Federated DeepSeek-Assisted Traffic Prediction and Anomaly Detection in Edge-Enabled Mobile Networks

Meaish Tandon
Department of Computer Science, George Mason University, Fairfax, VA, USA.
manishmail@gmu.edu

Abstract

The exponential growth of mobile data traffic and the proliferation of latency-sensitive applications demand intelligent, real-time traffic forecasting and anomaly detection mechanisms that operate close to the data source. Centralized cloud-based learning architectures incur prohibitive communication overhead, privacy vulnerabilities, and regulatory friction. This paper presents a federated learning framework that integrates DeepSeek, a large language model with strong sequence modeling and spatial-temporal reasoning capabilities, to enable collaborative traffic prediction and anomaly detection across edge nodes in mobile networks. The proposed system architecture distributes model fine-tuning across base stations and user equipment, employing parameter-efficient adaptation and communication-efficient aggregation to manage the large model footprint and heterogeneous client resources. We dissect the structural trade-offs between global prediction fidelity and communication cost, examining hierarchical aggregation, differential privacy, and resilience against adversarial perturbations. By coupling federated learning with DeepSeek, mobile operators can exploit complex long-range temporal dependencies and geographic correlations while preserving data locality, aligning with data governance frameworks, and enhancing infrastructure robustness. The analysis extends to fairness issues arising from non-identically distributed traffic patterns across diverse urban and rural regions, and discusses policy implications for cross-jurisdictional deployments. Deployment considerations address resource-constrained edge hardware, energy sustainability, and system reliability under dynamic network conditions. This study provides an interdisciplinary blueprint for a secure, scalable, and responsible socio-technical system that harnesses large-scale pre-trained models for next-generation mobile network intelligence, emphasizing the interplay between algorithmic innovation, governance, and operational sustainability.

Keywords

Federated learning, mobile edge computing, traffic prediction, anomaly detection, DeepSeek, spatial-temporal modeling, network governance.

1. Introduction

The relentless expansion of mobile broadband, the Internet of Things, and high-definition streaming services has generated an unprecedented surge in mobile data traffic, placing severe strains on cellular infrastructure and necessitating intelligent resource orchestration [1]. Accurate traffic prediction is foundational for proactive network management, including dynamic spectrum allocation, load balancing, and energy-aware base station activation, all of which directly impact quality of service and operational costs [2]. However, conventional centralized machine learning pipelines aggregate raw traffic logs from geographically

dispersed cells into cloud data centers for model training. This paradigm introduces significant latency, consumes core network bandwidth, and exposes sensitive user-mobility fingerprints to centralized storage, raising profound privacy and regulatory concerns. The emergence of strict data protection regimes such as the European Union's General Data Protection Regulation has further compelled network operators to seek architectures that preserve data locality and comply with cross-border data flow restrictions.

Federated learning (FL) has emerged as a transformative paradigm that enables collaborative model training without centralizing raw data, by instead aggregating locally computed model updates [3]. In the context of mobile networks, FL can be deployed directly on edge servers co-located with base stations or even on capable user equipment, thereby drastically reducing data movement and latency. Simultaneously, large language models such as DeepSeek have demonstrated extraordinary sequence modeling and generalization capabilities across diverse modalities, extending well beyond natural language [4]. Recent research has shown that pre-trained language models can be effectively repurposed for time series forecasting, leveraging their self-attention mechanisms to capture long-range temporal dependencies and complex non-linear interactions [5]. The convergence of FL and large pre-trained models, however, introduces a distinct set of systems challenges, including the need to adapt billion-parameter architectures to resource-limited edge devices, to compress gradient updates, and to maintain statistical efficiency under non-identically distributed local data partitions [6].

To address the communication bottleneck, hierarchical federated learning architectures that perform intermediate model aggregation at edge nodes before a final global synchronization have been proposed, effectively decoupling the wide-area network from local clusters [7]. These structures align naturally with the multi-tier topology of mobile networks, where baseband units, multi-access edge computing platforms, and regional data centers form a hierarchical aggregation tree. Within this setting, recent work integrating DeepSeek into spatial-temporal traffic prediction has achieved notable accuracy improvements over traditional graph neural network and recurrent neural network baselines, indicating that the model's multi-head attention and mixture-of-experts design can effectively capture the complex geographic and temporal correlations inherent in mobile traffic [8]. Building on these advances, this paper proposes a federated DeepSeek-assisted system in which a shared global backbone is collaboratively fine-tuned across edge nodes for both traffic forecasting and anomaly detection. The remainder of the paper presents an in-depth analysis of the architectural foundations, communication-efficient aggregation strategies, anomaly detection mechanisms, governance and fairness considerations, and deployment sustainability of such a system, offering a comprehensive interdisciplinary perspective.

2. Architectural Foundations of Edge-Enabled Mobile Traffic Intelligence

Mobile network architectures have evolved from monolithic core-centric designs to distributed, edge-enabled infrastructures where computing capabilities are pushed to the radio access network. Multi-access edge computing platforms, deployed in close proximity to base stations, provide low-latency, high-throughput environments suitable for hosting machine learning inference and lightweight training tasks [1]. Traffic prediction and anomaly detection, which historically relied on batch processing of call detail records in centralized operations support systems, can in this new paradigm be executed continuously on streaming telemetry data without violating subscriber privacy. The key architectural challenge lies in designing a learning topology that reflects the natural hierarchy of the mobile network: individual cells

generate fine-grained traffic measurements, clusters of cells are managed by edge aggregation points, and regional orchestration nodes coordinate policy enforcement.

The spatial-temporal dynamics of mobile traffic are governed by recurring diurnal patterns, mobility-induced handovers, and unpredictable events such as flash crowds or network failures. Traffic prediction models must therefore simultaneously capture local periodicity, spatial interactions among adjacent cells, and abrupt regime changes. Conventional approaches based on long short-term memory networks, temporal convolutional networks, and graph neural networks require substantial labeled data and careful feature engineering, and they often struggle to generalize across diverse geographies [2]. A centralized model trained on aggregated data from multiple regions can benefit from rich cross-domain patterns but exposes raw mobility traces, creating a tension between model accuracy and data governance. Federated learning resolves this tension by enabling each edge node to retain its data locally while participating in a collaborative training process that yields a shared global model.

The introduction of large language models like DeepSeek into this domain represents a paradigm shift. Unlike domain-specific architectures that are tailored to a fixed input dimensionality and feature set, DeepSeek's transformer backbone can ingest tokenized time series and contextual metadata, learning universal representations that transfer across cells with minimal fine-tuning [4]. By treating time-stamped traffic metrics as sequences and utilizing positional encodings, the model can learn to attend to both recent trends and distant historical patterns, a capability that is critical for predicting traffic surges and detecting subtle anomalies. However, the sheer parameter count of DeepSeek—measured in tens or hundreds of billions—makes full model federation impractical. Therefore, the system architecture must incorporate parameter-efficient fine-tuning methods such as low-rank adaptation or adapter modules, which add only a tiny fraction of trainable parameters on top of the frozen base model. These lightweight adaptations can then be aggregated across clients, drastically reducing per-round communication payloads while preserving the expressive power of the pretrained foundation.

3. Federated Learning Paradigm with DeepSeek Language Model Assistance

Federated learning operates on the principle that clients compute local model updates on their private data and transmit only these updates to a central server for aggregation, typically via the Federated Averaging algorithm [3]. In the context of mobile networks, clients can be edge servers collocated with base stations, each holding traffic volume time series for a set of cells. The non-IID nature of traffic data—where a downtown business district exhibits sharp morning and evening peaks while a suburban residential area shows flatter daytime usage—poses convergence challenges for standard FL algorithms. Model personalization techniques, such as Federated Multi-Task Learning and FedProx, mitigate these effects by allowing local models to diverge partially from the global consensus, but they do not address the communication overhead of transmitting large model parameters.

DeepSeek's architecture, which employs a Mixture-of-Experts design and multi-query attention, achieves state-of-the-art performance on sequence modeling tasks while maintaining computational efficiency relative to dense transformer models of comparable size [4]. The model's ability to be pretrained on vast, diverse corpora and then adapted to specific domains with limited data aligns with the needs of mobile traffic analysis, where labeled anomaly examples are scarce and traffic patterns vary across regions. Recent studies on adapting large language models to time series have demonstrated that a frozen pretrained

backbone augmented with a small trainable stem and head can outperform dedicated time series architectures, leveraging the rich temporal representations learned during language pretraining [5]. This observation motivates the use of DeepSeek as a shared foundation model in a federated setting, where the global backbone remains frozen and only lightweight adaptation layers and prediction heads are trained collaboratively.

The integration of federated learning with large language models introduces systemic trade-offs that must be carefully managed. Transmitting gradients or deltas for adapter layers reduces communication costs by orders of magnitude compared to full model fine-tuning, yet it introduces a new form of bias: the frozen backbone may encode language-centric priors that are suboptimal for traffic data, and the capacity of adaptation layers limits how much the model can specialize to fringe scenarios [6]. Hierarchical aggregation adds another layer of complexity. An edge aggregation node first consolidates updates from multiple base stations within a region, producing a regional model that is subsequently merged into the global model [7]. This two-tier process can reduce wide-area communication frequency and improve resilience to node churn, but it also risks amplifying regional biases if intermediate aggregation is performed over statistically homophilous groupings. The architectural choice between synchronous and asynchronous global aggregation further impacts convergence speed and robustness to stragglers.

The TIDES framework, a recent centralized spatial-temporal prediction system built on DeepSeek, demonstrated that incorporating auxiliary metadata such as cell tower coordinates, day-of-week embeddings, and public event calendars into the token stream can significantly boost prediction accuracy [8]. In a federated extension, each edge client would encode its local metadata and traffic readings into the same tokenized format, fine-tune decoder layers locally, and send the adapted weights upward. This design preserves the spatial generalization benefits of DeepSeek's attention mechanism while respecting data sovereignty. Moreover, the federated lifecycle can be orchestrated such that initial global pretraining occurs in a data center using publicly available traffic datasets or synthetic traces, after which the model is distributed to edge nodes for continual federated fine-tuning, realizing a virtuous cycle of central pretraining and decentralized adaptation.

4. Communication-Efficient Aggregation and Model Personalization

Communication efficiency is the linchpin of any practical federated learning deployment in mobile edge networks. While parameter-efficient fine-tuning techniques drastically shrink the per-client payload, the cumulative overhead of frequent synchronization rounds over constrained backhaul links remains nontrivial. Strategies such as gradient compression via sparsification or quantization, periodic aggregation with local fine-tuning between rounds, and the use of knowledge distillation to transfer knowledge from a larger teacher to a more compact student model are all applicable. In the case of DeepSeek, the inherent modularity of mixture-of-experts offers a unique compression route: active expert routing can be analyzed to identify rarely used experts, which can be pruned or frozen during federation, further reducing communication and local computation.

Model personalization is essential to handle the heterogeneous traffic profiles of different cells while still benefiting from the shared representation learned across the network. A purely global model would underfit tail distributions—such as those of stadiums during events or rural highways during holiday migrations—while a purely local model would miss the transferable patterns that characterize human mobility and network usage. Hybrid personalization schemes, wherein each client maintains a local copy of the prediction head

while the shared backbone aggregates only intermediate representations, offer a pragmatic middle ground. The DeepSeek-assisted system can exploit this by treating the deep attention layers as a shared representation encoder and fine-tuning low-rank adaptations that are client-specific, allowing the model to preserve common spatial-temporal patterns while accommodating local idiosyncrasies.

Hierarchical aggregation and partial model sharing also reduce the attack surface for privacy leakage. Even though FL does not expose raw data, model updates can inadvertently encode sensitive information that can be reconstructed through gradient inversion attacks. Incorporating differential privacy mechanisms at the edge aggregation level, where calibrated noise is added to the aggregated regional update before it is transmitted to the global coordinator, provides a formal privacy guarantee while balancing utility and communication cost [11]. This is particularly relevant when the same edge node serves multiple operators or tenants, as differential privacy bounds the information disclosure about any single client's contribution. The governance and risk posture of the entire federated ecosystem must be scrutinized to ensure that these technical mechanisms align with organizational data policies and regulatory frameworks.

5. Anomaly Detection in Dynamic Mobile Traffic Patterns

Beyond routine traffic prediction, the ability to detect anomalies—such as sudden flash crowds, distributed denial-of-service attacks, or infrastructure failures—is critical for network resilience and security. Anomaly detection in mobile traffic is inherently challenging because legitimate traffic patterns can exhibit high variance, and labeled anomalous samples are exceptionally rare. Federated learning offers a collaborative framework for anomaly detection where edge nodes can share insights about emerging threats without revealing sensitive operational data. Previous work on federated anomaly detection has shown that distributed models can achieve detection performance comparable to centralized counterparts while preserving confidentiality, provided that robust aggregation methods are employed to guard against malicious clients injecting false patterns [9].

DeepSeek's transformer architecture, with its capacity to capture long-range dependencies, is well-suited for distinguishing abnormal traffic bursts from normal cyclical variations. The self-attention mechanism can be leveraged to compute reconstruction errors or predictive residuals that serve as anomaly scores: when the model's forecast deviates significantly from observed traffic for a sustained window, an alert is triggered. In a federated setting, each edge node computes local anomaly scores and shares aggregate statistics, such as the distribution of reconstruction errors, without transmitting raw time series. The global model can then update its anomaly threshold adaptively based on the collective intelligence of the network, reducing false positives caused by localized non-stationarity.

Robustness against adversarial manipulation is a paramount concern. Federated learning systems are known to be susceptible to model poisoning attacks, where a compromised client submits manipulated updates that degrade global model accuracy or implant backdoors [10]. In anomaly detection, an adversary might craft updates that cause the model to ignore specific failure patterns, enabling stealthy attacks. Defensive aggregation rules, such as median-based or trimmed-mean aggregation, can mitigate the impact of Byzantine clients, but they must be carefully tuned to avoid suppressing genuinely rare but legitimate traffic surges. The integration of differential privacy, as previously noted, adds another layer of defense by limiting the influence any single participant can exert on the global model [11]. These robustness considerations intersect with fairness: overzealous defense mechanisms can

inadvertently penalize clients with unusual but valid traffic profiles, such as those serving culturally distinct city quarters or industrial zones, thereby reducing the model's predictive equity.

6. Governance, Fairness, and Policy Challenges

The federated DeepSeek-assisted system operates at the intersection of multiple governance domains, encompassing data protection law, algorithmic fairness, and telecommunications regulation. Federated learning is often presented as a privacy-enhancing technology, yet its compliance with global data protection frameworks depends on careful implementation. Although raw data never leaves the edge node, model updates can still embed personal information, and metadata such as client participation patterns may reveal sensitive operational details. Governance structures must therefore encompass model auditing protocols, consent mechanisms for data processing, and contractual clauses that define liability when federated models produce inaccurate predictions leading to service degradation or security breaches.

Fairness in federated learning is multidimensional. Statistical fairness concerns arise when the global model performs significantly worse for certain subpopulations of cells—such as those in low-income urban districts—due to underrepresentation in the training data or systematic differences in traffic patterns [12]. These disparities can be amplified in hierarchical aggregation if regional aggregators concentrate bias. Causal fairness adds another layer: a prediction model used for proactive resource allocation may systematically allocate fewer resources to areas where traffic is predicted to be lower, reinforcing digital divides. The federated framework, by keeping data local, offers a unique opportunity to implement fairness constraints that are tailored to each regional context while contributing to a global fairness objective.

The broader governance of federated large model ecosystems requires institutional innovation. Setting standards for update formats, aggregation protocols, and evaluation metrics is essential to enable interoperability across different operators and equipment vendors [13]. Regulatory bodies may need to articulate guidelines for the use of federated AI in critical infrastructure, balancing innovation incentives with resilience requirements. The model itself becomes a form of shared infrastructure, raising questions about ownership, versioning, and liability. If a poisoned update from one operator causes widespread misprediction that results in a network outage, apportioning responsibility between the aggregator, the originator, and the model provider is nontrivial. In the European context, the right to explanation and human oversight mandated by the General Data Protection Regulation compels deployers to design interpretable interfaces on top of transformer-based models that are inherently opaque [14]. These policy dimensions underscore that technical design choices in federated learning are deeply entangled with socio-legal norms.

7. Deployment, Sustainability, and Infrastructure Resilience

Deploying a federated DeepSeek system on real-world edge infrastructure involves confronting severe resource constraints. Edge servers integrated into base stations have limited GPU memory, power budgets, and cooling capacity, which restricts the feasible model size and the frequency of local training. Deep compression techniques, including pruning, quantization, and Huffman coding, can reduce the model's memory footprint and computational demands without substantially compromising performance [15]. For DeepSeek, applying structured pruning to its mixture-of-experts layers and quantizing attention weights

to 8-bit integers can enable inference and lightweight fine-tuning on commodity edge hardware. Nonetheless, the energy consumption of continuous federated training rounds must be measured against the sustainability goals of mobile operators, many of whom have committed to net-zero carbon targets. The concept of Green AI urges practitioners to account for the full life-cycle energy cost of model development and operation, favoring architectures that achieve high accuracy per joule [16].

Sustainability considerations extend to the entire federated lifecycle. A plausible operational model involves periodic retraining during off-peak hours, using renewable energy-powered edge data centers, and employing model caching and reuse to amortize costs. The federated architecture can also contribute to network sustainability indirectly: more accurate traffic prediction enables dynamic cell sleep modes and adaptive resource allocation, reducing the overall energy footprint of the radio access network. These systemic feedback loops between learning accuracy, communication cost, and energy consumption constitute a complex optimization landscape that demands holistic analysis rather than isolated metric tuning.

Infrastructure resilience is another first-order concern. Edge nodes are subject to intermittent connectivity, hardware failures, and natural disasters. Federated learning systems must be designed to gracefully degrade rather than catastrophically collapse when a subset of clients becomes unavailable. The hierarchical aggregation model provides natural fault isolation, as regional aggregators can continue to serve local predictions even if disconnected from the global coordinator [17]. Robust aggregation methods that tolerate Byzantine failures can also defend against equipment malfunction leading to corrupted updates [18]. From a deployment standpoint, adopting a DevOps-inspired MLOps pipeline for federated systems—encompassing canary testing of model updates, automated rollback, and continuous monitoring of prediction quality—is essential to ensure reliable operation at scale. Ultimately, the transition from laboratory benchmarks to production-grade federated infrastructure demands not only technical maturity but also a supportive ecosystem of standards bodies, certification authorities, and operator consortia willing to share the collective burden of governance.

8. Conclusion

This paper has presented a comprehensive analysis of a federated DeepSeek-assisted system for traffic prediction and anomaly detection in edge-enabled mobile networks. By integrating the powerful sequence modeling capabilities of DeepSeek with communication-efficient federated learning, the proposed architecture addresses the dual imperatives of predictive accuracy and data sovereignty. We examined the structural trade-offs among model expressiveness, communication overhead, and training efficiency introduced by parameter-efficient fine-tuning and hierarchical aggregation, and analyzed the robustness requirements to mitigate adversarial and Byzantine threats. The discussion extended beyond algorithmic considerations to encompass fairness, governance, and sustainability, highlighting that the successful deployment of such systems demands an interdisciplinary approach that reconciles technical performance with legal, ethical, and operational realities. Future research should focus on developing empirical benchmarks for federated large model performance on real mobile network data, designing interpretable anomaly scoring mechanisms, and establishing cross-operator governance frameworks that can accelerate responsible adoption. The federated paradigm, augmented by foundation models, holds the promise of a resilient, intelligent, and privacy-preserving mobile infrastructure, but only if its socio-technical dimensions are proactively addressed.

References

1. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322-2358.

2. Li, R., Zhao, Z., Zhou, X., Ding, G., Chen, Y., Wang, Z., & Zhang, H. (2017). The prediction analysis of cellular traffic based on big data: A deep learning approach. IEEE Access, 5, 17447-17457.

3. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 1273-1282.

4. DeepSeek-AI. (2024). DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434.

5. Zhou, T., Niu, P., Wang, X., Sun, L., & Jin, R. (2023). One fits all: Power general time series analysis by pretrained lm. arXiv preprint arXiv:2302.11939.

6. Xu, M., Cai, S., Li, D., & Song, L. (2024). Federated large language model: A comprehensive survey. arXiv preprint arXiv:2407.10612.

7. Liu, L., Zhang, J., Song, S. H., & Letaief, K. B. (2020). Client-edge-cloud hierarchical federated learning. IEEE International Conference on Communications (ICC), 1-6.

8. Zhang, C., Zhang, H., Qiao, J., Li, Z., & Alouini, M. S. (2025). TIDES: Traffic Intelligence with DeepSeek Enhanced Spatial Temporal Prediction. IEEE Journal on Selected Areas in Communications.

9. Preuveneers, D., Rimmer, V., Tsingenopoulos, I., Spooren, J., Joosen, W., & Ilie-Zudor, E. (2018). Chained anomaly detection models for federated learning: An intrusion detection case study. Applied Sciences, 8(12), 2663.

10. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. International Conference on Artificial Intelligence and Statistics (AISTATS), 2938-2948.

11. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. ACM SIGSAC Conference on Computer and Communications Security (CCS), 308-318.

12. Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2020). Fair resource allocation in federated learning. International Conference on Learning Representations (ICLR).

13. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1-210.

14. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76-99.

15. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations (ICLR).

16. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63.

17. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems (MLSys), 1, 374-388.

18. Shen, S., Tople, S., Saxena, P., & van der Meyden, R. (2020). Auror: Defending against poisoning attacks in collaborative deep learning systems. Proceedings of the 32nd Annual Computer Security Applications Conference (ACSAC), 508-519.