ELMİ TƏDQİQAT Oktyabr 2026, Cild: 6, Sayı: 10, s. 253−271
SCIENTIFIC RESEARCH October 2026, Volume: 6, Issue: 10, pp. 253−271
DOI: https://doi.org/10.36719/2789-6919/64/253-271

TEXNİKA ELMLƏRİ

TECHNICAL SCIENCES

Attention-Guided Sparse Aggregation and Hetero Geneity-Aware Optimization for Federated Wireless Traffic Prediction

Rajakani Velumani[1], N.Babu[2], K.Madhan[3*], and P.Maithili[4]

Abstract. *In cellular networks, wireless traffic forecasting is essential for proactive communication system adaptation. Wireless traffic prediction at the edge using Federated Learning (FL) has garnered a lot of attention due to its improved privacy protection and exemption from raw data transmission. However, existing FL-based traffic prediction methods still require substantial gradient exchange between local customers and a central server, and it remains unclear how to model the spatial dependencies among local customers within the FL architecture. To address this, we propose an FL framework, built on a Distributed Autonomous Network (DAN) architecture for 6G wireless communication, that proceeds through four stages, local modelling, gradient clipping and compression, path-level accumulation, and global optimization, to minimize communication load while maintaining forecasting accuracy. At each customer, an L2-norm gradient clipping step is followed by Top-K sparsification and 1-bit quantization to substantially shrink the size of the uploaded gradients. To offset the information lost through this compression, a Path Attention Aggregation Module learns customer-specific attention weights from each customer's upload trajectory, combining gradient-stability indicators with local performance measures, so that consistent, high-quality updates are weighted more heavily during aggregation; three customized aggregation strategies, all-correlated, δ-threshold, and k-relevant, further use gradient correlation to capture the spatial interdependence among customers. A Heterogeneity-Aware Adaptive Optimizer, built on Adam, additionally tunes each customer's learning rate, momentum, and weight-decay settings to its own traffic characteristics and enforces cross-round consistency, mitigating the effects of non-IID data across customers.*

[1]Department of Electronics and Communication Engineering, Anjalai Ammal Mahalingam Engineering College, Kovilvenni -614403, Assistant Professor, Thiruvarur District, Tamil Nadu, India.
[2]Department of Electrical and electronics engineering, J.J. College of Engineering and technology, Associate professor, Tiruchirapalli, India.
[3]Department of Electronics and Communication Engineering, Sri Ramakrishna college of engineering, Associate professor, Permabalur, India
[4.]II, Department of Electrical and Electronics Engineering, Kumaraguru College of Technology, Associate professor, Coimbatore, India
[*]Corresponding author. E-mail: journals856@rediffmail.com
Received: 26 June 2026; Accepted: 2 October 2026; Published online: 30 October 2026
© The Author(s) 2026. This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Experiments on two real-world wireless traffic databases show that the proposed approach outperforms modern algorithmic techniques and can increase communication efficacy by up to two orders of magnitude, without compromising forecasting precision.

Keywords: *cellular networks, traffic forecasting, autonomous network, compression, quantization, heterogeneity*

Federativ simsiz trafik proqnozlaşdırılması üçün diqqət mexanizminə əsaslanan seyrək aqreqasiya və heterogenliyi nəzərə alan optimallaşdırma

Rajakani Velumani[1], N.Babu[2], K.Madhan[3*], P.Maithili[4]

Xülasə. *Mobil rabitə şəbəkələrində simsiz trafikin proqnozlaşdırılması rabitə sisteminin qabaqlayıcı şəkildə uyğunlaşdırılması üçün vacibdir. Federativ öyrənmədən (FL) istifadə etməklə şəbəkənin kənarında simsiz trafikin proqnozlaşdırılması məxfiliyin daha yaxşı qorunması və ilkin məlumatların ötürülməsinə ehtiyacın aradan qaldırılması sayəsində böyük diqqət cəlb etmişdir. Bununla belə, mövcud FL əsaslı trafik proqnozlaşdırma metodları yerli müştərilərlə mərkəzi server arasında hələ də əhəmiyyətli həcmdə qradiyent mübadiləsi tələb edir və FL arxitekturası daxilində yerli müştərilər arasındakı məkan asılılıqlarının necə modelləşdirilməsi tam aydın deyil. Bu problemi həll etmək üçün 6G simsiz rabitəsi üzrə paylanmış avtonom şəbəkə (DAN) arxitekturasına əsaslanan FL çərçivəsi təklif edirik. Bu çərçivə proqnozlaşdırma dəqiqliyini qorumaqla rabitə yükünü minimuma endirmək məqsədilə dörd mərhələni əhatə edir: yerli modelləşdirmə, qradiyentlərin məhdudlaşdırılması və sıxılması, trayektoriya səviyyəsində toplanma və qlobal optimallaşdırma. Hər bir müştəridə L_2 normasına əsaslanan qradiyent məhdudlaşdırılmasından sonra yüklənən qradiyentlərin həcmini əhəmiyyətli dərəcədə azaltmaq üçün Top-K seyrəkləşdirməsi və 1 bitlik kvantlaşdırma tətbiq edilir. Sıxılma nəticəsində yaranan məlumat itkisini kompensasiya etmək üçün trayektoriya üzrə diqqətə əsaslanan aqreqasiya modulu hər bir müştərinin yükləmə trayektoriyasından həmin müştəriyə xas diqqət çəkilərini öyrənir. Modul qradiyent sabitliyi göstəricilərini yerli performans göstəriciləri ilə birləşdirərək ardıcıl və yüksək keyfiyyətli yeniləmələrə aqreqasiya zamanı daha böyük çəki verir; bütün korrelyasiyalı yeniləmələri, δ həddini və k uyğun yeniləməni əsas götürən üç fərdiləşdirilmiş aqreqasiya strategiyası isə müştərilər arasındakı məkan qarşılıqlı asılılığını nəzərə almaq üçün qradiyent korrelyasiyasından istifadə edir. Adam alqoritminə əsaslanan, heterogenliyi nəzərə alan adaptiv optimallaşdırıcı əlavə olaraq hər bir müştərinin öyrənmə sürətini, momentumunu və çəkilərin zəiflədilməsi parametrlərini onun trafik xüsusiyyətlərinə uyğunlaşdırır və təlim raundları arasında ardıcıllığı təmin edərək müştərilər üzrə müstəqil və eyni paylanmış olmayan məlumatların (non-IID) təsirini azaldır. Real simsiz trafikə aid iki verilənlər bazası üzərində aparılan eksperimentlər göstərir ki, təklif olunan yanaşma müasir alqoritmik üsulları üstələyir və proqnozlaşdırma dəqiqliyini qorumaqla rabitə səmərəliliyini 100 dəfəyədək artıra bilir.*

Açar sözlər: *mobil rabitə şəbəkələri, trafikin proqnozlaşdırılması, avtonom şəbəkə, sıxılma, kvantlaşdırma, heterogenlik*

[1]Elektronika və kommunikasiya mühəndisliyi kafedrası, Anjalai Ammal Mahalingam Mühəndislik Kolleci, assistent professor, Kovilvenni–614403, Tiruvarur rayonu, Tamil Nadu, Hindistan
[2]Elektrik və elektronika mühəndisliyi kafedrası, J. J. Mühəndislik və Texnologiya Kolleci, dosent, Tiruchirappalli, Hindistan
[3]Elektronika və kommunikasiya mühəndisliyi kafedrası, Sri Ramakrishna Mühəndislik Kolleci, dosent, Perambalur, Hindistan.
[4]Elektrik və elektronika mühəndisliyi kafedrası, Kumaraguru Texnologiya Kolleci, dosent, Koimbatur, Hindistan.
*Məsul müəllif. E-poçt: journals856@rediffmail.com

1. Introduction

One of the crucial features that present fifth-generation (5G) networks lack is edge intelligence, enabled by AI methods (such as ML and DL). As a result, both academic and commercial research communities have concluded that edge intelligence will be crucial to the development of sixth-generation (6G) communication networks (Al-Quraan et al., 2023). Shifting a portion of the information storage and computation of services from the central cloud to peripheral nodes that are both physically and cognitively close to the end customers can significantly reduce end-to-end latency and improve computing efficiency and communication effectiveness (Letaief et al., 2019). Since many intelligent network functions, like self-organisation and self-optimisation, rely significantly on edge computing to be realised, wireless traffic modelling at a distance is crucial for anticipated traffic situations. Even though a lot of research has been done on wireless traffic forecasting, nearly all of it falls under the category of centralized system training, which necessitates sending raw data from edge nodes (local customers) to the central cloud (Niknam et al., 2020). Because these edge nodes are geographically dispersed, sending raw data not only uses up bandwidth but also raises privacy concerns in the event that bad actors target them (Yuan et al., 2024). System training and forecasting may now be done at the edge because of the development of distributed ML, especially federated learning (FL). In the FL paradigm, a central cloud orchestrates the cooperative training of a global system by several edge nodes. Instead of sending raw data to the cloud, a single edge node only needs to transmit the most recent system or the gradient data to the centralized database (Habib et al., 2024).

Applying FL to wireless traffic presents numerous problems, despite its promising benefits in system forecasting. First, the wireless system incurs nontrivial transmission expenses due to the frequent transmission of gradients or attributes between nodes on the edge and the centralized server, particularly when the system is big, even though there is no transfer of raw traffic data involved in the system training (Xu et al., 2019). Second, due to user mobility, edge nodes exhibit distinct traffic patterns and intricate spatiotemporal connections. Traditional FL algorithms are not effective because they don't distinguish between the distinct contributions that each edge node makes to the eventual global system; instead, they just average the inclinations of those nodes. More precisely, the geographic locations and traffic patterns of three edge nodes are chosen from our Milan database (C. Zhang et al., 2019; C. Zhang et al., 2018; L. Zhang et al., 2022). Forecasting performance is enhanced by averaging the systems of two edge nodes with comparable traffic patterns. However, system aggregation might not be beneficial and might potentially worsen performance if their patterns are different (Zhao et al., 2020). This propose a unique federated learning is a standard technique based on association and gradient reduction to address the two issues mentioned above in wireless traffic

forecasting. In particular, we reduce the cost of sending messages between the server in the centre and customers in the area by including gradient sparsification in wireless traffic forecasting. Researchers also suggest employing error feedback and gradient tracking approaches to counteract the unfavorable impacts of gradient compression due to the lossy nature of sparsification. Additionally, in order to capture different geographical dependencies of local customers, we suggest three customized system aggregation procedures that are generated from the gradient correlation matrix. This is the first study to describe such a distributed architecture and its related techniques, to the best of the researchers' knowledge, for wireless traffic forecasting that incorporates spatiotemporal properties. The experimental findings show that the proposed technique with asymmetrical compression and association performs better than the state-of-the-art and can increase the effectiveness of communication by up to two orders of magnitude higher without compromising forecasting accuracy.

The remainder of this paper is organized as follows. Section 2 reviews existing studies on wireless traffic prediction, deep-learning-based forecasting, and federated learning approaches. Section 3 presents the proposed system model and methodology. Section 4 describes the experimental setup and evaluates the proposed framework. Finally, Section 5 concludes the paper by summarizing the major findings and contributions.

2. Related works

Wireless connectivity traffic forecasting plays a critical role in enhancing the cognitive capabilities of upcoming communication relationships, including intelligent network administration and efficiency has received a lot of attention lately (Y. Li et al., 2020). Previous research on wireless traffic forecasting has mostly concentrated on conventional statistical systems like entropy, α-stable, and ARIMA and its variants. However, these systems' inability to accurately depict wireless traffic is limited by the nonlinear characteristics of wireless traffic (Kunang et al., 2021). The advancement of wireless traffic forecasting has been significantly accelerated by the quick development of AI methods, particularly ML and DL. In recent years, a number of ML- and DL-based methods have emerged. To address wireless traffic forecasting issues, techniques based on graph neural networks, deep belief networks, long-short term relationships, and Bayesian processes have been examined (Sherubha & Mohanasundaram, 2019a; Sherubha & Mohanasundaram, 2019b; Sherubha et al., 2020). Additionally, the encoder-decoder architecture can capture spatiotemporal interdependence in wireless traffic forecasting. The researchers in (Wang et al., 2019) recommended employing convolutional neural networks, for instance can simultaneously encapsulate both temporal and spatial reliance for wireless traffic predictions at the city or network level. Additionally, they took into account several temporal connections, including periodicity and closeness. A three-dimensional convolutional network was used in (R. Li et al., 2017) to improve spatial-temporal learning over wireless communications. Additionally, cross-domain information, such as points of significance, has been shown to be beneficial in enhancing forecasting accuracy in (Shen et al., 2021). To lower system complexity and enhance representation learning, attention methods are later incorporated into wireless traffic forecasting.

The aforementioned studies primarily concentrate on centralized wireless traffic forecasting, which uses a lot of throughput since a strong node needs to receive raw wireless traffic information for development. To address this issue and train a forecasting system at the edge, FL-based approaches are beginning to appear. For the purpose of predicting traffic flow, as in (Tran et al., 2019) recommended a federated gated recurrent neural network. Additionally, the researchers in (T. Li et al., 2020) created FedDA, a dual-attention method that informs the method of learning with both past global expertise and local improvements in order to address the data heterogeneity problem that FL faces. FedDA achieves state-of-the-art performance and was

specifically intended for wireless traffic forecasting (Gao et al., 2024). Although our work takes into account using the FL paradigm to train a cell phone traffic estimation platform, it is different from all of the aforementioned efforts in two significant ways. To reduce the cost of sending messages between the server in the center and customers in the area, we first use gradient compression for wireless traffic forecasting. Second, in order to capture the geographical relationships among different local customers, we create customized gradient aggregation schemes based on gradient correlation (L. Li et al., 2023; Lin et al., 2020). Table 1 shows the comparison of existing approaches.

Table 1
Existing approach comparison

Ref.	Main Technique	Key Strength	Major Limitation / Research Gap
(Xu et al., 2019)	Scalable Gaussian Process	Provides a probabilistic framework for wireless traffic prediction	Scalability and modeling complexity can increase with large-scale, highly nonlinear traffic patterns
(C. Zhang et al., 2018)	Densely Connected CNN	Effectively captures spatial characteristics of citywide cellular traffic	Centralized learning requires collection of traffic data and introduces communication/privacy concerns
(C. Zhang et al., 2019)	Deep Transfer Learning	Exploits cross-domain information to improve cellular traffic prediction	Depends on centralized data processing and does not directly address federated communication overhead
(Zhao et al., 2020)	Spatial-Temporal Attention CNN	Jointly models spatial and temporal dependencies through attention and convolution	Primarily designed for centralized prediction rather than privacy-preserving distributed learning
(Wang et al., 2019)	Spatio-temporal Cellular Traffic Prediction	Models temporal and spatial relationships in metropolitan cellular traffic	Centralized architecture requires aggregation of raw traffic information
(R. Li et al., 2017)	Application-level Traffic Learning and Prediction	Learns application-level cellular traffic characteristics	Does not address privacy preservation or federated communication efficiency
(Shen et al., 2021)	Time-wise Attention CNN	Attention mechanism improves temporal feature representation	Does not explicitly solve non-IID client heterogeneity or FL communication cost
(L. Zhang et al., 2022)	Federated Meta-Learning	Enables edge-based wireless traffic prediction while reducing raw-data sharing	Frequent model/gradient communication can still impose substantial transmission overhead
(Tran et al., 2019)	Federated Learning over Wireless Networks	Provides an optimization framework for distributed FL over wireless networks	General FL framework rather than a traffic-specific spatial-correlation mechanism

Ref.	Main Technique	Key Strength	Major Limitation / Research Gap
(T. Li et al., 2020)	Federated Optimization for Heterogeneous Networks	Uses a proximal objective to improve training under non-IID data	Does not explicitly model spatial relationships among wireless traffic clients
(Gao et al., 2024)	FL with multi-time-scale information	Incorporates multiple temporal scales into federated traffic forecasting	Communication and client-dependency modeling remain challenges
(L. Li et al., 2023)	Gradient similarity-based FL	Exploits similarity among client gradients during aggregation	Does not jointly provide aggressive gradient compression and adaptive path-level attention
(Lin et al., 2020)	Gradient Sparsification / Compression	Significantly reduces communication bandwidth during distributed training	Lossy compression can discard useful gradient information and is not specifically designed for wireless traffic forecasting

3. System Model

In the context of distributed autonomous networks (DAN), a recommended sixth-generation (6G) cellular communication network framework, we conduct an investigation of wireless traffic forecasting. We abstract a rudimentary system model for wireless traffic forecasting after introducing the idea of DAN (Lu et al., 2019). To satisfy the many requirements of upcoming 6G networks, the DAN architecture is designed to enable plug-and-play, flexible deployment, and on-demand network customization through distributed and autonomous features. To accomplish underground, ground, satellite, and marine communications, DAN has numerous dispersed cloud units placed in various locations. Each cloud unit contains processing and networking features, namely a smart interface engine and baseband unit (BBU) pool (Nie et al., 2017), they are intelligent enough to assist the network data analytics function by gathering and analyzing data. Here, M satellite cloud units at the perimeter and one centralised cloud server in the mobile centralized network make up our wireless traffic forecasting system model (Ji et al., 2019). The number of BSs supported by each cloud unit varies based on coverage demands. Every BS's mobile traffic data is tracked, gathered, and stored in the local database in real-time. Employing these databases, the cloud perform load-aware traffic management and network control while training a forecasting system w in a federated manner to forecast future traffic quantity. For example, when the anticipated traffic flow is low, to save energy, some BBUs can be placed in an idle mode. A wireless traffic database $D_m = \{x_i^m, y_i^m\}_{i=1}^n, where\ x_i^m = [v_i^m, v_{i+1}^m, \ldots, v_{i+p-1}^m]\ and\ y_i^m = v_{i+p}^m$, is stored by each of the M cloud unit in the communication system. In particular, when creating training and test data samples, p is the chosen sliding window size and v_i^m represents the amount of wireless traffic during cloud unit $m's$ scheduled time slot $'i'$. To get the forecasting system w, leveraging FL or more specifically, cross-silo FL. This work can express our objective function as follows:

$$\arg\min_{w \epsilon R^d} \{f(w) = \frac{1}{M} \sum_{m=1}^{M} f_m(w)\} \quad (1)$$

Whereby, d is the system characteristic's dimensionality and $f_m(\cdot)$ is customer $m's$ local objective function. $f_m(\cdot)$ typically has the form of $(x_i^m, y_i^m; w)$, where $(\cdot)$ assesses the difference between the ground truth value and the forecasts. In the conventional FL paradigm, there are two basic phases to solve (1): global aggregation on the server side and local updates on the customer. In particular, the local customer m calculates the gradient information $g_t^m = \nabla f_m(w_t)$ at the t^{th} communication cycle and sends it along to the server. The server updates the global model in the second step employing the gradients it receives from local customers. FedAvg, the traditional weight update of FL algorithms can be expressed as:

$$\arg\min_{w \epsilon R^d} \{f(w) = \frac{1}{M} \sum_{m=1}^{M} f_m(w)\} \quad (2)$$

Where, M_t is a collection of local customers during transmission round t, and η is the server's standard rate of learning. Additionally, a set's cardinality is represented by $|.|$.

3.1. Recommended system

Four consecutive stages comprise the general architecture of the recommended FQL architecture, as depicted in Fig 1: local modelling, global optimization, path-level accumulation, and gradient processing. First, each financial institution's local strategy modelling networks receive locally generated market data that has been evaluated, as well as macroeconomic variables independently. Local strategy characteristics are generated at each customer employing a uniform forecasting backbone, such as a Transformer or LSTM version (Han et al., 2024). Customers do not send the full set of system characteristics after multiple rounds of local training. Rather, the Gradient Clipping method and Compression Module, which finds and quantizes the top$-K$ gradients to drastically lower transmission volume, are used to choose the most meaningful gradients. The compacted gradients are attached along with path-related information about them, such as indications like past gradient volatility and local strategy performance, before being sent to the central server. In order to improve the legitimacy and efficacy of the global update, the on the server-side Path Attention Aggregating Module uses unbalanced grouping along with attention score calculations to dynamically evaluate each customer's uploading channel, removing unstable or ineffective contributions. The next development cycle is subsequently started by broadcasting the final global system to every customer. Additionally, throughout local instruction, each client employs a Heterogeneity-Aware Optimizer to manage structural and stochastic variability across commodities. To ensure that the update procedure complies with region-specific financial behaviors, this module adaptively modifies optimizer settings based on local market dynamics. The module ensures that the FQL architecture concurrently operates within all rounds of neighborhood training and creates a locked evaluation loop with the global aggregates stage to ensure high-quality communication, privacy security, and outstanding strategy flexibility across universities (Jeleel-Ojuade, 2024).

3.2. Path Attention Aggregation Module

Vectors are used or sequences of image patches are commonly used as input in traditional self-attention methods. To create context-sensitive representations, the main goal is to calculate the pairwise interaction intensity among elements. These methods, which concentrate on intra-sample connections, are extensively used in computer vision and NPL. On the other hand, inter-customer path-level importance modelling is the focus of this article presents the Path Attention Aggregation Mechanism. It specifically seeks to dynamically assess how the release of this update trajectory of each consumer is based on its previous training performance and the communicated characteristic sequences contribute to the training of the global system (Chatzigiannis et al., 2023). This method is a federated-level focus mechanism that focuses on each customer's prior training

effectiveness; this revised trajectory is released. As a result, it is very different from traditional approaches that focus on input interpretation and normalization objectives, and modelling complexity. A simple three-layer completely interconnected neural network is used to build the Path Concentration component. An attribute vector for the uploading path of each consumer serves as the network's input. It consists of two main types of data: (1) advanced gradient sustainability indicators, which document how consistently local changes have occurred over the course of previous rounds. Among these is the mean interpersonal cosine equivalent between contemporary slopes, calculated as $\sigma_i = std\ (|\left|\nabla\theta_i^{(t-j)}\right||)$, for $j = 1, \ldots, k$ which measures the directional stability of local communications; and the subject of study is strategy measures that demonstrate the monetary quality of a consumer's local ecosystem. These consist of the average validation return $r_i = 1/T \sum_{t=1}^{T} R_i^{(t)}$ and the Pearson correlation coefficient ($Corr_i = corr(\hat{\hat{R}}_i, \hat{R}_{global})$ between the global system's recommendations over a held-out set and the customer's expected returns. To create customer-specific weights for aggregation, the interest system is fed these concatenated characteristics. Three layers make up the network structure: a layer for entering data, a layer that is concealed with 64 units activated by ReLU, and a layer called output that creates the interest weight αi for each path, standardized using the Softmax algorithm with dimensions [$batch_size$,8]. This is the definition of the mathematical flow:

$$h_i^{(1)} = ReLU\ (W_1 x_i + b_1); W_1 \epsilon R^{8*64} \tag{3}$$

$$h_i^{(2)} = W_2 h_i^{(1)} + b_2;\ \ W_2 \in R^{64*1} \tag{4}$$

$$\propto_i = \frac{\exp(h_i^2)}{\sum_j \exp(h_j^2)} \tag{5}$$

In this case, the vector of path attributes for customer I is denoted by xi, and the appropriate attention weight employed in gradient aggregation is represented by αi. A normalized total of the customer grades based on their attentiveness ratings is used to calculate the global constant update:

$$\theta^{t+1} = \sum_{i=1}^{N} \alpha_i . \nabla\theta_i^{(t)} \tag{6}$$

This module works in tandem with the Heterogeneity-Aware Optimizer (Kang et al., 2024). The latter enhances the attention mechanism's ability to discriminate while assessing customer inputs by improving the consistency and accuracy of local elevation variations. Consequently, during aggregation, low-quality pathways are efficiently suppressed. This creates a positive feedback loop at the system level: the path attention module improves the global aggregation process, while the heterogeneity-aware optimization increases the quality of local training. In the end, this cooperative dynamic improves the FQL architecture's resilience and strategic performance. This integrated mechanism is represented by the structure diagram's aggregate branch.

3.3. Gradient Clipping and Compression

The recommended Gradient Clipping and Compression Module is intended to improve training robustness and convergence stability while addressing communication bottlenecks brought on by inconsistent network resources across borders in federated schooling. Following local training, each customer initially completes gradient standardization and trimming using $L2$-norm, as illustrated in Figure 3. The gradients are then sent to the main location for worldwide aggregation after quantized compression and sparse selection. This module functions as a gradient-path reduction component in accordance with the component diagram's gateway and

rescaler architecture, rather than adhering to the structure of traditional deep networks. Let $\nabla\theta i_i \in R^d$ be an inclination vector at customer i. The following is how $L2$-norm clipping is used to stop a typical gradient explosion or instability:

$$\widetilde{\nabla}\theta_i = \frac{\nabla\theta_i}{\max(1, \frac{|\nabla\theta_i|_2}{C})} \tag{7}$$

Where, C is a predefined clipping threshold that is either dynamically changed or often set to $C = 1.0$ according to the customer's past volatility. By ensuring that $\left|\widetilde{\nabla}\theta_i\right|_2 \leq C$, this process keeps any one customer from having an unreasonable effect on the global system. After the trimming, the device performs Top-K sparse picking. The sparse operator of selection $\mathcal{T}(\cdot)$, which keeps the $K\%$ items with the definition of the most significant absolute integer that sets all other values to zero, is as follows. Let K represent the upload ratio.

$$\widetilde{\nabla}\theta_i^{(K)} = T_k(\widetilde{\nabla}\theta_i) \tag{8}$$

Here, 1-bit quantization procedure is provided to further minimize transmission overhead. This enables the server to reconstitute the compressed gradient by transmitting the average magnitude of the non-zero items while retaining only their sign. The following is the definition of the quantized representation:

$$q_i = sign\left(\widetilde{\nabla}\theta_i^K\right).\tilde{m}_i;\ \ \tilde{m}_i = \frac{1}{K}\sum_{j\in supp\ (\widetilde{\nabla}\theta_i^K)} |\widetilde{\nabla}\theta_{i,j}^K| \tag{9}$$

To achieve high compression speeds, the final upload is made up of the mean, the mean amplitude m and the sparse index set $supp(\cdot)$. The global gradient is rebuilt at the server in question using linear translation. This method is consistent with the diagram's router-rescaler architecture, in which rescalers carry out normalization and de-quantization while routers feed gradients into compression paths depending on Top-K significance. While the effectiveness of communication is significantly increased by $1-$bit compression, it intrinsically loses detailed information regarding gradient magnitudes, which could result in biased updates or impede convergence in real-world scenarios. To mitigate this, the condensed values are balanced by a score for attention taken from the Path Attention Module presented by the FQL framework, based on each customer's past performance and stability rather than being aggregated uniformly. By ensuring that more dependable and consistent updates have a bigger impact during accumulation, this approach successfully offsets the loss of sharpness caused by aggressive digitization. Quality-weighted aggregation and sparsity-aware compression work together to develop a "high-quality, high-weight" minimization approach that preserves robustness and fulfilment. The global aggregate can be stated analytically as follows:

$$\theta^{(t+1)} = \sum_{i=1}^{N} \alpha_i . q_i \tag{10}$$

Where, q_i is the condensed gradient, and α_i is the path module's concentration weight. In theory, this combined approach builds a compressed potential space adjusted minimum-bias estimation with the subsequent maximum error requirement:

$$E[|\theta^{(t+1)} - \theta^* \begin{bmatrix}2\\2\end{bmatrix} \leq O\left(\frac{1}{KN}\sum_{i=1}^{N} \propto_i^2 . |\nabla\theta_i - \theta^*|_2^2\right) \quad (11)$$

This shows that in addition to lowering communication costs, the compressed module maintains convergence durability and robustness across heterogeneity market data distributions when combined with path-aware weighted.

3.4. Optimizer

There are significant structural variations between commercial at the national and regional levels in federated quantitative strategy modelling. These variables include variances in trading systems, commodity liquidity, fluctuation trends, and patterns influenced by legislation. These result in customer data that is identically dispersed (non-IID) and very non-independent. Local over-fitting or failure to converge may occur when a uniform optimization method is applied across such contexts. A Heterogeneity-Aware Adaptive Optimizer is presented as a solution to this problem. While preserving convergence and consistency during global aggregation, through local instruction, this component changes dynamically optimization variables to match the unique market features of every consumer. The recommended method integrates a programmable controller network and expands on the Adam optimizer. The controller provides the optimizer's speed of learning ηi, tempo parameter β_i, and weight decomposition factor λ_i for this particular round depending on common measures of monetary volatility, including skewness of return patterns of distribution, volume fluctuation values, and historical variability. Considering an input width of $[batch_size,4]$, LeakyReLU-controlled invisible layer of width 32, and an outcome layer of length 3 that corresponds to the three adjustable hyper-characteristics, the controller network is made up of two completely connected layers. The following is how the resulting results are directly integrated into the standard Adam upgrade rules:

$$m_t = \beta_i m_{t+1} + (1-\beta_i)\nabla\theta_t; \; v_t = \lambda_i v_{t+1} + (1-\lambda_i)(\nabla\theta_t)^2 \quad (12)$$

$$\theta_{t+1} = \theta_t - \eta_i . \frac{m_t}{\sqrt{v_t} + \varepsilon} \quad (13)$$

As a result, each customer's characteristics $\eta_i, \beta_i, and\ \lambda_i$ are adaptively computed allowing each to adhere to a locally tailored upgrade trajectory based on its unique market features. Customers can avoid poor convergence or divergence caused by structural mismatches by employing this design. The server calculates a cross-round consistency factor c_i based on each customer's path variation during the preceding three rounds in order to guarantee consistency inside the federated system. Before aggregation, learning rates are modified employing this feedback in accordance with the following:

$$\eta_i^{new} = \eta_i^{old} . (1 - \gamma . \mathrm{c_i}) \quad (14)$$

Where, the scaling coefficient is γ. Greater divergence from the global strategy is indicated by a larger c_i, which calls for a lower learning rate. This approach is especially useful when combined with the Path Sensitivity Module: the adaptive optimizer lessens the impact of a customer's notable deviation caused by structural heterogeneity, preventing negative impacts on the global system. Consequently, the FQL architecture improves the durability and scalability of cross-market strategies for trading by the concurrent use of path concentration and heterogeneity-aware methods of optimization to increase stability and velocity of convergence in heterogeneous settings.

4. Numerical results

Researchers conduct tests on two real-world wireless traffic databases. First, we provide a brief description and analysis of the two databases. Next, we give a thorough description of the evaluation measures, characteristic values, and associated baselines. Finally, we compare our recommended algorithm's forecasting ability with the most advanced systems. Researchers conduct investigations on two real-world wireless communications datasets in this part. We first discuss and assess the two collections briefly. Next, we give a thorough description of evaluation measures, characteristic settings, and the associated baselines. Finally, we compare our suggested technique's forecasting effectiveness with that of cutting-edge systems.

4.1. Baseline algorithms

Algorithms for Baselines and Assessment Measures: The following cutting-edge algorithms are compared with our approach.

FedAvg: Federated learning from decentralized data is pioneered by this work. After performing local SGD several times, the customers communicate the gradients or updated characteristics to the server, which aggregates them to produce the final system.

FedProx is one example. By including a proximal term in the objective function of local customers, it was recommended to stabilize FedAvg's training on non-IID data. Federal Att. Instead of employing a simple average, the server uses a weighted aggregation with weights derived from the Euclidean distance between the local system and the global system.

FedDA. This approach is an example of a federated optimization procedure based on dual attention. The Euclidean distances between local systems, the worldwide system, and the quasi-global system are taken into account by the server when updating the worldwide system.

FedCOMGATE. This is the most advanced FL algorithm that uses gradient compression methods. When gradient compression is introduced, it offers theoretical guarantees regarding the convergence of FL. FedCOMGATE is not intended for wireless traffic forecasting, but rather for general ML challenges. In the subsequent experiments, we employ three commonly used metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) score to objectively assess the effectiveness of several wireless traffic forecasting systems. The discrepancy between forecasts and ground reality is indicated by RMSE and MAE. Better forecasting outcomes are shown by smaller values. Based on the percentage of total variation in forecasts that the system can account for, the R-squared score gives an indication of how well the system replicates forecasts. The better the forecasting outcomes, the higher the values. Furthermore, we compare the quantity of communicated gradients (ΔC) between various algorithms' central servers and local customers.

4.2. Experimental setup

The following characteristics are used to achieve the experiment results. First, we create training and test databases employing a sliding window method. In particular, the system is trained employing data from the first seven weeks, and forecasting performance is evaluated employing all remaining data. The window dimension p is set to 6, and we anticipate the traffic amount of the following time slot using one-hour web browsing as feature information. To standardize all data specimens subtract the mean from each cloud and divide the result by the standard deviation to speed up training. For edge customers, a relatively light forecasting system is desirable due to the need to reduce the additional computation and computer training overhead related to communication in traditional wireless networks. Consequently, a three-layer MLP

architecture with 128, 128, and 1 hidden neurons is created. We adhere to the guidelines in earlier studies when it comes to setting the hyper characteristic values involved in FL. In particular, we train our system with 200 rounds of communication employing stochastic gradient descent (SGD) as the optimizer. Employing a multi-step learning rate scheduler with checkpoints of 100 and 150, we begin with 0.1 and decay it by 10 for the learning rate of local updates. The server's learning rate, η, is set to 1.0. We update the neighborhood system employing five steps of SGD with a batch size of twenty and a compression ratio of 0.01 before transmitting the gradients to the server for aggregating. Unless otherwise noted, we use $k = 4\ and\ \delta = 0.5$ in our k-relevant and δ-threshold methods. It should be noted that all forecasting systems including our system and all baseline systems use precisely the same network design and characteristic settings to enable fair performance assessments. This work use the PyTorch library to implement all algorithms, and we execute the script on a CentOS workstation.

4.3. Overall performance analysis

The observed forecasting capabilities of several algorithms. FedProx has a characteristic μ that balances the proximal term's importance. In a similar vein, FedDA maintains a characteristic ϕ to regulate the number of enhanced data samples that are sent to the server to train quasi-global systems. This table makes it clear that our recommended method outperforms all of the baselines in terms of forecasting accuracy. More precisely, on the Milan database, several algorithms perform quite similarly, while on the Trentino database, they differ significantly. Employing the Milan database as an example, the highest-quality prediction result was obtained using our technique paired with the k-relevant consolidation approach of 0.0788 for the MAE metric. FedDA transfers everything of the server-side enhanced data specimens ($\phi = 100$) and obtains the second-best MAE score, 0.0790. This suggests that our approach achieves around the same level of effectiveness as the best benchmark technique. Our highest result for the R^2 score on the Milan database is 0.9501, which is similar to FedDA's second-best result of 0.9493. In the meanwhile, our approach significantly outperforms baselines on the Trentino database. For example, under the k-relevant aggregation method, our algorithm's best MAE result is 0.3621. This demonstrates that the best foundation MAE result, which has the value 0.3849 produced using the proposed technique, has been improved by at least 5.9%. Similarly, our methodology raises the R^2 score from the proposed method's 0.9273 to the all-correlated aggregation strategy's 0.9499 as in Fig 4. More specifically, FedProx only need to communicate gradient values of each cloud to the computer, Att and FedAvg have the identical interactions; no more correspondence is required. Even though FedDA outperforms all baselines in terms of forecasting performance, transferring a portion of the enhanced data specimens to the host to train a quasi-global algorithm necessitates a little more transmission. Compared to FedDA, our method is more communication-efficient because it just needs to communicate the associated gradient coefficients and the condensed value. Furthermore, compared to the traditional model, a networked decompression technique that additionally necessitates the movement of controlling variables for slope updating, our algorithm has a reduced communication complexity as in Fig 5. Despite having the lowest communication complexity of all baselines, the models forecasting performance is still worse than ours because our method is able in order to record the irregular spatial connections of various SCUs thanks to tailored aggregation strategies. As a result, our recommended algorithm outperforms the proposed model in terms of forecasting performance.

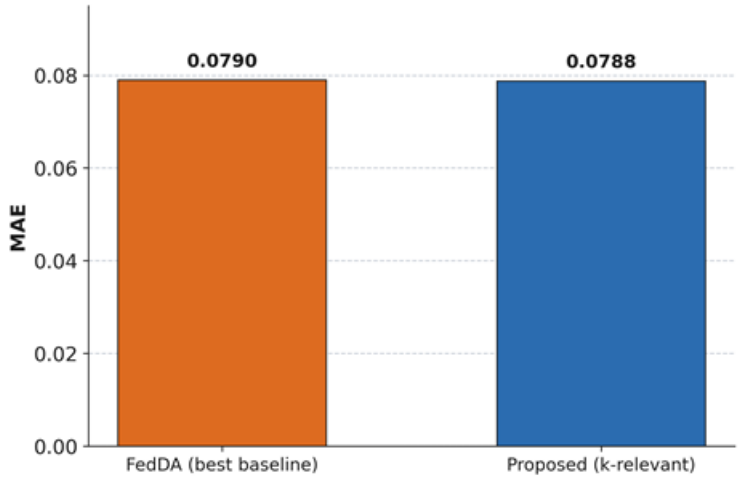


Figure 2
MAE prediction with dataset **1**

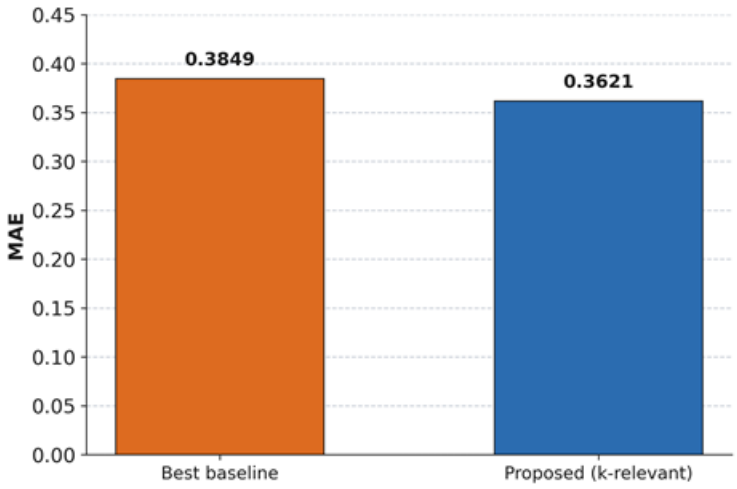


Figure 3
MAE prediction with dataset 2

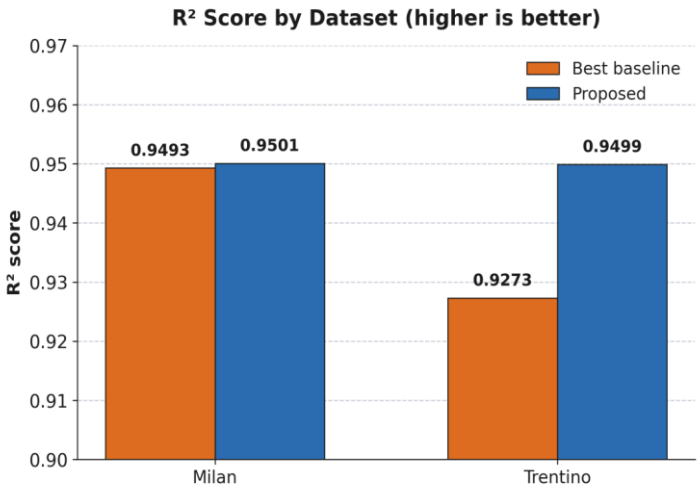


Figure 4
R^2 score based on given dataset

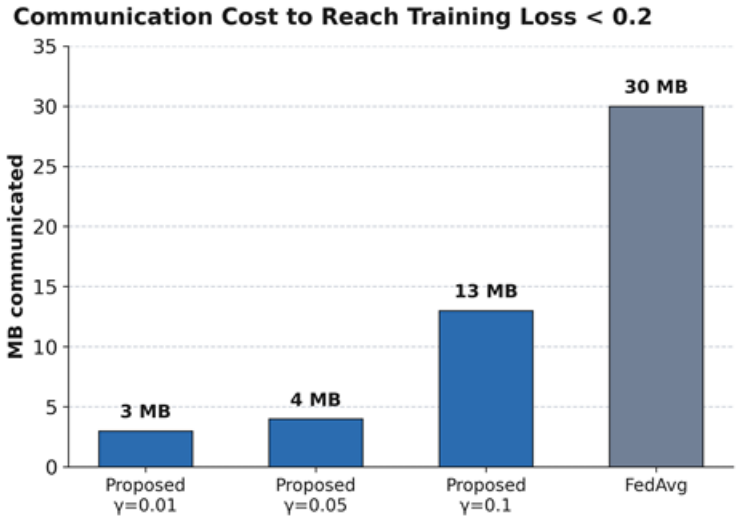


Figure 5
Communication cost

4.4. Efficiency analysis

The following section examines our algorithm's convergence and communication effectiveness in terms of transmitted gradient size as well as training loss, correspondingly. We just compare our algorithm with FedAvg. They present standardized communication inefficiencies at varying compression ratios ($\gamma \in \{0.01, 0.05, 0.1\}$) since the compression ratio has the greatest impact on our algorithm's performance. The model show that during the early stages of system training (less than 50 cycles), our method produces a greater loss than FedAvg. This is because FedAvg does not use gradient compression; therefore, all of the gradient data is sent to the server for federated optimization, which leads to a smaller training loss and faster convergence. But as instruction goes on, our approach gets closer to FedAvg and then surpasses it because our customized aggregation mechanism captures the spatial correlations of various SCUs in addition to

sending all gradient information to the server. Our method may save a significant amount of bandwidth during system training under specific training loss limitations employing the Trentino database as an illustration. To achieve a loss of less than 0.2, our method requires roughly $3\,MB\ (\gamma = 0.01), 4\,MB\ (\gamma = 0.05), and\ 13\,MB\ (\gamma = 0.1)$ communications. However, FedAvg needs more than 30 MB of bandwidth to achieve the same performance. The Milan database exhibits a similar phenomenon; the specific causes are not included here.

4.5. Predicting analysis

Our recommended algorithm's advantages in terms of both communication complexity and forecasting performance. Two SCUs are chosen from each database without sacrificing generality. Meanwhile, the two chosen cloud unit for the Trentino database are situated in Bocenago and Andalo, respectively. These cloud units are chosen at random considering their geographic locations or commonalities. It also includes the overall prediction-inaccuracy Cumulative Distribution Functions (CDFs) for different cloud units. For the sake of clarity, we solely compare our algorithm with FedAvg and FedDA. Additionally, the k-relevant aggregation strategy is used to collect all findings related to the recommended method. All three methods function remarkably well for the two cloud units since there is good agreement between the physical values of truth and the forecasts during throughout peak and off-peak hours are shown. The percentiles of errors in forecasting that are fewer than 100 are 66%, 70%, and 72% for FedAvg, FedDA, and our system, respectively. For the two chosen cloud unit that our method correctly forecasts the traffic levels, particularly for the peak traffic values. Because FedAvg cannot handle the issue of data heterogeneity, its result is unsatisfactory. The data dispersion and partial resolution of the data heterogeneity issues, the FedDA algorithm outperforms FedAvg in predicting traffic during peak hours. However, FedDA is still not as good as our method since it could not capture the unique spatial correlations between various cloud unit which are crucial for predicting spatial temporal wireless traffic.

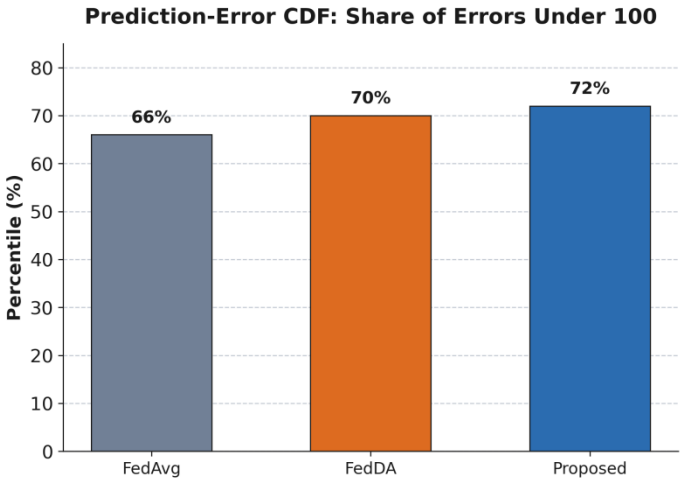


Figure 6
Prediction-error

4.6. Sensitivity analysis

Our approach's final forecasting success will depend on a number of factors, most notably the input feature size and the typical compression ratio. To make it easier to comprehend these characteristics, we only present

the RMSE results for the Trento database in Fig 2 and Fig 3. MAE and R^2 score exhibit RMSE values on both sets of databases showed remarkably comparable patterns.

1) Compression Ratio: Considering it shows how quickly the system gathers all of the local clients' gradient data, the compression ratio has the biggest impact on forecasting performance. Various compression ratios and present the outcomes of our three customized aggregation algorithms. This graph demonstrates how the RMSE effectiveness quickly increases when the server receives an increasing amount of gradient data from the local client as the compression ratio increases, for instance, from 0.01 to 0.1. Whenever the server gets sufficient gradient information, it can learn the forecasting system effectively. However, since the forecasting system's capacity for generalization does not increase, maintaining an increase in the compression ratio does not always result in more performance increases. In experiments, a compression ratio of 0.1 yields comparatively decent results as in Fig 6.

2) Aggregated Customers k and Threshold δ: The work display the outcomes of various $k\ and\ \delta$ selections when the compression ratio is set to 0.1. Both statistics show that performance deteriorates when too many customers participate in system aggregation. For instance, the RMSE findings are worse when $k = 6\ or\ \delta = 0.4\ compared\ to\ k = 3\ and\ \delta = 0.7$, respectively. On the other hand, since very little information is used for system personalization, too few customers are chosen, and system effectiveness is not enhanced by aggregated ($k = 1\ or\ \delta = 0.8$). Consequently, there is a trade-off between customer involvement and profitability.

3) Input Feature Size: Our team of algorithms' forecasts are also influenced by the size of the input feature; the outcomes are shown. This image makes it clear that input durations for continuous training that are either inadequate or too lengthy is inappropriate. The relationships between the input and the goal may not be accurately reflected if the input duration is too short. On the other hand, an input length that is excessively long may introduce significant noise, making the system difficult to learn. Using traffic for an hour ($n = 6$) as input to forecast the traffic for the following time slot, the forecasting performance is comparatively better.

4) Number of Local Updates τ and Local Batch Size: The RMSE results for regional batch size and the quantity of local repetitions τ are displayed. A trade-off between computation and communication is managed by these two factors. There are fewer communications between local customers and the server when there are more calculations at the local customers due to a large τ or batch size. However, if each customer goes too far in the direction of its own gradient update, an excessive amount of local processing could slow down convergence or stop the model from converging at all. This could lead to the distribution of a subpar aggregated system for training at nearby customers on the following grounds. For instance, our system's forecasting performance deteriorates when raising τ from 6 to 8 or batch size from 20 to 40. It is evident from the aforementioned considerations that choosing characteristics for the recommended FL architecture and the pertinent algorithms always involves trade-offs. Grid search methods can be used to find the ideal values of the aforementioned characteristics in practical tests and implementations.

5. Conclusion

In this study, we investigated wireless traffic forecasting at the edge for future 6G communication infrastructures and proposed an FL framework, built on a Distributed Autonomous Network architecture that combines gradient compression with attention-based, correlation-aware aggregation to reduce communication cost without sacrificing forecasting accuracy. At the client side, the framework applies L2-norm gradient clipping followed by Top-K sparsification and 1-bit quantization to substantially shrink the gradients transmitted to the central server. To offset the resulting loss of information, a Path Attention

Aggregation Module assigns each customer's upload an attention weight derived from its gradient-stability and performance history, so that dependable, consistent updates dominate the global aggregation; this is complemented by a Heterogeneity-Aware Adaptive Optimizer that adapts each customer's learning rate, momentum, and weight-decay settings to its local traffic characteristics and enforces cross-round consistency across customers. Additionally, we developed three customized aggregation techniques, all-correlated, δ-threshold, and k-relevant, to capture the spatial dependencies of various customers, employing gradient correlation to depict the links between dispersed local customers' traffic patterns. This work tested our recommended algorithm on two real-world databases, and the outcomes showed that it outperformed the most advanced methods for wireless traffic forecasting in terms of both accuracy and communication.

References

Al-Quraan, M., Mohjazi, L., Bariah, L., Centeno, A., Zoha, A., Arshad, K., Assaleh, K., Muhaidat, S., Debbah, M., & Imran, M. A. (2023). Edge-native intelligence for 6G communications driven by federated learning: A survey of trends and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence, 7*(3), 957–979.

Chatzigiannis, P., Gu, W. C., Raghuraman, S., Rindal, P., & Zamani, M. (2023). *Privacy-enhancing technologies for financial data sharing* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.10200

Gao, X., Zhao, Y., Wang, J., & Zhang, C. (2024). Federated learning cellular traffic prediction based on multi-time scale information. In D.-S. Huang, Z. Si, & Y. Pan (Eds.), *Advanced intelligent computing technology and applications* (pp. 276–289). Springer Nature Singapore.

Habib, M. A., Iturria-Rivera, P. E., Ozcan, Y., Elsayed, M., Bavand, M., Gaigalas, R., & Erol-Kantarci, M. (2024). *Transformer-based wireless traffic prediction and network optimization in O-RAN* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2403.10808

Han, W., Xu, J., Cheng, Q., Zhong, Y., & Qin, L. (2024). Robo-advisors: Revolutionizing wealth management through the integration of big data and artificial intelligence in algorithmic trading strategies. *Journal of Knowledge Learning and Science Technology, 3*(3), 33–45. https://doi.org/10.60087/jklst.vol3.n3.p33-45

Jeleel-Ojuade, A. (2024). *The role of information silos: An analysis of how the categorization of information creates silos within financial institutions, hindering effective communication and collaboration* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4881342

Ji, S., Pan, S., Long, G., Li, X., Jiang, J., & Huang, Z. (2019). Learning private neural language modeling with attentive aggregation. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Kang, R., Li, Q., & Lu, H. (2024). Federated machine learning in finance: A systematic review on technical architecture and financial applications. *Applied and Computational Engineering, 102*, 61–72. https://doi.org/10.54254/2755-2721/102/20241017

Kunang, Y. N., Nurmaini, S., Stiawan, D., & Suprapto, B. Y. (2021). Attack classification of an intrusion detection system using deep learning and hyperparameter optimization. *Journal of Information Security and Applications, 58*, Article 102804. https://doi.org/10.1016/j.jisa.2021.102804

Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y.-J. A. (2019). The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine, 57*(8), 84–90. https://doi.org/10.1109/MCOM.2019.1900271

Li, L., Zhao, Y., Wang, J., & Zhang, C. (2023). Wireless traffic prediction based on a gradient similarity federated aggregation algorithm. *Applied Sciences, 13*(6), Article 4036. https://doi.org/10.3390/app13064036

Li, R., Zhao, Z., Zheng, J., Mei, C., Cai, Y., & Zhang, H. (2017). The learning and prediction of application-level traffic data in cellular networks. *IEEE Transactions on Wireless Communications, 16*(6), 3899–3912.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems, 2*, 429–450.

Li, Y., Xu, Y., Liu, Z., Hou, H., Zheng, Y., Xin, Y., Zhao, Y., & Cui, L. (2020). Robust detection for network intrusion of industrial IoT based on multi-CNN fusion. *Measurement, 154*, Article 107450. https://doi.org/10.1016/j.measurement.2019.107450

Lin, Y., Han, S., Mao, H., Wang, Y., & Dally, W. J. (2020). *Deep gradient compression: Reducing the communication bandwidth for distributed training* (Version 3) [Preprint]. arXiv. https://arxiv.org/abs/1712.01887v3

Lu, S., Zhang, X., Sun, H., & Hong, M. (2019). GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)* (pp. 315–321). IEEE.

Nie, L., Jiang, D., Yu, S., & Song, H. (2017). Network traffic prediction based on deep belief network in wireless mesh backbone networks. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 1–5). IEEE. https://doi.org/10.1109/WCNC.2017.7925498

Niknam, S., Dhillon, H. S., & Reed, J. H. (2020). Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine, 58*(6), 46–51. https://doi.org/10.1109/MCOM.001.1900461

Shen, W., Zhang, H., Guo, S., & Zhang, C. (2021). Time-wise attention aided convolutional neural network for data-driven cellular traffic prediction. *IEEE Wireless Communications Letters, 10*(8), 1747–1751.

Sherubha, P., & Mohanasundaram, N. (2019a). An efficient intrusion detection and authentication mechanism for detecting clone attack in wireless sensor networks. *Journal of Advanced Research in Dynamical and Control Systems, 11*(5), 55–68.

Sherubha, P., & Mohanasundaram, N. (2019b). An efficient network threat detection and classification method using ANP-MVPS algorithm in wireless sensor networks. *International Journal of Innovative Technology and Exploring Engineering, 8*(11), 1597–1606. https://doi.org/10.35940/ijitee.K3958.0981119

Sherubha, P., Sasirekha, S. P., Manikandan, V., Gowsic, K., & Mohanasundaram, N. (2020). Graph based event measurement for analyzing distributed anomalies in sensor networks. *Sādhanā, 45*, Article 212. https://doi.org/10.1007/s12046-020-01451-w

Tran, N. H., Bao, W., Zomaya, A., Nguyen, M. N. H., & Hong, C. S. (2019). Federated learning over wireless networks: Optimization model design and analysis. In *2019 IEEE Conference on Computer Communications (INFOCOM)* (pp. 1387–1395). IEEE.

Wang, X., Zhou, Z., Xiao, F., Xing, K., Yang, Z., Liu, Y., & Peng, C. (2019). Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing, 18*(9), 2190–2202. https://doi.org/10.1109/TMC.2018.2870135

Xu, Y., Yin, F., Xu, W., Lin, J., & Cui, S. (2019). Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification. *IEEE Journal on Selected Areas in Communications, 37*(6), 1291–1306. https://doi.org/10.1109/JSAC.2019.2904330

Yuan, X., Chen, J., Zhang, N., Ye, Q., Li, C., Zhu, C., & Shen, X. (2024). Low-cost federated broad learning for privacy-preserved knowledge sharing in the RIS-aided Internet of vehicles. *Engineering, 33*, 178–189. https://doi.org/10.1016/j.eng.2023.04.015

Zhang, C., Zhang, H., Qiao, J., Yuan, D., & Zhang, M. (2019). Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE Journal on Selected Areas in Communications, 37*(6), 1389–1401. https://doi.org/10.1109/JSAC.2019.2904363

Zhang, C., Zhang, H., Yuan, D., & Zhang, M. (2018). Citywide cellular traffic prediction based on densely connected convolutional neural networks. *IEEE Communications Letters, 22*(8), 1656–1659.

Zhang, L., Zhang, C., & Shihada, B. (2022). Efficient wireless traffic prediction at the edge: A federated meta-learning approach. *IEEE Communications Letters, 26*(7), 1573–1577. https://doi.org/10.1109/LCOMM.2022.3167813

Zhao, N., Ye, Z., Pei, Y., Liang, Y.-C., & Niyato, D. (2020). Spatial temporal attention-convolution network for citywide cellular traffic prediction. *IEEE Communications Letters, 24*(11), 2532–2536. https://doi.org/10.1109/LCOMM.2020.3012279