applied sciences

MDPI

Article

Wireless Traffic Prediction Based on a Gradient Similarity Federated Aggregation Algorithm

Luzhi Li [1], Yuhong Zhao [1,*], Jingyu Wang [1,*] and Chuanting Zhang [2]

[1] School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China; liluzhilm@163.com
[2] Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, UK
* Correspondence: zhaoyuhong35@163.com (Y.Z.); 13734728816@126.com (J.W.)

Abstract: Wireless traffic prediction is critical to the intelligent operation of cellular networks, such as load balancing, congestion control, value-added service promotion, etc. However, the BTS data in each region has certain differences and privacy, and centralized prediction needs to transmit a large amount of traffic data, which will not only cause bandwidth consumption, but may also cause privacy leakage. Federated learning is a kind of distributed learning method with multi-client joint training and no sharing between clients. Based on existing related research, this paper proposes a gradient similarity-based federated aggregation algorithm for wireless traffic prediction (Gradient Similarity-based Federated Aggregation for Wireless Traffic Prediction) (FedGSA). First of all, this method uses a global sharing enhanced data strategy to overcome the data heterogeneity challenge of multi-client collaborative training in federated learning. Secondly, the sliding window scheme is used to construct the dual channel training data to improve the feature learning ability of the model; In addition, to improve the generalization ability of the final global model, a two-layer aggregation scheme based on gradient similarity is proposed. The personalized model is generated by comparing the gradient similarity of each client model, and the central server aggregates the personalized model to finally generate the global model. Finally, the FedGSA algorithm is applied to wireless network traffic prediction. Experiments are conducted on two real traffic datasets. Compared with the mainstream Federated Averaging (FedAvg) algorithm, FedGSA performs better on both datasets and obtains better prediction results on the premise of ensuring the privacy of client traffic data.

Keywords: wireless traffic prediction; federal learning; FedAvg; deep learning; gradient similarity

Citation: Li, L.; Zhao, Y.; Wang, J.; Zhang, C. Wireless Traffic Prediction Based on a Gradient Similarity Federated Aggregation Algorithm. *Appl. Sci.* **2023**, *13*, 4036. https://doi.org/10.3390/app13064036

Academic Editor: Christos Bouras

Received: 22 February 2023
Revised: 14 March 2023
Accepted: 17 March 2023
Published: 22 March 2023

1. Introduction

According to the State Ministry of Industry and Information Technology, as of the end of August 2022, 1,854,000 5G base stations have been completed and opened nationwide. Indeed, 5G networks play an essential role in realizing application scenarios such as AR, Telematics, and 4K TV, but 5G base station construction also faces problems such as difficulty and long investment cycles, etc. If we can accurately understand the different demands and growth trends of network traffic in each region, we can allocate network resources and reasonably plan the construction of 5G base stations. Meanwhile, the development and application of big data and artificial intelligence technologies are very effective in improving the quality of service (QoS) of access and core networks [1]. The application of artificial intelligence in the convergence of communication networks is significant for the accurate prediction of wireless traffic. Wireless traffic prediction estimates future traffic data volumes based on historical data and provides a decision basis for communication network management and optimization [2], and, based on the predicted traffic data, proactive measures can be taken to alleviate network congestion and improve network operation performance. In addition, common heterogeneous service requirements can be well met in future 6G communication networks by wireless traffic prediction at a lower cost [3].

Existing centralized traffic prediction methods require large and frequent interactions to share data from various regions for learning prediction. In real-world applications, it is difficult to achieve sufficient data sharing across enterprises due to multi-level privacy factors, i.e., the existence of data silos. A large number of data interactions also poses a huge communication overhead and risk of privacy leakage, and the centralized training data model poses a huge challenge to the computational and storage capacity of the central server. The implementation and application of Federal Learning (FL) provide a new way of thinking for traffic prediction models. Specifically, FL provides a distributed training architecture that can be jointly applied with many machine learning algorithms, especially deep neural networks, based on which local data can be effectively learned and global models can be obtained by iteratively aggregating local training models, which can also share the data information of the clients while protecting the privacy of the training client data, thus obtaining more accurate prediction results.

FL has been studied in the field of wireless traffic prediction, but there are still many challenges and problems. First, the client data of collaborative learning in FL has certain heterogeneity, i.e., Non-Independent Identically Distribution (Non-IID) characteristics, and the effective solution of the data heterogeneity problem is a prerequisite for the effective execution of the federal learning algorithm. In addition, the generalization performance of the global model generated by the final aggregation in FL traffic prediction largely determines the model prediction capability. In 2017, the federal average [4] algorithm proposed by Google uses the average aggregation approach for the integration of model parameters across edge nodes, but the strategy does not consider the differences between edge computing nodes, and the average global aggregation weights will undoubtedly reduce the generalization effect of the global model. Based on the above problems, this paper proposes a new wireless network traffic prediction method, called Federated Gradient Similarity-based Aggregation Algorithm for Wireless Traffic Prediction (FedGSA), which can collaboratively train multiple base stations and provide them with high-quality prediction models, including an enhanced data strategy based on global incremental integration, a two-channel training data scheme using sliding window construction, and a gradient aggregation mechanism to cope with data heterogeneity and global model generalization in FL.

Paper Organization and Contribution:

In this paper, we study the application of federation learning in wireless network traffic prediction. To achieve this goal, this paper addresses research-related issues in the following article sections. In Section 2, we discuss recent developments and applications of federation learning and wireless network traffic prediction. Then, in Section 3, we discuss the specific implementation details of the methodological techniques used in our proposed framework. Subsequently, in Section 4, we show the details of the experiments and the conclusions of the comparative analysis with existing methods.

Based on the proposed FedGSA framework and the descriptions in previous articles, the contributions of this paper can be summarized as follows:

- To balance the individuality of clients and the correlation characteristics among multiple clients to obtain a global model with better generalization capability, we propose a two-layer global aggregation scheme based on gradient similarity, which quantifies the client similarity relationship by calculating the Pearson correlation coefficient of each client's gradient to guide the weighted aggregation on the server side;
- To address the problem of statistical heterogeneity between traffic patterns collected by different clients, which can lead to difficulties in generalizing the global model, we introduce a quasi-global model and use it as an auxiliary tool in the model aggregation process;
- Considering the time-dependent characteristics of base station network traffic, we use a sliding window strategy here to represent the traffic of each time slot as a two-channel Tensor matrix, and divide the historical traffic data into adjacent time traffic data and periodic daily historical traffic data;

- We conducted validation experiments on two publicly available real datasets and compared and analyzed the experimental results with existing experimental methods.

2. Related Work

As the present work is closely related to wireless traffic prediction and FL, we re-view the most related achievements and milestones of these two research topics in this section.

2.1. Federated Learning

Federated Learning (FL), first proposed by Google in 2016, provides a collaborative training architecture based on deep learning. FL is a distributed learning framework in which raw data is collected and stored on multiple edge clients, and model training is locally performed on the clients, and then the models are progressively optimized to learn the models through client interaction with a central server. Its classical architecture diagram is shown in Figure 1.

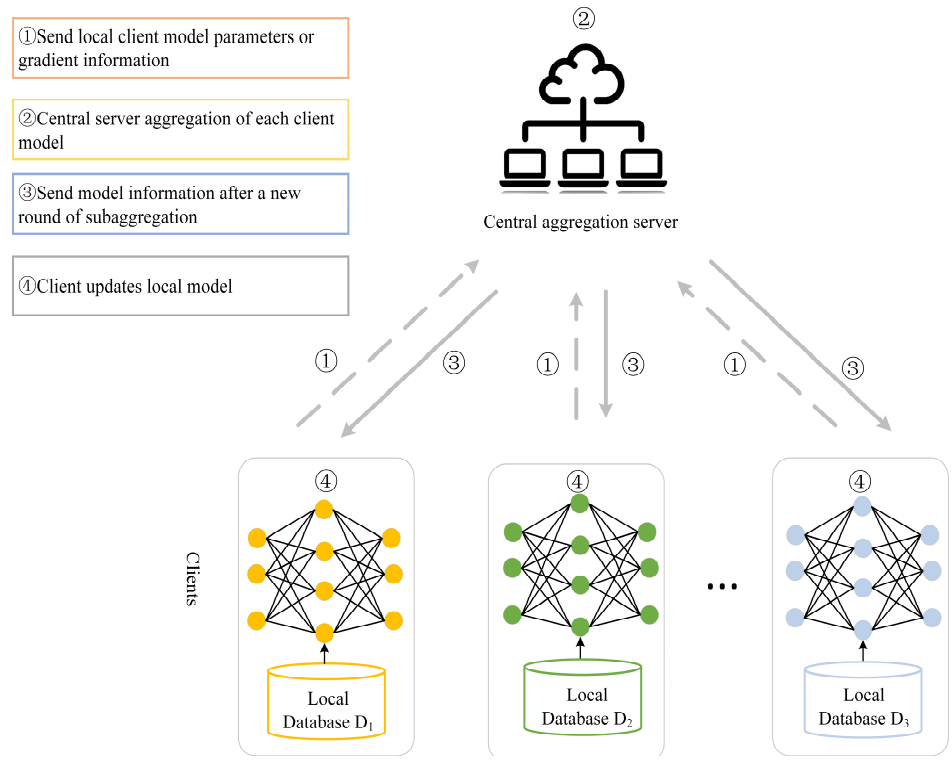


Figure 1. Classic architecture of federated learning.

As shown in Figure 1, the global model is initialized by the client application and trained based on local data, and the local model is obtained after the local training is completed, and its parameters or gradient information is uploaded to the central server, which aggregates the local client model based on the aggregation algorithm to generate a new global model, and then sends the new round of global models to each edge client again to iterate the above process until the final global model is obtained.

It is shown that the performance of federated learning is similar to that of centralized learning when the client data have Independent Identically Distributed (IID) characteristics. However, when the multi-client data are Non-IID, the performance of the federation learning algorithm is significantly reduced. Therefore, solving the problem of statistical heterogeneity of data is an urgent prerequisite to be addressed before deploying federated learning algorithms. To address this issue, a data-sharing strategy is proposed in the literature [5] for creating a globally shared data subset to integrate the local data features of the participating training clients to overcome the data heterogeneity challenge faced by FL. In addition, how to aggregate each client model to the global model while ensuring its

generalization capability is a critical issue in federation learning. FedAvg is currently the mainstream federation aggregation scheme, the core idea of which is to weight the average of each local model participating in the aggregation according to the ratio of the amount of data each client has to the total training data, and its process can be described as:

Assuming K clients participate in federation training, each client has multiple training data volumes n_k and local model weights w_k^{t+1} in the $(t+1)$st global iteration, and the FedAvg aggregation approach can be expressed as Equation (1):

$$w_G^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{t+1} \tag{1}$$

where w_G^{t+1} is the global model parameters after the $t+1$ st communication aggregation, n denotes the total amount of data for K clients and $\sum_{k=1}^{K} n_k = n$, and w denotes the parameters of the local model of the kth client at $t+1$ communication rounds.

However, in the federation learning aggregation process, FedAvg's aggregation approach by averaging the model parameters or gradients across clients is difficult to guarantee the generalization ability of the global model generated by the final aggregation, in addition to the fact that federation learning cannot observe the amount of data from the clients of edge computing nodes. Therefore, the aggregation weight assignment using the actual data volume is difficult to achieve, and the quantity of data does not represent the quality of data. To address this problem, this paper proposes a federated aggregation scheme based on gradient similarity, which considers the similarity of gradient information among individual clients and performs two-level aggregation of client models based on similarity knowledge. The simulation experiments show that the scheme can achieve better results.

2.2. Wireless Traffic Prediction

Accurate traffic modeling and forecasting capabilities play an important role in wireless services, and research related to wireless traffic forecasting has received significant attention. Wireless traffic prediction is essentially a time series prediction problem. The solution methods can be broadly classified into three categories, namely, simplex methods, parametric methods, and non-parametric methods.

The historical average method and the simplex method are the representatives of the first type of method. The two methods use the average value of historical data and the last observation as the prediction value, respectively. This type of prediction method does not require complex calculations and is easy to implement, but it cannot capture the hidden patterns of the wireless traffic and the prediction performance is relatively poor.

For the second category, i.e., parametric methods, tools based on statistics and probability theory are used to model and forecast wireless services, among which the most classical method is the Auto-Regressive Integrated Moving Average (ARIMA). In order to characterize the self-similarity and burstiness of wireless traffic, the authors explored ARIMA and its variants in the literature [6,7]. In addition to ARIMA models, literature [8–10] explored alpha-stable models, entropy theory, and covariance functions to perform wireless traffic forecasting, respectively.

With the rapid development of machine learning and artificial intelligence techniques, nonparametric methods have become a strong contender among wireless traffic prediction methods. In particular, research on wireless traffic prediction based on deep neural networks has attracted great attention. In the literature [11], the authors designed a traffic prediction model based on a multi-channel sparse long-term short-term memory network to capture multi-source network traffic information and improve the ability of deep neural network models to capture important features. In [12], the authors designed a Generative Adversarial Network (GAN) traffic prediction method and separately captured traffic spatio-temporal features and base-station-type features, input the spliced features into

the composite residual module to generate predicted traffic, judge the generated traffic by the discriminative network, and then generate highly accurate predicted traffic by the generative network after the game confrontation between the generative network and the discriminative network.

To effectively extract spatial and temporal features, a joint spatio-temporal prediction model based on neural networks has been proposed in the literature [13], which uses graph convolutional networks to extract complex topological spatial features in a targeted manner, while using gated cyclic units to extract temporal features of the traffic. City-scale wireless traffic forecasting is also studied in the literature [14], where the authors introduce a new forecasting framework by modeling the spatio-temporal correlation of cross-domain datasets.

The above work mainly uses centralized wireless traffic prediction, and to address the problems of communication overhead, privacy leakage, and data silos in centralized prediction schemes, this paper implements wireless traffic prediction through distributed architecture and federated learning.

3. Proposed Framework and Methods

This section may be divided into subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

3.1. Overview

In this section, we describe the proposed FedGSA framework in detail. Figure 2 shows the overall model framework of FedGSA; specifically, FedGSA has the following steps.

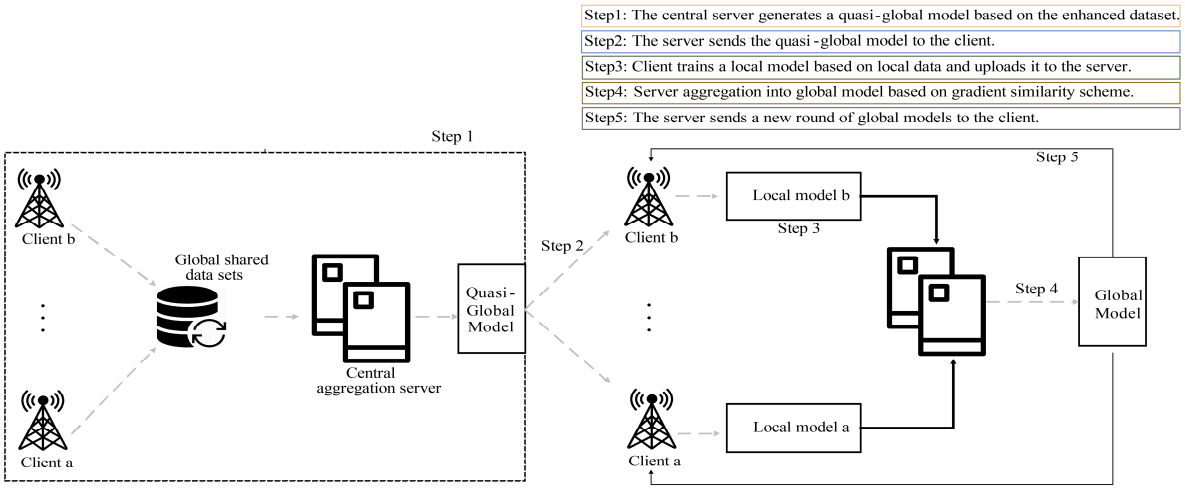


Figure 2. FedGSA Overall Architecture.

1. First, clients share local augmented data to form an augmented dataset based on global incremental integration, and a central aggregation server is trained to generate a quasi-global model based on this dataset and apply it to each client.
2. After each client applies the quasi-global model locally, a sliding window scheme is used to generate local two-channel network traffic data for each client, and then the client executes a local training procedure and passes the local model parameter information to the central aggregation server after the local training is completed.
3. Finally, the central server performs a two-level weighted aggregation of each client's network model based on the gradient similarity of each client, and finally generates a global model.

3.2. Enhanced Data Strategy Based on Global Incremental Integrations

The variability of base station traffic patterns and the mobile characteristics and communication behaviors of users within the base station range further expand the model

diversity of wireless services, and the wireless traffic data from different base stations are highly heterogeneous and non-IID in nature. It is shown that Non-IID client-side data leads to a degradation of the performance of the federation learning algorithm, since the weight differences of the client-side model parameters should be considered when performing model aggregation at the server side. Therefore, this paper uses an augmented data strategy based on global incremental integration to overcome the traffic data heterogeneity challenge by creating a small augmented dataset using the original wireless traffic dataset and generating a global shared dataset.

The augmentation strategy in this paper is as follows. The dataset is first partitioned into weekly slices based on temporal indexes. For weekly traffic, statistical averages are calculated for each time point and the obtained results are considered as augmented data, and finally, the augmented data are normalized as shown in Figure 3.

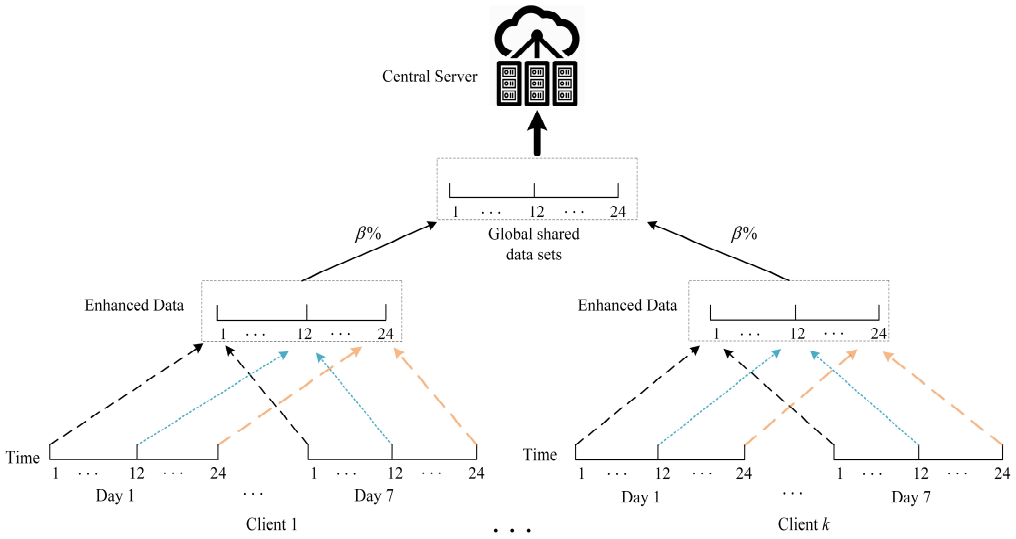


Figure 3. Enhanced data strategy.

As can be seen from Figure 2, the employed enhancement strategy is easier to implement and generate enhanced data than the traditional time-domain or frequency-domain time-series data enhancement strategies [15]. It has been experimentally proven to provide an effective solution to the problem of data heterogeneity.

During the training process, each base station sends a small fraction of its augmented dataset, say $\beta\%$, to the central server to eventually generate a global dataset that obeys the original client data distribution. The size of the augmented data is much smaller compared to the size of the original data. Based on this augmented dataset, a quasi-global model can be trained and used as prior knowledge for all clients, and the model is trained using the augmented data for all clients rather than the original data. Even so, due to the high similarity between the augmented data and the original data, the model can still be used as prior knowledge for all clients.

3.3. Constructing Two-Channel Sliding Window Training Data

The wireless traffic prediction service in general is: given K base stations, the local wireless traffic data of each base station can be represented as a dataset, as in Equation (2):

$$d^k = \left\{ d_1^k, d_2^k, d_3^k, \cdots, d_z^k \right\} \tag{2}$$

where the total time interval is Z; and assuming that $\hat{d}_z^k$ is the target service volume to be predicted, then the wireless service prediction problem can be described as Equation (3).

$$\hat{d}_z^k = f(d_{z-1}^k, d_{z-2}^k, \cdots, d_1^k; w) \tag{3}$$

where $f(\cdot)$ denotes the chosen prediction model and w denotes the corresponding parameter. The prediction model $f(\cdot)$ can be in linear form (e.g., linear regression) or in nonlinear form (e.g., deep neural network).

For wireless network traffic prediction techniques, to reduce data complexity, partial historical traffic data are usually used as input features, and considering that the base station network traffic is time-dependent, the traffic of each time slot can be represented as a two-channel Tensor matrix [16,17].

Therefore, based on the wireless traffic dataset d^k, a set of input-output pairs $\left\{x_i^k, y_i^k\right\}_{i=1}^n$, where x_i^k denotes the historical traffic data associated with y_i^k, can be obtained using a sliding window scheme, and x_i^k is partitioned into two time channels, i.e., adjacent time and cycle time channels, which represent the predicted target time adjacent time traffic and cycle time traffic for the corresponding time points, respectively, as in Figure 4.

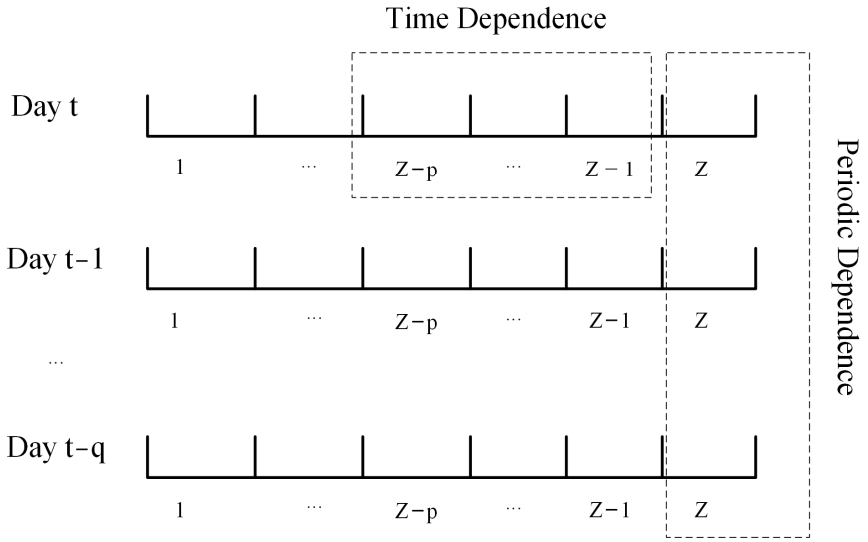


Figure 4. Two-channel training data.

Defining p as the adjacent time point sequence dependence length, the flow of the adjacent time series can be expressed as $\left\{d_{z-1}^k, \cdots, d_{z-p}^k\right\}$, q as the periodic time dependence length, and the periodic historical sampling flow can be expressed as b, μ is periodic. The flow prediction target of this paper is the flow value at the next time point, so Equation (3) can be described as Equation (4):

$$\hat{y}_i^k = f\left(x_i^k; w\right) \tag{4}$$

The objective of the experiment is to minimize the prediction error on K clients, so the objective of the traffic prediction can be described as solving for the parameter w under the optimal solution in Equation (5):

$$\min_w \updownarrow(w) = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \updownarrow\left(f\left(x_i^k; w\right), y_i^k\right) \tag{5}$$

where $\updownarrow$ is the loss function, which can be expressed as $\left|f\left(x_i^k; w\right) - y_i^k\right|$.

Long Short-Term Memory (LSTM)) has the powerful ability to model time series datasets, so this paper selects a LSTM Long Short-Term Memory network as the network model, sets two LSTM network layers, corresponding to the input adjacent time point dependent sequence traffic data and periodic time series dependent data in turn, after which the output data features of each channel are spliced, and finally, the features are mapped to the prediction by a linear layer.

3.4. Global Aggregation Based on Gradient Similarity

The aggregation process in FL is a key part of model training, and the quality of aggregation directly affects the strength of the generalization ability of the final generated global model. The goal of central server-side model aggregation is to obtain a global model with strong generalization capability across all clients, which should balance client personalization and correlation characteristics across multiple clients. To achieve this, the global model should find a balance between capturing the personalized traffic patterns of the clients and the public shared traffic patterns.

In literature research, it is found that similarity-based weighted fusion schemes have a wide range of applications in machine learning, such as natural language processing and transformers in image vision [18], where similarity knowledge can tap potential correlations among different clients, and FedGSA quantifies client similarity relationships by calculating Pearson correlation coefficients for each client gradient to guide the server-side weighting of client models for aggregation.

The Pearson correlation coefficient is used to describe the degree of linear correlation between two variables, i.e., the larger the absolute value of the correlation, the stronger the correlation, and the value is $[-1, 1]$. The Pearson correlation coefficient is the ratio of the covariance to the standard deviation, and the Pearson correlation coefficient for a set of data (x, y) is calculated as:

$$\rho_{x,y} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{\sqrt{(\Sigma x^2 - \frac{(\Sigma x)^2}{N})(\Sigma y^2 - \frac{(\Sigma y)^2}{N})}} \tag{6}$$

where N denotes the number of values of the variable.

Here, we use gradient information to measure the similarity between individual client models, rather than based on the original traffic data of each client itself. The central server uses the similarity relationship between the gradients of the individual client models to guide the clients in generating personalized models, thus helping to moderate their impact on global aggregation (i.e., reduce variance), and the aggregation principle of FedGSA can be described as follows:

Assuming K clients involved in training, after t rounds of global iterations, then in $t+1$ rounds, each client is trained based on the quasi-global model obtained under T rounds using local data to obtain its local model parameters $\left\{w_k^{t+1}\right\}_{k=1}^{K}$, and the central server in the aggregation phase has two layers of aggregation for the client models:

The first layer of aggregation aims to capture the similarity relationship between each client and quantify the impact of each client on the global model by assigning weights to its Pearson correlation coefficients among the clients, and for each client, a personalized model is formed based on its gradient similarity relationship with other clients using Equation (7):

$$\widetilde{w}_k^{t+1} = \sum_{r=1}^{K} \rho_{k,r} w_r^{t+1} \tag{7}$$

The second layer performs aggregation among the personalized models: the central server generates a new round of quasi-global models based on the final aggregation of Equation (8):

$$w_G^{t+1} = \frac{1}{K} \sum_{k=1}^{K} \widetilde{w}_k^{t+1} \tag{8}$$

where $\rho_{k,r}$ denotes the Pearson correlation coefficient between two models of K clients, $\widetilde{w}_k^{t+1}$ denotes the new personalized model parameters of each client obtained in $t+1$ round after the weighting operation of comparing the gradient similarity of each client in the round, and w_G^{t+1} denotes the quasi-global model parameters finally generated in the round.

Algorithm 1 describes the execution process of FedGSA:

Algorithm 1: FedGSA Implementation Process

Input: The wireless traffic data $\{x^k, y^k\}_{k=1}^K$; the clients share the enhanced data and form the globally shared dataset D_s; the percentage of selected clients is δ, and the client learning rate is η.

Output: Global Model w_G.

Obtain w_Q by using D_s//Get the quasi-global model w_Q and send it down to the client

for each round $t = 1,2, ...,$ do

  $m \leftarrow \max(K \cdot \delta, 1)$

  $S_t \leftarrow$Randomly selected m-base stations

  for each client $k \in S_t$ do

    $w_k^{t+1} \leftarrow w_k^t - \eta w^t$//Client local update model

  end for

    Obtain w_k^{t+1}//Get the client local model

  Obtain $\widetilde{w}_k^{t+1}$ by using Equation (5)//First aggregation to obtain personalized models for each client

end for

Obtain w_G^{t+1} byusing Equation (6)//Second aggregation, get the global model

Return w_G

4. Experiments and Conclusions

In this paper, two real datasets were selected to learn and train clients by combining federal learning mechanisms with LSTM long and short-term memory networks. To verify the feasibility and effectiveness of the method in this paper, some traditional network models based on LSTM, Lasso, Support Vector Regression (SVR), and FedAvg traffic prediction methods were selected for comparative analysis. Except for the shallow learning algorithm, the FedAvg algorithm and the structure of this experimental network remained consistent.

4.1. Dataset and Evaluation Metrics

This paper used the Trento and Milano telecommunication activity datasets provided by Telecom Italia in the European "Big data challenge" [19,20], and used the network traffic records of these two regions as the raw data for traffic prediction. The cellular networks for cellular user activity recorded traffic every ten minutes for two months from 11 January 2013 to 1 January 2014. For the experiments in the following subsections, the network traffic was resampled to hourly to avoid data sparsity issues.

To evaluate prediction performance, three widely used regression metrics were adopted in this paper, i.e., mean squared error (*MSE*), mean absolute error (*MAE*), and R-squared score:

1. Mean Absolute Error (*MAE*): Is the average of the absolute error, which can better reflect the actual situation of the prediction value error. The range is $[0, +\infty)$, as in Equation (9):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

2. Mean Square Error (*MSE*): Is the square of the difference between the true value and the predicted value, and then the average of the summation is used to detect the deviation between the predicted and true values of the model, and its range is as in Equation (10):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10}$$

where, in *MAE* and *MSE*, $\hat{y}_i$ denotes the predicted value of wireless traffic at the time i and y_i denotes the true value at the corresponding time.

3. R-squared score: the R-squared score is applied to regression problems with values between 0 and 1. The closer to 1 indicates a better fit and is generally expressed as R^2, as in Equation (11):

$$R^2 = 1 - \frac{\left(\sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2\right) \Big/ m}{\left(\sum_{i=1}^{m} \left(y^{(i)} - \overline{y}\right)^2\right) \Big/ m} \tag{11}$$

The numerator represents the Residual Sum Of Squares (RSS) and the denominator represents the Total Sum of Squares (TSS).

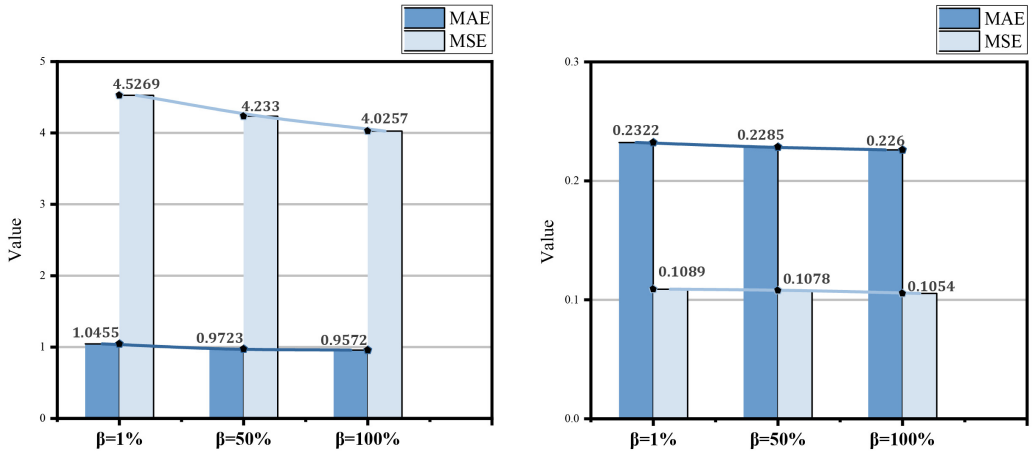
4.2. Experimental Settings and Overall Results

Experiments were conducted with 100 randomly selected cells from each dataset, and eight weeks of traffic data were randomly selected for the experiments, where the traffic of the first seven weeks was used for training the prediction model and the traffic of the last week was used for testing. In constructing the two-channel training samples using the sliding window scheme, both the temporal channel dependency length and the periodic channel dependency length were set to 3. A total of 100 rounds of communication were conducted between the local client and the central server, and the initial learning rate was set to 0.001, the local training batch size was set to 20, and 10% of the total samples were randomly selected in each round for the client samples to locally participate in the training and report the results of the last round, and had different results according to the different data sharing ratios in the shared data strategy; see Table 1. It can be seen from Table 1 that even if only 1% of the augmented data were shared, the performance of FedGSA, the method proposed in this paper, still outperformed other baseline methods in both datasets.

Table 1. Comparison of MSE and MAE prediction performance of different methods on two datasets.

Methods	Trento		Milano	
	MAE	**MSE**	**MAE**	**MSE**
Lasso	1.5391	5.9121	0.5475	0.4380
SVR	1.0470	5.9080	0.2220	0.1036
LSTM	1.1193	4.6976	0.2936	0.1697
FedAvg	1.0668	4.7988	0.2319	0.1096
FedGS A ($\beta = 1\%$)	1.0455	4.5269	0.2322	0.1089
FedGS A ($\beta = 50\%$)	0.9723	4.2330	0.2285	0.1078
FedGS A($\beta = 100\%$)	0.9572	4.0257	0.2260	0.1054
Improve	↑10%	↑16%	↑3%	↑4%

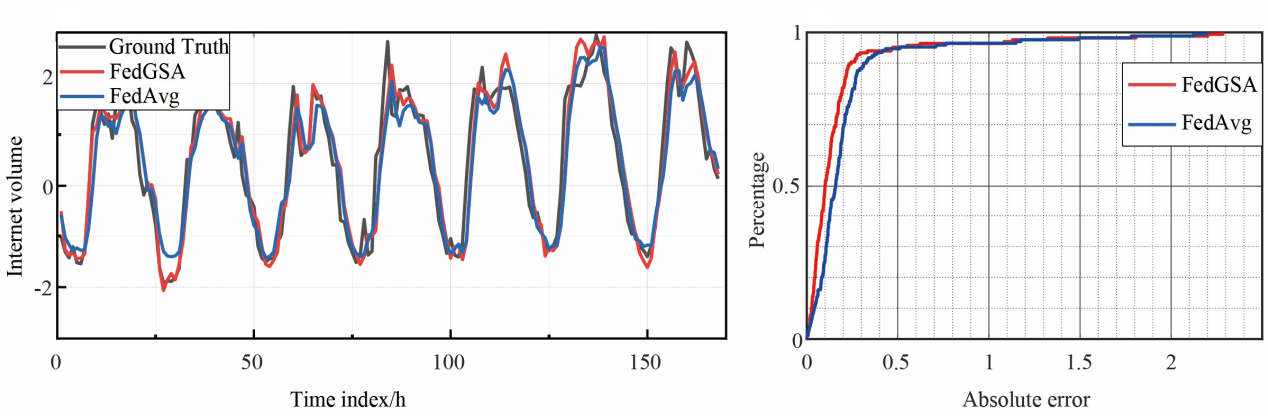
Specifically, for the results on the Trento dataset, the present experimental algorithm (FedGSA) provides MAE and MSE gains of 10% and 16%, respectively, compared to the best performing method in the baseline (i.e., FedAvg). Similarly, for the Milano dataset, the FedGSA performance gains (MAE, MSE) are 3% and 4%, respectively. Furthermore, observing Table 1, it can be found that the prediction performance of FedGSA keeps improving with the increase in the shared enhanced data size, i.e., as shown in Figure 5a,b. This is because the initialized quasi-global model can better capture the traffic patterns when more data samples are available. Unless otherwise stated, the following experimental results in the article default to the results at $\beta = 100\%$.



(**a**) (**b**)

Figure 5. (**a**) describes the performance indicators on the Trento dataset; (**b**) describes the performance indicators on the Milano dataset.

To further evaluate the prediction capabilities of different algorithms, comparisons between the predicted and real network traffic values derived using different prediction algorithms for randomly selected base stations on the Trento and Milano datasets are presented in Figures 6 and 7, respectively, which include the Cumulative Distribution Function (CDF) results of the absolute prediction errors, and this experiment chooses FedAvg as the benchmark for performance comparisons because it achieves the best performance among all the baseline methods in Table 1. As can be seen in Figures 6 and 7, the FedGSA prediction capability outperforms the popular FedAvg algorithm.



(**a**) (**b**)

Figure 6. (**a**) describes the traffic comparison on the Trento dataset; (**b**) describes the results of the on cumulative distribution function (CDF) on the Trento dataset.

For the prediction error, in the Trento dataset, for example, FedGSA has about 95% errors less than 0.3, while FedAvg has about 89%, and FedGSA outperforms FedAvg in predicting the peak fluctuation of flow values. Based on the above evaluation, it can be concluded that the algorithm of this experiment can obtain more accurate prediction results than the baseline method.

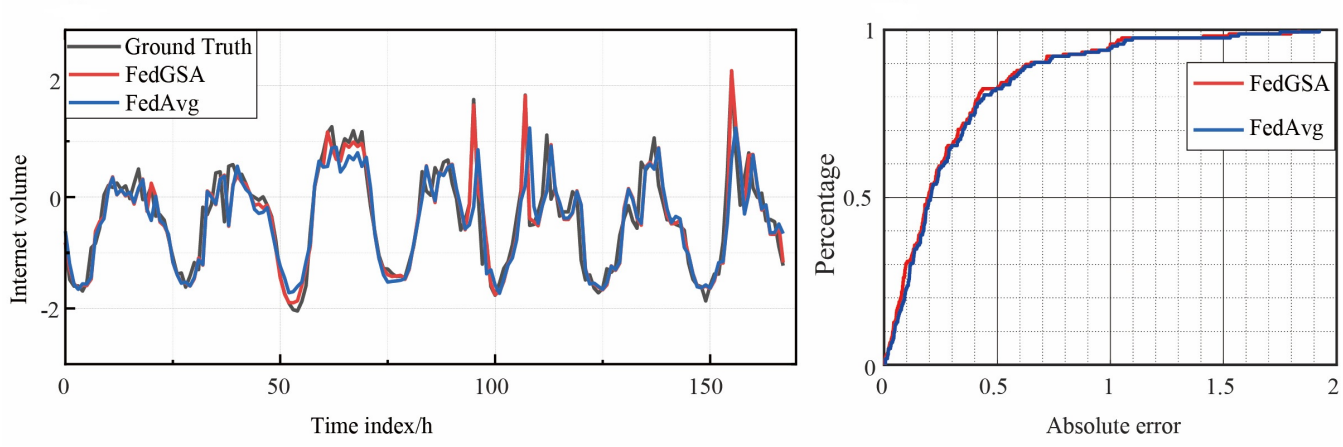


Figure 7. (**a**) describes the traffic comparison on the Milano dataset; (**b**) describes the results of the cumulative distribution function (CDF) on the Milano dataset.

4.3. Communication Rounds versus Prediction Accuracy

In FL or any other distributed learning framework, communication resources are often more valuable than computational resources, and fewer communications are preferred. Therefore, in this subsection, we report the prediction accuracy along with each communication cycle (epoch) and use the R-squared fraction to indicate the accuracy as it reflects how well the model predicts the true value of the network traffic [21]. As shown in Figure 8, we can observe that FedGSA achieves a higher prediction accuracy on both datasets, in addition to the fact that FedGSA requires fewer communication rounds to reach a certain accuracy; for example, for the Milano dataset, after 30 communication rounds, FedGSA achieves an accuracy of about 82% for wireless network traffic, and as for FedAVG, the prediction accuracy is about 75% Therefore, we consider that our proposed method has higher communication efficiency, which is also one of the important metrics for evaluating federation learning methods.

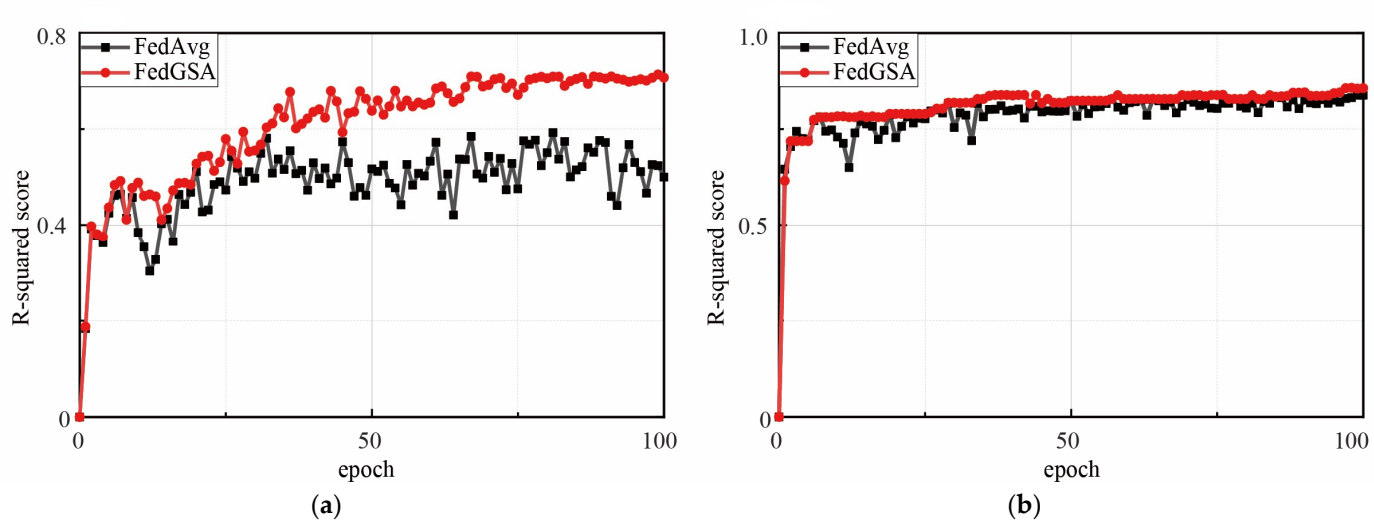


Figure 8. (**a**) Prediction accuracy and communication rounds on the Trento dataset. (**b**) Prediction accuracy and communication rounds on the Milano dataset.

5. Conclusions

In this paper, we propose a model gradient-based similarity aggregation scheme for federation learning and wireless network traffic prediction and name the framework

FedGSA, which uses similarity knowledge to construct individualized models for each client and realize model aggregation for each client to improve the generalization ability of the final global model. Experiments are conducted to predict the base station network traffic based on LSTM long short-term memory network on two real network traffic datasets, and the enhanced data scheme and sliding window strategy are combined to further overcome the problem of high data heterogeneity during FL training and improve the prediction capability. Compared with the current mainstream federal average algorithm, the method proposed in this paper achieves good results in simulation experiments.

Author Contributions: L.L.: methodology, formal analysis, data curation and validation and writing—original draft. Y.Z.: methodology, writing—review and editing and project administration. J.W.: Funding acquisition and Supervision. C.Z.: conceptualization, methodology and formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Natural Science Foundation of Inner Mongolia Province of China (Grant No. 2022MS06006, 2020MS06009) and supported by Basic research funds for universities of Inner Mongolia Province of China (Grant Name. Application of Federated Learning for Privacy protection in wireless traffic prediction).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in the study are from the literature, which is publicly accessible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yao, M.; Sohul, M.; Marojevic, V.; Reed, J.H. Artificial intelligence defined 5G radio access networks. *IEEE Commun. Mag.* **2019**, *57*, 14–20. [CrossRef]
2. Xu, Y.; Yin, F.; Xu, W.; Lin, J.; Cui, S. Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1291–1306. [CrossRef]
3. Kato, N.; Mao, B.; Tang, F.; Kawamoto, Y.; Liu, J. Ten challenges in advancing machine learning technologies toward 6G. *IEEE Wirel. Commun.* **2020**, *27*, 96–103. [CrossRef]
4. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
5. Zhang, C.; Dang, S.; Shihada, B.; Alouini, M.-S. Dual attention-based federated learning for wireless traffic prediction. In Proceedings of the IEEE INFOCOM 2021-IEEE Conference on Computer Communications, Virtual, 10–13 May 2021; pp. 1–10.
6. Yantai, S.; Minfang, Y.; Oliver, Y.; Jiakun, L.; Huifang, F. Wireless traffic modeling and prediction using seasonal arima models. *IEICE Trans. Commun.* **2005**, *88*, 3992–3999.
7. Zhou, B.; He, D.; Sun, Z. Traffic predictability based on ARIMA/GARCH model. In Proceedings of the 2006 2nd Conference on Next Generation Internet Design and Engineering, Valencia, Spain, 3–5 April 2006; pp. 8–207.
8. Li, R.; Zhao, Z.; Zheng, J.; Mei, C.; Cai, Y.; Zhang, H. The learning and prediction of application-level traffic data in cellular networks. *IEEE Trans. Wirel.* **2017**, *16*, 3899G–3912. [CrossRef]
9. Li, R.; Zhao, Z.; Zhou, X.; Palicot, J.; Zhang, H. The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice. *IEEE Commun. Mag.* **2014**, *52*, 234–240. [CrossRef]
10. Chen, X.; Jin, Y.; Qiang, S.; Hu, W.; Jiang, K. Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 3585–3591.
11. Zhang, Z.; Wu, D.; Zhang, C. Study of Cellular Traffic Prediction Based on Multi-channel Sparse LSTM. *Comput. Sci.* **2021**, *48*, 296–300.
12. Gao, Z.; Wang, T.; Wang, Y.; Shen, H.; Bai, G. Traffic Prediction Method for 5G Network Based on Generative Adversarial Network. *Comput. Sci.* **2022**, *49*, 321–328.
13. Song, Y.; Lyu, G.; Wang, G.; Jia, W. SDN Traffic Prediction Based on Graph Convolutional Network. *Comput. Sci.* **2021**, *48*, 392–397.
14. Zhang, C.; Zhang, H.; Qiao, J.; Yuan, D.; Zhang, M. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1389–1401. [CrossRef]
15. Wen, Q.; Sun, L.; Song, X.; Gao, J.; Wang, X.; Xu, H. Time series data augmentation for deep learning: A survey. *arXiv* **2020**, arXiv:2002.12478.
16. Zhang, C.; Zhang, H.; Yuan, D.; Zhang, M. Citywide cellular traffic prediction based on densely connected convolutional neural networks. *IEEE Commun. Lett.* **2018**, *22*, 1656–1659. [CrossRef]

17. Zhang, J.; Zheng, Y.; Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1843–1846.
19. Telecom Italia. *Telecommunications-SMS, Call, Internet-TN*; TIM: Rome, Italy, 2015. [CrossRef]
20. Telecom Italia. *Telecommunications-SMS, Call, Internet-MI*; TIM: Rome, Italy, 2015. [CrossRef]
21. Cameron, A.; Windmeijer, F. R-squared measures for count data regression models with applications to health-care utilization. *J. Bus. Econ. Stat.* **1996**, *14*, 209–220.