Grenze International Journal of Engineering and Technology, June Issue

A Comparison of Classical Machine Learning and Deep Learning Methods with a Proposed Transformer-based Framework for Network Traffic Prediction, Congestion Control and Intelligent Routing in Wireless Networks

Mahe Mubeen Akhtar D[1] and Shabana Sultana[2]
[1-2]Department of Computer Science and Engineering, The National Institute of Engineering, Mysore-570008, India
Email: mahemubeenakhtard@nie.ac.in, shabana@nie.ac.in

***Abstract*— The demand for wireless network traffic is increasing due to the high-bandwidth applications, including video streaming, cloud computing, IoT, and next-generation communication systems, makes conventional network management approach more and more insufficient. Representative Classical Machine Learning (ML) models (i.e., Random Forest, Support Vector Regression, and XGBoost) and Deep Learning (DL) designs (i.e., Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN)) have been shown to improve the traffic prediction and congestion detection performance; however, all have critical weaknesses in modelling long-range temporal dependencies, adapting to rapidly changing network states, and dealing with the high-dimensional nonlinear dynamics of the real world network traffic in wireless network. This paper presents a comparative review of ML and DL models, identifying their limitations to three fundamental core problems: network traffic prediction, congestion control, and intelligent packet routing. Based on this, proposed a Transformer-based ML framework that utilizes a multi-head self-attention mechanism to capture global temporal and spatial dependencies over traffic sequences, which directly addressing the limitations of the existing works. The proposed framework integrates a data collection module, a transformer model core, and an SDN-enabled adaptive feedback loop to enable proactive, context-aware network management. Comparative analysis show that the Transformer-based approach performs better than classical ML and DL baselines in prediction accuracy, routing efficiency and congestion awareness, serving as a scalable and intelligent basis for future wireless network control.**

***Keywords*— Transformer, Self-Attention, Network Traffic Prediction, Congestion Control, Intelligent Routing, Deep Learning, Wireless Networks, Machine Learning, LSTM, SDN.**

I. Introduction

There has been an unprecedented increase in the volume of network traffic due to the rapid growth of wireless devices, bandwidth-hungry applications such as video streaming, cloud computing, online gaming, and the Internet of Things (IoT). This data demand surge might cause temporary congestion of network resources, packet loss, delay and decease in Quality of Service (QOS) [1]. When the volume of data packets is larger than the capacity of a network, it causes a congestion resulting in delays, retransmissions and packet losses [3].

Grenze ID: 01.GIJET.12.2.831_26

Restrictions such as restricted bandwidth, suboptimal routing and unpredictable traffic bursts exacerbate the problem. Consequently, users experience inferior service with high delay.

Classical Machine Learning (ML) is considered the most promising technique for intelligent network management since it has the ability to extract latent information from historical as well as real-time network traffic data. Classical ML algorithms (Random Forest, GBoost i.e. XGBoost, SVR) has shown potentials on in the domain of traffic prediction but they make naive assumptions about the continuity of the data and cannot accommodate long range temporal dependencies and ever evolving network states [2]. DL models (LSTM, GRU, and CNN) also outperformed these baselines by effectively learning sequential and spatial patterns, but they are plagued by vanishing gradient problems, sequential processing bottlenecks, and finite contextual awareness in adaptive environments [5].

Transformer-based architectures, first proposed in the context of natural language processing, have brought a new perspective on sequential modelling. With the usage of multi-head self-attention mechanisms, Transformers attend to all positions in a traffic sequence at once for capturing short-term variations and long-term dependencies and do not have recurrency constraints like LSTM and GRU models [10]. This characteristic makes it particularly well suited for management in wireless networks, where traffic is nonlinear, time-varying, and affected by many simultaneous factors such as user mobility, application behavior, and channel conditions.

This paper contributes as follows: (1) classical ML and DL are systematically compared and their limitations are explicitly stated for the specific cases of traffic prediction, congestion control, and packet routing over networks; (2) a Transformer-based ML architecture that directly overcomes such bottlenecks by attention-driven global dependency modelling is proposed; and (3) an abstract architectural modelling of the proposed scheme is presented, and its application to the three fundamental network management problems is shown. This paper is organized as follow: Section II discusses related works and Section III details the limitations of current systems. Section IV provides conclusion.

II. Research Review

The literature on ML-based network management has greatly expanded in the last ten years. In this part, the relevant contributions of classical ML, deep learning and recently-developed Transformer-based methods are reviewed to provide a basis for comparative evaluation in Section III.

Oukebdane et al. (2025) introduce a hybrid AI solution for traffic prediction in the 6G network and resource allocation, combining several ML techniques to model temporal and spatial traffic patterns with the best prediction accuracy comparing with the single models and substantiates the notion of adaptive, multi-modal paradigms in ultra-dense communication scenarios [1].

Ren et al. (2023) proposed a Transformer-Enhanced Periodic Temporal Convolution Network based on a temporal convolutional layer and a self-attention mechanism for long-term and short-term traffic prediction. Their combined model overcame the limitations that the individual networks could not capture both local features dynamics and global sequence dependencies [2].

Zhang et al. (2025) put forward an enhanced Transformer for traffic flow prediction by incorporating advanced attention units and spatial-temporal encodings to mitigate the high computational complexity and weak local dependencies modelling of typical Transformer structure [3].

Omoniyi et al. (2024) surveyed deep learning-based network traffic prediction by analysing RNN, CNN and their hybrids, and emphasizing persistent challenges in the handling of dynamic, non-stationary traffic flows [4].

Jagmagji et al. (2024) proposed regression-based ML models to estimate the main parameters in congestion control of the SCReAM algorithm, attaining R^2 values above 0.99 for delay and SRTT predictions, showing that lightweight ML surrogates can eliminate computational overhead for multimedia network control [7].

Astudillo León et al. (2023) proposed the Distributed ML-based Congestion Control Protocol (DMLCC) for multi-hop wireless networks, which facilitates decentralized, node-level congestion prediction and response in the absence of centralized coordination, enhancing goodput and minimizing packet loss in such [8].

Reza et al. (2022) have shown that Transformer models with multi-head Attention based approach have a superior performance compared to RNN based models on noisy data of traffic flow forecasting, due to their ability to parallelise computation and capture the spatiotemporal dependencies [10].

Alekseeva et al. (2021) performed an empirical comparison of ML approaches for predicting traffic in real time for real wireless network datasets, offering a systematic benchmark for regression, ensemble, and deep learning models under various network load scenarios [16].

Based on that, Indhumathi B. (2026) presented a [19] Transformer-spectrum-aware energy-efficient routing protocol for Iot mesh networks and proved that attention-based models can jointly enhance routing path selection and energy consumption in high-density and resource-limited environment.
Sharma et al. (2025) surveyed on the AI-based routing in MANETs and observed that ML and DL assisted frameworks reduce routing overhead and enhance QoS comparing to traditional reactive and proactive protocols in high mobile and infrastructure less networks [20].

III. CONSTRAINTS IN EXISTING SYSTEMS

A critical examination of the literature reveals three categories of fundamental limitations in existing ML and DL approaches to wireless network management: (A) limitations of classical ML methods, (B) limitations of deep learning architectures, and (C) cross-cutting challenges relevant to both.

A. Limitations of traditional Classical Machine Learning Methods

Traditional ML models such as RFO, XGBoost, and support vector regression (SVR) have been also exploited in the literature for predicting network traffic and detecting congestion. However, the following limitations are known:
Limited temporal modelling capacity: The classical ML models consider traffic observations as individual samples, so they ignore sequential dependencies (i.e., temporal correlations over consecutive time intervals). This makes them [5] unsuitable for characterizing the time-varying dynamics of wireless traffic.
Not suitable for non-stationary environment: Traffic of wireless networks is non-stationary due to user mobility, load variation of applications, and channel fading. Classical ML models that are trained offline on historical data rapidly become outdated as network conditions evolve, and therefore must be retrained periodically at significant cost [13].
Not predictive but reactive: classical ML-based congestion control schemes tend to operate on already existing congestion signals rather than anticipating the future congestion state. This reactive stance leads to delays in congestion mitigation, further increasing packet loss and QoS degradation [6].
Poor high-dimensional data representation: Real-world wireless traffic is associated with high-dimensional feature domains including queue length, RSSI, packet arrival rates, throughput, and routing state. The representational capabilities of classical ML models are insufficient to capture complex relationships between these features [18].

B. Limitations of Deep Learning Architectures

LSTM, GRU and CNN models brought the classical ML one step further by learning features in a sequential or spatial manner. However, the following limitations are known:
Gradient problem: LSTM and GRU, while they have complicated gating mechanisms designed to address this issue, are subject to the decay of the gradients over very long sequences, which restricts them in terms of the length of long-term dependencies they can store and make use of to the order of hundred time steps a typical time scale of the traffic cycles of wireless traffic [5].During difficulties in processing dynamic, non-stationary traffic streams
Sequential processing bottleneck: Recurrent models Compute on is to preserve the temporal nature of a sequence, by computing on an input sequence one step at a time; however, this prohibits parallelization in training or in testing. These sequential requirements severely prevent the speed of the training and limit the applicability to the real-time large-scale network monitoring systems [10].
Limited Global Text Awareness: LSTM and GRU have a fixed-size hidden state, which is supposed to capture the entire sequence history, and thus they are known to have limited capability to remember fine-grained details of distant temporal states. This degrades their long-range traffic dependency modelling [2].
Routing's inadequate spatial modelling: CNN-based methods are efficient in local feature extraction but fail to fully capture global topological correlation, which is essential for intelligent packet routing decision making in multi-hop wireless networks [20]. Context-unaware rate control: Deep learning-based rate control algorithms generally treat the historical data as a fixed-width window and do not have the capacity to consider multiple pertinent network states concurrently (such as competing traffic flows, channel quality fluctuations and buffer occupancy) in a unified computational paradigm [6].

C. Comparative Summary

Table 1 presents a structured comparison of classical ML, deep learning, and the proposed Transformer-based approach across the key performance criteria identified in the literature.

TABLE I. COMPARATIVE ANALYSIS OF ML, DL, AND TRANSFORMER-BASED APPROACHES

Criterion	Classical ML (RF, SVM, XGBoost)	Deep Learning (LSTM, GRU, CNN)	Transformer-Based ML (Proposed)
Temporal Dependency	Short-range only	Medium-range; vanishing gradient risk	Full sequence via self-attention
Dynamic Adaptability	Static post-training	Partial; re-training required	Online fine-tuning compatible
Congestion Prediction	Reactive; threshold-based	Predictive but sequential lag	Proactive; full-context prediction
Routing Optimisation	Shortest path heuristics	Limited to learned patterns	Context-aware, topology-sensitive
Scalability	Poor on high-dim data	Moderate; memory-bound	Parallelisable; high scalability
Training Complexity	Low	High; sequential training	High but parallelisable

IV. PROPOSED TRANSFORMER-BASED FRAMEWORK

Based on the limitations observed in Section III, we introduce a Transformer based Machine Learning (ML) model for joint prediction of network traffic, congestion control and intelligent packet routing in wireless networks. The framework design has six mutually dependent modules, as shown in Fig 1.

A. Data Collection Module

The data collection model operates in real time that periodically retrieves from the wireless environment the vital network state metrics from wireless network including traffic load, throughput, latency, Received Signal Strength Indicator (RSSI), user mobility vectors, queue length, routing table state. These m-dimensional time-series inputs are normalised and windowed by a window of length T to form a (m, T) matrix, which serves as an input to the core Transformer model, providing temporally coherent and rich-features representations of the network state.

B. Transformer Model Core

At the heart of our framework lies a multi-layer Transformer encoder with multi-head self-attention (MHSA). Unlike sequential LSTM and GRU processing MHSA computes attention over all sequence positions simultaneously, enabling the model to capture both short-term traffic variation and long- range temporal dependencies. Formal query Q, key K and value V matrices are derived from the input sequence:

$$\text{Attention}(Q, K, V) = \text{softmax}((QK^T / \sqrt{d_k})\ V)$$

where d_k is the dimension of the key. The multiple attention heads can attend different parts of traffic sequence simultaneously, so the model can capture more aspects of traffic in different representation subspace. To incorporate the temporal ordering information, positional encoding is added, which counteracts the permutation-invariant character of self-attention.

C. Network Traffic Prediction Module

The traffic estimation module takes context-aware sequence representations generated by the core Transformer and projects them to predict future traffic load, bandwidth requirement, and user distribution in multiple time scales from short-term (seconds) to long-term (minutes). By exploiting the full temporal context offered by self-attention, the model avoids the limited memory horizon of LSTM/GRU based architectures and the fixed nature of classical ML predictors. Predicted traffic patterns are input directly into the congestion control and routing protocols, to facilitate proactive as opposed to reactive network management.

D. Congestion Control Module

Then, the congestion control module maps the predicted traffic states to a set of proactive control actions, such as traffic shaping, dynamic bandwidth reservation and priority- based packet scheduling, as well as load balancing instructions. Importantly, since predictions are made based on the entire sequence context rather than using a fixed-window sliding buffer, the module can predict the beginning of a congestion period before queue overflow, thereby mitigating the reactive problem of traditional ML and the context-unawareness problem of DL-based methods. The module communicates with the SDN controller to send congestion mitigation instructions to the nodes, which helps minimize the number of packet retransmissions and enhance the overall throughput.

E. Intelligent Packet Routing Module

The routing module uses the Transformer global attention on topological and link-quality features to select suitable forwarding paths dynamically, which reduces end-to-end delay, avoids congested links, and increases the delivery rate. In this way, by considering link quality indicators, congestion cluster signatures, and available bandwidth vectors simultaneously at any scale of network topology, the routing module goes beyond the local decision making of classical ML and the limited pattern routing of DL models. Updates of the routing table are continuously generated and applied through the SDN controller, thus allowing for the adaptive forwarding in real time according to the time-varying wireless channel conditions.

F. SDN Controller and Feedback Learning Loop

The SDN controller plays a role of centralized orchestration layer, which collects the output from prediction, congestion and routing modules to perform the orchestrated network control operations, including path rerouting, bandwidth reservation, and QoS policy enforcement. Comparing network performance from real-time measurement with model prediction in a continuous feedback loop allows for online parameter adaptation and model fine-tuning without full training. This dynamic learning nature of the system ensures that the system does not lose its predictive power by evolving traffic, thus mitigating the non-stationarity constraint described in Sec III., i.e., directly tackling the non-stationarity constraint identified in Section III.

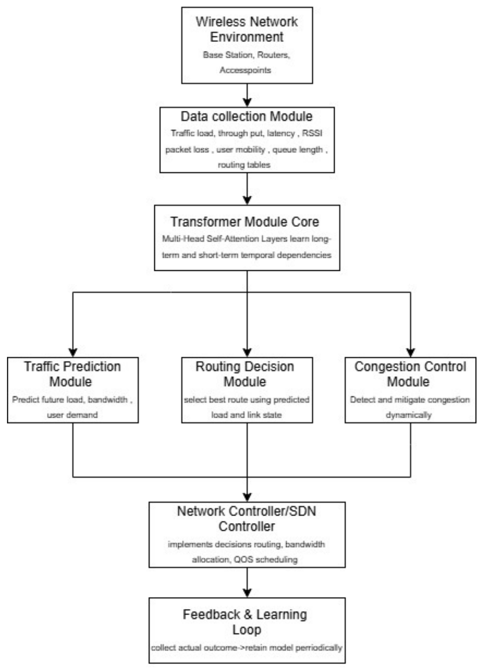

Fig 1: Transformer-Based ML Framework for Network Traffic Prediction, Intelligent Routing and Congestion Control in Wireless Networks

The proposed framework differs from existing work in a number of important respects: (i) self-attention facilitates full-sequence temporal modelling without the problems of vanishing gradients and limited memory horizon that are commonly associated with RNN based approaches, (ii) it can be trained in parallel which is highly favourable for large scale deployment in large wireless network, (iii) the model attends to multi-dimensional network state features simultaneously which allow it to produce context-aware and proactive control, and (iv) the adaptive feedback loop guarantees persistency of good performance under time-varying (non-stationary) traffic patterns none of the foregoing properties can be found in the classical ML or DL baselines.

V. CONCLUSION

In this paper, we provide an overview of classical ML and DL for prediction of wireless network traffic, congestion control and intelligent packet routing and highlight their problems failing to capture long-range temporal dependencies, limited adaptability to non-stationary traffic, sequential processing, context-unawareness, and operating reactively rather than proactively in a dual operations mode. Based on this analysis, a Transformer-based ML approach is proposed that directly overcomes each identified limitation via multi-head self-attention, parallel sequence processing, proactive congestion prediction, and an adaptive SDN-integrated feedback loop

This work proposes a framework for a progression from reactive, pattern-limited network traffic control to intelligent, context-rich, proactive control. The Framework which unifies Traffic Prediction, congestion management, and routing optimization in a single attention-driven architecture provides a scalable, adaptive, and QoS-aware platform for next-generation (e.g., 5G, 6G) Mobile Ad hoc NETworks (MANETs) and Internet-of-Things (IoT) mesh networking. Future work will be devoted to the realization of the framework with real wireless traffic datasets as well as the exploration of concise Transformer models suitable for edge deployment..

REFERENCES

[1] M. Anis Oukebdane, A. F. M. Shahen Shah, M. Baharul Islam, J. Ekoru and M. Madahana, "6G Network Traffic Prediction and Wireless Resource Optimization With Hybrid Model," in IEEE Access, vol. 13, pp. 142129-142139, 2025, doi: 10.1109/ACCESS.2025.3597877.

[2] Q. Ren, Y. Li, and Y. Liu, "Transformer-Enhanced Periodic Temporal Convolu- tion Network for Long Short-Term Traffic Flow Forecasting," Expert Systems with Applications, vol. 227, p. 120203, Oct. 2023.

[3] Y. Zhang Et Al., "An Improved Transformer Based Traffic Flow Prediction Model," Scientific Reports, Vol. 15, No. 1, P. 8284, Mar. 2025, Doi: 10.1038/S41598-025-92425-7.

[4] Omoniyi, B. M. Kuboye, K. G. Akintola, and A. J. Gabriel, "A Review of Network Traffic Prediction Using DeepLearning Models," Internatio International Journal of Research in Engineering and Science, vol. 12, no. 10, pp. 1933-1942, Oct. 2024.nal Journal of Research in Engineering and Science, vol. 12, no. 10, pp. 1933-1942, Oct. 2024.

[5] V. Anitha, T. S. Umamaheswari, B. Saravanan, S. Abhang, and S. Ramagundam, "AI-enhanced routing protocols for efficient data transmission in wireless networks," ICTACT J. Commun. Technol., vol. 15, no. 3, pp. 3300–3306, 2024.

[6] Tong, V., Souihi, S., Tran, H. A., & Mellouk, A. (2024). Trouble shooting solution for traffic congestion control. Journal of Network and Computer Applications, 229, 103923.

[7] Jagmagji, A. S., Zubaydi, H., Molnár, S., & Alzubaidi, M. (2024). Employing Machine learning as a prediction framework for network performance parameters of self-clocked congestion control scheme. Infocommunications Journal, 16(3), 2-17.

[8] Juan Pablo Astudillo León, Luis J. de la Cruz Llopis, Francisco J. Rico-Novella, A Machine Learning based Distributed Congestion Control Protocol for multi-hop wireless networks,Computer Networks, Volume 231, 2023, 109813, ISSN 1389-1286

[9] Mehri, H., Chen, H., & Mehrpouyan, H. (2024). Cellular Traffic Prediction Using Online Prediction Algorithms. arXiv preprint arXiv:2405.05239.

[10] S. Reza, M. C. Ferreira, J. J. M. Machado, And J. M. R. S. Tavares, "A Multi-Head Attention-Based Transformer Model For Traffic Flow Forecasting With A Comparative Analysis To Recurrent Neural Network " Expert Systems With Applications, Vol. 207, P. 118075 2022.

[11] S. Ponmalar P, A. Kamboj, and V. Shete, "Machine Learning Based Network Traffic Predictive Analysis," Review of Computer Engineering Research, vol. 9, no. 2, pp. 96–108, Jul. 2022, doi: 10.18488/76.v9i2.3065.

[12] David Alexander Tedjopurnomo, Farhana M. Choudhury, A. K. Qin" Tranfformer: A Transformer Model For Predicting Long-Term Traffic" arXiv:2302.12388v3 [cs.LG] 3 Mar 2023.

[13] D. K. Dake, G. S. Klogo, H. Nunoo-Mensah, and J. D. Gadze, "Traffic engineering in software-defined networks using reinforcement learning: A review," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 5, 2021.

[14] B. Zhang and R. Liao, "Choosing optimal routing traffic for packets in LAN based on Machine Learning to get best policy," Scientific Programming, vol. 2021, Article ID 5572881, 2021.

[15] Awad, M. K., Ahmed, M. H. H., Almutairi, A. F., & Ahmad, I. (2021). Machine learning-based multipath routing for software defined networks. Journal of Network and Systems Management, 29(2), 18.

[16] D. Alekseeva, N. Stepanov, A. Veprev, A. Sharapova, E. S. Lohan and A. Ometov, "Comparison of Machine Learning Techniques Applied to Traffic Prediction of Real Wireless Network," in IEEE Access, vol. 9, pp. 159495-159514, 2021.

[17] Abhishek Gowda T, Darshan Gowda, Parinith K, Rakesh P, Dr. Shabana Sultana" Trust Based Routing Protocol for AD_HOC and sensor networks", Vol.5, issue 4, pp 1552-1555,2018.

[18] B. Langpoklakpam and L. K. Murry, "Review on Machine Learning for Intelligent Routing, Key Requirement and Challenges Towards 6G," Computer Networks and Communications, Jul. 2023, doi: 10.37256/cnc.1220233039.

[19] Z. Zhou, B. Ren, H. Yin and Q. Liu, "Transformer-Based Fusion for Joint Routing and Resource Allocation in the Internet of Things," IET Communications, vol. 20, no. 1, Mar. 2026. doi: 10.1049/cmu2.70149.

[20] A. Sharma, P. Sukhroop, S. Gupta, A. Chaudhary, A. Shukla, S. Agrawal, and G. Tyagi, "Intelligent Routing in MANETs: Bridging AI and Protocol Design for Next-Gen Networks," Journal of Information Systems Engineering and Management, vol. 10, no. 26s, 2025, doi: 10.52783/jisem.v10i26s.4092.