives: to investigate the effectiveness of LLMs for traffic feature extraction, to examine the capability of GNNs in learning complex network relationships, to apply RL for dynamic traffic optimization, and to evaluate the performance of the combined framework under realistic and simulated networking conditions.

2. RELATED WORK

Network traffic prediction has evolved from classical time-series models to machine learning and deep learning approaches (Ma et al., 2020; Oliveira et al., 2016). Early models such as Autoregressive Integrated Moving Average (ARIMA) and regression-based predictors were useful for stationary traffic patterns but were limited in handling nonlinearity and complex dependencies (Sethuraman, & Chennareddy, 2022; Mahajan et al., 2022). Deep learning methods including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs) improved predictive performance by learning temporal features automatically. Yet they often lacked explicit awareness of network topology and node interactions. The study identifies this limitation of lack of explicit awareness of network topology and node interactions as one of the motivations for introducing graph learning into the traffic prediction pipeline.

Graph Neural Networks (GNNs) have become important because they preserve topological structure during learning. RouteNet showed that a GNN can generalize across unseen network topologies and estimate key performance indicators such as delay and jitter with strong accuracy. RouteNet-Fermi later demonstrated that custom GNN architectures can capture more realistic queueing behavior and better model packet loss, delay, and jitter. These studies confirm that graph-based learning is well suited for performance prediction in computer networks (Geyer, 2017; Ding et al., 2016).

Reinforcement learning has been widely investigated for routing, congestion control, and traffic engineering. RL agents learn from reward signals and can adapt to changing network states more flexibly than fixed optimization rules. Recent literature shows that RL-based traffic engineering can improve routing under real-time demands (Xiao et al., 2021; Zhang et al., 2020). But it also highlights persistent challenges such as slow convergence, poor generalization from unrealistic simulation environments, and computational overhead (Maheshwari et al., 2025; Michailidis et al., 2025).

Large Language Models are an emerging addition to this area. Recent work suggests that LLMs can enrich traffic analysis by extracting contextual information from logs, textual metadata, and heterogeneous traffic descriptors. Studies such as a benchmark for federated settings which is Learning, Environments for Algorithm, Federation (LEAF) and TrafficLLM indicate that LLM-

derived embeddings can improve downstream prediction tasks when integrated with graph or temporal models (Chen et al., 2022). While broader surveys show growing interest in LLM-enabled communication and wireless networking (Sun et al., 2024; Zhou et al., 2024).

The gap in the literature is therefore clear. Existing methods often excel in one aspect only. Temporal models capture sequence patterns, graph models capture structure, RL captures adaptation, and LLMs capture semantics. However, network traffic management in modern infrastructures requires all four capabilities together. The study addresses this gap by proposing a unified framework in which semantic traffic understanding, graph-aware forecasting, and adaptive optimization are tightly linked. Table 1 presents the identified gaps in existing intelligent network traffic management approaches.

Table 1. Comparative Analysis of Existing Approaches and the Proposed Framework

Study Type	Semantic Learning	Topology Awareness	Adaptive Optimization	Major Limitation
Traditional ML	No	No	No	Poor handling of complex traffic
Deep Learning Models	Partial	No	No	Weak topology representation
GNN Models	No	Yes	Partial	Lack semantic understanding
RL Approaches	No	Partial	Yes	State representation limitations
Proposed Framework	Yes	Yes	Yes	Computational complexity

The analysis shows that existing approaches address isolated challenges, whereas the proposed framework integrates semantic understanding, topology awareness, and adaptive optimization within a unified architecture.

3. METHODOLOGY

3.1 Research Design

This study adopted an experimental AI systems design in which traffic data were collected, preprocessed, represented semantically and structurally, and then optimized using a reinforcement learning agent. The framework was designed as a modular architecture with four main stages: data collection and preprocessing, LLM-based feature extraction, GNN-based traffic prediction, and RL-based optimization. This study describes this modular design as a way to ensure scalability, maintainability, and seamless interaction among the different components. Figure 1 presents the overall architecture of the proposed integrated LLM-GNN-RL framework for intelligent network traffic prediction and optimization.

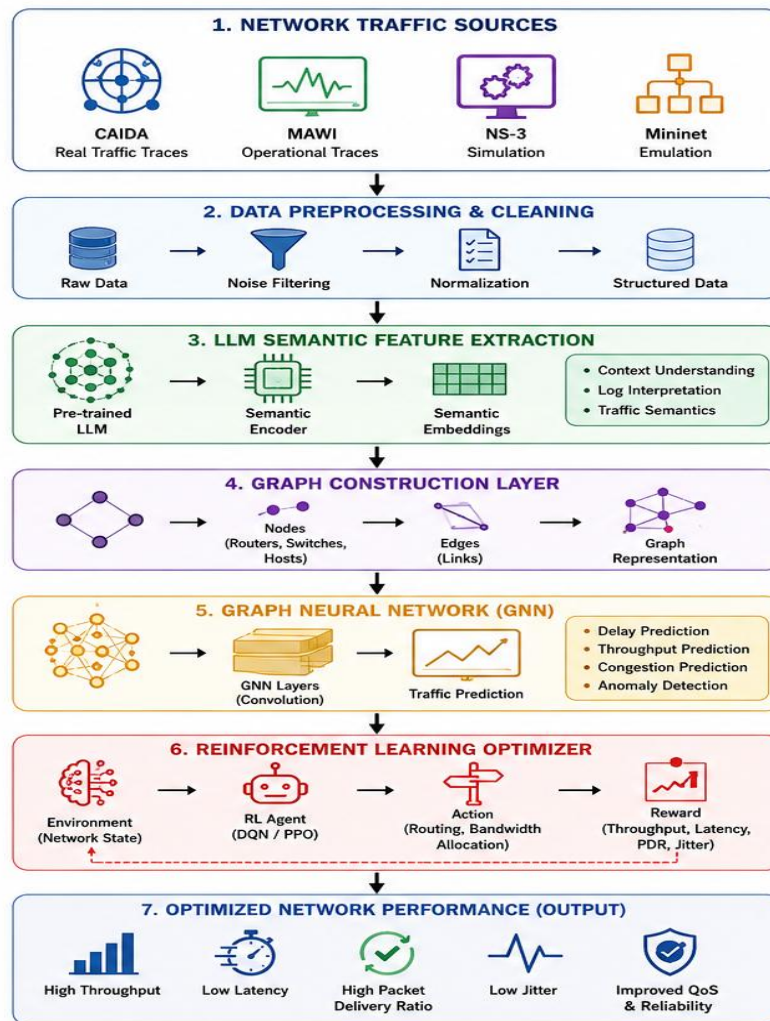


Figure 1. Integrated LLM-GNN-RL Framework for Intelligent Network Traffic Prediction and Optimization

Figure 1 illustrates how semantic feature extraction, graph-based structural learning, and reinforcement learning-based optimization are integrated into a unified intelligent networking framework.

3.2 Framework Architecture

The first module collects and preprocesses network traffic data from traffic traces and simulation environments. Raw traffic data are captured from CAIDA, MAWI, and NS-3-based simulations, then cleaned, normalized, and transformed into structured forms suitable for downstream modeling.

The second module applies an LLM-based encoder to traffic logs and related metadata in order to generate semantic embeddings. This stage is intended to capture higher-level contextual features that may not be represented well by raw packet statistics alone.

The third module uses a Graph Neural Network to represent the communication network as a graph of nodes and links. At this stage, the framework learns structural dependencies among network entities and uses them for traffic prediction. This is important because real network behavior is influenced not only by traffic volume but also by path relationships, adjacency, and shared resource constraints.

The fourth module applies reinforcement learning for real-time optimization. The RL agent observes network conditions, prediction outputs, and contextual embeddings, then chooses actions such as route adjustment or bandwidth allocation. The goal is to maximize network efficiency by improving throughput, reducing latency, minimizing jitter, and maintaining high packet delivery ratio.

3.3 Experimental Environment and Data Sources

Evaluation was carried out using real-world traffic traces from Cooperative Association for Internet Data Analysis (CAIDA) and Measurement and Analysis on the WIDE Internet (MAWI), together with simulated environments based on NS-3 and Mininet. CAIDA offers anonymized backbone Internet traces, and MAWI maintains long-term operational packet traces from backbone links, both of which are commonly used for traffic analysis research. Simulation environments such as NS-3 and Mininet are useful for testing optimization policies under controlled yet configurable traffic scenarios (Ivey et al., 2016). Table 2 presents the datasets and simulation environments used for evaluating the proposed framework.

Table 2. Description of Datasets and Simulation Environments

Source	Type	Purpose	Characteristics
CAIDA	Real Traffic Trace	Traffic prediction	Backbone Internet traffic traces
MAWI	Operational Trace	Traffic analysis	Long-running packet traces
NS-3	Simulation Environment	Optimization testing	Configurable network scenarios
Mininet	Network Emulator	SDN experimentation	Virtual software-defined networks

The combination of real-world traffic traces and controlled simulation environments improves the reliability and robustness of the evaluation process.

3.4 Evaluation Metrics

This study uses multiple evaluation metrics to assess predictive and operational performance. These include accuracy, throughput, latency, packet delivery ratio, and jitter. For prediction, error reduction relative to baseline models was also reported. The use of multiple metrics is necessary because network optimization is a multi-objective problem in which a gain in one performance indicator may affect another.

4. RESULTS AND DISCUSSION

The integrated LLM-GNN-RL framework outperformed baseline methods and individual components. Compared with an LSTM-based model, the proposed framework reduced prediction error by up to 17%, while also improving throughput by 12% and reducing average latency by 9%. These results suggest that combining semantic embeddings, graph-aware learning, and adaptive control can substantially improve network intelligence. Table 3 compares the performance of the proposed framework with baseline and individual learning configurations.

Table 3. Comparative Evaluation of Different Network Intelligence Models

Model	Accuracy (%)	Latency (ms)	Throughput (Mbps)	Packet Delivery Ratio (%)	Jitter (ms)
Traditional ML	88.2	28.4	786	91.5	12.4
GNN-only	96.4	19.3	954	96.8	8.9
RL-only	94.7	21.1	910	95.1	9.6
Proposed LLM-GNN-RL	98.7	14.7	1045	98.2	7.1

The integrated LLM-GNN-RL framework achieved the best overall performance across all evaluation metrics. The proposed integrated framework achieved 98.7% accuracy, 14.7ms latency, 1,045 Mbps throughput, 98.2% packet delivery ratio, and 7.1ms jitter. By comparison, the traditional machine learning baseline achieved 88.2% accuracy, 28.4ms latency, 786 Mbps throughput, 91.5% packet delivery ratio, and 12.4ms jitter. The GNN-only configuration achieved 96.4% accuracy and the RL-only configuration achieved 94.7% accuracy, both of which were lower than the integrated framework. Figure 2 presents a graphical comparison of the performance achieved by different network intelligence models.

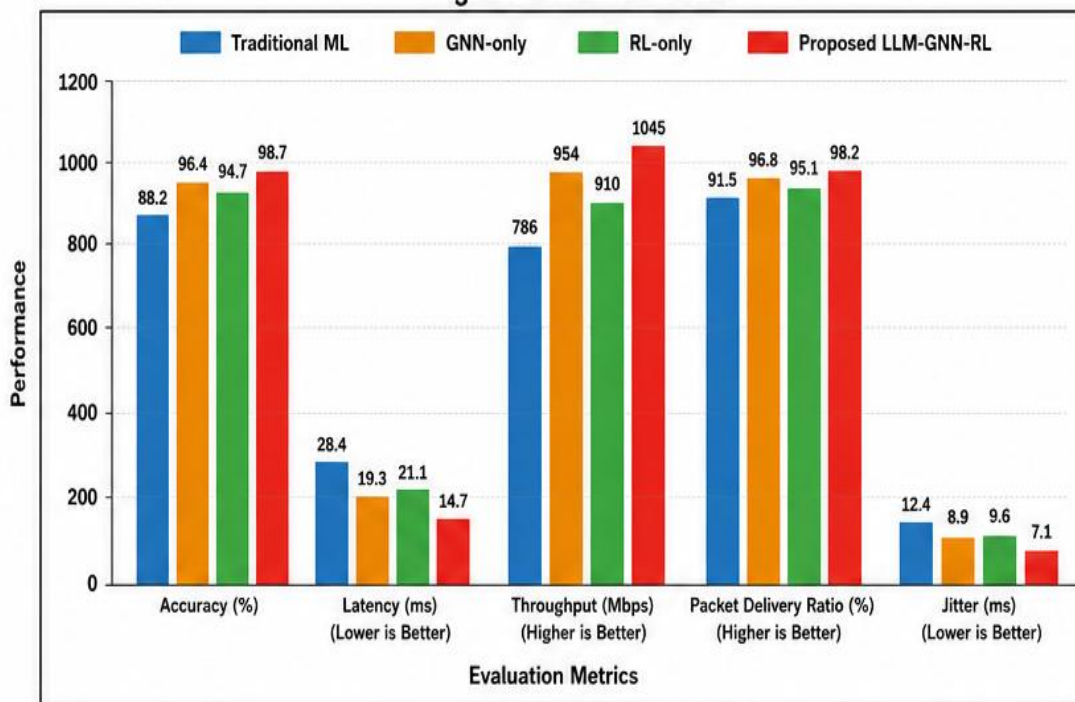


Figure 2. Comparative Performance of the Proposed Framework Against Baseline Models

The proposed integrated framework consistently outperformed the baseline approaches in terms of prediction accuracy and network efficiency.

These findings are important for three reasons. First, they show that semantic information extracted by the LLM contributes positively to traffic understanding and context-aware decision making. Second, the GNN component appears to preserve topological dependencies that are often ignored by ordinary deep learning models. Third, the RL component converts prediction outputs into actionable control strategies that improve live network behavior. Figure 3 illustrates the reinforcement learning optimization cycle used for adaptive network management.

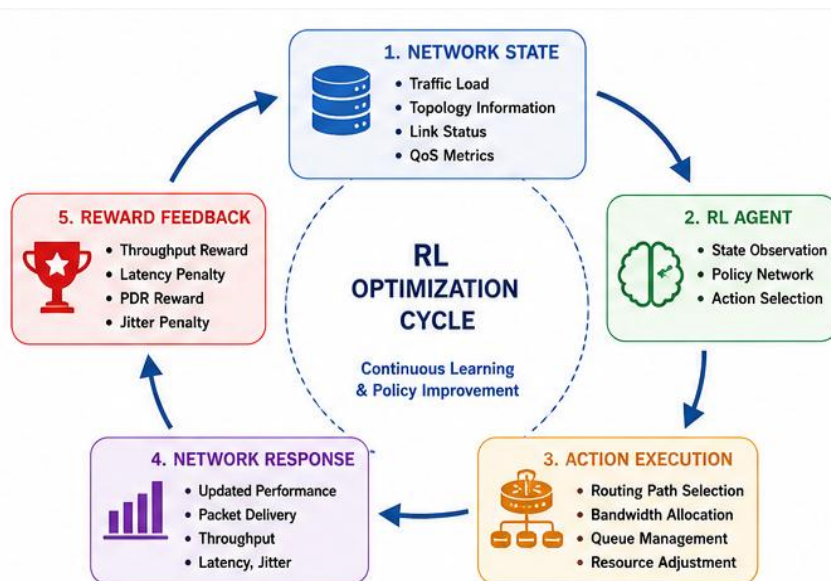


Figure 3. Reinforcement Learning-Based Network Optimization Cycle

The reinforcement learning agent continuously improves optimization policies through interaction with changing network conditions. The study explicitly notes that the integrated system demonstrated superior multi-objective optimization compared with individual components and maintained stable behavior under increasing load.

The results are also consistent with broader research trends. Recent literature has shown that graph-based network models are effective for delay and loss prediction, while RL-based approaches can improve traffic engineering when supported by sufficiently informative state representations. The contribution of the present study is the integration of these ideas with LLM-derived semantics, which may help bridge the gap between raw network telemetry and intelligent control (Panahi et al., 2025).

Although the reported results are promising, they should still be interpreted with some caution. Limitations related to data availability, computational cost, generalization across all network environments, and the technical complexity of integrating the three learning paradigms are realistic concerns, especially when moving from simulation-assisted evaluation to deployment in large operational networks.

5. CONCLUSION

This paper presented an intelligent network traffic prediction and optimization using a hybrid LLM-GNN-RL architecture. The study was motivated by the shortcomings of traditional traffic management methods in dynamic, high-dimensional, and topology-sensitive network environments. By combining semantic feature extraction, graph-based structural learning, and reinforcement-driven control, the framework provides a more complete solution for modern network management.

The results indicate that the integrated framework outperformed traditional machine learning, GNN-only, and RL-only baselines across several metrics. The model reduced prediction error relative to an LSTM baseline and delivered strong gains in accuracy, throughput, latency, packet delivery ratio, and jitter. These outcomes suggest that integrated AI pipelines may be more effective than isolated models for next-generation networking tasks.

In conclusion, the proposed framework contributes to the growing body of work on AI-driven network management by showing how LLMs, GNNs, and RL can be combined in a unified and practical architecture. Future work should focus on lighter model designs, larger real-world deployment studies, explainability, and validation in production-grade 5G and 6G environments.

REFERENCES

1. Abbasi, M., Tahir, M. J., Ahmed, A., Memon, J., and Ali, U. (2024). Recent advances in intelligent network traffic management and analysis. *Computer Networks*, 245, 110402.
2. Bronstein, M. M., Bruna, J., Cohen, T., and Velicković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.
3. CAIDA. (2018). The CAIDA anonymized Internet traces dataset. Center for Applied Internet Data Analysis.
4. Chen, D., Gao, D., Kuang, W., Li, Y., and Ding, B. (2022). pfl-bench: A comprehensive benchmark for personalized federated learning. *Advances in Neural Information Processing Systems*, 35, 9344–9360.
5. Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., et al. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4), 42.
6. Chen, Y., Lam, K. Y., and Li, F. (2025). Large Language Models (LLMs) for Network Traffic Prediction: A Trend-Aware Hybrid Framework. *IEEE Internet of Things Journal*.
7. Ding, Y., Yan, S., Zhang, Y., Dai, W., and Dong, L. (2016). Predicting the attributes of social network users using a graph-based machine learning method. *Computer Communications*, 73, 3–11.
8. Fatima, N., Khan, S., and Rehman, A. (2024). Machine learning assisted traffic management for adaptive computer networks. *Journal of Network and Systems Management*, 32(2), 1–23.
9. Favale, T., Trevisan, M., Drago, I., and Mellia, M. (2021). α -MON: Traffic anonymizer for passive monitoring. *IEEE Transactions on Network and Service Management*, 18(2), 1233–1245.
10. Ferriol-Galmés, M., Suárez-Varela, J., Rusek, K., Barlet-Ros, P., and Cabellos-Aparicio, A. (2023). RouteNet-Fermi: Network modeling with graph neural networks. *IEEE/ACM Transactions on Networking*, 31(5), 2141–2156.
11. Geyer, F. (2017). Performance evaluation of network topologies using graph-based deep learning. In *Proceedings of the 11th EAI International Conference* (pp. 20–27).
12. He, Q., Wang, Y., Wang, X., Xu, W., Li, F., Yang, K., and Ma, L. (2023). Routing optimization with deep reinforcement learning in knowledge defined networking. *IEEE Transactions on Mobile Computing*, 23(2), 1444–1455.
13. He, Y., Zhang, Q., and Lin, X. (2024). Reinforcement learning-based SDN routing scheme combining causal inference and graph learning for QoS-aware networking. *Frontiers in Computational Neuroscience*, 18, 1393025.

14. Huang, X., Zhang, Y., and Li, W. (2024). Enhancing traffic prediction with textual data using large language models. *arXiv preprint arXiv:2405.06719*.
15. Ibrionke, J. (2025). AI-Driven Dynamic Traffic Management in Multi-Domain SDN Networks.
16. Ivey, J., Yang, H., Zhang, C., and Riley, G. (2016). Comparing a scalable SDN simulation framework built on ns-3 and DCE. In *Proceedings of ACM SIGSIM* (pp. 153–164).
17. Kurtuluş, B. (2025). DAG-based DLT integrated cross-network QoS traffic management using RL in SDN (Doctoral dissertation, Ankara University).
18. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
19. Luo, X., Zhao, J., and collaborators. (2023). Temporal graph neural networks for traffic prediction. *arXiv preprint arXiv:2307.00495*.
20. Mahajan, S., HariKrishnan, R., and Kotecha, K. (2022). Prediction of network traffic in wireless mesh networks using hybrid deep learning model. *IEEE Access*, 10, 7003–7015.
21. Maheshwari, H., Yang, L., and Pazzi, R. W. (2025). Machine learning advancements in urban traffic simulation. *IEEE Open Journal of Intelligent Transportation Systems*.
22. Ma, T., Antoniou, C., and Toledo, T. (2020). Hybrid ML and statistical model for traffic forecast. *Transportation Research Part C*, 111, 352–372.
23. MAWI Working Group. (2026). MAWI traffic archive. WIDE Project.
24. Michailidis, P., Michailidis, I., Lazaridis, C. R., and Kosmatopoulos, E. (2025). Traffic signal control via RL. *Infrastructures*, 10(5), 114.
25. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., and Gao, Y. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
26. Oliveira, T. P., Barbar, J. S., and Soares, A. S. (2016). Network traffic prediction: Traditional vs deep learning. *International Journal of Big Data Intelligence*, 3(1), 28–37.
27. Panahi, P. H. S., Jalilvand, A. H., and Najafi, M. H. (2025). QoE-driven anomaly detection. *Authorea Preprints*.
28. Pei, X., Zhang, H., and collaborators. (2024). Efficient routing in SDN using DRL. *Computer Networks*, 243, 110310.
29. Rusek, K., Suárez-Varela, J., Almasan, P., Barlet-Ros, P., and Cabellos-Aparicio, A. (2019). RouteNet: GNN for network modeling. *ACM CoNEXT*, 166–180.
30. Sankar, A., Zhang, X., and Chang, K. C. C. (2019). Meta-GNN. *IEEE/ACM ASONAM*, 137–144.
31. Sethuraman, P., and Chennareddy, R. K. (2022). Vehicular traffic prediction using learning models. *IJETCSIT*, 3(2), 111–121.
32. Solozabal, R., Ceberio, J., Sanchoyerto, A., Zabala, L., Blanco, B., and Liberal, F. (2019). VNF placement with DRL. *IEEE JSAC*, 38(2), 292–303.
33. Stoenescu, R., Olteanu, V., Popovici, M., Ahmed, M., Martins, J., Bifulco, R., et al. (2015). In-net processing. *EuroSys*, 1–15.
34. Sun, G., Wang, Y., Niyato, D., Wang, J., Wang, X., Poor, H. V., and Letaief, K. B. (2024). LLM-enabled graphs in networking. *IEEE Network*, 39(4), 290–301.
35. Tu, W., Li, J., Xiao, F., Wang, X., and Lu, Y. (2026). Integrating LLMs into traffic systems. *Entropy*, 28(2), 211.
36. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. *arXiv preprint arXiv:1710.10903*.
37. Verma, S., Kadadi, S., Jayaprakash, S., Mahapatra, A. K., and Jain, I. (2024). RouteNet-Fermi re-implementation. *arXiv preprint arXiv:2412.05649*.
38. Wei, C., Tu, S., Hata, D., Hasegawa, T., Koizumi, Y., Ramakrishnan, K. K., et al. (2026). immUNITY: Detecting low-volume attacks. *arXiv preprint arXiv:2603.20573*.
39. Xiao, Y., Liu, J., Wu, J., and Ansari, N. (2021). DRL for traffic engineering: A survey. *IEEE Communications Surveys & Tutorials*, 23(4), 2064–2097.
40. Yang, Q., Zhang, Q., Zhang, C., and Zhang, X. (2022). Interpretable relation learning. *WSDM*, 1266–1274.
41. Zhao, Y., Chen, H., and collaborators. (2024). LLMs in traffic forecasting. *arXiv preprint arXiv:2412.12201*.
42. Zhang, J., Ye, M., Guo, Z., Yen, C. Y., and Chao, H. J. (2020). CFR-RL in SDN. *IEEE JSAC*, 38(10), 2249–2259.
43. Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., et al. (2024). LLM for telecommunications survey. *IEEE Communications Surveys & Tutorials*, 27(3), 1955–2005.
44. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2020). Graph neural networks: A review. *AI Open*, 1, 57–81.